THE 2025 INTERNATIONAL CONFERENCE OF THE KOREAN SOCIAL SCIENCE RESEARCH COUNCIL (KOSSREC)

KOSSREC
(사)한국사회과학협의회

FROM DIGITAL TRANSFORMATION (DX) TO AI TRANSFORMATION (AX): EXPLORING NEW PARADIGMS IN THE SOCIAL SCIENCES

May 27(Tue) - May 28(Wed)

Venue

- SEOUL DRAGON CITY HOTEL, SEOUL, REPUBLIC OF KOREA

Hosted by

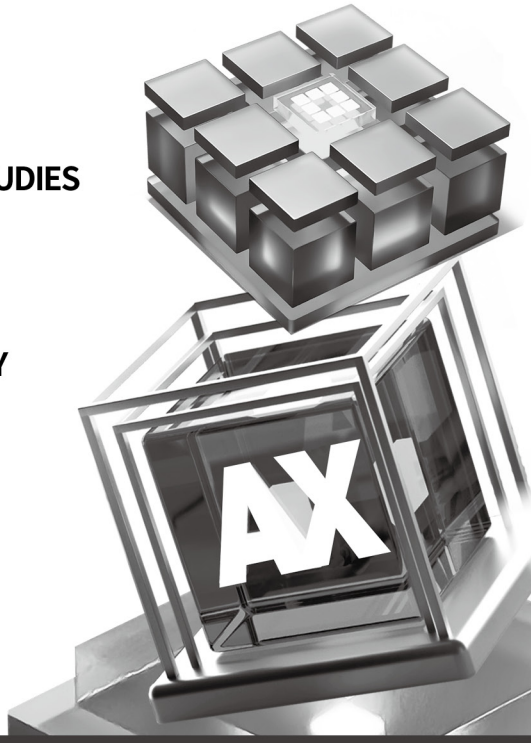
- KOREAN SOCIAL SCIENCE RESEARCH COUNCIL (KOSSREC)
- MINISTRY OF SCIENCE AND ICT(MSIT)

Organized by

- THE KOREA ASSOCIATION OF INTERNATIONAL DEVELOPMENT AND COOPERATION
- THE KOREAN ACADEMIC SOCIETY OF BUSINESS ADMINISTRATION
- THE KOREAN ACADEMY OF SOCIAL WELFARE
- THE KOREAN ASSOCIATION FOR PUBLIC ADMINISTRATION
- THE KOREAN ASSOCIATION OF WOMEN'S STUDIES
- THE KOREAN ECONOMIC ASSOCIATION
- THE KOREAN GEOGRAPHICAL SOCIETY
- THE KOREAN POLITICAL SCIENCE ASSOCIATION
- THE KOREAN SOCIETY FOR JOURNALISM & COMMUNICATION STUDIES
- THE KOREAN SOCIOLOGICAL ASSOCIATION

Supported by

- INSTITUTE OF INFORMATION & COMMUNICATIONS TECHNOLOGY PLANNING & EVALUATION (IITP)
- NATIONAL INFORMATION SOCIETY AGENCY (NIA)
- HUMANITIES UTMOST SHARING SYSTEM (HUSS)
- GLOBAL DEVELOPMENT COOPERATION INSTITUTE (GDCI)
- NATIONAL RESEARCH FOUNDATION OF KOREA(NRF)
- SOCIAL SCIENCE KOREA (SSK) RESEARCH SUPPORT OFFICE



IITP Institute of Information & Communications Technology Planning & Evaluation
NIA NATIONAL INFORMATION SOCIETY AGENCY
HUSS HUMANITIES UTMOST SHARING SYSTEM
GDCI GLOBAL DEVELOPMENT COOPERATION INSTITUTE
NRF National Research Foundation of Korea
SSK SOCIAL SCIENCE KOREA

CONTENTS

keynote

1

The Future of Artificial Intelligence: Impacts on Society, Human Values, and Ethics Keynote speech by Jong Sup Jun at the international conference on From Digital Transformation to AI Transformation in the Social Sciences, Seoul, Korea, May 27-28, 2025.

A short introduction about me.

As I am not a software or hardware designer specializing in AI machine learning, I will present my views on Artificial Intelligence from the perspective of a social scientist trained in public administration, political science, and public policy.

Introduction

Artificial Intelligence (AI) has been a leading technology that has significantly influenced our lives in recent years, continually growing and becoming increasingly powerful. And the rapid transformation of economies, businesses, societies, and human values is evident today. We see the use of AI in various applications, including schools, banks, hospitals, smartphones, self-driving cars, chatbots, weather forecasts, robots, medical surgery equipment, and warfare. AI scientists and salespeople present videos to senior executives and managers to enhance productivity and customer service. Top executives and managers may be convinced to learn the benefits of AI technology when their organizations adopt it. However, some AI researchers argue that AI programs may work well in some cases but do not work as intended to solve the original problem (Narayanan & Kapoor, 2024).

My presentation addresses three main points: First, to ensure that the benefits of AI align with human values and ethical principles; Second, AI programs must be designed to consider specific cultures and social contexts; Finally, to ensure human control of AI technology requires a close interface between new technology and human intelligence.

What is Artificial Intelligence?

In 1956, John McCarthy, a computer scientist, coined the term *"Artificial Intelligence" at* the Dartmouth Conference*,* which he organized, marking a significant milestone in the development of AI. This conference, held at Dartmouth College as a workshop, invited a group of computer scientists. The participants had agreed to accept Artificial Intelligence as a field of study.

I believe it is necessary to clarify the difference between Artificial Intelligence and human intelligence. Among many explanations, John Lennox provides a good description of AI for us in his book, 2084 and the AI Revolution: “AI is the application of mathematical software code to teach computers how to understand, synthesize, and generate knowledge in ways similar to how people do it. AI is a computer program like any other—it runs, takes inputs, processes, and generates outputs...”. (John Lennox, 2024, p.15). Thus, AI refers to making computers and machines think and perform. Scientists teach machines to learn, make decisions, and solve problems, just like people do.

AI generally deals with the analysis of typical human behavior, as presupposed and designed by an AI designer. Unlike AI, human interest tends to reflect people’s actions, feelings, emotions, consciousness, values, and intuitions. These points will be discussed further in this presentation. To ensure justification for my claims about AI, I have searched literature on AI subjects. I found that it is impossible to know how many articles, books, papers, and even TV podcasts on social media, and not even try to sort out which ones are more important than others among the hundreds of materials written and spoken in the media.

The Impact on Society and the Global Economy

There have been numerous discussions about the impact of AI on society and the world. These discussions can be classified into two leading positions: the pros and cons of AI. Both arguments are valid because they are interrelated. In the following, I list some of the expected benefits and risks that are discussed in the literature:

Benefits (positive aspects) of AI:

. *Impacts on the global economy*. One of the rising interests in AI technology can be seen in investors' interest in the stock market. People have invested in technological companies like Nvidia, Apple, Alphabet (Google), Microsoft, Meta, Tesla, OpenAI, Palantir AI Technologies, Tempus Technologies, Samsung Electronics, Taiwan Semiconductor, Global X AI, etc. These companies develop powerful computer chips designed to improve productivity, such as in business organizations, factories, hospitals, defense management, etc.

According to a recent survey of executives and AI experts, the impact of AI, particularly the development of generative AI, is expected to increase global GDP. The McKinsey Global Institute estimates that "AI will add between $2.6 and $4.4 trillion in annual value to the global economy, increasing the economic impact of AI by 15 to 40%" (MIT Technology Review Insights, 2024). Furthermore, these companies are investing billions of dollars in developing the infrastructure to handle massive data collection centers, increasing the capacity for electrical power use, among other initiatives.

. *Increasing productivity and problem-solving.* To solve a targeted problem, AI scientists design a software program that predicts the outcome, such as increasing

productivity. In the process of creating software, they use a mathematical code system, such as assigning numbers and code systems for language usage, so-called algorithms.

. *Speedy decision-making*. The basic assumption of relying on AI machines is to make faster decision-making among alternatives derived from a rational analysis of massive data stored in the computer. Thus, an AI machine aims to improve human decision-making and increase the output performance. A machine will make a final choice, not a human.

. *Improving customer service.* Organizations have adopted AI to enhance communication and interaction with clients by understanding human language and responding to their questions, such as widely used technologies like chatbots and ChatGPT, which translate the language of a customer that a machine can understand. AI machines are supposed to learn selected languages chosen by the AI experts.

. *Increasing use of robots*. As I know, the rise of robotics began in the 1950s. Today, robots with new technology are used in industrial manufacturing to perform routine tasks, such as assembly, welding, and packaging. And the robots perform some medical surgeries and are engaged in military functions.

. Speeding medical research: AI has been changing the speed of medical research to discover unknown factors in the human body. According to **Harvard Health Letter** (vol. 49, no.11, 2024), in July 2021, the DeepMind research company, owned by Google, was able to determine the “shape of nearly all the human proteins (98%). Overnight, we went

from knowing the shape of about 30% of human proteins to knowing the shape of nearly all of them."

In summary, many applications of AI technology aim to improve predictive outcomes, as AI experts continue to generate new insights.

Risks (negative aspects) of AI :

We are increasingly adopting AI technologies in various aspects of our lives. However, there are many unintended consequences of AI. For example, the widespread use of smartphones has enhanced our communication with others, but at the same time, it tends to reduce direct interaction between people as we spend more time looking at our phones.

While we hear many positive things about AI, there are also significant potential risks to people, primarily related to the protection of human values.

In the following, some risks stem from the AI application:

. *Privacy Issue:* When a person purchases something online, she or he discloses personal data to a company. Although many companies have security systems in place to protect personal data and ensure privacy. However, due to the misuse of AI data, such as sharing with other companies, security risks are always possible.

. . *Job Losses and Job Displacement:* The efficient use of AI-powered automation to increase productivity tends to reduce the number of workers and assign them to untrained positions in the company.

. *Robotics and the Existential Threat to Humanity*. Science fiction and movies depict the killing of people, and even the master who commands the robots. They show a possibility of dismantling humanity, societies, cultures, and the world.

. *Lack of Transparency and Accountability*: When the software designers develop a specific model for a machine, they apply the code system using algorithms. Likely, people in the company do not understand the process and the meanings of the codes. When the application doesn't work correctly, who is responsible for correcting the error? In the case of a lawsuit, is the machine responsible for causing the problem, the software designer, or the company?

. *Discrimination and Bias in Applying Data:* AI designers might continue to practice using existing biased information that results in discrimination in recruitment and social injustice.

. *AI-powered Voice Cloning*: AI-powered voice replication has been beneficial to entertainment organizations. They can replicate or alter the original voice of actors and use it for multiple situations to meet the commercial demands. When voice cloning is done without actors' permission, knowledge, or compensation, it leads to the unethical practice of business. This occurrence certainly is a wake-up call for voice actors to protect their safety and rights. There is a movement to influence the political process to make regulations to protect their rights.

. *Imposters Deceive People:* In the world of AI, imposters are everywhere. When someone sends you an email or a phone message to offer you gifts, loan forgiveness, free trips, or money coupons at a big discount price, it may not be genuine. To get this one-time deal, this person asks you to charge it to your credit card soon. Once we provide our personal information, such as our credit card number with a security code, the scammer may make unauthorized charges. So, anyone

who asks you for this kind of deal might be a scammer generated by AI. If something seems off or too good to be true, it may be fake, and we should not reply.

We cannot predict any other potential negative issues that may emerge in the future. The above problems serve as severe warnings regarding the misuse of AI technology.

Values and Ethical Challenges

To develop and implement responsible AI projects, I believe two potential issues need to be seriously considered: technical and ethical issues. The technical elements include how to collect, evaluate, and utilize data effectively, as well as how to code algorithms that apply a set of rules. AI program designers, at some point, decide to stop collecting data to be used as input to the computer. This data includes past and present statistical information, which may reflect inequality and prejudice toward marginalized groups in society. If these kinds of data are used in the computer analysis and problem-solving process, the results of implementing an AI project could perpetuate discriminatory practices. Another technical program arises from overreliance on the set of procedures used in algorithms that deal with a set of instructions designed to solve a problem. The above illustration suggests that a biased way of collecting data can inevitably lead to a biased result.

As AI programs focus on increasing organizational efficiency and productivity as their primary goal, they tend to overlook ethics and morality because these are qualitative and even irrational elements in their rational and scientific analysis. Therefore, in the process of analyzing data as an input to solving organizational problems and making decisions, ethical considerations are viewed as interfering with their scientific endeavors.

In contrast to AI, human values hold that ethical justification is unavoidable. People working in organizations, whether they are executives, managers, or lower-level employees, are

expected to perform their roles responsibly. They have the experience of judging what is the right thing to do, both individually and collectively, for their job and the organization.

In contrast to the objective approach of determining data and following algorithmic procedures to solve a specific problem and task, ethical requirements support the subjective aspect of human values. People ask questions that reflect on their work experiences, enabling them to contribute to complementing the machine learning process and improving the problem-solving process.

Philosophical Issues:

. *Epistemological Bias:* Epistemology is a theory of knowledge that is concerned with the investigation of the origin, structure, methods, and validity of knowledge. The AI experts' approach to understanding the problematic situation is objective from their points of view in the sense that they tend to focus on the externally observed behavior of people and collect available quantitative data to analyze, such as targeting a problem of low productivity in business. Their understanding of the behavior of employees and ways of increasing productivity leads to applying a set of procedures to the problem-solving process.

In contrast to the scientist's approach to understanding the social situation, the humanities and many social scientists attempt to comprehend a phenomenon of productivity through the experiences of people and their actions. They do consider a method of explaining human activity from a psychological or other qualitative point of view. They try to understand why people experience problems with low productivity and what employees think about improving the situation.

. *Need for Critical Reflection*: One of the problems of AI is overreliance on behavior and objective data in designing software programs, which discounts people's ability to reflect, raise critical questions, and act accordingly. AI machines cannot perform a critical evaluation of their operations as humans can reflect on their actions and behaviors. A critical perspective of humans can make a significant contribution to the improvement of AI implementation. It could offer ways of improving the problems occurring in the process of its operation, reflecting on their experiences and practices.

The capability of critical reflection enables people to understand and interpret existing behavior and actions in terms of their ethical implications and the degree of responsibility involved. The meaningful transformation of a new technology can occur through a dialectical process in which artificial intelligence and human intelligence interact, resulting in a creative synthesis to accomplish organizational change. Thus, a critical perspective provides a framework for discovering alternatives and exploring possibilities for action, as well as identifying inadequacies in AI operations that are more congruent with organizational problem-solving.

. Do *AI machines have consciousness?* A software designer programs AI machines to choose how to solve a problem. A decision made by a machine is an unconscious choice. However, a designer instructed the computer to follow the procedures to execute a program consciously at the time of selecting certain information. Machine learning is not equivalent to human learning because people make their choices based on their intuition, experience, and personal knowledge. Thus, it is fair to say that machines do not have absolute consciousness and cognitive ability that reflects feelings, emotions, and intuition.

The Future of Artificial Intelligence:

The AI revolution is viewed as a significant transformation in scientific and technological development within modern industrial society. It is considered a shift in scientific knowledge that is valued by society and that enables control over the natural environment, the global economy, and enhances human experiences.

To summarize my presentation, I would like to say that the future of Artificial Intelligence is a challenging subject to comprehend because human behavior, consciousness, and thinking ability are evolving in response to this technological change. Scientists who have great expectations for the future of AI advocate that by 2035, people might expect AI machines to be developed to achieve the "so-called super-intelligence". I believe that knowledge based on AI technology has limits that cannot exceed human intelligence in some elements, such as creativity and innovation. AI machines cannot experience emotions such as sorrow, excitement, love, or passion. Kurzweil and many scholars have a great face in AI believe that "we are fast approaching a 'technological singularity', a point at which AI far surpasses human intelligence and can solve problems we were not able to solve before, with unpredictable consequences for civilization and human nature" (Kurzweil, 1999; Schneider, 2019, p. 10).

Another group of scholars advocates the idea of "transhumanism. "Transhumanists aim to redesign the human condition, striving for immortality and synthetic intelligence, all in hopes of improving our overall quality of life." (Schneider, 2919, p. 73) For example, the implantation of sophisticated microchips in the brain may be able to cure diseases, such as mental illness. They are "techno-optimists" who believe in the possibility of "synthetic consciousness." What if a microchip is implanted in a patient's brain, and that person has severe side effects that require

immediate medical treatment? Suppose this person is permanently disabled. If he or she files a malpractice suit, who is responsible for the case: a chip designer, an AI machine, or the doctor?

Thus, there might be unpredictable episodes that we cannot know in advance. Of course, I doubt that an AI machine can predict and prevent undesirable elements from far away in the future. To predict future events, AI experts invest vast amounts of resources in developing extensive infrastructure to store and process information. Quantum computing, as the main technological strategy, aligns with the complexity of nature by storing vast amounts of data and manipulating it for the benefit of business and human needs. Today, we do not know how effectively quantum computing may work.

Finally, I would like to repeat that programs used in machine learning should be aligned with human values to increase the so-called "super intelligence" or "transhumanism," which might be achieved through a high level of interaction between humans and machines. Furthermore, aligning human learning with AI machine learning would be highly challenging, if not impossible. I also want to emphasize that while humans take advantage of AI technology to enhance our ability to meet human needs, as well as to enhance our consciousness, humans should control evolving technologies and not let machines control our destiny.

References

Gambelin, Olivia. (2024). Responsible AI: Implement an Ethical Approach in Your Organization. London: Kogan Page Limited.

Jun, Jong S. (2006). The Social Construction of Public Administration: Interpretive and Critical Perspectives. Albany, New York: State University of New York Press.

Kissinger, Henry. A., Schmidt, Eric, Huttenlocher, Daniel. (2021). The Age of AI and our Human Future. New York, NY: Little, Brown, and Company.

Kurzweil, R. (1999). Age Spiritual Machines: When Computers Exceed Human Intelligence. New York: Penguin.

Lennox, John C. Updated and expanded edition. (2024). 20284 and the AI Revolution: How Artificial Intelligence Informs Our Future. Grand Rapids, Mich: Zondervan.

MIT Technology Review Insights (2024).

Narayanan, A., & Kapoor, Sayash. (2024). AI Snake Oil: What Artificial Intelligence Can Do, What It Cannot, and How to Tell the Difference. Princeton, NJ: Princeton University Press.

Schneider, S. (2019). Artificial You: AI and the Future of Your Mind. Princeton, New Jersey: Princeton University Press.

Comedians have said the following jokes about Artificial Intelligence:

. Why did the AI break up with its girlfriend? It could not find a compatible algorithm for love.

. Why was AI bad at stand-up comedy? Because the jokes were predictable.

. Parents ask a child: If all your friends jumped into the well, would you?

Machine learning: Yes

Kid: No

. How many AI researchers does it take to change light bulbs?

None, they simulate the light bulb and leave the room dark.

. Why did the AI go on a diet? It had too many bites.

. Why don’t AI researchers like nature? It has too many bugs.

. Scientists predict human-level AI by 2030.

Maybe sooner if the bar keeps dropping!

United Nations University

Institute for the Advanced Study of Sustainability

Education and Research in an Era of AI:
What Should Remain and What Should Change

Prof. Shinobu Yume Yamaguchi
Director, UNU-IAS
27 May 2025
The 2025 International Conference of the Korean Social Science Research Council (KOSSREC)

UNU IAS

United Nations University

- In December 1972, the General Assembly of UN adopted the decision to establish the UNU.
- With the contribution from the Government of Japan, headquarters facilities in Tokyo and US$100million to establish an endowment fund, UNU launch its academic work in September 1975.

Mission

- Serves as a bridge between the United Nations and the international academic community
- Provides opportunities for global and local dialogues and sharing of creative new ideas
- Contributes to capacity building in developing and transitional countries

UNU IAS

UNU System

A global system of research and training institutes coordinated by UNU Centre in Tokyo

12 Countries

13 Institutes

UNU IAS

UNU Institute for the Advanced Study of Sustainability (UNU-IAS)

UNU-IAS is a research and teaching institute dedicated to realizing a sustainable future for people and our planet.

Established in 2014 through the consolidation of two previous UNU institutes

Mission

"to meet the pressing challenges for achieving sustainability that are of concern to the United Nations and its Member States"

- Inform policymaking for sustainability by producing and disseminating solution-oriented research
- Promote interdisciplinary understanding and approaches
- Develop future generations of policymakers and researchers

UNU IAS

Outline

1. Three dilemmas in higher education
2. Promising practices: AI, sustainability and higher education
3. What UNU does
4. Some food for thoughts

Dilemma 1:
To use vs. Not to use

The Guardian Int

Maya Yang

Fri 6 Jan 2023 18.19 GMT

New York City schools ban AI chatbot that writes essays and answers prompts

NBC News

https://www.nbcnews.com/tech/chatgpt-ban-dropped-new...

New York City public schools remove ChatGPT ban

WEB May 18, 2023 · New **York City's Department of Education** will rescind its popular chatbot ChatGPT — which some worried could inspire more student

Leaders are pushing to ban cellphones in schools. Are they a distraction or necessary tool?

Kayla Jimenez

USA TODAY

Published 5:04 a.m. ET Aug. 24, 2024 | **Updated 1:41 p.m. ET Aug. 30, 2024**

香 港 大 學

THE UNIVERSITY OF HONG KONG

HKU introduces new policy to fully integrate GenAI in Teaching

03 Aug 2023

BBC

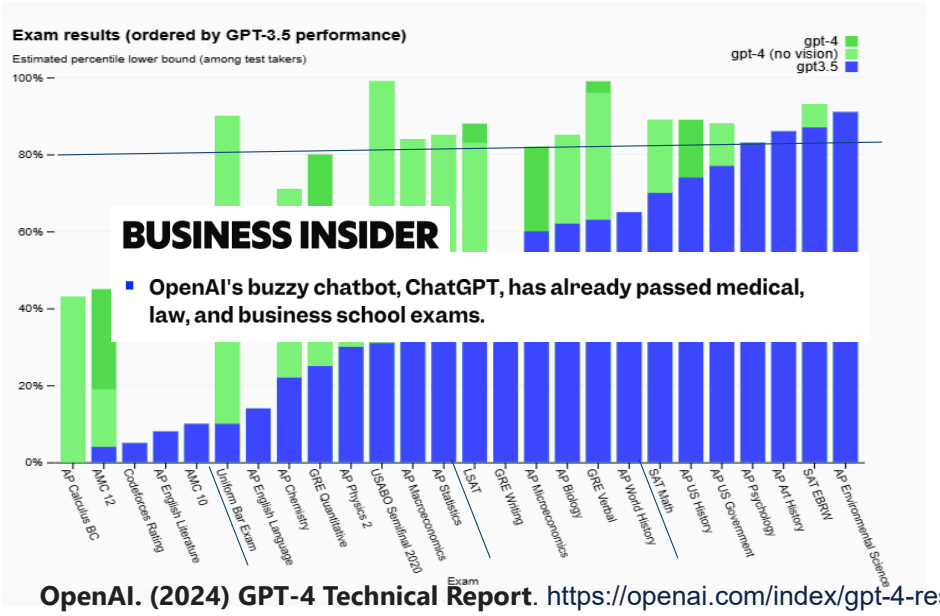
Home News US Election Sport Business Innovation Culture Arts Travel E

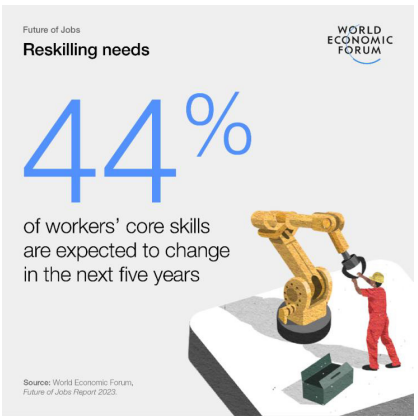
Fight begins to make mobile-free schools law

15 October 2024

UNU IAS

Dilemma 2:
Future competencies we don't know



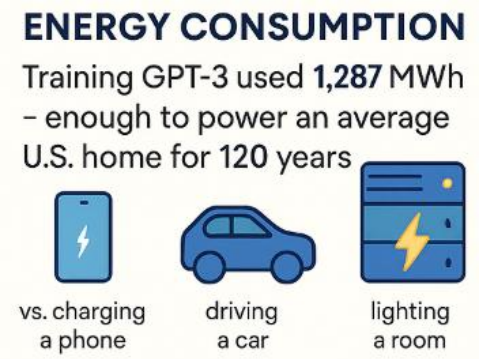


OpenAI. (2024) GPT-4 Technical Report. https://openai.com/index/gpt-4-research/

Dilemma 3:
AI: Enablers or barriers to sustainable development?

1. Preservation of indigenous language vs. acceleration of extinction of minor language

2. Environmental footprints:
   - Promoting data transparency (e.g. usage of natural resources etc.)
   - Excessive energy consumption
   - A single ChatGPT conversation uses about 500ml of water, equivalent to one plastic bottle. (Gordon, 2024)



UNU IAS

Outline

1. Three dilemmas in higher education
2. Promising practices: AI, sustainability and higher education
3. What UNU does
4. Some food for thoughts

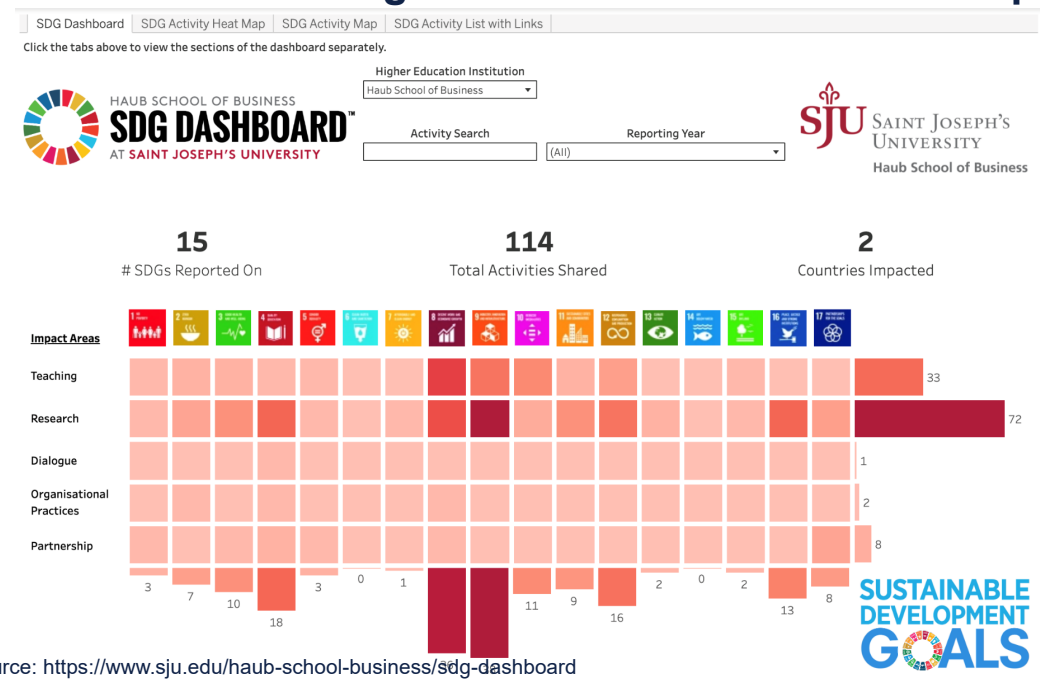
Global Digital Compact (Sept 2024)
A comprehensive global framework for digital cooperation and governance of artificial intelligence

5 Objectives

1. Close all **digital divides**
2. Expand **inclusion in digital economy**
3. Foster an **inclusive, open, safe, and secure digital space,** respecting **human rights**
4. Advance responsible, equitable and inter-operatable **data governance**
5. Strengthen international **governance of AI for humanity**

Source: https://www.un.org/global-digital-compact/en#:~:text=Close%20all%20digital%20divides%20and%20deliver%20an,digital%20public%20goods%20and%20digital%20public%20infrastructure.

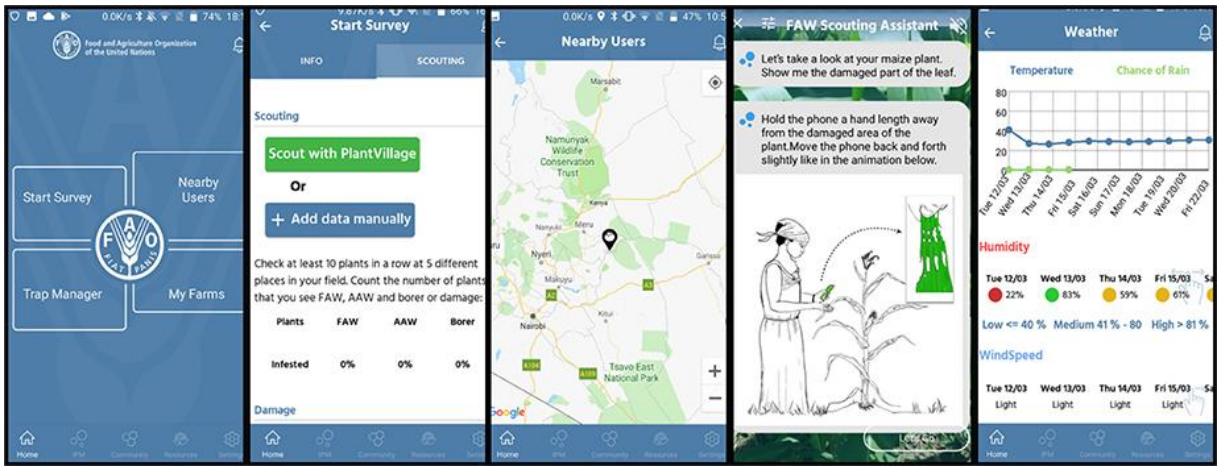
ChatSDG: Harnessing AI for Measurable Societal Impact



Source: https://www.sju.edu/haub-school-business/sdg-dashboard
https://www.sju.edu/news/university-report/living-the-mission/chatsdg-harnessing-ai-measurable-societal-impact

Food and Agriculture Organization of the United Nations

Global Action for Fall Armyworm Control



Source: https://www.fao.org/fall-armyworm/monitoring-tools/famews-mobile-app/en/
https://www.fao.org/fall-armyworm/background/en/

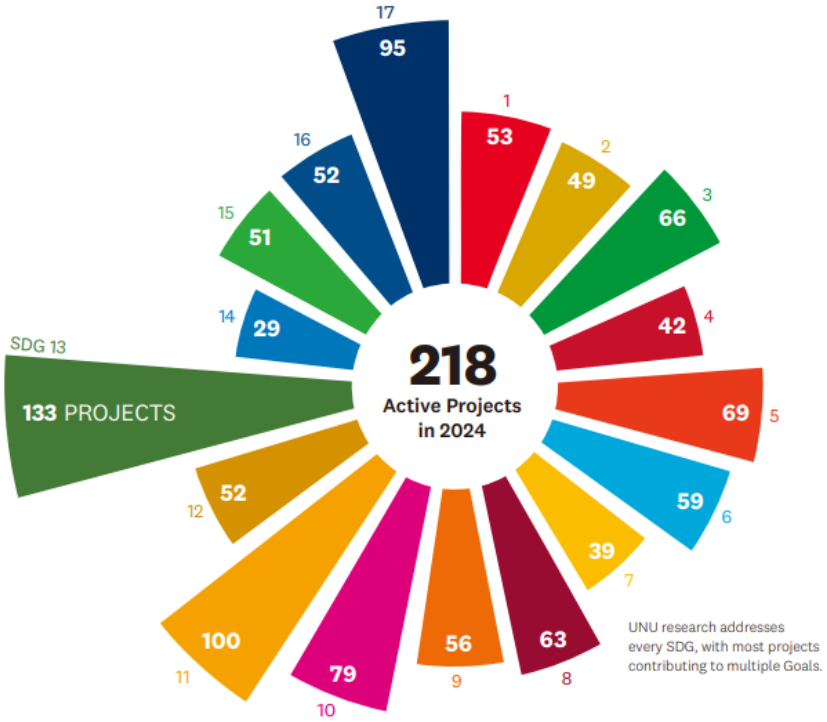
1 NO POVERTY
2 ZERO HUNGER
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION

Outline

1. Three dilemmas in higher education
2. Promising practices: AI, sustainability and higher education
3. What UNU does
4. Some food for thoughts

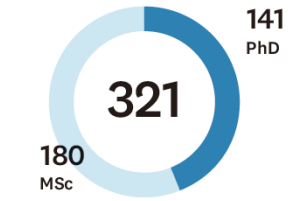
UNU Research 2024

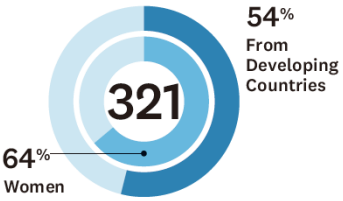
UNU research addresses every SDG, with most projects contributing to multiple Goals.



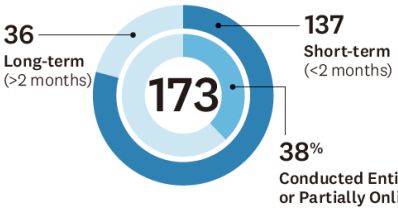
UNU Education 2024

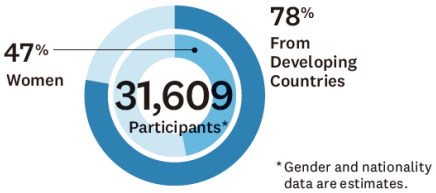
Postgraduate Students





Non-degree Training Programmes and Activities





IAS

UNU Macau AI Conference 2025

Oct. 23-24, 2025 | Macau SAR, China

- Theme: **AI for Humanity: Building an Equitable Digital Future**
- Three pillars:
  - AI Research for Narrowing the Digital Divide
  - Inclusiveness and Capacity Building in the AI Era
  - Navigating the AI Ecosystem through Synergies and Interdisciplinary Innovation

UNU IAS

UNU Global AI Network

To become a member: AINetwork@unu.edu

Higher Education Sustainability Initiative

Launched in the lead-up to the Rio+20 Conference in 2012 as an open partnership between several UN entities and the higher education community

AIM

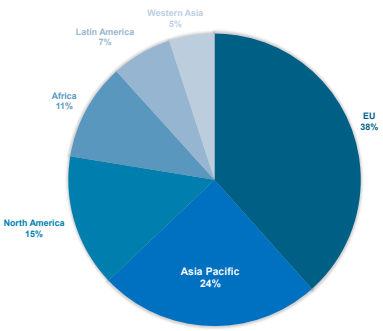
Enhance the role of the higher education sector in advancing sustainable development through multi-stakeholder discussions, actions, and the dissemination of best practices.

LEADERSHIP

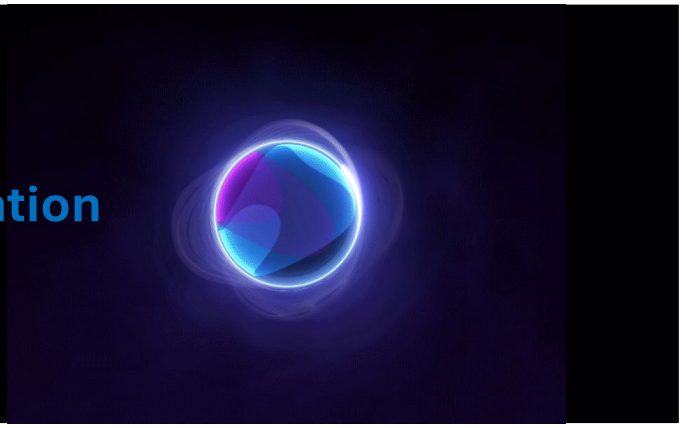
1,257 Members
(and growing)

Any higher education institution or interested organization may join HESI.

HESI Members Regional Distribution

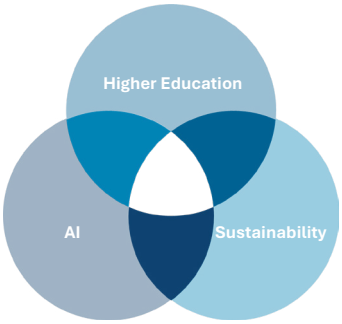


HESI Action Group on
The Futures of Higher Education and Artificial Intelligence



Aim: Explore the impact of AI to the higher education sector through the lens of sustainability.

UNU IAS



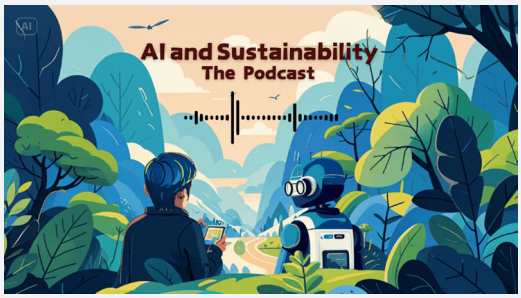
Opportunities for collaboration

Action Group Implementation Teams

- AI in research
- AI in teaching & learning
- AI in uni management
- Ethics, safety, inclusivity

For more info:
sdgs.un.org/HESI/Futures

AI & Sustainability Podcast Series



Apple Podcast | Spotify | YouTube

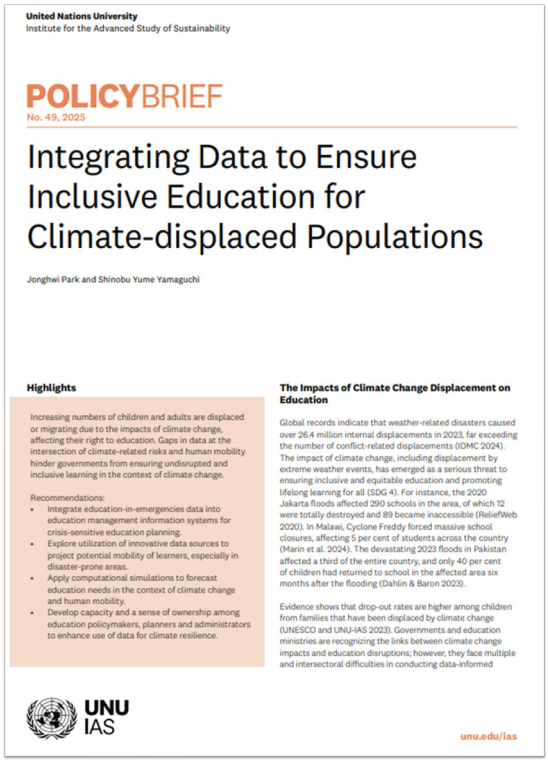
Become a partner:

UNU IAS

Research

- Education as fundamental human right in the context of climate change displacement
- Policy Recommendations:
  - Integrate education-in-emergencies data into education management information systems for crisis-sensitive education planning.
  - Explore utilization of innovative data sources to project potential mobility of learners, especially in disaster-prone areas.
  - Apply computational simulations to forecast education needs in the context of climate change and human mobility.
  - Develop capacity and a sense of ownership among education policymakers, planners and administrators to enhance use of data for climate resilience.

UNU IAS



United Nations University
Institute for the Advanced Study of Sustainability

POLICYBRIEF
No. 49, 2025

Integrating Data to Ensure Inclusive Education for Climate-displaced Populations

Jonghwi Park and Shinobu Yume Yamaguchi

Highlights

Increasing numbers of children and adults are displaced or migrating due to the impacts of climate change, affecting their right to education. Gaps in data at the intersection of climate-related risks and human mobility hinder governments from ensuring undisrupted and inclusive learning in the context of climate change.

Recommendations:
- Integrate education-in-emergencies data into education management information systems for crisis-sensitive education planning.
- Explore utilization of innovative data sources to project potential mobility of learners, especially in disaster-prone areas.
- Apply computational simulations to forecast education needs in the context of climate change and human mobility.
- Develop capacity and a sense of ownership among education policymakers, planners and administrators to enhance use of data for climate resilience.

The Impacts of Climate Change Displacement on Education

Global records indicate that weather-related disasters caused over 26.4 million internal displacements in 2023, far exceeding the number of conflict-related displacements (IDMC 2024). The impact of climate change, including displacement by extreme weather events, has emerged as a serious threat to ensuring inclusive and equitable education and promoting lifelong learning for all (SDG 4). For instance, the 2020 Jakarta floods affected 290 schools in the area, of which 12 were totally destroyed and 89 became inaccessible (ReliefWeb 2020). In Malawi, Cyclone Freddy forced massive school closures, affecting 5 per cent of students across the country (Marin et al. 2024). The devastating 2023 floods in Pakistan affected a third of the entire country, and only 40 per cent of children had returned to school in the affected area six months after the flooding (Dahlin & Baron 2023).

Evidence shows that drop-out rates are higher among children from families that have been displaced by climate change (UNESCO and UNU-IAS 2023). Governments and education ministries are recognizing the links between climate change impacts and education disruptions; however, they face multiple and intersectoral difficulties in conducting data-informed

UNU IAS

unu.edu/ias

RCE

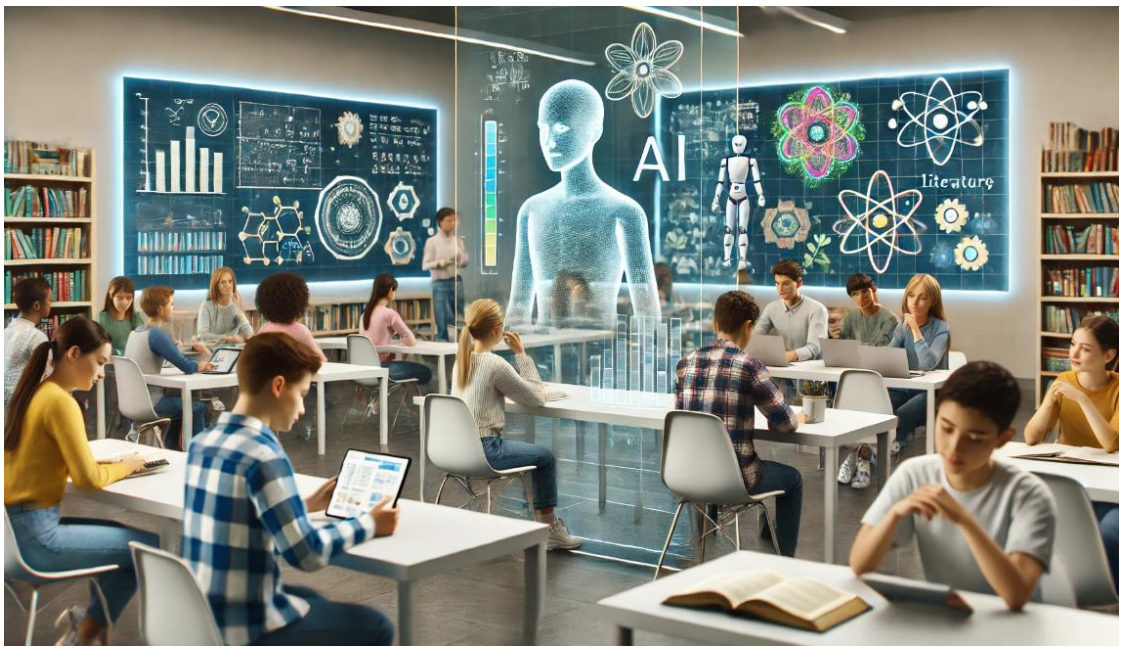
Outline

1. Three dilemmas in higher education
2. Promising practices: AI, sustainability and higher education
3. What UNU does
4. Some food for thoughts

Final thoughts: what are we aiming for?

By DALL-E (generated on 11 Nov 2024)

Create an image of AI assisting education



UNU IAS

Final thoughts: what are we aiming for?

By DALL-E (generated on 11 Nov 2024)

Create an image of technology in education

UNU IAS

UNU IAS

Thank you!

Shinobu Yume Yamaguchi
Institute for the Advanced Study of Sustainability
United Nations University
yamaguchi@unu.edu

UNU IAS

Artificial Intelligence Governance: The Magic Bullet for Transformation of Government?

Kim Normann Andersen
Professor, Head of Studies (Graduate IT Study Programs)
Copenhagen Business School, Denmark

2025 International Conference
Korean Social Science Research Council

From Digital Transformation (DX) to AI Transformation (AX): Exploring New Paradigms in the Social Sciences

May 27-28, 2025

FRUSTRATION

What will the future landscape of AI entail?

To what extent will AI and robotics redefine and displace functional and creative job tasks and processes?

Will government as we know it be radically transformed?

What is the right governance response?

How to balance the need for national competitiveness with other policy concerns (security, privacy, accountability etc.)

My Three Propositions

Rethinking and repositioning of governance structures and the traditional boundaries of government roles.

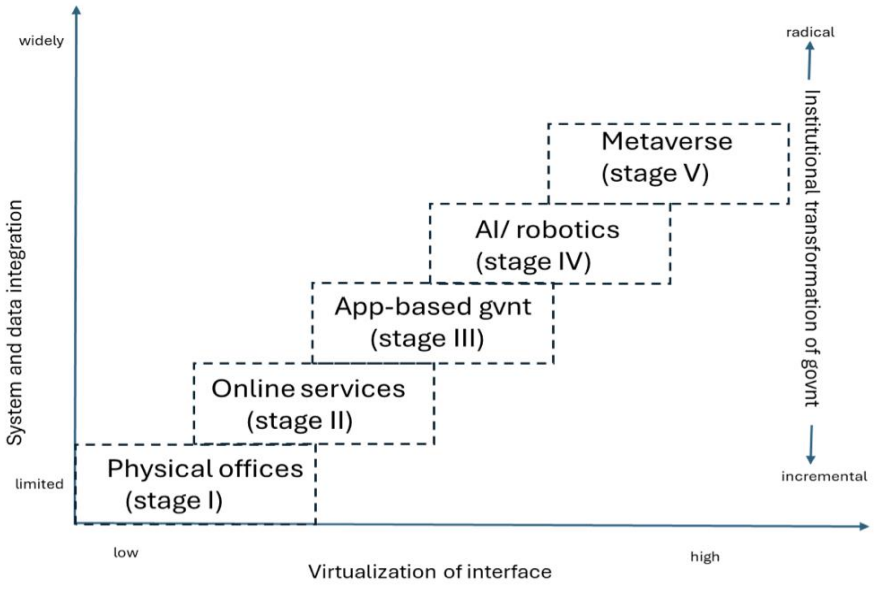
Deployment of an agile governance approach.

Mandated integration of AI across government, supported by a comprehensive, system-wide approach.

Proposition 1:

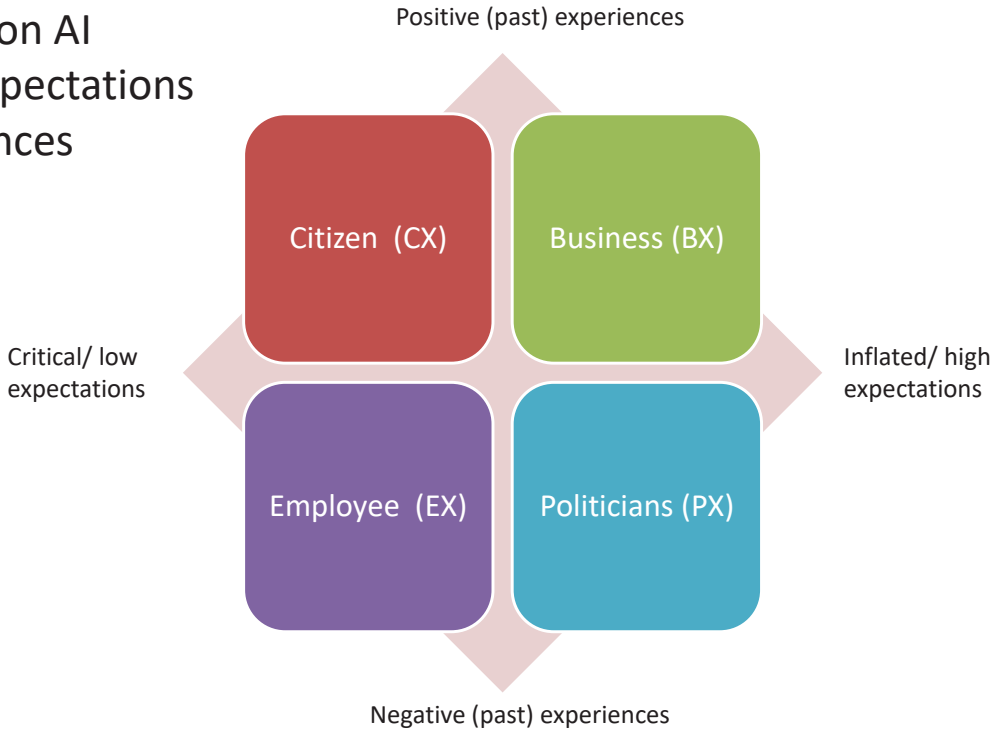
WHILE DIGITAL-ENABLED AUTOMATION AND TRANSFORMATION OF GOVERNMENT PROCESSES HAVE BEEN AN ONGOING PART OF THE GOVERNANCE AGENDA SINCE THE 1950S, THE COCKTAIL OF AI, QUANTUM, IOT, ROBOTICS, ETC., WILL REQUIRE A RADICAL SHIFT IN GOVERNANCE AND THE DEFINITION OF BOUNDARIES OF GOVERNMENT

Stagemodel of Interface Levels to Government (SMILE)



Source. Andersen, K. N., Lee, J., & Kim, S. (2024). Metaverse+ in South Korea and Denmark: Snapshots. In 25th Annual International Conference on Digital Government (DGOV). Taiwan, June.

Perspectives on AI Transformation: Expectations and Experiences



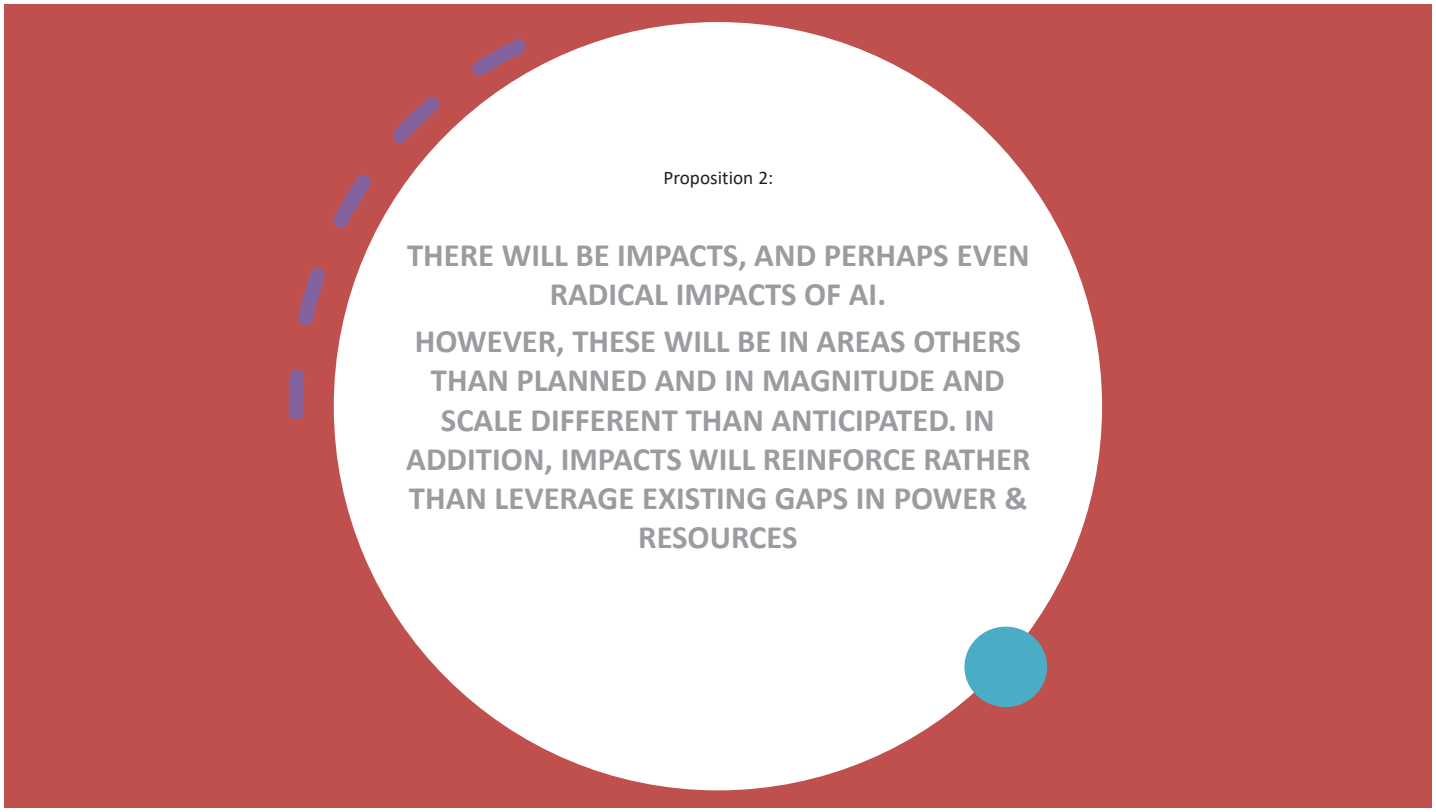
Concerns

Data Privacy and Security: Given that public sector deals with sensitive employee information, ensure that any AI tool complies with relevant data protection regulations

Ethical AI Use: Prioritize tools that focus on fairness, transparency, and inclusivity to avoid biases in AI-based decision-making (**Dandi, Entelo**: **BM Watson AI OpenScale; FairHire**)

Integration with Existing Systems: Ensure that AI tools can integrate with existing platforms and software for smoother transitions and workflows.

Organizational change management: cultural issues,
mindset, rules, behavioral factors
at individual, teams, and organizational level



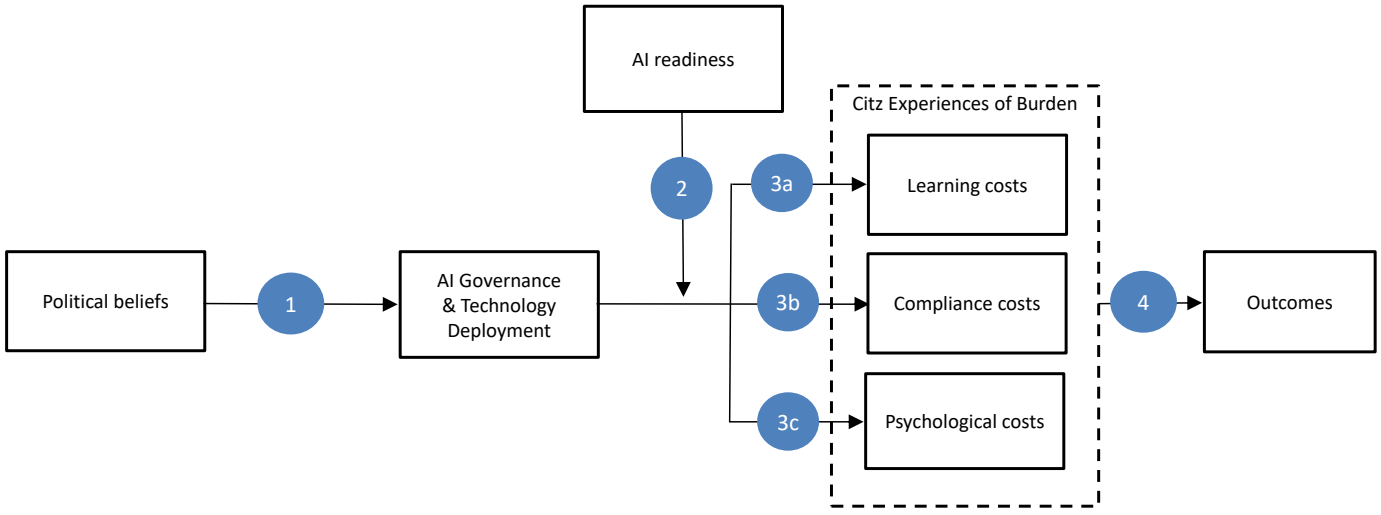
Waves of transformation

- Inhouse adoption (Fiscal impact budgetting systems, physical planning, demographic forecasting, word processing, DSS, automation)
- Data exchanges with other government institutions
- Asynchronous selfservices to business and citizens
- Synchronous, interactive services & e-democracy/ participation/voting
- Digital transformation (DX)
- AI Transformation (AX)

AI Transformation Trivials

- Implementing AI tools such as ChatGPT, Copilot, or generative AI in workflows
- Restructuring business strategies to leverage the potential of AI
- Training employees to use AI effectively and responsibly
- Creating new data-driven products and services based on AI

AI and Experiences of Administrative Burden

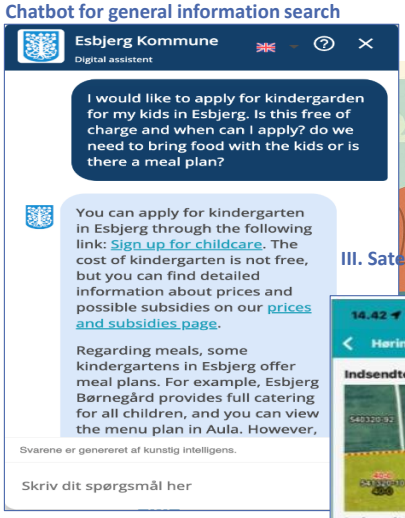


Page 12

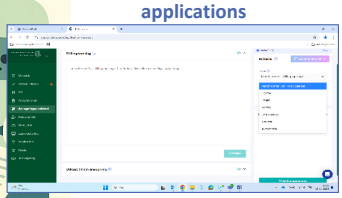
FOUR CASES ON AI AND EXPERIENCED ADMINISTRATIVE BURDEN

FOUR CASES ON AI AND EXPERIENCED ADMINISTRATIVE BURDEN

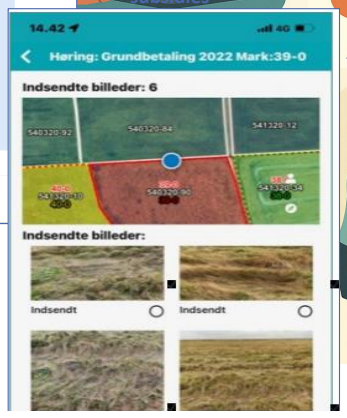
I. Chatbot for general information search



II. Automated drafting of job seekers' applications

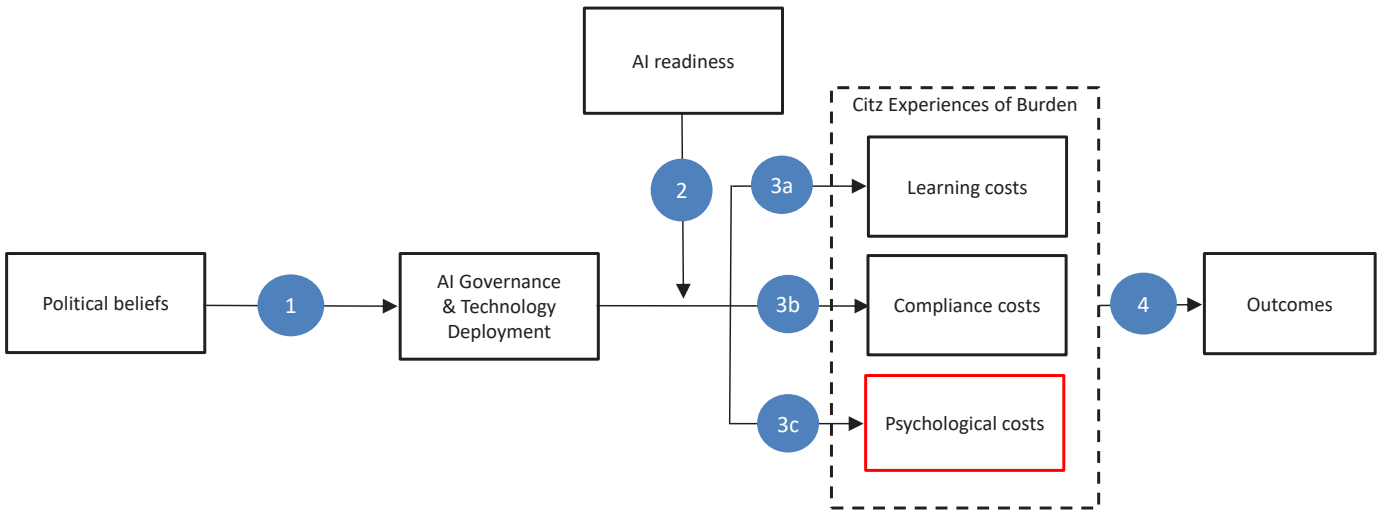


III. Satellite monitoring agricultural subsidies



IV. Behavioural training for citizens with profound social disabilities

AI and Experiences of Administrative Burden



Page 15

Proposition 3:

THE BLEND OF HARD REGULATION AND SOFT GOVERNANCE IS UNDER FIRE. ENFORCED USE OF AI IN GOVERNMENT AND A GOVERNMENT WIDE-AND-DEEP APPROACH IS THE WAY FORWARD.

Classical Governance Themes

Ethical Principles: Ensuring AI systems are fair, transparent, and free from bias.

Data Privacy and Security: Protecting personal and sensitive data used in AI systems.

Accountability: Defining responsibilities for AI-related outcomes.

Compliance and Regulation: Adhering to laws and standards related to AI use.

Risk Management: Identifying and mitigating potential risks associated with AI.

Transparency: Making AI processes and decision-making understandable to stakeholders.

Human Oversight: Ensuring human intervention and control over AI systems when necessary.

Externalities: Identifying risk on the global climate impacts and find ways to factor these impacts in use of AI

Adoption rates

- Slow adoption of "real" AI in government
- Companies are lagging far behind the policy ambitions

- Governance responses are limited and difficult to base on research-based knowledge

Example III: "AI Robotics and the Revitalization of HR"

Opportunity to solve multi-faceted problem

Demand-side

Shrinking workforce

Cost-cutting

Innovation services (availability, content, scope, format)

Maximize the return of investment in the "new technologies"

Enhancing Talent

Global Competitiveness

Liase with industry partners

System integration

Mitigate the downside

Ethical dilemmas

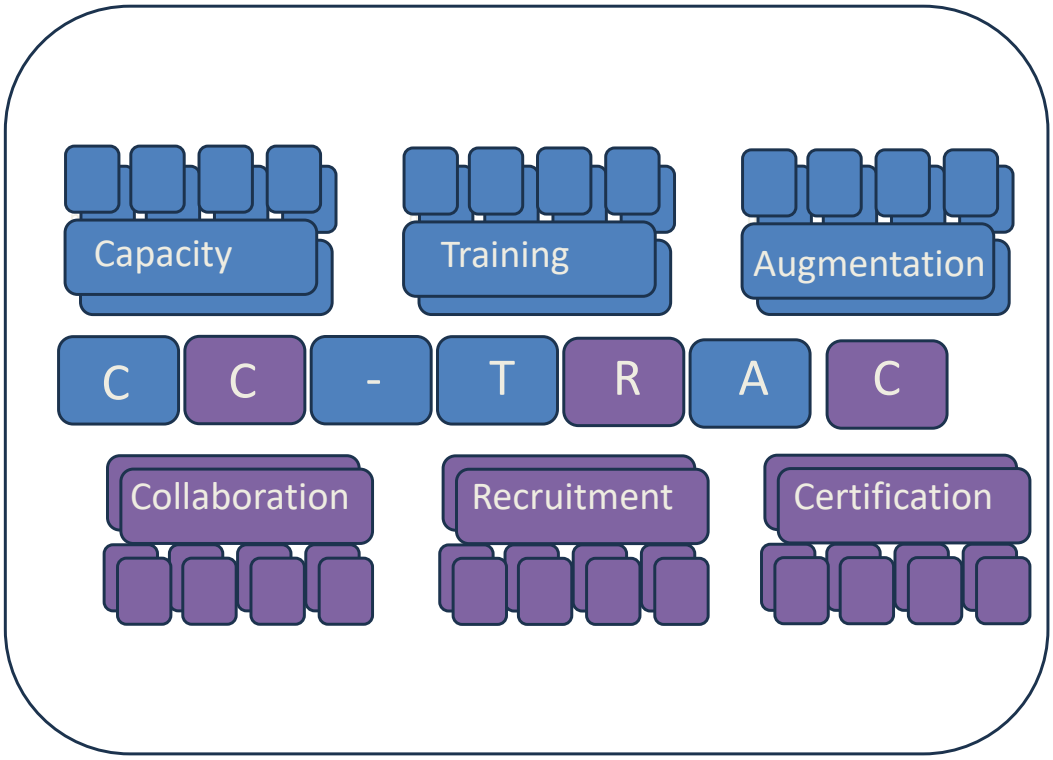
Legal challenges

Security

The Challenge of Predicting Impact of AI on HR

- Technology
- Structures
- Tasks
- People

KakaoWork, SAP, Slack, Classum, Pymetrics



Summary Three propositions

1. Although governments have pursued digital automation and transformation since the 1950s, the convergence of emerging technologies—particularly the growing dominance of AI alongside quantum computing, IoT, and robotics—demands a fundamental rethinking of governance structures and the traditional boundaries of government roles.
2. AI is likely to produce significant—and potentially transformative—impacts, though often in unexpected areas and at scales that differ from initial predictions. Moreover, rather than closing existing gaps in power and resources, AI may amplify and entrench them. This will call for a more agile governance approach.
3. The current mix of hard regulation and soft governance is facing increasing scrutiny. A more effective path forward may involve the mandated integration of AI across government, supported by a comprehensive, system-wide approach.

Summary of the three propositions

1. Rethinking and repositioning of governance structures and the traditional boundaries of government roles.

2. Deployment of an agile governance approach.

3. Mandated integration of AI across government, supported by a comprehensive, system-wide approach.

THANK YOU!

Kim Normann Andersen

Professor, Head of Studies (IT and communication graduate programs)
Department of Digitalization, Copenhagen Business School, Denmark

Whats App: +45 24794328 E-mail: andersen@cbs.dk

25

session

1

AI and Human Capital Accumulation: Aggregate and Distributional Implications*

Yang K. Lu[1] and Eunseong Ma[2]

[1]HKUST

[2]Yonsei

May 7, 2025

Abstract

This paper develops a model to analyze the effects of AI advancements on human capital investment and their impact on aggregate and distributional outcomes in the economy. We construct an incomplete markets economy with endogenous asset accumulation and general equilibrium, where households decide on human capital investment and labor supply. Anticipating near-term AI advancements that will alter skill premiums, we analyze the transition dynamics toward a new steady state. Our findings reveal that human capital responses to AI amplify its positive effects on aggregate output and consumption, mitigate the AI-induced rise in precautionary savings, and stabilize the adjustments in wages and asset returns. Furthermore, while AI-driven human capital adjustments increase inequalities in income, earnings, and consumption, they unexpectedly reduce wealth inequality.

Keywords: AI, Job Polarization, Human Capital, Inequality

*Author emails: yanglu@ust.hk; masilver@yonsei.ac.kr

1 Introduction

The distinctive nature of AI advancements lies in their ability to perform cognitive, non-routine tasks that previously required significant education and expertise, fundamentally differentiating its impact on the labor market and economy from that of general automation. For example, AI tools in medical diagnostics now assist radiologists in analyzing medical images, potentially reducing demand for entry-level radiologists while simultaneously increasing the productivity of senior professionals. More generally, AI could shift the premium associated with various skills levels, devaluing middle-level skills while increasing the demand for high-level expertise. In anticipation of these changes, households are likely to adjust their human capital investments.

According to the National Center for Education Statistic,[1] college enrollment in the U.S. has been declining since 2010. The National Student Clearinghouse Research Center reports that the undergraduate college enrollment decline has accelerated since the pandemic began, resulting in a loss of almost 6% of total enrollment between fall 2019 to fall 2023, while graduate enrollment has risen by about 5%.[2] These shifts, regardless of their causes, highlight evolving patterns in human capital investment.

This paper develops a model to study the effects of AI advancements on human capital investment and their subsequent impact on aggregate and distributional outcomes of the economy. We posit an economy consisting of three sectors, requiring low, middle and high levels of skill (human capital) with increasing sectoral labor productivity. Households can invest in their human capital to move up to more productive sectors. But if they do not invest, their human capital depreciates and, over time, they will move down to less productive sectors. We model human capital investment at two levels, a low level attainable on the job and a high level requiring full-time commitment, such as pursuing higher education. Households are subject to uninsurable idiosyncratic risk in terms of productivity shocks that affect both labor productivity and effectiveness in human capital investment.

The interaction between human capital investment and labor supply presents a tradeoff at the household level between current wage earning and future wage gains. At aggregate level, the interaction implies that when individuals transition from the middle to the high sector, they may temporarily exit the workforce to upskill, reducing immediate labor supply but improving future labor productivity.

We model AI advancements as increasing the productivity for the low and high sectors but not for the middle sector so that the skill premium of the middle sector decreases and the skill premium of the high sector increases. Allowing for human

[1] https://nces.ed.gov/programs/digest/d22/tables/dt22_303.70.asp

[2] https://public.tableau.com/app/profile/researchcenter/viz/CTEEFall2023dashboard/CTEEFall2023

capital adjustments not only alters AI's economic implications quantitatively, it also makes a qualitative difference.

If the skill distribution is fixed, AI will unambiguously improve the labor productivity of the whole economy. However, allowing human capital to adjust enables workers to upskill or downskill. The response of overall labor productivity could be enhanced, or dampened, or even reverted depending on whether workers move to more or less productive sectors.

Using a two-period model, we show how households' labor supply and human capital investment are affected by their productivity shocks, asset holdings and stocks of human capital. The effects of AI, in this partial equilibrium analysis, are shown to discourage human capital investment for households in the low sector and encourage human capital investment for households in the middle sector, thereby increasing human capital inequality. In addition, AI worsens consumption inequality for households with low levels of human capital and reduces consumption inequality for those with high levels of human capital.

At the economy level, the effects of AI advancements depend on the sectoral distribution of households and the general equilibrium effects via wage and capital return responses. We quantify these effects using a fully-fledged dynamic quantitative model that incorporates an infinite horizon, endogenous asset accumulation, and general equilibrium. The model is calibrated to reflect key features of the U.S. economy, capturing realistic household heterogeneity. The steady state distribution of human capital without AI advancements pins down the sectoral distribution of households. We then introduce fully anticipated AI advancements happening in the near future and study the transition dynamics from the current state of the economy to the eventual new steady state.

We find that aggregate human capital rises sharply even before AI introduction, indicating that a substantial portion of workers, anticipating changes in skill premium, leave the labor force early to accumulate human capital. The economy also experiences AI-induced job polarization, with a notable reallocation of workers from the middle sector to either low or high sectors.

Building on these labor dynamics, our model examines how AI influences both the aggregate and distributional outcomes of the economy, including output, consumption, investment, employment, income inequality, consumption inequality, and wealth inequality. Our focus is on how human capital adjustments reshape AI's effects on each of these outcomes. Specifically, we examine two primary channels through which human capital adjustments operate: the redistribution channel, which reallocates workers across skill sectors, and the general equilibrium channel, which operates through wages and capital return changes.

Our findings reveal that human capital responses to AI amplify its positive effects on aggregate output and consumption, mitigate the AI-induced rise in precautionary

savings, and stabilize the adjustments in wages and asset returns. Furthermore, while AI-driven human capital adjustments increase inequalities in income, earnings, and consumption, they unexpectedly reduce wealth inequality. We also show that the redistribution channel is the dominant factor in the effects of human capital adjustments, whereas the general equilibrium channel, via wage and capital return changes, plays a comparatively minor role.

This paper relates to the literature examining how technological advancements, including AI, have significantly contributed to job polarization. Goos and Manning (2007) show that since 1975, the United Kingdom has experienced job polarization, with increasing employment shares in both high- and low-wage occupations. Autor and Dorn (2013) expanded on this by providing a unified analysis of the growth of low-skill service occupations, highlighting key factors that amplify polarization in the U.S. labor market. Empirical evidence from Goos *et al.*, (2014) further confirms pervasive job polarization across 16 advanced Western European economies. In the U.S., Acemoglu and Restrepo (2020) show that robots can reduce employment and wages, finding robust negative effects of automation on both in various commuting zones.

The introduction of AI and robotics has had adverse effects on labor markets, with significant implications for employment and labor force participation. Lerch (2021) highlights that the increasing use of robots not only displaces workers but also negatively impacts overall labor force participation rates. Similarly, Faber *et al.*, (2022) demonstrate that the detrimental effects of robots on the labor market have resulted in a decline in job opportunities, particularly in sectors where automation is prevalent. These findings suggest that while technological advancements bring productivity gains, they simultaneously reduce employment prospects and participation in the labor market, exacerbating economic challenges for certain groups of workers.

The introduction of AI and robotics also influences human capital accumulation as workers respond to technological disruption. Faced with the employment risks brought about by automation, many exposed workers may invest in additional education as a form of self-insurance, rather than relying on increases in the college wage premium (Atkin, 2016; Beaudry *et al.*, 2016). Empirical evidence supports this response. Di Giacomo and Lerch (2023) find that for every additional robot adopted in U.S. local labor markets between 1993 and 2007, four individuals enrolled in college, particularly in community colleges, indicating a rise in educational investments triggered by automation. Similarly, Dauth *et al.*, (2021) show that within German firms, robot adoption has led to an increase in the share of college-educated workers, as firms prioritize higher-skilled employees over those with apprenticeships.

The response of human capital accumulation to technological disruption could also go to the other extreme. A 2022 report by Higher Education Strategy Associates

finds that following decades of growth, dropping student enrollment has become a major trend in higher education in the Global North.[3] In the U.S., the public across the political spectrum has increasingly lost confidence in the economic benefits of a college degree. Pew Research Center reports that about half of Americans say having a college degree is less important today than it was 20 years ago in a survey conducted in 2023.[4] A 2022 study from Public Agenda, a nonpartisan research organization, shows that young Americans without college degrees are most skeptical about the value of higher education.

The rise of AI and automation also plays a significant role in exacerbating general inequality, particularly through its impact on education and wealth distribution. Prettner and Strulik (2020) present a model showing that innovation-driven growth leads to an increasing proportion of college graduates, which in turn drives higher income and wealth inequality. As technology advances, workers with higher educational attainment benefit disproportionately, widening the gap between those with and without advanced skills. Sachs and Kotlikoff (2012) also explore this dynamic, providing a model within an overlapping generations framework that examines the interaction between automation and education. They demonstrate how automation can further entrench inequality by favoring workers with higher levels of education, as those without adequate skills are more likely to be displaced or see their wages stagnate. This interaction between technological change and educational attainment not only amplifies economic inequality but also perpetuates disparities in wealth across generations.

The rest of the paper is organized as follows. Section 2 describes the model environment. Section 3 solves the household's problem using a two-period version of the model. Section 4 solves the fully-fledged quantitative model and calibrates it to fit key features of the U.S. economy, including employment rate, human capital investment, and household heterogeneity. Section 5 incorporates AI into the quantitative model and examines its economic impact on both aggregate and distributional outcomes. Section 6 analyzes how human capital adjustments change the economic impact of AI advancements. Section 7 concludes.

2 Model Environment

Time is discrete and infinite. There is a continuum of households. Each household is endowed with one unit of indivisible labor and faces idiosyncratic productivity shock, z, that follows an AR(1) process in logs:

$$\ln z' = \rho_z \ln z + \varepsilon_z, \varepsilon_z \overset{\text{iid}}{\sim} N(0, \sigma_z^2) \tag{1}$$

[3] https://higheredstrategy.com/world-higher-education-institutions-students-and-funding/

[4] https://www.pewresearch.org/social-trends/2024/05/23/public-views-on-the-value-of-a-college-degree/

The asset market is incomplete following Aiyagari (1994), and the physical capital, a, is the only asset available to households to insure against this idiosyncratic risk. Households can also invest in human capital, h, which allows them to work in sectors with different human capital requirement.

2.1 Production Technology

The production technology in the economy is a constant-returns-to-scale Cobb-Douglas production function:

$$F(K, L) = K^{1-\alpha} L^{\alpha} \tag{2}$$

K is the aggregation of all physical capital held by the households. L is the aggregation of effective labor supplied by the households and employed in three sectors: low, middle, and high.

These sectors differ in their technologies for converting labor into effective labor units and in the levels of human capital required for employment. The middle sector employs households with human capital above h_M and converts one unit of labor to one effective labor unit. The high sector, requiring human capital above h_H, converts one unit of labor to $1 + \lambda$ effective units, while the low sector, with no human capital requirement, converts one unit into $1 - \lambda$ effective units. This implies a sectoral labor productivity $x(h)$ that is a step function in human capital:

$$x(h) = \begin{cases} 1 - \lambda & \text{low sector if } h < h_M \\ 1 & \text{middle sector if } h_M < h < h_H \\ 1 + \lambda & \text{high sector if } h > h_H \end{cases} \tag{3}$$

A household i who decides to work thus contributes $z_i x(h_i)$ units of effective labor, where z_i is his idiosyncratic productivity. Denote $n_i \in \{0, 1\}$ as the indicator that takes one if the household works and zero if the household does not. The aggregate labor is

$$L = \int n_i z_i x(h_i) di, \tag{4}$$

assuming perfect substitutability of effective labor across the three sectors.

2.2 Household's Problem

Households derive utility from consumption, incur disutility from labor and effort of human capital investment. A household maximizes the expected lifetime utility by optimally choosing consumption, saving, labor supply and human capital investment

each period, based on his idiosyncratic productivity shock z_t:

$$\max_{\{c_t, a_{t+1}, n_t, e_t\}_{t=0}^{\infty}} E_0 \left[\sum_{t=0}^{\infty} \beta^t (\ln c_t - \chi_n n_t - \chi_e e_t) \right] \tag{5}$$

where c_t represents consumption, a_{t+1} represents saving, $n_t \in \{0, 1\}$ is labor supply, and e_t is the effort of human capital investment.

If a household decides to work in period t, he will be employed into the appropriate sector according to his human capital h_t and receive labor income $w_t z_t x(h_t)$, where w_t is the economy-wide wage rate of effective labor unit.

Denote r_t as the interest rate on the physical capital a_t. The household's budget constraint is

$$c_t + a_{t+1} = n_t(w_t z_t x(h_t)) + (1 + r_t)a_t \tag{6}$$

$$c_t \geq 0 \text{ and } a_{t+1} \geq 0 \tag{7}$$

We prohibit households from borrowing $a_{t+1} \geq 0$ to simplify analysis.[5]

Human capital investment can take three levels of effort: $\{0, e_L, e_H\}$. A non-working household is free to choose any of the three effort levels but a working household cannot devote the highest level of effort e_H, reflecting a trade-off between working and human capital investment. Hence:

$$e_t \in \{0, e_L, (1 - n_t)e_H\}. \tag{8}$$

Its contribution to next-period human capital is subject to the productivity shock:

$$h_{t+1} = z_t e_t + (1 - \delta)h_t \tag{9}$$

where δ is human capital's depreciation rate.

3 Household Decisions in a Two-Period Model

In this section, we solve the household's problem with two periods to gain intuition.

3.1 Period-2 Decisions

Households do not invest in human capital or physical capital in the last period. The only relevant decision is whether to work.

[5] According to Aiyagari (1994), a borrowing constraint is necessarily implied by present value budget balance and nonnegativity of consumption. Since the borrowing limit is not essential to our analysis, we set it to zero for simplicity.

The household works $n = 1$ if and only if $z \geq \overline{z}(h, a)$, with $\overline{z}(h, a)$ defined as

$$\ln(w\overline{z}(h, a)x(h) + (1 + r)a) - \chi_n = \ln((1 + r)a) \tag{10}$$

The left-hand-side is the utility from working and the right-hand-side is the utility from not working.

Using the sector-specific productivity $x(h)$ specified in (3), the cutoff of idiosyncratic productivity $\overline{z}(h, a)$ takes three possible values given the capital holding a:

$$\overline{z}(h, a) = \begin{cases} \overline{z}(a)\frac{1}{1-\lambda} & \text{if } h < h_M \\ \overline{z}(a) & \text{if } h_M \leq h < h_H \\ \overline{z}(a)\frac{1}{1+\lambda} & \text{if } h > h_H \end{cases} \tag{11}$$

$$\text{where } \overline{z}(a) := \frac{(\exp(\chi_n) - 1)(1 + r)a}{w} \tag{12}$$

Households with higher human capital is more likely to work, whereas households with higher physical capital is less likely to work.

In addition to labor supply, period-1 decisions include saving and human capital investment, both of which are forward-looking and affected by the idiosyncratic risk associated with the productivity shock z'. Our model also features a trade-off between human capital investment and labor supply as a working household cannot devote the highest level of effort e_H in human capital investment. Therefore, human capital investment grants households the possibility of a discrete wage hike in the future but may entail a wage loss in the current period.

To see the implication of this trade-off and how it interacts with uninsured idiosyncratic risk, we proceed in two steps. We first derive the period-1 decisions without uncertainty by assuming that z' is known to the household at period 1 and z' is such that the household will work in period 2. We then reintroduce uncertainty in z' and compare the decision rules with the case without uncertainty.

3.2 Period-1 Consumption and Saving

The additive separability of household's utility implies that labor supply n and human capital investment e enters in consumption and saving choices only via the intertemporal budget constraint:

$$c + \frac{c'}{1 + r'} = (1 + r)a + n(wzx(h)) + \frac{w'z'x(h')}{1 + r'}$$
$$\text{with } h' = ze + (1 - \delta)h.$$

The log utility in consumption implies the optimality condition:

$$c' = \beta(1 + r')c. \tag{13}$$

Combining it with the budget constraint, we obtain the optimal consumption as a function of labor supply n and human capital investment e:

$$c(n,e) = \frac{1}{1+\beta}\left[(1+r)a + n(wzx(h)) + \frac{w'z'x\left(h' = ze + (1-\delta)h\right)}{1+r'}\right]. \tag{14}$$

3.3 Period-1 Labor Supply and Human Capital Investment

The optimal consumption conditions (14) and (13) yield a convenient objective function for the households to optimize by choosing their labor supply n and human capital investment e:[6]

$$\max_{n,e}(1+\beta)\ln c(n,e) - \chi_n n - \chi_e e \tag{15}$$

It is useful to partition households according to their human capital into three ranges: $h < h_M(1-\delta)^{-1}$, $h_M(1-\delta)^{-1} \leq h < h_H(1-\delta)^{-1}$, and $h \geq h_H(1-\delta)^{-1}$. We derive the decision rules for households with $h < h_M(1-\delta)^{-1}$ in details, as households in the other two ranges have similar decision rules.

For households with $h < h_M(1-\delta)^{-1}$, we define two cutoffs in z:

$$\underline{z}_M(h) := \frac{h_M - (1-\delta)h}{e_H}; \overline{z}_M(h) := \frac{h_M - (1-\delta)h}{e_L} \tag{16}$$

These cutoffs divide households into three groups based on their ability to be employed in the middle sector in the next period.

The non-learners are households with $z < \underline{z}_M(h)$. They cannot achieve $h' > h_M$ with either e_L or e_H level of human capital investment today. As a result, they will choose not to invest in human capital, $e = 0$, and their future sectoral productivity will be $x(h') = 1 - \lambda$.

These non-learners work $n = 1$ if and only if $z \geq \overline{z}_{non}(h,a)$, with $\overline{z}_{non}(h,a)$ taking two possible values given the capital holding a:

$$\overline{z}_{non}(h,a) = \begin{cases} \overline{z}^L_{non}(a)\frac{1}{1-\lambda} & \text{if } h < h_M \\ \overline{z}^L_{non}(a) & \text{if } h_M \leq h < h_M\frac{1}{1-\delta} \end{cases} \tag{17}$$

$$\text{where } \overline{z}^L_{non}(a) := \frac{(\exp(\frac{\chi_n}{1+\beta}) - 1)[(1+r)a + \frac{w'z'(1-\lambda)}{1+r'}]}{w} \tag{18}$$

The slow learners are households with $z \in (\underline{z}_M(h), \overline{z}_M(h))$. They can achieve $h' > h_M$ in the next period only if they invest $e = e_H$ today. Households' choices are between $e = 0$ and $e = e_H$, because choosing $e = e_L$ will only entail utility cost but bring no future benefit.

[6] This is because $c' = \beta(1+r')c$, so that $\ln c' = \ln\beta + \ln(1+r') + \ln c$.

The slow learners face the trade-off between working and human capital investment: choosing $e = e_H$ implies no working today $n = 0$. Alternatively, they can choose to work but not to invest in human capital $(n = 1, e = 0)$.[7]

The slow learners prefer $(n = 1, e = 0)$ to $(n = 0, e = e_H)$ if and only if $z \geq \overline{z}_{slow}(h, a)$, with $\overline{z}_{slow}(h, a)$ taking two possible values given the capital holding a:

$$\overline{z}_{slow}(h,a) = \begin{cases} \overline{z}^L_{slow}(a)\frac{1}{1-\lambda} & \text{if } h < h_M \\ \overline{z}^L_{slow}(a) & \text{if } h_M \leq h < h_M\frac{1}{1-\delta} \end{cases} \tag{19}$$

$$\text{where } \overline{z}^L_{slow}(a) := \frac{(\exp(\frac{\chi_n - \chi_e e_H}{1+\beta}) - 1)[(1+r)a + \frac{w'z'}{1+r'}] + \lambda\frac{w'z'}{1+r'}}{w} \tag{20}$$

The fast learners are households with $z > \overline{z}_M(h)$. They can achieve $h' > h_M$ in the next period if they invest $e = e_L$ today. In this case, there is no need to exert high effort e_H in human capital investment. The fast learners choose among three options: $(n = 1, e = 0)$, $(n = 1, e = e_L)$, and $(n = 0, e = e_L)$.[8]

The fast learners prefers $(n = 1, e = 0)$ to $(n = 1, e = e_L)$ if and only if $z \geq \overline{z}_{fast}(h, a)$, where

$$\overline{z}_{fast}(h,a) = \begin{cases} \overline{z}^L_{fast}(a)\frac{1}{1-\lambda} & \text{if } h < h_M \\ \overline{z}^L_{fast}(a) & \text{if } h_M \leq h < h_M\frac{1}{1-\delta} \end{cases} \tag{21}$$

$$\text{and } \overline{z}^L_{fast}(a) := \frac{\left\{\exp(\frac{\chi_e e_L}{1+\beta})\lambda\left[\exp(\frac{\chi_e e_L}{1+\beta}) - 1\right]^{-1} - 1\right\}\frac{w'z'}{1+r'} - (1+r)a}{w} \tag{22}$$

The fast learners prefers $(n = 1, e = e_L)$ to $(n = 0, e = e_L)$ if and only if $z \geq \underline{z}_{fast}(h, a)$, where

$$\underline{z}_{fast}(h,a) = \begin{cases} \underline{z}^L_{fast}(a)\frac{1}{1-\lambda} & \text{if } h < h_M \\ \underline{z}^L_{fast}(a) & \text{if } h_M \leq h < h_M\frac{1}{1-\delta} \end{cases} \tag{23}$$

$$\text{and } \underline{z}^L_{fast}(a) := \frac{(\exp(\frac{\chi_n}{1+\beta}) - 1)[(1+r)a + \frac{w'z'}{1+r'}]}{w} \tag{24}$$

We set up our model so that $\overline{z}^L_{fast}(a) > \underline{z}^L_{fast}(a)$.[9] The decision rule for the fast

[7] The choice between $(n = 0, e = e_H)$ and $(n = 0, e = 0)$ does not depend on z. To make e_H relevant, λ needs to be large enough so that $(n = 0, e = e_H)$ dominates $(n = 0, e = 0)$. See Appendix for the details on the lower bound of λ.

[8] Similar to the case of slow learners, the choice between $(n = 0, e = e_L)$ and $(n = 0, e = 0)$ does not depend on z. Moreover, since our model is set up so that $(n = 0, e = e_H)$ dominates $(n = 0, e = 0)$, it implies that $(n = 0, e = e_L)$ dominates $(n = 0, e = 0)$.

[9] Appendix provides the parameter restrictions such that the condition for $(n = 0, e = e_H)$ to dominate $(n = 0, e = 0)$ is sufficient for $\overline{z}^L_{fast}(a) > \underline{z}^L_{fast}(a)$.

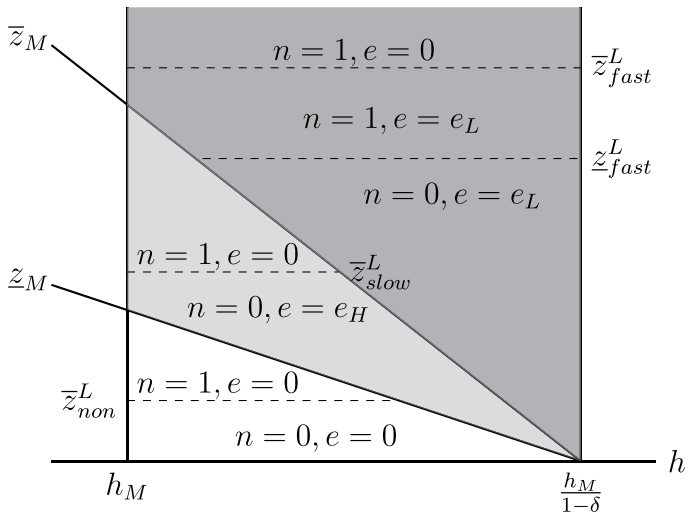


Figure 1: Decision Rule Diagram for $h_M \leq h < h_M(1-\delta)^{-1}$

The human capital h changes along the horizontal line and the idiosyncratic productivity z changes along the vertical line. The two diagonal lines, $\overline{z}_M(h)$ and $\underline{z}_M(h)$, separate the state space into three areas: the unshaded area represents the non-learners, the lightly-shaded area represents the slow learners, and the darkly-shaded area represents the fast learners. The areas are divided by four dashed horizontal lines associated with cutoffs $\overline{z}^L_{non}$, $\overline{z}^L_{slow}$, $\underline{z}^L_{fast}$, and $\overline{z}^L_{fast}$ that are functions of capital holding a.

learners are as follows:

$$n(z,h,a), e(z,h,a) = \begin{cases} n=1, e=0 & \text{if } z \geq \overline{z}_{fast}(h,a) \\ n=1, e=e_L & \text{if } \underline{z}_{fast}(h,a) \leq z < \overline{z}_{fast}(h,a) \\ n=0, e=e_L & \text{if } z < \underline{z}_{fast}(h,a) \end{cases} \tag{25}$$

Decision rule diagram for households: Figure 1 illustrates the decision rule (n,e) as a function of states (z,h,a) for households with $h_M \leq h < h_M \frac{1}{1-\delta}$. The human capital h changes along the horizontal line and the idiosyncratic productivity z changes along the vertical line. The two diagonal lines, $\overline{z}_M(h)$ and $\underline{z}_M(h)$ defined in (16), separate the state space into three areas: the unshaded area represents the non-learners, the lightly-shaded area represents the slow learners, and the darkly-shaded area represents the fast learners. The areas are divided by four dashed horizontal lines associated with cutoffs $\overline{z}^L_{non}(a)$, $\overline{z}^L_{slow}(a)$, $\underline{z}^L_{fast}(a)$, and $\overline{z}^L_{fast}(a)$ that are functions of capital holding a and defined in (18), (20), (24), and (22).

This decision rule diagram is representative for households with other levels of human capital. For households with $h < h_M$, $\overline{z}_M(h)$ and $\underline{z}_M(h)$ continue to be the boundaries that separate non-learners, slow learners and fast learners, but the four cutoffs are $\overline{z}^L_{non}\frac{1}{1-\lambda}$, $\overline{z}^L_{slow}\frac{1}{1-\lambda}$, $\underline{z}^L_{fast}\frac{1}{1-\lambda}$, and $\overline{z}^L_{fast}\frac{1}{1-\lambda}$.

For households with $h_M \frac{1}{1-\delta} \leq h < h_H \frac{1}{1-\delta}$, the boundaries for state space division

change to $\overline{z}_H(h)$ and $\underline{z}_H(h)$:

$$\underline{z}_H(h) := \frac{h_H - (1-\delta)h}{e_H}; \overline{z}_H(h) := \frac{h_H - (1-\delta)h}{e_L} \tag{26}$$

If $h_M \frac{1}{1-\delta} \leq h < h_H$, the four cutoffs for households are:[10]

$$\overline{z}^M_{non}(a) := \frac{(\exp(\frac{\chi_n}{1+\beta}) - 1)[(1+r)a + \frac{w'z'}{1+r'}]}{w} \tag{27}$$

$$\overline{z}^M_{slow}(a) := \frac{(\exp(\frac{\chi_n - \chi_e e_H}{1+\beta}) - 1)[(1+r)a + \frac{w'z'(1+\lambda)}{1+r'}] + \lambda\frac{w'z'}{1+r'}}{w} \tag{28}$$

$$\underline{z}^M_{fast}(a) := \frac{(\exp(\frac{\chi_n}{1+\beta}) - 1)[(1+r)a + \frac{w'z'(1+\lambda)}{1+r'}]}{w} \tag{29}$$

$$\overline{z}^M_{fast}(a) := \frac{\left\{\lambda\left[\exp(\frac{\chi_e e_L}{1+\beta}) - 1\right]^{-1} - 1\right\}\frac{w'z'}{1+r'} - (1+r)a}{w} \tag{30}$$

If $h_H \leq h < h_H \frac{1}{1-\delta}$, the cutoffs are $\overline{z}^M_{non}\frac{1}{1+\lambda}$, $\overline{z}^M_{slow}\frac{1}{1+\lambda}$, $\underline{z}^M_{fast}\frac{1}{1+\lambda}$, and $\overline{z}^M_{fast}\frac{1}{1+\lambda}$.

All households with $h \geq h_H \frac{1}{1-\delta}$ are non-learners because their current human capital is enough for employment in the high sector next period even without any human capital investment. The only relevant cutoff for them is $\overline{z}^H_{non}(a)\frac{1}{1+\lambda}$ where

$$\overline{z}^H_{non}(a) := \frac{(\exp(\frac{\chi_n}{1+\beta}) - 1)[(1+r)a + \frac{w'z'(1+\lambda)}{1+r'}]}{w} \tag{31}$$

3.4 Comparative Statics

The decision rules derived in the previous section imply that the fast learners invest in human capital if $z < \overline{z}_{fast}(h,a)$ and the slow learner invest in human capital if $z < \overline{z}_{slow}(h,a)$. The close form expressions of the cutoffs allow us to compare human capital investment between groups of households with different levels of human capital and physical capital.

Effect of human capital h on human capital investment:

Lemma 1 *Both the fast learners and the slow learners with $h < \frac{h_M}{1-\delta}$ invest more in human capital than their counterparts with $h > \frac{h_M}{1-\delta}$:*

$$\frac{\overline{z}^L_{fast}}{1-\lambda} > \overline{z}^M_{fast} \; ; \overline{z}^L_{fast} > \frac{\overline{z}^M_{fast}}{1+\lambda}$$
$$\frac{\overline{z}^L_{slow}}{1-\lambda} > \overline{z}^M_{slow} \; ; \overline{z}^L_{slow} > \frac{\overline{z}^M_{slow}}{1+\lambda}$$

Figure 2 provides an illustration to this proposition. The striped areas indicate the state space for positive human capital investment. The darkly-shaded areas

[10] Appendix provides parameter restrictions for $\overline{z}^M_{fast}(a) > \underline{z}^M_{fast}(a)$.

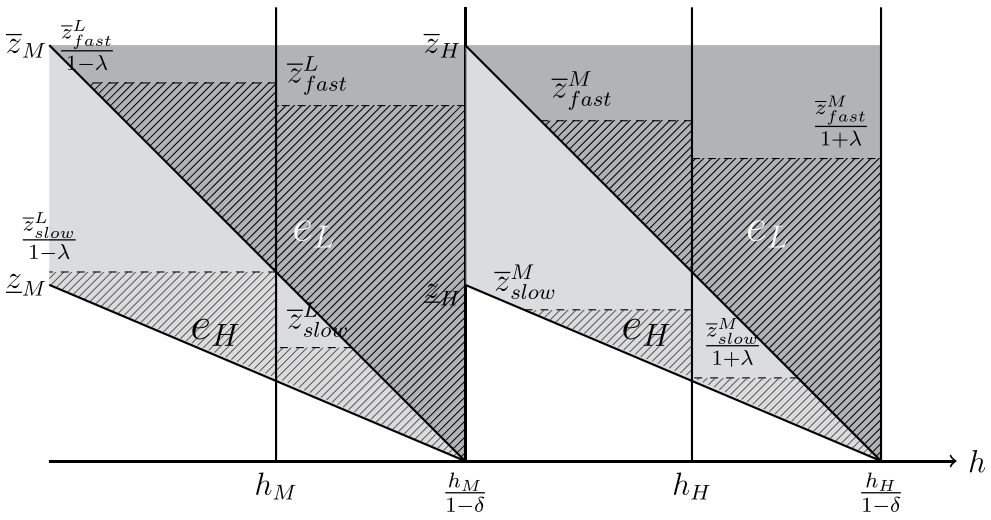


Figure 2: State Space for Human Capital Investment

The darkly-shaded striped areas indicate the state space for human capital investment equal to e_L by the fast learners. The lightly-shaded striped areas indicate the state space for human capital investment equal to e_H by the slow learners.

correspond to the fast learners. The lightly-shaded areas correspond to the slow learners. Let us take the slow learners as an example. Those with $h < \frac{h_M}{1-\delta}$ need to invest e_H to either stay in or to move up to the middle sector next period. Those with $h > \frac{h_M}{1-\delta}$ need to invest e_H to either stay in or to move up to the high sector. The most productive households does not invest in human capital because it requires giving up their labor earning. This productivity cutoff is lower for those with higher human capital, meaning that their investment in human capital is lower than those with lower human capital.

Effect of physical capital a on human capital investment:

Lemma 2 *The fast learners with lower asset holding invest more in human capital:*

$$\frac{\partial \overline{z}^L_{fast}(a)}{\partial a} < 0 \ ; \ \frac{\partial \overline{z}^M_{fast}(a)}{\partial a} < 0$$

The slow learners with lower asset holding invest more in human capital if and only if $\chi_n < \chi_e e_H$*:*

$$\frac{\partial \overline{z}^L_{slow}(a)}{\partial a} < 0 \ \textit{and} \ \frac{\partial \overline{z}^M_{slow}(a)}{\partial a} < 0 \ \textit{iff} \ \chi_n < \chi_e e_H$$

3.5 The Effects of an Anticipated Period-2 AI Shock

Suppose that an AI shock is anticipated to occur in period 2 and to increase the labor productivity for the low sector and the high sector but not the middle sector.

The effect of AI shock on the sectoral productivity is captured by γ with $0 < \gamma < 1$:

$$x(h') = \begin{cases} 1 - \lambda + \gamma\lambda & \text{low sector if } h' < h_M \\ 1 & \text{middle sector if } h_M < h' < h_H \\ 1 + \lambda + \gamma\lambda & \text{high sector if } h' > h_H \end{cases} \tag{32}$$

In other words, the AI shock increases average labor productivity, reduces the earnings premium for the middle sector, and enlarges the earnings premium for the high sector relative to the middle sector.

The non-learners: The AI shock increases the labor income of households who work in the low sector or the high sector in period 2, i.e, those with $h < h_M \frac{1}{1-\delta}$ or $h > h_H \frac{1}{1-\delta}$. The positive income effect makes them work less in period 1 so that $\overline{z}^L_{non}(a)$ and $\overline{z}^H_{non}(a)$ increases in γ:

$$\overline{z}^i_{non}(a;\gamma) = \overline{z}^i_{non}(a;\gamma=0) + \gamma\lambda \frac{w'z'}{w(1+r')} \left[\exp(\frac{\chi_n}{1+\beta}) - 1\right] \text{ for } i = L, H$$

The slow learners: The AI shock reduces the incentive to work in the middle sector in period 2, i.e., $\overline{z}^L_{slow}(a)$ is decreasing and $\overline{z}^M_{slow}(a)$ is increasing in γ:

$$\begin{aligned} \overline{z}^L_{slow}(a;\gamma) &= \overline{z}^L_{slow}(a;\gamma=0) - \gamma\lambda \frac{w'z'}{w(1+r')} \\ \overline{z}^M_{slow}(a;\gamma) &= \overline{z}^M_{slow}(a;\gamma=0) + \gamma\lambda \frac{w'z'}{w(1+r')} \exp(\frac{\chi_n - \chi_e e_H}{1+\beta}) \end{aligned}$$

Therefore, those with $h < h_M \frac{1}{1-\delta}$ invest less human capital and work more in period 1 and those with $h > h_M \frac{1}{1-\delta}$ invest more human capital and work less.

The fast learners: Similar to the slow learners, the AI shock reduces households' incentive to work in the middle sector in period 2. As a result, human capital investment is lower for those with $h < h_M \frac{1}{1-\delta}$, and is higher for those with $h > h_M \frac{1}{1-\delta}$. The effects of AI shock γ on the cutoff governing human capital investment are:

$$\begin{aligned} \overline{z}^L_{fast}(a;\gamma) &= \overline{z}^L_{fast}(a;\gamma=0) - \gamma\lambda \frac{w'z'}{w(1+r')} \frac{\exp(\frac{\chi_e e_L}{1+\beta})}{\exp(\frac{\chi_e e_L}{1+\beta}) - 1} \\ \overline{z}^M_{fast}(a;\gamma) &= \overline{z}^M_{fast}(a;\gamma=0) + \gamma\lambda \frac{w'z'}{w(1+r')} \frac{1}{\exp(\frac{\chi_e e_L}{1+\beta}) - 1} \end{aligned}$$

Conditional on the human capital investment being e_L, the fast learners' labor supply decision is affected by the AI shock via the future earning increase if the households will work in the high sector in period 2. That is, those with $h > h_M \frac{1}{1-\delta}$ work less

in period 1, i.e., $\underline{z}^L_{fast}$ increases in γ:

$$\underline{z}^M_{fast}(a;\gamma) = \underline{z}^M_{fast}(a;\gamma=0) + \gamma\lambda\frac{w'z'}{w(1+r')}\left[\exp(\frac{\chi_n}{1+\beta}) - 1\right]$$

3.5.1 AI effect on human capital inequality

Recall from Lemma 1 that, without the AI shock, households with low h invest more in human capital than households with high h. The analysis above shows that the AI shock discourages human capital investment for those with low h but encourages it for those with high h. Therefore, a small AI shock reduces human capital investment disparity between groups with different levels of h, and a large AI shock could lead to a reversal in the comparison, making households with high h invest more in human capital than households with low h. The human capital distribution will be more unequal due to the AI shock.

Proposition 1 *AI shock increases human capital inequality.*

3.5.2 AI effect on consumption inequality

According to the optimal consumption rule in (13) and (14), consumption is proportional to the present value of household incomes in two periods. The AI shock increases the period-2 labor income of the low and high sectors, and in turn increases the consumption of households who would have worked in the low or the high sector in period 2 without the AI shock.

For households with $h < h_M\frac{1}{1-\delta}$, the affected groups are those whose human capital investment would be zero without the AI shock. In Figure 2, they are the unstriped areas to the left of the vertical line $\frac{h_M}{1-\delta}$. Within the fast learners, it is the households with higher current z that are affected by the AI shock and have their consumption increased. Since higher current z is associated with a higher consumption, the AI shock increases the consumption inequality within the fast learners. The same argument applies to the slow learners.

By contrast, the affected groups for households with $h_M\frac{1}{1-\delta} < h < h_H\frac{1}{1-\delta}$ are those whose human capital investment would be positive without the AI shock. In Figure 2, they are the striped areas to the right of the vertical line $\frac{h_M}{1-\delta}$. Within the fast learners, the AI shock increases the consumption of the households with lower current z, therefore reducing the consumption inequality. The same argument applies to the slow learners.

Proposition 2 *AI shock increases consumption inequality within the fast (slow) learners of low human capital, $h < h_M\frac{1}{1-\delta}$. AI shock reduces consumption inequality within the fast (slow) learners of high human capital, $h > h_M\frac{1}{1-\delta}$.*

For the non-learners, the AI shock only affects those with $h_M \frac{1}{1-\delta} < h < h_H \frac{1}{1-\delta}$, moving their consumption closer to those with lower h and lower consumption, but away from those with higher h and higher consumption.

3.6 The Effects of Uninsured Idiosyncratic Risk

We now reintroduce the idiosyncratic risk to households in period 1 by assuming that z' follows a log-normal distribution with mean $\overline{z}'$ and variance σ_z^2.

Our previous analysis without uncertainty is a special case with $\sigma_z^2 = 0$. The effects of uninsured idiosyncratic risk can be thought as how households' decisions change when the distribution of z' undergoes a mean-preserving spread in the sense of second-order stochastic dominance.

From a consumption-saving perspective, the uncertain z' is associated with future labor income risk. It is well understood in the literature that idiosyncratic future income risk raises the expected marginal utility of future consumption for households with log utility and makes them save more. In our model, households can also supply more labor to mitigate the effect of idiosyncratic income risk on the marginal utility of consumption.

From the perspective of human capital investment, the uncertain z' is associated with risk in the return to human capital. Conditional on working, households' income increases with z': $c' = (1+r')a' + w'x(h')z'$. $\ln(c')$ is increasing and concave in z', and a higher $x(h')$ increases the concavity.[11] Consider two levels of h', $\overline{h'} > \underline{h'}$, a mean-preserving spread of z' distribution reduces the expected utility at both levels of h' but the reduction is larger for the higher level $\overline{h'}$. Hence, the expected utility gain of moving from $\underline{h'}$ to $\overline{h'}$ is smaller due to the idiosyncratic risk. Human capital investment is discouraged.

Taking into account endogenous labor supply reinforces the discouragement of human capital investment by the idiosyncratic risk. Recall from Section 3.1 that households with z' lower than a cutoff do not work. The endogenous labor supply therefore provides insurance against the lower tail risk of the idiosyncratic z'. Moreover, the cutoff in z' is lower for those with higher human capital h'. This makes households with higher h' more exposed to the lower tail risk than those with lower h', further reducing the gain of human capital investment.

[11] The marginal effect of z' on $\ln(c')$ is

$$\frac{\partial \ln(c')}{\partial z'} = \frac{w'x(h')}{(1+r')a' + w'x(h')z'} > 0$$

The second derivative is

$$\frac{\partial^2 \ln(c')}{(\partial z')^2} = -\left[\frac{w'x(h')}{(1+r')a' + w'x(h')z'}\right]^2 < 0$$

and is more negative if $x(h')$ is higher.

Table I: Parameters for the Calibration

Parameter	Value	Description	Target or Reference
β	0.9535	Time discount factor	Annual interest rate
ρ_z	0.94	Persistence of z shocks	See text
σ_z	0.287	Standard deviation of z shocks	Earnings Gini
$\underline{a}$	0	Borrowing limit	See text
χ_n	2.39	Disutility from working	Employment rate
χ_e	1.20	Disutility from HC effort	See text
$\overline{n}$	1/3	Hours worked	Average hours worked
e_H	1/3	High level of effort	Average hours worked
e_L	1/6	Low level of effort	See text
h_H	0.61	Human capital cutoff for H	See text
h_M	1.56	Human capital cutoff for M	See text
λ	0.2	Skill premium	Income Gini
α	0.36	Capital income share	Standard value
δ	0.1	Capital depreciation rate	Standard value

Proposition 3 *The uninsured idiosyncratic risk in z' makes households in period 1 save more, work more and invest less in human capital.*

Limitations to the two-period model: In the two-period model, we take the period-1 asset holding as exogenous. In the full model, the idiosyncratic risk increases households saving and leads to more asset holding. According to Lemma 2, more asset holding reduces human capital investment for the fast learners and reduces human capital investment for the slow learners if and only if $\chi_n < \chi_e e_H$.

4 A Quantitative Model

We now solve the full dynamic model with infinite horizon, endogenous asset accumulation, and general equilibrium. We calibrate the model to reflect key features of the U.S. economy, capturing reasonable household heterogeneity.

4.1 Calibration

We calibrate the model to match the U.S. economy. For several preference parameters, we adopt values commonly used in the literature. Other parameters are calibrated to align with targeted moments. The model operates on an annual time period. Table I summarizes the parameter values used in the benchmark model.

The time discount factor, β, is calibrated to match an annual interest rate of 4 percent. We set χ_n to replicate an 80 percent employment rate. We calibrate χ_e to match the fact that around 30 percent of the population invests in human capital. The borrowing limit, $\underline{a}$, is set to 0.

Table II: Key Moments

Moment	Data	Model
Employment rate	0.80	0.80
New Investment Pop ratio	0.29	0.30
Gini coefficient for wealth	0.78	0.74
Gini coefficient for earnings	0.63	0.64
Gini coefficient for income	0.57	0.59

Note: The Gini coefficient is sourced from Diaz-Gimenez *et al.*, (1997), while the new investment-to-population ratio is obtained from OECD (1998).

We calibrate parameters regarding labor productivity process as follows. We assume that x follows the AR(1) process in logs: $\log z' = \rho_z \log z + \epsilon_z$, where $\epsilon_z \sim N(0, \sigma_z^2)$. The shock process is discretized using the Tauchen (1986) method, resulting in a transition probability matrix with 9 grids. The persistence parameter $\rho_z = 0.94$ is chosen based on estimates from the literature. The standard deviation σ_z, is is chosen to ensure that the steady-state earnings Gini coefficient matches the data, which is 0.64.

We deviate from the two-period model by assuming that the labor supply is a discrete choice between 0 and $\overline{n} = 1/3$. This change only rescales the two-period model without altering the trade-off facing the households. But such rescaling facilitates the interpretation that households are deciding whether to allocate one-third of their fixed time endowment to work. The high-level human capital accumulation effort, e_H is assumed to equal $\overline{n}$. The low-level effort, e_L is set to half of e_H. The skill premium across sectors, λ, is set at 0.2 to match the income Gini coefficient. Human capital cutoffs, h_H and h_M, are set so that 40 percent of the population falls into the middle sector, with the remaining 60 percent split between other sectors.

On the production side, we set the capital income share, α, to 0.36, and the depreciation rate, δ, to 0.1.

4.2 Key Moments: Data vs. Model

In Table II, we present a comparison of key moments between the model and the empirical data. The model does an excellent job of replicating the 80% employment rate observed in the data. In this context, employment is defined as having positive labor income in the given year, consistent with the common approach used in the literature. According to OECD (1998), the share of the population investing in human capital—those who are actively engaged in skill acquisition or education—is approximately 30%, a figure well matched by the model's predictions. This is an important metric because it reflects the model's capacity to capture the dynamics of human capital formation, which plays a critical role in shaping long-run earnings and income inequality. Additionally, the model accurately captures the distribution of income and earnings, aligning closely with observed data. This suggests that the model effectively incorporates the key mechanisms driving labor market outcomes

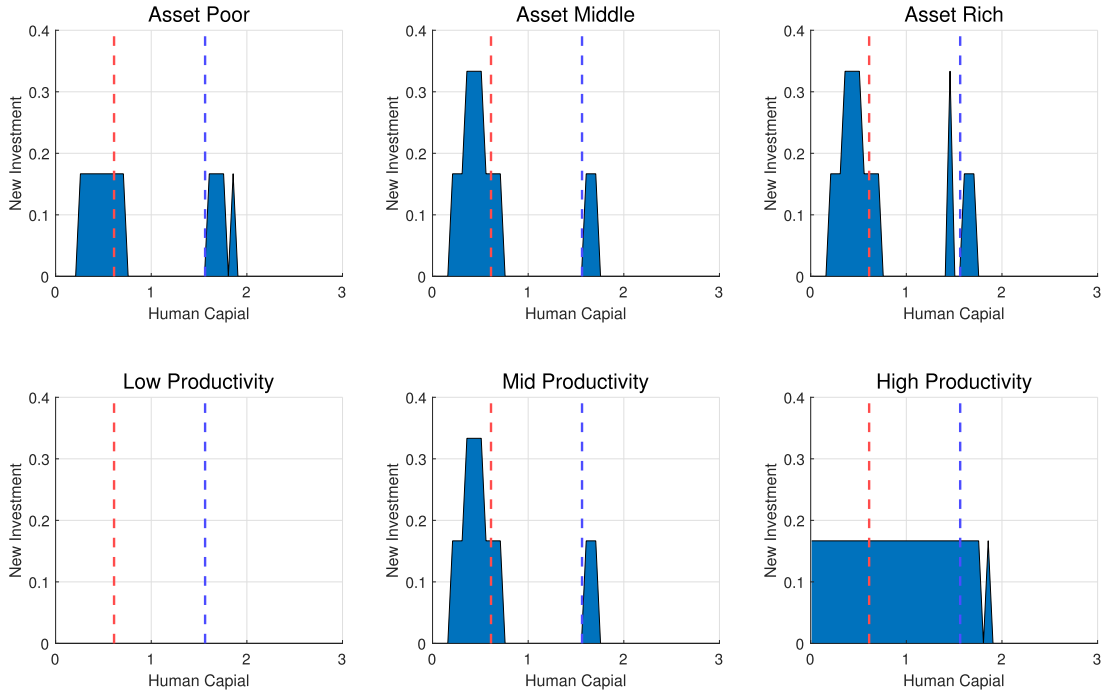


Figure 3: Human Capital Effort Decisions

Note: Human capital effort decisions by productivity and asset levels. The y-axis represents human capital effort, e_t, while the x-axis denotes the level of human capital, h_t. A red line marks the threshold between low and middle sectors, and a blue line indicates the cutoff between middle and high sectors. The upper panel displays human capital effort by asset distribution, holding average productivity constant. Specifically, households are grouped into asset-poor (with assets at half of the average), asset-middle (with assets at the average), and asset-rich (with assets at twice the average). The bottom panel illustrates human capital effort by productivity distribution, controlling for average assets. Households are categorized by productivity: low productivity (lowest productivity grid), middle productivity (average productivity), and high productivity (highest productivity grid).

and the corresponding distributional aspects of earnings. While the model does not explicitly target the wealth Gini coefficient, it performs reasonably well in approximating it. In the data, the wealth Gini is measured at 0.78, indicating significant wealth inequality in the economy. The model, however, produces a slightly lower wealth Gini of 0.74.

4.3 Decision Rules for Human Capital

Given our focus on the impact of AI introduction on human capital accumulation and inequality, it is essential to understand how workers make decisions regarding human capital investment. Figure 3 presents the policy function for human capital effort across two dimensions: the asset distribution (upper panel) and the productivity distribution (lower panel). In the figure, we mark the sectoral cutoffs with two dashed lines—red for the cutoff between the low and middle sectors, and blue for the cutoff between the middle and high sectors.

In general, workers' efforts toward human capital accumulation tend to increase around the sectoral cutoffs. What is particularly interesting is that workers in the low-income sector near the cutoff show more active human capital investment. This behavior suggests that the potential for upward mobility incentivizes greater effort in accumulating human capital. Assets also play a significant role in shaping these

decisions, as workers with more assets are better positioned to invest in human capital. By comparing the left and middle panels of the upper section in Figure 3, we see that workers in the low-income sector with fewer assets have a lower level of effort. However, workers in the same sector with higher assets are able to exert more effort in accumulating human capital. This suggests that assets act as a form of insurance, allowing workers to buffer against income fluctuations and take on greater human capital investment.

Productivity levels are crucial in determining the efficiency of human capital accumulation. According to the bottom section in Figure 3, workers with low productivity levels do not accumulate human capital, as the returns on investment are minimal. On the other hand, workers with higher productivity levels tend to invest more in human capital across most ranges of effort. However, an intriguing observation is that even these high-productivity workers do not choose the highest possible level of effort. This is likely because their returns are already sufficiently high, and the opportunity cost of being non-employed for extended periods—required for the highest levels of human capital effort—becomes too great. Overall, the model highlights the complex interactions between assets, productivity, and sectoral cutoffs in shaping workers' human capital investment decisions.

Table III shows the steady state distribution of population, employment, and assets across sectors. The population share of the middle sector is calibrated to approximately 40% by adjusting the sector cutoffs for human capital. The employment share and assets share are determined by households' labor supply and saving decisions, respectively. In the steady state, the high sector, while having the smallest population and employment shares, holds the largest share of physical capital.

Table III: Distribution of Population, Employment and Assets

Sectors	Pop. Share (%)	Emp. Share (%)	Assets Share (%)
Low	35.35	35.30	6.86
Middle	37.37	33.65	27.44
High	27.28	31.05	65.70

Note: Human capital cutoffs, h_H and h_M, determine the population share across sectors. Employment share and assets share are implied by households labor supply decisions and saving decisions.

5 The Economic Impact of AI

To analyze the economic impact of AI introduction, we focus on the transition dynamics between the current state of the economy and the eventual new steady state. We assume that AI technology will be introduced in 10 years, with households possessing full information about this forthcoming development. This setup enables us to examine the economic adjustments both in anticipation of and in response to the arrival of AI.

5.1 Introducing AI in the Quantitative Model

The effect of AI on the sectorial productivity is modeled as in (32) with $\gamma = 0.3$. That is, AI boosted the productivity of the low sector workers by 7.5% and the productivity of the high sector workers by 5%, leaving the middle sector intact. It captures the key idea that AI increases average labor productivity (Acemoglu and Restrepo, 2019), but reduces the earning premium for the middle sector, and enlarges the earning premium for the higher sector relative the middle sector.

Given the employment distribution in the initial steady state, AI is projected to increase the economy's labor productivity by 4% on average, assuming households do not alter their decisions in response. However, changes in earning premiums incentivize households to adjust their human capital investments. Households closer to acquiring high-sector human capital may intensify their efforts in human capital accumulation, transitioning into high-sector roles. Conversely, those positioned to become middle-sector workers might reduce their human capital investments, shifting to lower sectors. These adjustments not only amplify AI's positive impact on labor productivity but also contribute to job polarization.

5.2 Aggregate Implications: Anticipation and Post-AI Adjustments

5.2.1 Aggregate variables and factor prices

The transition dynamics of key aggregate variables, such as output, consumption, capital, employment, and factor prices, reflect how households optimally adjust their decisions regarding consumption, investment, labor supply, and human capital accumulation in anticipation of and in responds t AI introduction. Figure 4 illustrates these transition paths.

The introduction of AI results in a long-term increase in output, consumption, and investment, but its short-term anticipation has the opposite effect. Before the AI shock, output and investment decline, while consumption remains stable. This negative anticipation effect arises from a sharp decline in employment as some households exit the labor market to invest in human capital in preparation for the post-AI economy.[12] During this period, labor productivity remains unchanged, so the employment decline leads to a loss in output. The standard consumption-smoothing motive dictates that the reduction in output is absorbed primarily by a decline in investment. Meanwhile, the drop in employment drives up wages and reduces capital returns in the lead-up to AI implementation.

Following the introduction of AI, sectoral labor productivity increases, offsetting the negative impact of employment declines on output. Output rebounds to a higher level than the initial steady state immediately after the AI introduction and

[12] Empirical studies, such as Lerch (2021) and Faber *et al.*, (2022), support the adverse short-term effects of AI adoption on the labor market.

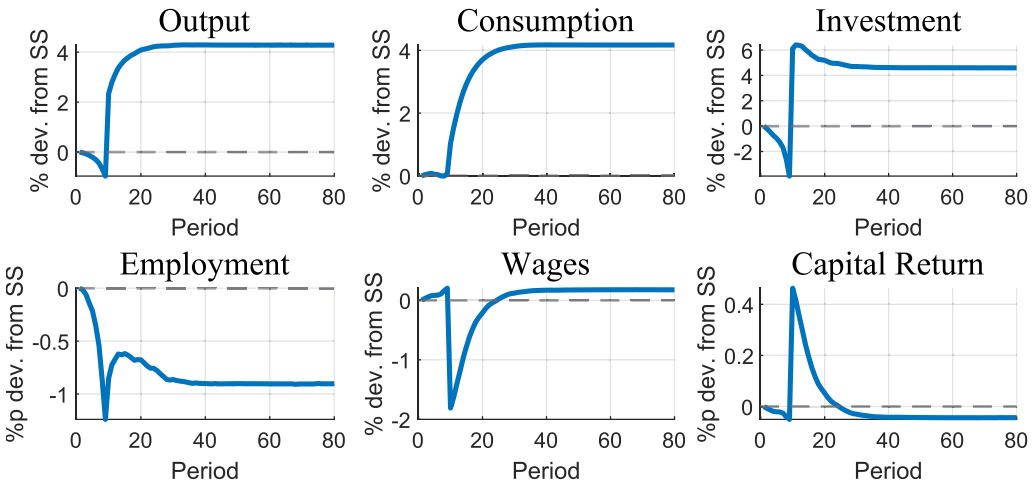


Figure 4: Transition Path: Aggregate Variables

Note: The transition paths of key aggregate variables. The x-axis represents years, and the y-axis shows the percentage (or percentage point) deviation from the initial steady state. AI introduction is assumed to occur in period 10.

continues to rise toward a new steady state. Investment follows a similar trajectory. Consumption also increases gradually over time. The higher sectoral labor productivity modeled here implies that each unit of labor supply translates into more efficient labor units. As a result, AI expands the supply of efficient labor, leading to a significant decline in the marginal product of efficient labor (wages) and a sharp increase in the marginal product of capital (capital returns).

The lower wage rate does not necessarily reduce labor earnings, as each unit of labor supplied by workers in the low and high sectors generates more efficient labor units. This explains why employment rebounds after AI implementation, even though it never fully returns to its initial steady-state level.

In the new steady state, output, consumption, and investment are around 4% higher than their initial steady-state levels, while employment is 1% lower. Wages and capital returns converge to levels similar to the initial steady state, as they are largely determined by households' preferences.

5.2.2 Human capital accumulation

A key focus of this paper is to analyze how the introduction of AI influences human capital accumulation. Given that sector-specific technologies demand varying levels of human capital and that AI alters the earning premiums in the middle and high sectors, understanding this mechanism is essential for assessing the broader implications of AI on both aggregate and sectoral economies.

Figure 5 illustrates the transition dynamics of aggregate human capital, which ultimately increases by approximately 2%. The figure also depicts its extensive margin (the share of households investing in human capital) and intensive margin (average human capital per household among those who invest in it). Several noteworthy insights emerge regarding the relationship between AI and human capital

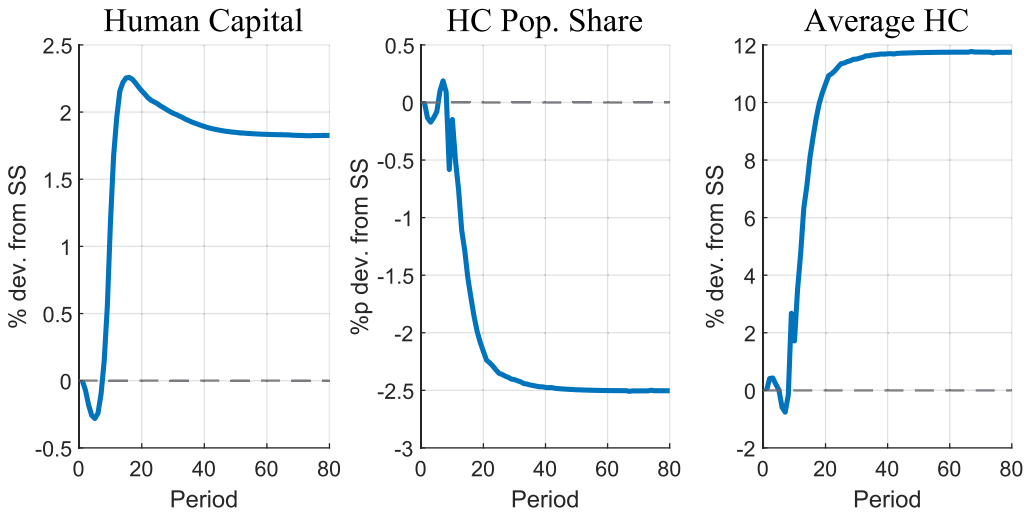


Figure 5: Transition Path: Human Capital

Note: The transition paths of human capital. The x-axis represents years, and the y-axis shows the percentage (or percentage point) deviation from the initial steady state. AI introduction is assumed to occur in period 10. "HC Pop. Share" represents the share of households who invest in human capital, while "Average HC" denotes average human capital per household among those who invest in it.

accumulation.

First, aggregate human capital initially declines but then rises sharply during the 10 years leading up to the AI introduction. This pattern reflects the coexistence of both disinvestment and investment in human capital by households in anticipation of changes in sectoral earning premiums. A reduction in the middle-sector earning premium discourages households with relatively low human capital from investing in the additional human capital needed to qualify for middle-sector employment. Conversely, an increase in the high-sector earning premium (relative to the middle sector) motivates households with relatively high human capital to intensify their human capital accumulation, sometimes even opting out of the labor force temporarily to focus on more intensive investment in their skills.[13] This dynamic reflects workers' forward-looking behavior in response to labor market changes, as suggested by Di Giacomo and Lerch (2023).

Second, the introduction of AI triggers an overshoot in human capital accumulation. Specifically, the sharp increase in human capital in anticipation of AI is followed by a gradual decline after its implementation. This pattern underscores the distinction between anticipating AI and responding to it in human capital accumulation. As shown in the middle panel of Figure 5, following the AI introduction, the share of the population investing in human capital experiences a significant and persistent decline, ultimately stabilizing at a level 2.5% lower than the initial steady state.

Third, post-AI human capital accumulation becomes increasingly concentrated among a smaller fraction of the population. Accompanying the decline in the population share investing in human capital is a sustained rise in per-household human

[13]Employment dynamics (Figure 4) shows a portion of the labor force exiting to pursue further skills training or education in preparation for the upcoming technological shift.

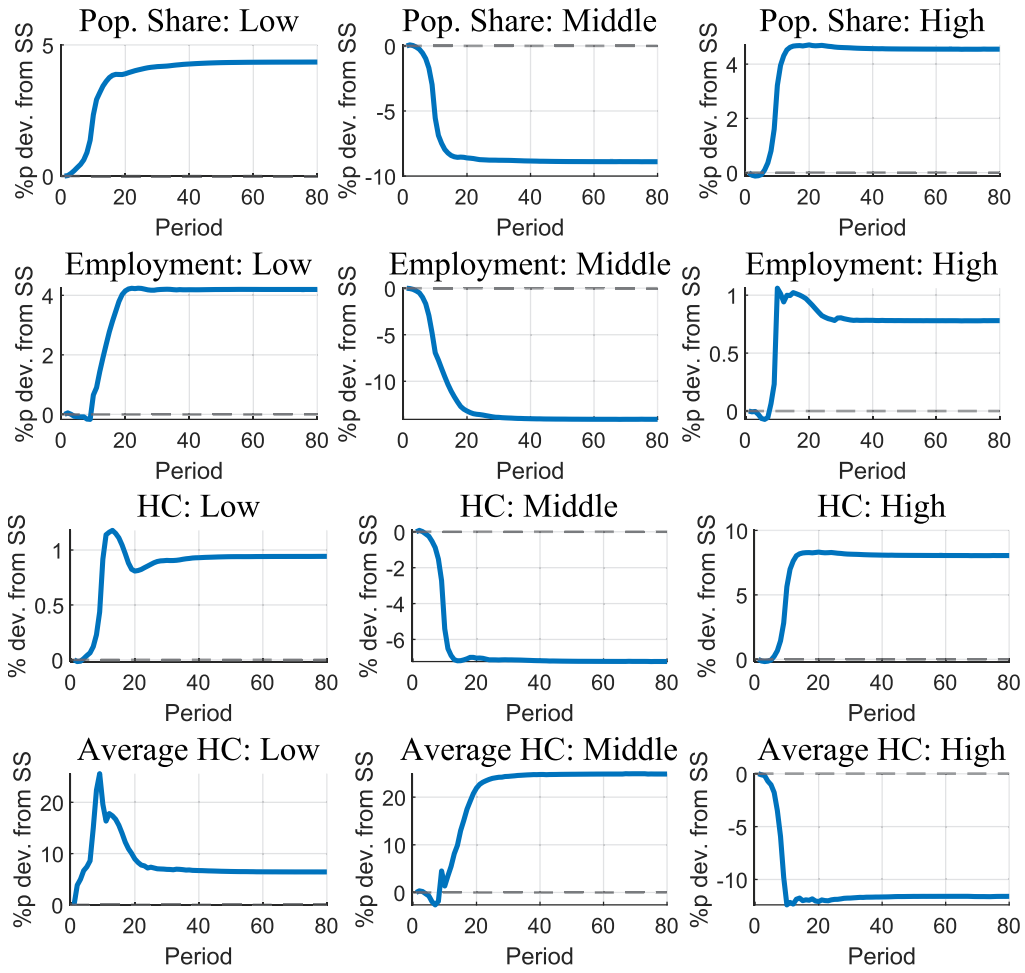


Figure 6: Sectoral Transition Path

Note: The transition paths within each sector. The x-axis represents years, and the y-axis shows the percentage (or percentage point) deviation from the initial steady state. AI introduction is assumed to occur in period 10. "Pop. Share" denotes the population share within each sector. "Employment" is the percentage of households who are employed in each sector. "HC" represents the level of human capital in each sector, while "Average HC" denotes the average human capital per worker among those investing in it within each sector.

capital, as shown in the right panel of Figure 5. In the new steady state, the average human capital among those who invest in it is 12% higher than in the initial steady state.

5.3 Sectoral Implications: Job Polarization

In this section, we further explore the sectoral implications of AI introduction, particularly in relation to job polarization. These sectoral implications help us to understand what is behind the aforementioned aggregate transition dynamics. They also have profound distributional effects across the three main sectors of the labor market.

5.3.1 Redistribution of population and employment

The first and second rows of Figure 6 illustrate the transition paths of population shares and employment rates in each sector. Notably, the middle sector experiences

a significant decline, with its population share decreasing by approximately 9%. Additionally, employment within this sector plummets to a level 14% lower than the initial steady state. In contrast, both the low and high sectors see increases in their population shares and employment rates. These dynamics indicate a reallocation of workers from the middle sector to the low and high sectors following the introduction of AI.

This worker reallocation aligns with the phenomenon of "job polarization"(Goos *et al.*, 2014), where AI and automation technologies disproportionately replace tasks commonly performed by middle-skilled workers. However, our model introduces a complementary mechanism to the conventional understanding of this reallocation. Specifically, households in our model voluntarily exit the middle sector even before AI implementation by adjusting their human capital investments – many middle-sector workers opt for non-employment to invest in skills that will better position them for the post-AI labor market. To emphasize this key difference, our model deliberately abstracts from any direct negative effect of AI on middle-sector workers.

Another intriguing finding in our model is the more pronounced employment effect in the low sector compared to the high sector. In the new steady state, the employment rate in the low sector increases by 4%, whereas in the high sector, it rises by only 0.8%. Given that the population shares in both sectors increase by similar margins, this asymmetry in employment rate changes suggests an unbalanced reallocation of workers from the middle sector, with a greater flow toward the low sector.

This disparity arises from two key factors. First, AI enhances the productivity of low-sector workers by 7.5% and high-sector workers by 5%. However, this productivity differential alone does not fully account for the significant asymmetry. The second factor is the variation in labor supply elasticity across sectors. Compared to the high sector, the low sector exhibits higher labor supply elasticity, meaning that the same change in labor earnings triggers larger labor supply responses. This is because households in the low sector have lower consumption levels, making their marginal utility of consumption more sensitive to changes in their budget. Consequently, a greater proportion of households in the low sector are at the margin between employment and non-employment (Chang and Kim, 2006).

5.3.2 Sectoral human capital adjustments

We now turn to the sectoral human capital adjustments – the key mechanism behind the job polarization in our model.

The last two rows of Figure 6 illustrate the dynamics of human capital adjustments in the low, middle and high sectors. The third row shows the transition paths of the total level of human capital in each sector, while the bottom row displays human capital per household actively investing in skills. As the population reallo-

cates from the middle sector to the other two sectors, total human capital naturally decreases in the middle sector and increases in the low and high sectors.

One particularly interesting finding is that the average human capital per household (among those actively investing in their skills) in the middle sector increases significantly, despite the decline in its total human capital. This outcome is driven by a selection effect in human capital adjustments among middle-sector households.

Households seeking to exit the middle sector follow two distinct paths: they may either transition downward to the low sector by ceasing human capital investment and allowing their skills to depreciate, or transition upward to the high sector by intensifying their human capital investment. Those who opt not to invest in human capital are typically individuals with relatively lower human capital, while those who choose to increase their investment tend to have relatively higher initial human capital. Consequently, the composition of middle-sector households that continue to invest in human capital becomes increasingly concentrated among individuals with higher human capital levels, driving the observed rise in the sector's average human capital.

The outflow of households from the middle sector also explains the distinctive patterns in the dynamics of average human capital in the low and high sectors. In the high sector, average human capital declines because households transitioning from the middle sector tend to have lower human capital levels than the average high-sector household. Conversely, in the low sector, the average human capital increases, as households moving from the middle sector possess higher levels of human capital compared to the average low-sector household.

Finally, the overshoot of human capital accumulation observed in the aggregate dynamics in Figure 5 is also evident in the low sector. This occurs because the initial influx of households from the middle sector is concentrated at the higher end of the human capital distribution among low-sector households. As AI reduces the incentives for these households to move upward to the middle sector, they gradually redistribute across other parts of the low-sector human capital distribution. Recall that households with relatively high human capital are more likely to invest in their skills, the gradual dissolution of this high end of the human capital distribution leads to a persistent decline in the share of the low-sector population that invests in human capital.

5.4 Effects on Inequality

The findings above highlight the nuanced effects of AI on human capital adjustments within each sector, as households seek to adapt to a rapidly changing labor market. This dynamic is particularly important for understanding long-term economic inequality. The polarization of the labor market is not only about job displacement but also about the differential ability of households to upskill or reskill in response

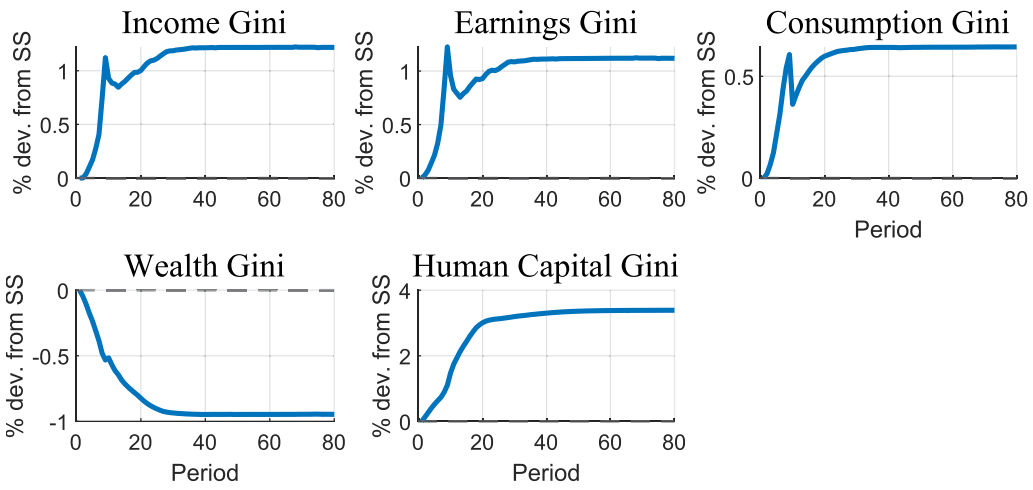


Figure 7: Transition Path: Inequality Measures

Note: The transition paths of Gini coefficients for income, earnings, consumption, wealth, and human capital. The x-axis represents years, and the y-axis shows the percentage deviation from the initial steady state. AI introduction is assumed to occur in period 10. Earnings include only labor income, while income includes both capital and labor incomes.

to technological changes. Households who invest in human capital are better positioned to move up the skill ladder, while those unable to do so may face downward mobility, contributing to greater inequality.

Figure 7 illustrates the transition paths of Gini coefficients for earnings (labor income), income (capital and labor incomes), consumption, wealth (physical capital), and human capital. The introduction of AI leads to a notable increase in inequality for income, earnings, consumption, and human capital, while surprisingly reducing wealth inequality.

AI-induced job polarization directly translates into polarization in human capital, as households previously qualified for middle-sector jobs transition to positions in either the low or high sector. Those moving to the low sector face a decline in labor earnings, while those transitioning to the high sector experience increased earnings. Consequently, earnings and income inequality widen. As income disparities grow, consumption inequality also increases, though its growth is tempered by households' consumption smoothing behavior.

To cope with changes in their labor earnings, households transitioning to the low sector draw down their savings to offset the negative impact of reduced earnings on consumption. Conversely, those moving to the high sector also reduce their savings, but for a different reason: expecting a sustained increase in future labor income, they find it optimal to increase current consumption.

For households remaining in the middle sector, the situation is markedly different. They anticipate a sharp decline in wages — and consequently in earnings and income — after AI implementation, as the middle sector does not directly benefit from AI-driven productivity gains (Figure 4). Although wages eventually recover to a level slightly above the initial steady state, the temporary slump in earnings prompts these households to increase precautionary savings.

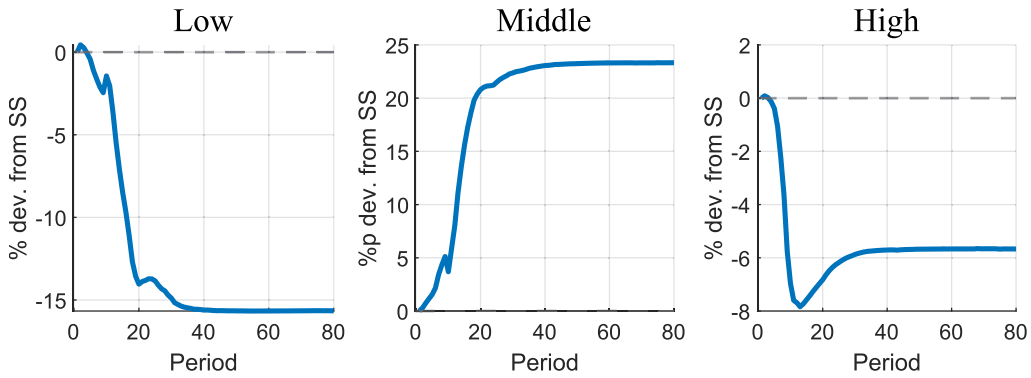


Figure 8: Sectoral Transition Path: Average Capital

Note: The transition paths of average capital within each sector. The x-axis represents years, and the y-axis shows the percentage deviation from the initial steady state. AI introduction is assumed to occur in period 10. "Average Capital" denotes the physical assets per household in each sector.

These effects on household savings are illustrated in Figure 8, which shows the transition paths of average capital holdings per household in each sector. In the new steady state, per-household capital holdings decrease by approximately 16% in the low sector and 6% in the high sector, while increasing by around 23% in the middle sector. These changes reallocate capital from households with high levels of human capital (and thus high income) to those with intermediate levels of human capital, leading to a reduction in wealth inequality.

5.5 Welfare Implications

We now analyze the welfare effects of AI introduction across different sectors of the economy. Welfare effects are evaluated by comparing households' welfare during the transition period — accounting for the entire path to the new steady state — with their welfare in the initial steady state. Table IV presents the welfare effects in consumption-equivalent terms. On average, the introduction of AI benefits households, yielding a welfare gain of approximately 1.62% in consumption-equivalent terms.

The low sector experiences the greatest welfare gain, as its productivity increase from AI is the highest among the three sectors. More unexpectedly, the middle sector achieves a larger welfare gain than the high sector, despite not benefiting directly from AI-induced productivity improvements. This surprising result stems from the reallocation of physical capital from the high sector to the middle sector, driven by the precautionary saving behavior described in the previous subsection.

Table IV: Welfare Effect across Sectors

Low	Middle	High	Average
2.06	1.40	1.33	1.62

Note: The welfare effect in consumption-equivalent terms. It is calculated by comparing welfare during the transition period, accounting for the entire path to the new equilibrium, with welfare in the initial steady state.

6 Human Capital Adjustment in AI's Economic Impact

The aggregate and distributional effects of AI are jointly determined by its direct impact on sectoral productivity and the endogenous response of human capital accumulation to these changes. In this section, we investigate the importance of endogenous human capital adjustments in shaping both the transitional and long-run effects of AI on the economy.

To this end, we analyze an alternative economy where households maintain their initial steady-state level of human capital throughout the AI implementation until the new steady state is reached. This scenario, referred to as the "No HC model," still allows households to make endogenous decisions regarding consumption, savings, and labor supply. We then compare the transition dynamics of this economy with those of our benchmark economy to isolate the role of human capital accumulation in driving aggregate and distributional outcomes.

6.1 Human Capital Adjustment in Aggregate Outcomes

Figure 9 compare the transition paths for key macroeconomic variables in models with and without endogenous human capital accumulation.

Reducing labor supply: The most notable effect of human capital adjustment is observed in employment dynamics. In the benchmark economy, where households can accumulate human capital, some middle-sector workers opt for voluntary non-employment to invest in skill enhancement, leading to reduced employment both before and after the implementation of AI. In contrast, when human capital is assumed to be fixed (the No HC model), employment increases because households no longer need to choose between working and investing in human capital.

Increasing long-run output: In the long run, output in the benchmark model surpasses that in the No HC model, even with lower employment levels. This is due to human capital adjustments reallocating more workers to sectors that benefit from AI advancements.

Lowering precautionary saving: The benchmark model achieves higher consumption and lower investment relative to the No HC model. This occurs because human capital adjustments offer households an alternative means to hedge against idiosyncratic productivity risk, thereby reducing their reliance on physical capital.

Stabilizing factor prices: These differences in employment and investment responses to AI result in significantly divergent dynamic effects on wages and capital returns. Before AI implementation, employment dynamics lead to opposite movements in wages and capital returns between the benchmark model and the No HC

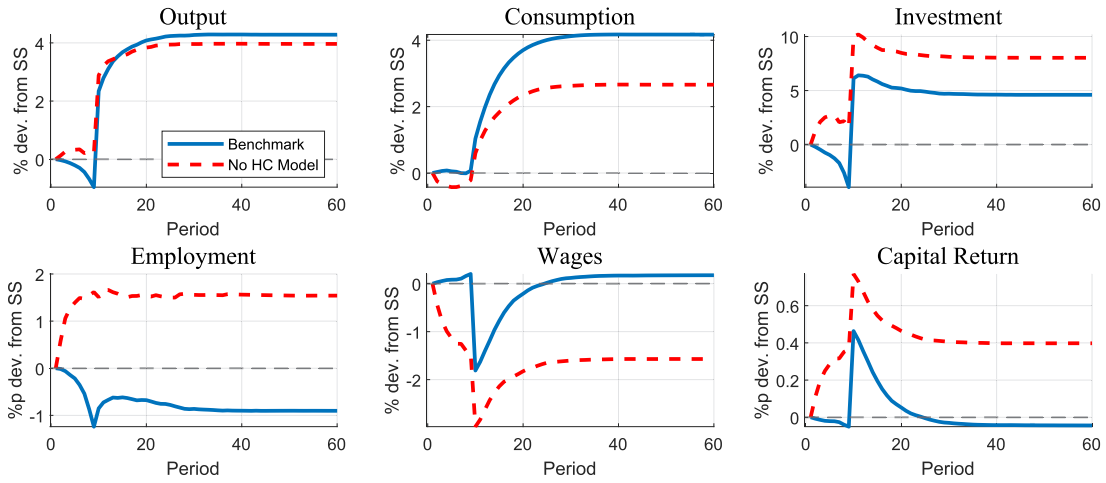


Figure 9: Transition Path of Aggregate Variables: Benchmark vs. No HC Models.

Note: The transition paths of aggregate variables: benchmark vs. No HC models. The x-axis represents years, and the y-axis shows the percentage deviation from the initial steady state. AI introduction is assumed to occur in period 10. The No HC model is an economy in which workers maintain their initial steady-state level of human capital throughout the AI implementation until the new steady state is reached.

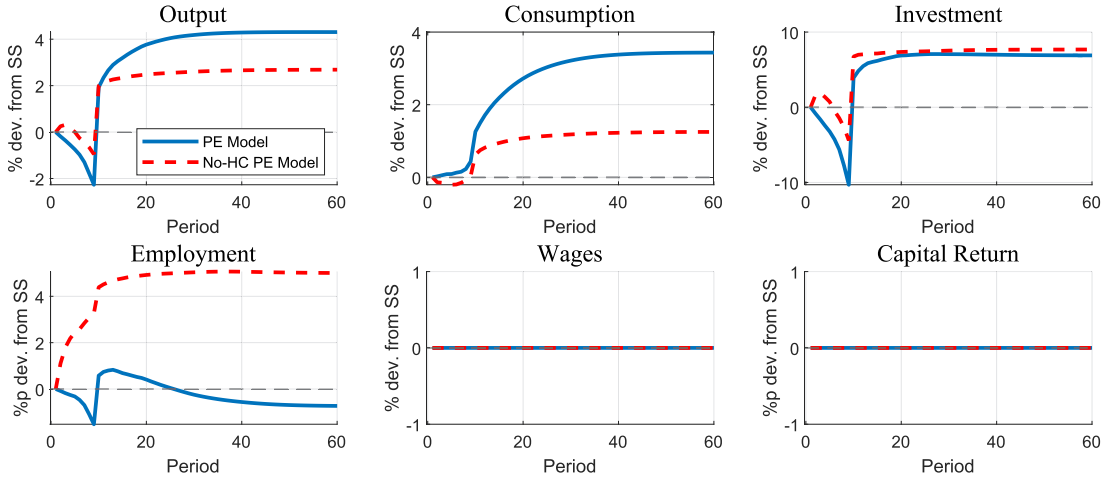


Figure 10: Transition Path of Aggregate Variables: PE vs. No-HC PE Models.

Note: The transition paths of aggregate variables: PE vs. No-HC PE models. The x-axis represents years, and the y-axis shows the percentage deviation from the initial steady state. AI introduction is assumed to occur in period 10. The PE model is an economy in which wages and real interest rates remain fixed at the initial steady state during the transition to the new steady state. The No-HC PE model is an economy in which workers maintain their initial steady-state level of human capital throughout the AI implementation until the new steady state is reached, and wages and capital returns remain fixed at the initial steady state during the transition to the new steady state.

model. In the long run, AI substantially increases capital returns and decreases wages when human capital remains fixed. However, endogenous human capital adjustments largely neutralize these changes in factor prices, stabilizing wages and capital returns over time.

Redistribution versus general equilibrium effects: The effects of human capital adjustments on AI's aggregate impacts operate through two primary channels: the *redistribution channel*, which reallocates households across skill sectors, and the *general equilibrium (GE) channel*, which operates through changes in wages and capital returns. We now assess the relative importance of these channels in shaping economic outcomes.

Figure 10 compares the transition dynamics between scenarios with and without human capital adjustments, while holding wages and capital returns fixed at their initial steady-state levels to eliminate GE effects. We refer to the former as the PE Model" and the latter as the "No-HC PE Model." The difference between the solid blue line and the dashed red line isolates the effect of redistribution channel. Comparing this difference to the gap between the benchmark model and the No HC model in Figure 9 enables us to evaluate the importance of the redistribution channel relative to the GE channel. Two key observations emerge.

First, the *redistribution channel* alone accounts for all the *qualitative effects* of human capital adjustments on AI's aggregate impacts. Redistribution of human capital increases consumption, even before AI implementation, as more households shift to the high sector. It also reduces investment by mitigating precautionary savings and lowers employment as middle-sector workers leave the labor market to invest in human capital. In the long run, redistribution amplifies AI's positive impact on output by reallocating more workers to sectors that benefit most from AI advancements.

Second, the *GE channel* primarily affects the *quantitative magnitude* of human capital adjustments' impact on AI's aggregate outcomes. When the GE channel is included, the differences in output, consumption, and employment between models with and without human capital adjustments are smaller compared to when the GE channel is excluded. In contrast, and somewhat unexpectedly, the difference in investment is larger when the GE channel is included. This indicates that allowing capital returns to adjust amplifies the impact of human capital accumulation on how household savings respond to AI.

6.2 Human Capital Adjustment in Distributional Outcomes

Figure 11 contrasts the transition dynamics of inequality measures in the No HC economy to those in the benchmark economy to isolate the role of human capital accumulation. Figure 12 extends this comparison by fixing wages and capital returns,

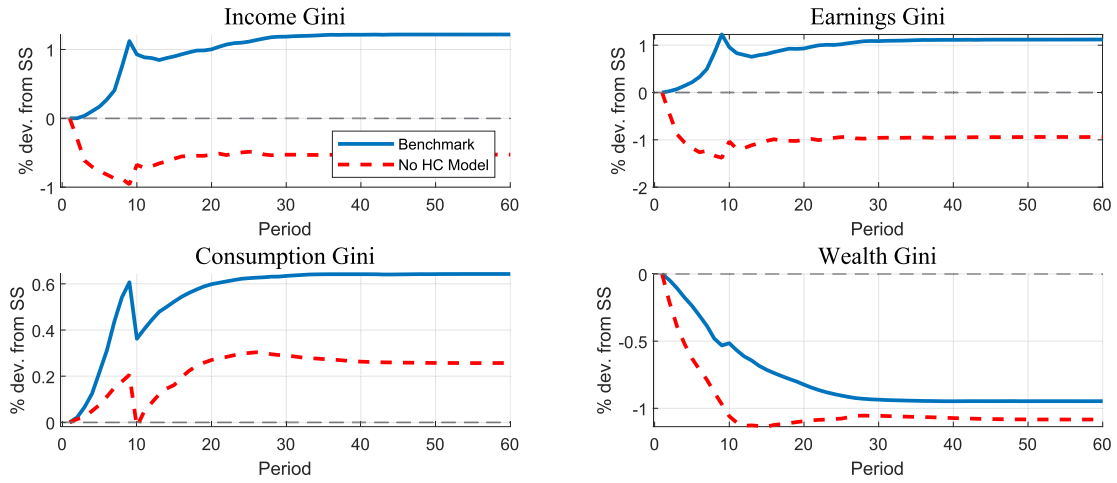


Figure 11: Transition Path of Inequality Measures: Benchmark vs. No HC Models.
Note: The transition paths of inequality measures: benchmark vs. No HC models. The x-axis represents years, and the y-axis shows the percentage deviation from the initial steady state. AI introduction is assumed to occur in period 10. The No HC model is an economy in which workers maintain their initial steady-state level of human capital throughout the AI implementation until the new steady state is reached.

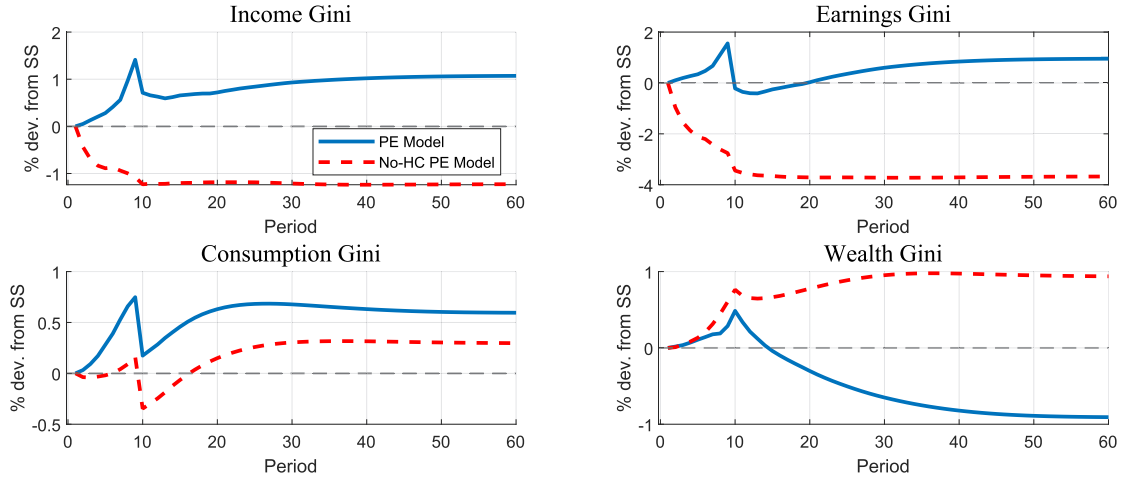


Figure 12: Transition Path of Inequality Measures: PE vs. No-HC PE Models.
Note: The transition paths of inequality measures: PE vs. No-HC PE models. The x-axis represents years, and the y-axis shows the percentage deviation from the initial steady state. AI introduction is assumed to occur in period 10. The PE model is an economy in which wages and real interest rates remain fixed at the initial steady state during the transition to the new steady state. The No-HC PE model is an economy in which workers maintain their initial steady-state level of human capital throughout the AI implementation until the new steady state is reached, and wages and capital returns remain fixed at the initial steady state during the transition to the new steady state.

in order to assess the relative importance of the two primary channels through which human capital adjustments operate, the *redistribution channel* and the *GE channel*.

Income, earnings, and consumption inequalities: Without any human capital adjustments, both income and earnings inequalities decrease, while consumption inequality increases, though not as much as in the benchmark economy. These findings reaffirm that the impact of AI on income, earnings, and consumption inequalities operates primarily through AI-induced human capital adjustments, which drive job polarization.

When the *GE channel* is disabled, as shown in Figure 12, there is no qualitative change in how human capital adjustments influence inequalities in income, earnings, and consumption. Thus, it is once again the *redistribution channel* that accounts for the qualitative effects of human capital adjustments on AI's distributional impacts across these dimensions.

The gap between the transition paths in Figures 11 and 12 captures the magnitude of the effects of human capital adjustments. Disabling the *GE channel* in Figure 12 widens this gap, suggesting that flexible factor prices help mitigate the negative impact of human capital adjustments on income, earnings, and consumption inequalities.

Wealth inequality: An interesting interaction exists between the *redistribution channel* and the *GE channel* in determining how human capital adjustments influence AI's impact on wealth inequality.

When the *GE channel* is active (Figure 11), AI reduces the wealth Gini, but the *redistribution channel* moderates this effect. However, when the *GE channel* is disabled (Figure 12), AI increases wealth inequality in the long run without the *redistribution channel* from human capital adjustment. In contrast, with the *redistribution channel* active, AI reduces wealth inequality.

These observations lead to two key conclusions:

First, the *redistribution channel* alone introduces a qualitative shift in AI's long-run impact on the wealth Gini (as shown in Figure 12).

Second, the *GE channel*, when combined with human capital adjustment, qualitatively alters the effect of anticipating AI on the wealth Gini (as shown by comparing the blue lines in Figures 11 and 12).

Policy implications: The impact of human capital adjustments on AI's distributional outcomes, along with the roles of the *redistribution channel* and *GE channel*, provides valuable insights for policy discussions on how to address the challenges posed by AI shocks.

In particular, government interventions aimed at stabilizing wages in response to AI-induced economic shocks may unintentionally worsen wealth inequality. Our

analysis indicates that if wages are prevented from adjusting to reflect productivity differences, this distorts households' incentives to adjust their human capital and precautionary savings—both of which play a critical role in mitigating wealth inequality.

7 Conclusion

Recent studies on AI suggest that advancements are likely to reduce demand for junior-level positions in high-skill industries while increasing the need for roles focused on advanced decision-making and AI oversight. We demonstrate how human capital investments are expected to adapt in response to these shifts in skill demand, highlighting the importance of accounting for these human capital responses when assessing AI's economic impact.

Our work points to several promising directions for future research on the economic impacts of AI. First, while general equilibrium effects—such as wage and capital return adjustments—have a limited role in our model, further research could examine how these effects might vary under different economic conditions or policy environments. Second, if governments implement redistribution policies to address AI-induced inequality, understanding how these policies influence human capital accumulation, and thus their effectiveness, would be valuable. Finally, our model assumes households have perfect foresight when making human capital investments. Relaxing this assumption could reveal new insights into the economic trajectory of AI advancements and offer important policy implications.

References

Acemoglu, Daron and Restrepo, Pascual (2019). "8. Artificial Intelligence, Automation, and Work", *The Economics of Artificial Intelligence*. Ed. by Agrawal, Ajay, Gans, Joshua, and Goldfarb, Avi. University of Chicago Press: Chicago, pp. 197–236.

— (2020). "Robots and Jobs: Evidence from US Labor Markets", *Journal of Political Economy*, Vol. 128 No. 6, pp. 2188–2244.

Aiyagari, S. Rao (1994). "Uninsured Idiosyncratic Risk and Aggregate Saving", *The Quarterly Journal of Economics*, Vol. 109 No. 3. Publisher: Oxford University Press, pp. 659–684.

Atkin, David (2016). "Endogenous Skill Acquisition and Export Manufacturing in Mexico", *American Economic Review*, Vol. 106 No. 8, pp. 2046–85.

Autor, David H. and Dorn, David (2013). "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market", *American Economic Review*, Vol. 103 No. 5, pp. 1553–97.

Beaudry, Paul, Green, David A., and Sand, Benjamin M. (2016). "The Great Reversal in the Demand for Skill and Cognitive Tasks", *Journal of Labor Economics*, Vol. 34 No. S1, S199–S247.

Chang, Yongsung and Kim, Sun-Bin (2006). "From Individual to Aggregate Labor Supply: A Quantitative Analysis based on a Heterogeneous Agent Macroeconomy", *International Economic Review*, Vol. 47 No. 1, pp. 1–27.

Dauth, Wolfgang *et al.*, (2021). "The Adjustment of Labor Markets to Robots", *Journal of the European Economic Association*, Vol. 19 No. 6, pp. 3104–3153.

Di Giacomo, Giuseppe and Lerch, Benjamin (2023). "Automation and Human Capital Adjustment", *Journal of Human Resources*,

Diaz-Gimenez, Javier, Quadrini, Vincenzo, and Rios-Rull, Jose-Victor (1997). "Dimensions of inequality: facts on the U.S. distributions of earnings, income, and wealth", *Quarterly Review*, Vol. 21 No. Spr, pp. 3–21.

Faber, Marius, Sarto, Andres P, and Tabellini, Marco (2022). *Local Shocks and Internal Migration: The Disparate Effects of Robots and Chinese Imports in the US*, Working Paper No. 30048. National Bureau of Economic Research.

Goos, Maarten and Manning, Alan (2007). "Lousy and Lovely Jobs: The Rising Polarization of Work in Britain", *The Review of Economics and Statistics*, Vol. 89 No. 1, pp. 118–133.

Goos, Maarten, Manning, Alan, and Salomons, Anna (2014). "Explaining Job Polarization: Routine-Biased Technological Change and Offshoring", *American Economic Review*, Vol. 104 No. 8, pp. 2509–26.

Lerch, Benjamin (2021). *Robots and Nonparticipation in the US: Where Have All the Workers Gone?*, IdEP Economic Papers No. 2003. USI Università della Svizzera italiana.

OECD (1998). *Human Capital Investment*, p. 111.

Prettner, Klaus and Strulik, Holger (2020). "Innovation, automation, and inequality: Policy challenges in the race against the machine", *Journal of Monetary Economics*, Vol. 116, pp. 249–265.

Sachs, Jeffrey D. and Kotlikoff, Laurence J. (2012). *Smart Machines and Long-Term Misery*, NBER Working Papers No. 18629. National Bureau of Economic Research, Inc.

A Parameter Restrictions for the Two-Period Model

To guarantee that $(n = 0, e = e_H)$ dominates $(n = 0, e = 0)$, we need a lower bound for λ. The slow learners prefer $(n = 0, e = e_H)$ if and only if

$$(1 + \beta) \ln c(n = 0, e = e_H) - \chi_e e_H \geq (1 + \beta) \ln c(n = 0, e = 0)$$

or equivalently:

$$\lambda \geq \underline{\lambda}_1 := \frac{(1+r)a + \frac{w'z'}{1+r'}}{\frac{w'z'}{1+r'}} \left(1 - \frac{1}{\exp(\frac{\chi_e e_H}{1+\beta})}\right) \text{ if } h < h_M \frac{1}{1-\delta} \tag{33}$$

$$\lambda \geq \underline{\lambda}_3 := \frac{(1+r)a + \frac{w'z'}{1+r'}}{\frac{w'z'}{1+r'}} \left(\exp(\frac{\chi_e e_H}{1+\beta}) - 1\right) \text{ if } h \geq h_M \frac{1}{1-\delta} \tag{34}$$

To avoid $(n = 1, e = e_L)$ from being a dominated choice, we need another lower bound for λ. To see it, recall that $(n = 1, e = 0)$ is better than $(n = 1, e = e_L)$ if $z > \overline{z}_{fast}$, and $(n = 1, e = e_L)$ is better than $(n = 0, e = e_L)$ if $z > \underline{z}_{fast}$. $(n = 1, e = e_L)$ is therefore the best choice over the interval $(\underline{z}_{fast}, \overline{z}_{fast})$. For such an interval to exist, it must be the case that when $z = \underline{z}_{fast}$, $z < \overline{z}_{fast}$.

$z = \underline{z}_{fast}$ means that the fast learners are indifferent between $(n = 1, e = e_L)$ and $(n = 0, e = e_L)$ so that

$$(1+r)a + wzx(h) + \frac{w'z'}{1+r'} = \exp(\frac{\chi_n}{1+\beta}) \left[(1+r)a + \frac{w'z'}{1+r'}\right] \text{ if } h < h_M \frac{1}{1-\delta} \tag{35}$$

$$(1+r)a + wzx(h) + \frac{w'z'(1+\lambda)}{1+r'} = \exp(\frac{\chi_n}{1+\beta}) \left[(1+r)a + \frac{w'z'(1+\lambda)}{1+r'}\right] \text{ if } h \geq h_M \frac{1}{1-\delta} \tag{36}$$

For the fast learners to prefer $(n = 1, e = e_L)$ over $(n = 1, e = 0)$, we need

$$(1+\beta) \ln \frac{c(n=1, e=e_L)}{c(n=1, e=0)} \geq \chi_e e_L \tag{37}$$

If $h < h_M \frac{1}{1-\delta}$, this inequality is:

$$(1+\beta) \ln \frac{(1+r)a + wzx(h) + \frac{w'z'}{1+r'}}{(1+r)a + wzx(h) + \frac{w'z'(1-\lambda)}{1+r'}} \geq \chi_e e_L$$

Evaluating the left-hand-side at $z = \underline{z}_{fast}$ yields:

$$\lambda \geq \underline{\lambda}_2 := \frac{(1+r)a + \frac{w'z'}{1+r'}}{\frac{w'z'}{1+r'}} \left(1 - \frac{1}{\exp(\frac{\chi_e e_L}{1+\beta})}\right) \exp(\frac{\chi_n}{1+\beta}) \tag{38}$$

If $h > h_M \frac{1}{1-\delta}$, inequality (37) is:

$$(1+\beta) \ln \frac{(1+r)a + wzx(h) + \frac{w'z'(1+\lambda)}{1+r'}}{(1+r)a + wzx(h) + \frac{w'z'}{1+r'}} \geq \chi_e e_L$$

Evaluating the left-hand-side at $z = \underline{z}_{fast}$ yields:

$$\lambda \geq \underline{\lambda}_4 := \frac{(1+r)a + \frac{w'z'}{1+r'}}{\frac{w'z'}{1+r'}} \frac{\left(\exp(\frac{\chi_e e_L}{1+\beta}) - 1\right)\exp(\frac{\chi_n}{1+\beta})}{\exp(\frac{\chi_e e_L}{1+\beta}) + \exp(\frac{\chi_n}{1+\beta}) - \exp(\frac{\chi_e e_L + \chi_n}{1+\beta})} \tag{39}$$

We have that $\underline{\lambda}_1 > \underline{\lambda}_2$ and $\underline{\lambda}_3 > \underline{\lambda}_4$ if

$$\exp(\frac{\chi_e e_H}{1+\beta}) > \frac{\exp(\frac{\chi_e e_L}{1+\beta})}{\exp(\frac{\chi_e e_L}{1+\beta}) + \exp(\frac{\chi_n}{1+\beta}) - \exp(\frac{\chi_e e_L + \chi_n}{1+\beta})} \tag{40}$$

Therefore, the inequality above implies that the conditions (33) and (34) are sufficient for the conditions (38) and (39). Furthermore, $\lambda_3 \geq \lambda_1$ so that the condition (34) is sufficient for the condition (33).

We can then conclude that the conditions (34) and (40) are sufficient for 1) the slower learners always prefers $(n = 0, e = e_H)$ over $(n = 0, e = 0)$, and 2) $\overline{z}_{fast} > \underline{z}_{fast}$.

B Cutoffs ranking for the Two-Period Model

For the fast learners, their cutoffs rank as follows

$$\frac{\overline{z}^L_{fast}(a)}{1-\lambda} > \overline{z}^L_{fast}(a) > \overline{z}^M_{fast}(a) > \frac{\overline{z}^M_{fast}(a)}{1+\lambda} \tag{41}$$

$$\frac{\underline{z}^L_{fast}(a)}{1-\lambda} > \underline{z}^M_{fast}(a) > \underline{z}^L_{fast}(a) > \frac{\underline{z}^M_{fast}(a)}{1+\lambda} \tag{42}$$

For the slow learners, the rank of their cutoffs is

$$\frac{\overline{z}^L_{slow}(a)}{1-\lambda} > \overline{z}^M_{slow}(a) > \overline{z}^L_{slow}(a) > \frac{\overline{z}^M_{slow}(a)}{1+\lambda} \tag{43}$$

For the non-learners, the rank of their cutoffs is

$$\frac{\overline{z}^L_{non}(a)}{1-\lambda} > \overline{z}^M_{non}(a) > \frac{\overline{z}^H_{non}(a)}{1+\lambda} > \frac{\overline{z}^M_{non}(a)}{1+\lambda} \tag{44}$$

$$\overline{z}^M_{non}(a) > \overline{z}^L_{non}(a) \tag{45}$$

C Computational Procedure for the Quantitative Model

C.1 Steady-state Equilibrium

In the steady-state, the measure of households, $\mu(a, h, x)$, and the factor prices are time-invariant. We find a time-invariant distribution μ. We compute the households' value functions and the decisions rules, and the time-invariant measure of the households. We take the following steps:

1. We choose the number of grid for the risk-free asset, a, human capital, h, and the idiosyncratic labor productivity, x. We set $N_a = 151$, $N_h = 151$, and $N_x = 9$ where N denotes the number of grid for each variable. To better incorporate the saving decisions of households near the borrowing constraint, we assign more points to the lower range of the asset and human capital.

2. Productivity x is equally distributed on the range $[-3\sigma_x/\sqrt{1-\rho_x^2}]$. As shown in the paper, we construct the transition probability matrix $\pi(x'|x)$ of the idiosyncratic labor productivity.

3. Given the values of parameters, we find the value functions for each state (a, h, x). We also obtain the decision rules: savings $a'(a, h, x)$, and $h'(a, h, x)$. The computation steps are as follow:

4. After obtaining the value functions and the decision rules, we compute the time-invariant distribution $\mu(a, h, x)$.

5. If the variables of interest are close to the targeted values, we have found the steady-state. If not, we choose the new parameters and redo the above steps.

C.2 Transition Dynamics

We incorporate the transition path from the status quo to the new steady state. We describe the steps below.

1. We obtain the initial steady state and the new steady state.

2. We assume that the economy arrives at the new steady state at time T. We set the T to 100. The unit of time is a year.

3. We initialize the capital-labor ratio $\{K_t/L_t\}_{t=2}^{T-1}$ and obtain the associated factor prices $\{r_t, w_t\}_{t=2}^{T-1}$.

4. As we know the value functions at time T, we can obtain the value functions and the decision rules in the transition path from $t = T - 1$ to 1.

5. We compute the measures $\{\mu_t\}_{t=2}^{T}$ with the measures at the initial steady state and the decision rules in the transition path.

6. We obtain the aggregate variables in the transition path with the decision rules and the distribution measures.

7. We compare the assumed paths of capital and the effective labor with the updated ones. If the absolute difference between them in each period is close enough, we obtain the converged transition path. Otherwise, we assume new capital-labor ratio and go back to 3.

Scenarios for the Transition to AGI

Anton Korinek* Donghyun Suh**

*University of Virginia, Brookings, GovAI

**Bank of Korea

2025 International Conference of the KOSSREC

Motivation

- AI is advancing rapidly, and the pace has been accelerating
- Recent progress has brought the potential for Artificial General Intelligence (AGI) within tangible reach:
  - Def: AGI is the ability of machines to perform all work tasks that humans can do
- AGI promises significant productivity gains while substituting for labor

Questions:

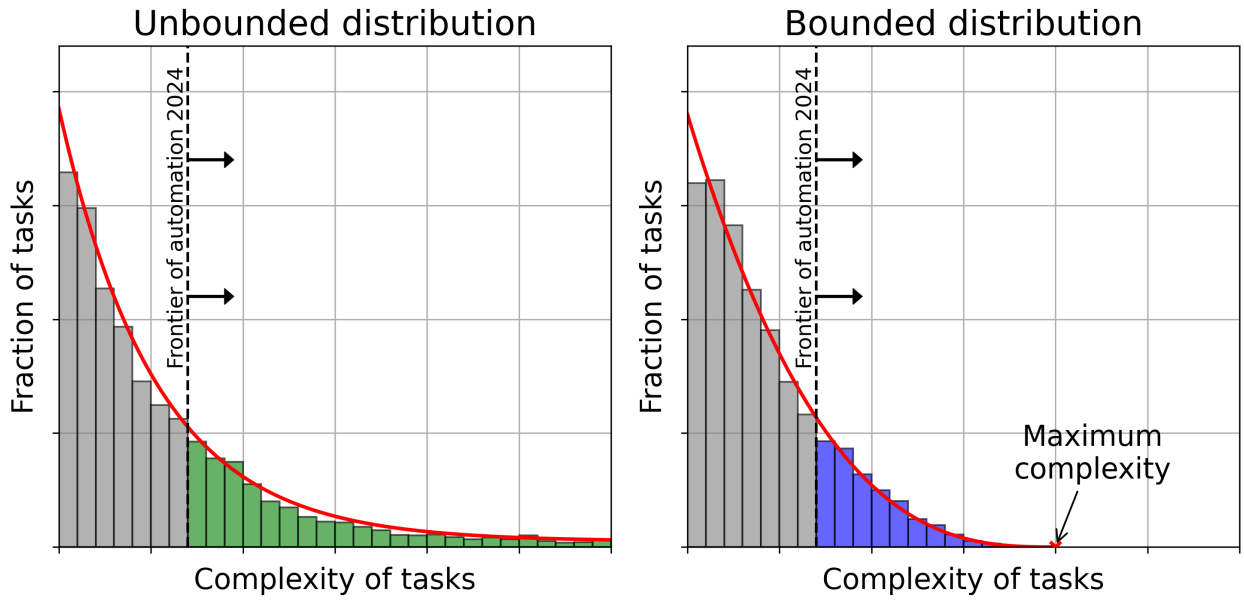
- What would the transition to AGI look like?
- What would AGI imply for output and wages?
- Under what conditions would wages collapse?

Key Assumptions

- Work is made up of elementary tasks (far smaller than O*Net tasks)
- Main distinguishing characteristic of tasks: complexity, which depends to an important extent on their "compute intensity"

- Consider different scenarios for the distribution of tasks in complexity space ("tasks in compute space"):
  - Unbounded distributions (e.g., Pareto): some human tasks will always remain too complex for machines
  - Bounded distributions (e.g., beta): there is a maximum complexity level for human tasks

Scenarios for Task Complexity



A Model of Automation

Static Environment (a la Zeira, 1998; Acemoglu-Restrepo, 2018)

- There is a representative household endowed with $L = 1$ units of labor and $K > 0$ units of capital.
- There is a continuum of tasks differing in computational complexity ($i \geq 1$) with an associated CDF $\Phi(i)$.
- Aggregate output Y is

$$Y = A\left[\int_i y(i)^{\frac{\sigma-1}{\sigma}} d\Phi(i)\right]^{\frac{\sigma}{\sigma-1}}$$

  where $y(i)$ is the amount of type i tasks.
- Each task is performed according to a production function

$$y(i) = \begin{cases} k(i) + \ell(i) & \text{for } i < I \\ \ell(i) & \text{for } i \geq I \end{cases}$$

  where I is an automation index.

Scarcity of Labor

Lemma (Scarcity of Labor): The capital intensity K/L of the economy defines a threshold $\hat{I}$ given by

$$\Phi\left(\hat{I}\right) = \frac{K/L}{1 + K/L},$$

such that there are two regions:

Region 1 (Abundant K, Scarce L) for $I < \hat{I}$

- Wages $w > R$ returns on capital
- Labor used only for unautomated tasks, production is CES

Region 2 (Equally Scarce K and L): for $I \geq \hat{I}$

- Wages $w = R$ returns on capital
- Labor and capital effective substitutes
- Production is linear: $Y = A(K + L)$

Two Regions for the Scarcity of Labor

Region 1 (low I): many tasks are unautomated so labor remains scarce $\rightarrow w > R$

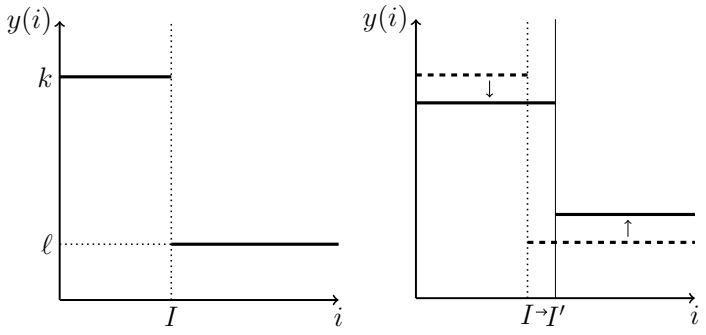


Figure: Automation relieves the scarcity of labor ($k/\ell \downarrow$)

Region 2 (high I): most tasks are unautomated so labor has lost its relative scarcity value $\rightarrow w = R$

Dynamics

- We assume exponential growth in the maximum automatable task complexity (automation index I growing at rate g)
  $\implies$ reflects Moore's law & its cousins
- Dynamics of output and wages depend on the interplay between
  - task automation $d\Phi(i)$ – which depends on the scenario $\Phi(I)$
  - capital accumulation

Consumer Problem

The representative household solves:

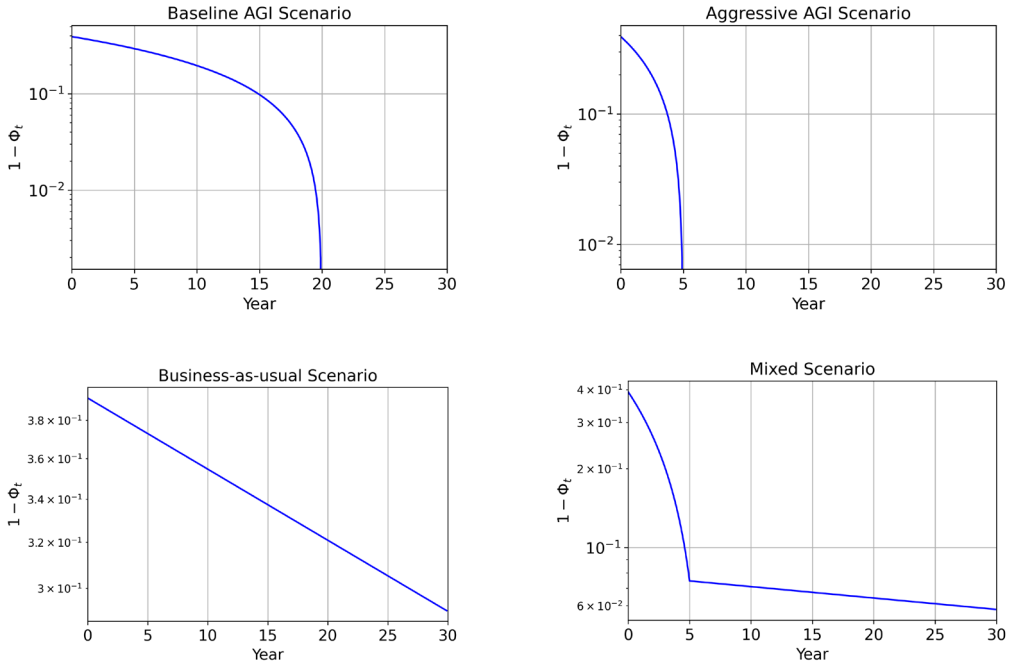
$$\max_{\{C_t\}} U = \int_0^{\infty} e^{-\rho t} u(C_t) dt$$

subject to the law of motion for capital:

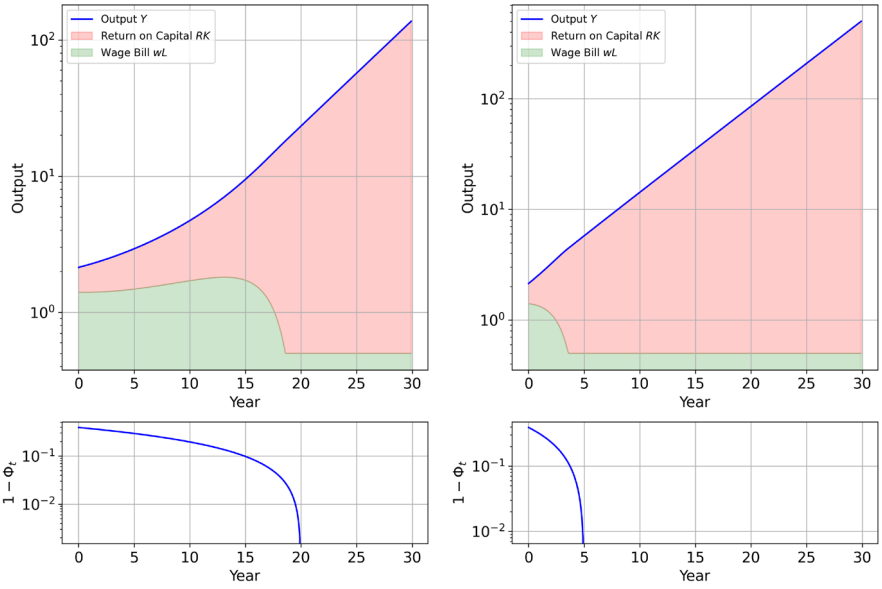
$$\dot{K}_t = F(K_t, L_t; I_t) - \delta K_t - C_t$$

for given K_0.

Four Scenarios (Log of Mass of Unautomated Tasks)



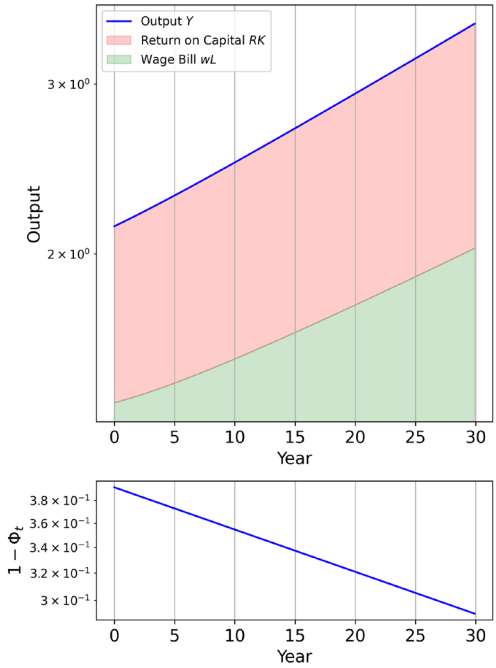
Paths of Wages Under the AGI Scenarios



Baseline AGI scenario

Aggressive AGI scenario

Path of Wages Under the Business-as-usual Scenario

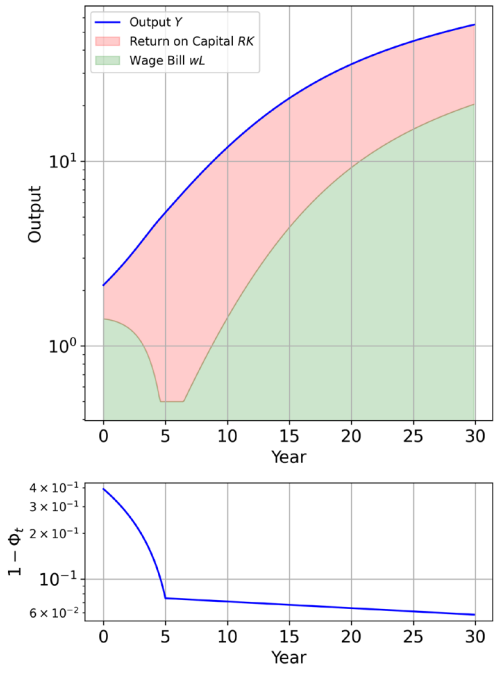


Business-as-usual Scenario: Long-run Dynamics of Wages

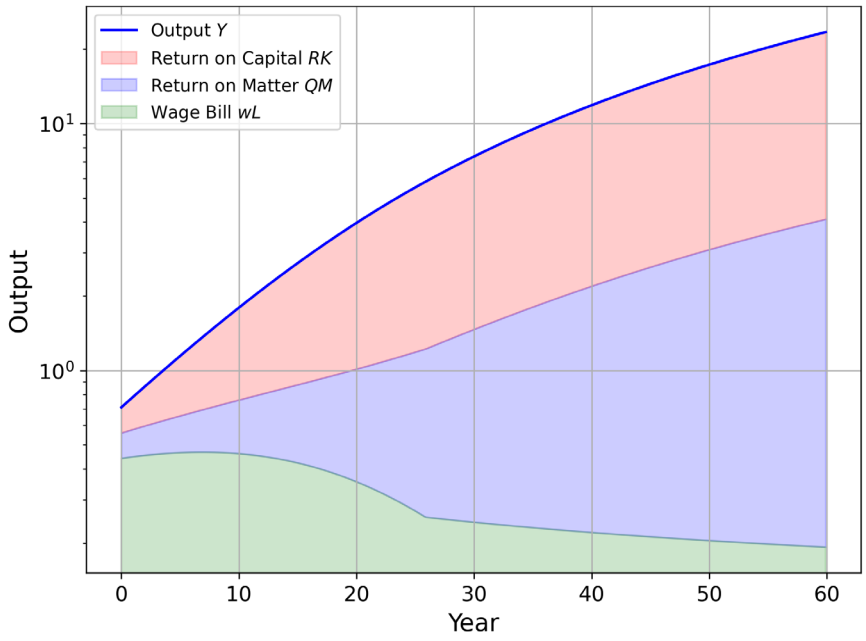
Proposition: Suppose the complexity distribution of tasks is Pareto and that the economy starts in region 1, i.e., $I_0 < \hat{I}_0$. Then the growth of wages is characterized by two thresholds on the rate of automation λg:

1. If $\lambda g \leq \frac{A-\rho-\delta}{\eta} \cdot (1-\sigma)$ then wages grow exponentially at an asymptotic rate $\frac{\lambda g}{1-\sigma}$.
2. If $\frac{A-\rho-\delta}{\eta} \cdot (1-\sigma) < \lambda g \leq \frac{A-\rho-\delta}{\eta}$ then wages grow exponentially at an asymptotic rate $\frac{1}{\sigma}\left(\frac{A-\rho-\delta}{\eta} - \lambda g\right)$.
3. Lastly, if $\lambda g > \frac{A-\rho-\delta}{\eta}$ then $\lim_{t\to\infty} w_t = A$.

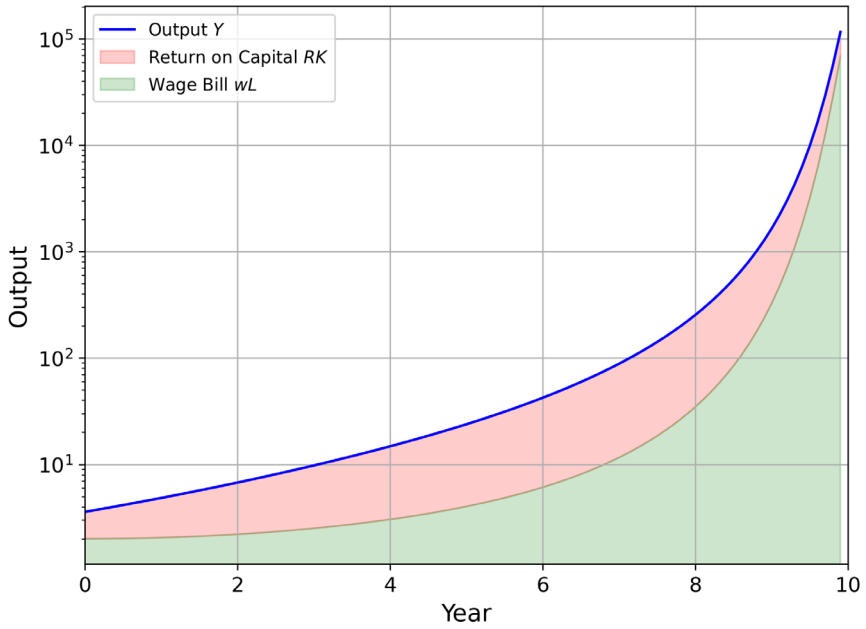
Path of Wages Under the Mixed Scenario



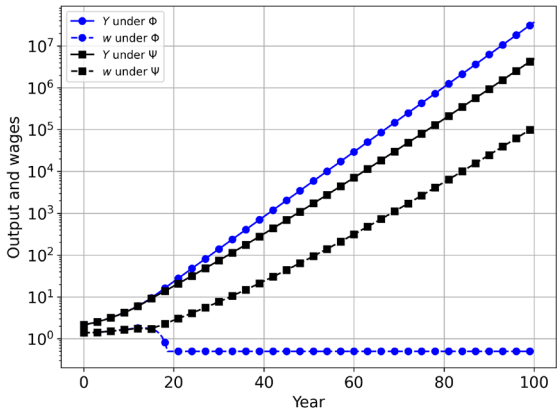
Extension 1: Fixed Factors and the Return of Scarcity

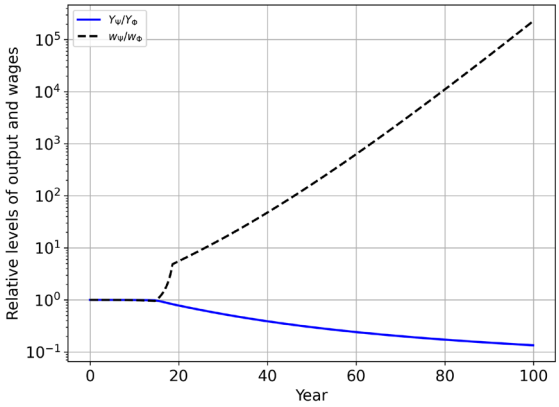


Extension 2: Automating Technological Progress



Extension 3: Nostalgic Jobs and Holding Back Automation





Extension 4: Heterogeneous Worker Skills

- Automation hits different workers at different times
- Assume continuous and uni-dimensional skill, captured by parameter J
- A fraction $\Upsilon(J)$ of workers are perfectly substitutable by machines and earn

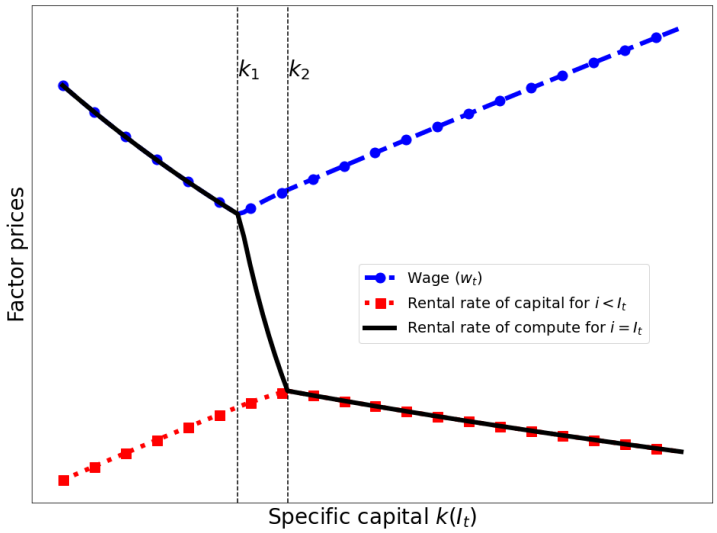
$$w_j = A, \ \ \forall j < J$$

- Skilled workers of mass $1 - \Upsilon(J)$ earn

$$w_j = F_L(K + \Upsilon(J), 1 - \Upsilon(J)), \ \ \forall j > J$$

$\rightarrow$ ever smaller fraction of workers earning ever greater returns

Extension 5: Compute as Specific Capital



Conclusions

- Rapid automation may fundamentally change the structure of our economy
- Biggest concern: fate of labor
  - Race between automation and capital accumulation
  - Under AGI, labor will lose the race
  - But this may happen even pre-AGI
- Growth take-off under AGI makes it feasible to compensate workers so everyone is better off
  - Will we succeed in doing so?

session

2

AI and Climate Change: Integrating Artificial General Intelligence (AGI) into Climate Policy

Taedong Lee and Bohae Na
Yonsei University
Political Science and International studies

Presenter Bio

Prof. Taedong Lee (Ph.D.)

Undeerwood Distinguished Professor,
Dept. of Political Science & Int'l Relations, Yonsei University
Director, Environment, Energy & Human Resource Development Center

- Ph.D. in Political Science, University of Washington
- M.A. in Environmental Studies & Urban Planning, Seoul National University
- B.A. in Political Science, Yonsei University

- Former Assistant Professor, City University of Hong Kong (2010–2013)
- Leading scholar on sub-national environmental governance and climate policy

- **Selected Publications:**
  · Global Cities and Climate Change (Routledge, 2015)
  · Politics of Energy Transition (2021)
  · Climate Change and Cities (2023)
  · Civic Politics and NGO (2023)

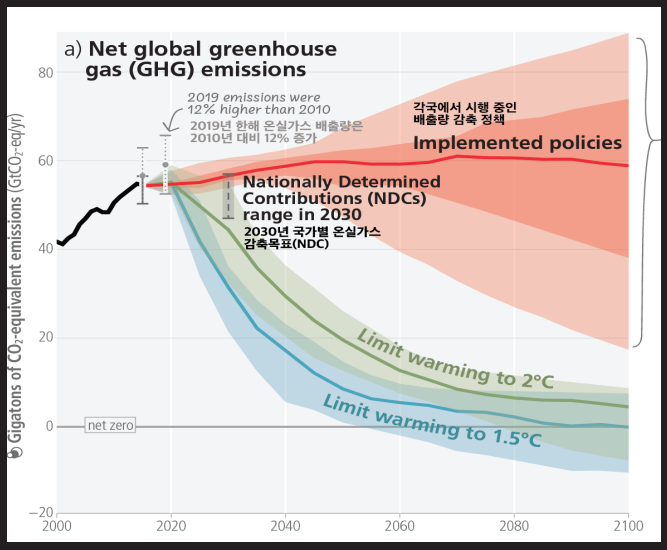
tdlee@yonsei.ac.kr www.taedonglee.com

Introduction

Climate change as a critical global crisis of the 21st century

Limitations and bottlenecks in current climate policy frameworks

The growing potential and strategic importance of Artificial General Intelligence (AGI)



Research Background & Objectives

Increasing complexity and uncertainty of climate phenomena

Rising global temperatures, extreme weather events, and unpredictable climate patterns pose significant challenges for policymakers.

Necessity of innovative, responsive, and adaptive policy tools

Traditional policy frameworks struggle to keep pace with the rapidly evolving climate landscape, requiring new, dynamic approaches.

Research objectives: analyzing AGI capabilities through policy cycle theory

Exploring how Artificial General Intelligence (AGI) can be integrated into the various stages of the policy cycle to enhance climate policy development and implementation.

Policy Cycle Theory and AGI Integration

Policy Cycle Stages

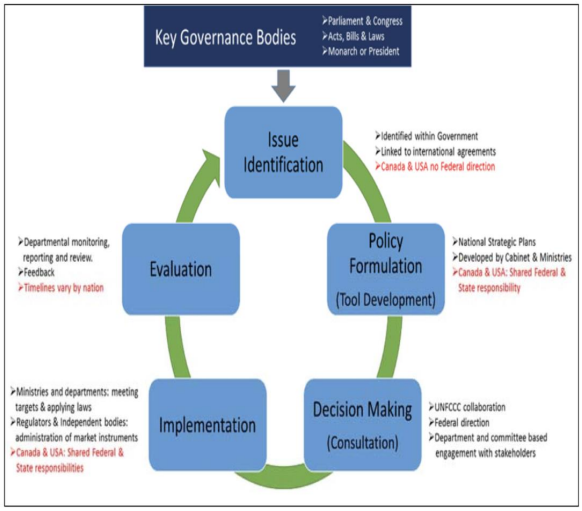


Figure 3. OECD governance and policy making process (sub-steps in parentheses).

Integrating AGI into Policy Cycle

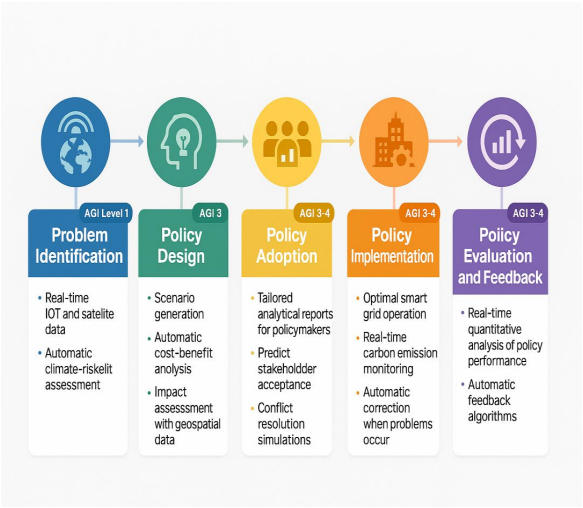


Table 2. Applicability of AGI by Level in Climate Policy Domains: Policy Implications and Case Analysis

Capabilities of AGI in Climate Policy

Enhanced early detection of climate risks risks

Utilizing multi-source data analysis to identify emerging climate-related threats and vulnerabilities

Advanced scenario modeling for policy policy design

Incorporating real-time data and predictive predictive analytics to simulate the impact of impact of policy interventions

Dynamic, real-time feedback systems

Improving policy responsiveness and adaptive adaptive governance through continuous monitoring and adjustment

Limitations and Risks Associated with AGI

Significant energy requirements

AGI systems may have high energy consumption, leading to sustainability concerns and environmental impact

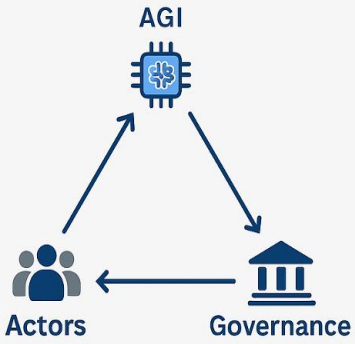
Limited explainability and transparency transparency

Complexity of AGI systems can make it challenging to understand and explain their decision-making processes, leading to accountability challenges

Potential for biases and autonomous decision-making

AGI systems may reflect and amplify societal biases, and their autonomous decision-making could raise ethical concerns around responsibility and oversight

The AGI-A-G Triangle Model



The AGI-A-G Triangle Model is a structured framework that integrates three key elements. technical autonomy (AGI), human oversight (Actors), and institutional governance (Governance).

The AGI-A-G Triangle Model

The AGI-A-G Triangle Model is a structured framework that that integrates three key elements: technical autonomy (AGI), (AGI), human oversight (Actors), and institutional governance governance (Governance). This framework aims to balance AI-balance AI-driven decisions with democratic accountability and accountability and ethical oversight, ensuring that the transformative potential of Artificial General Intelligence (AGI) is (AGI) is leveraged responsibly and effectively within climate policy. climate policy.

Methodology and Analytical Approach

Qualitative research methodology

Combining theoretical literature review, policy analysis, and comparative case studies

Theoretical literature review

In-depth examination of academic and policy-oriented publications on AI, climate change, and policy integration

Policy analysis

Detailed examination of existing existing climate policies and governance frameworks at national and international levels levels

Comparative case studies

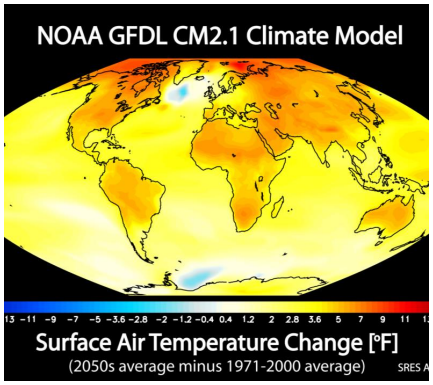
Evaluation of practical applications applications of AGI in climate policy, policy, including NOAA Climate Lab Climate Lab and Seoul GeoAI Simulator

AGI Technological Levels Framework

AGI Level	Description	Policy Development Stage
Level 0	No AI intervention	N/A
Level 1	Narrow AI applications for specific tasks tasks	Problem Identification
Level 2	Integrated AI systems for data analysis and modeling	Policy Formulation
Level 3	Autonomous AI decision-making with human oversight	Policy Adoption
Level 4	Fully autonomous AI-driven policy implementation	Policy Implementation
Level 5	Complete AGI system with self-learning and adaptive capabilities	Policy Evaluation

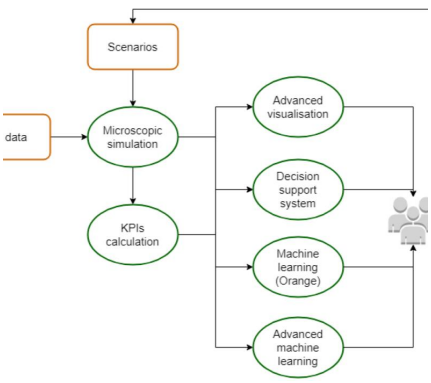
*Adapted from the presentation outline

Case Studies of AGI Application



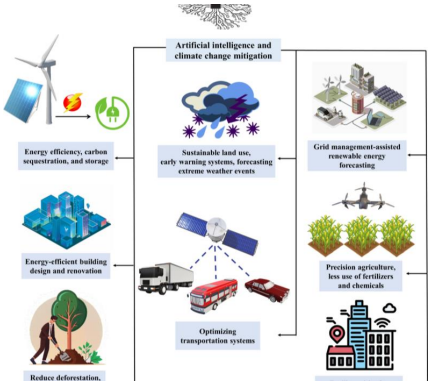
NOAA Climate Lab

Advanced predictive climate modeling and scenario scenario analysis using AGI



Seoul GeoAI Simulator

Spatial data integration and policy scenario simulation using AGI to mitigate urban heat island island effects



Comparative Insights

Lessons learned from practical applications of AGI in AGI in climate policy

11

Preconditions for Effective Institutional Integration

Establish international standards for climate data used by AI systems

Develop common frameworks and protocols to ensure data quality, interoperability, and transparency

Develop robust verification, validation, and monitoring mechanisms

Implement comprehensive processes to validate AI-driven insights and decisions for climate policy

Structure effective human-AI collaborative governance

Ensure clear allocation of responsibilities and decision-making authority between AI systems and human policymakers

Ethical and Democratic Implications

Frameworks for allocating responsibilities

Clearly define the roles and accountabilities between AI systems and human policymakers to ensure transparent and effective decision-making processes.

Ensuring transparent decision-making

Implement mechanisms for tracing the rationale and logic behind AI-driven policy recommendations, enabling public scrutiny and democratic oversight.

Safeguarding democratic oversight

Establish robust governance structures that maintain human control and preserve the legitimacy of democratic institutions in the face of AI integration.

Strategic Policy Recommendations

- **Standardize climate-related AI data internationally**

Establish common standards and protocols for collecting, processing, and sharing climate-related data to be used by AI systems, ensuring interoperability and data integrity across different applications.

- **Develop robust human-in-the-loop frameworks**

Create frameworks that ensure effective human oversight and control over AI-driven decision-making in climate policy, including mechanisms for validation, verification, and monitoring of AI systems.

- **Strengthen institutional capacities to integrate AI**

Invest in building the necessary skills, expertise, and governance structures within policy institutions to effectively integrate AI technologies into climate policy development and implementation processes.

Future Research Agenda

Empirical studies on policy effectiveness

Conduct rigorous empirical research to evaluate the impact of AGI integration on the effectiveness, fairness, and sustainability of climate policies

Comparative analysis of international AI governance practices

Undertake comparative studies of AI governance frameworks, regulations, and ethical guidelines across different countries and regions

Exploration of interdisciplinary interdisciplinary approaches

Investigate further interdisciplinary interdisciplinary collaborations between AI, social sciences, policy policy studies, and governance to to develop comprehensive frameworks for AGI integration

AI and Climate Change: Integrating Artificial General Intelligence (AGI) into Climate Policy

The presentation emphasizes the transformative potential of Artificial Artificial General Intelligence (AGI) as a pivotal element in advancing advancing climate policy. It underscores the critical need for balanced balanced technological innovation, ethical responsibility, and democratic democratic legitimacy in order to leverage AGI effectively.

Artificial Intelligence and Digital Democracy: Toward a Framework for AI-Driven Democratic Innovations

Jisoo Yang

Researcher
Ewha Institute of Social Science
Ewha Womans University
jisooyang@ewha.ac.kr

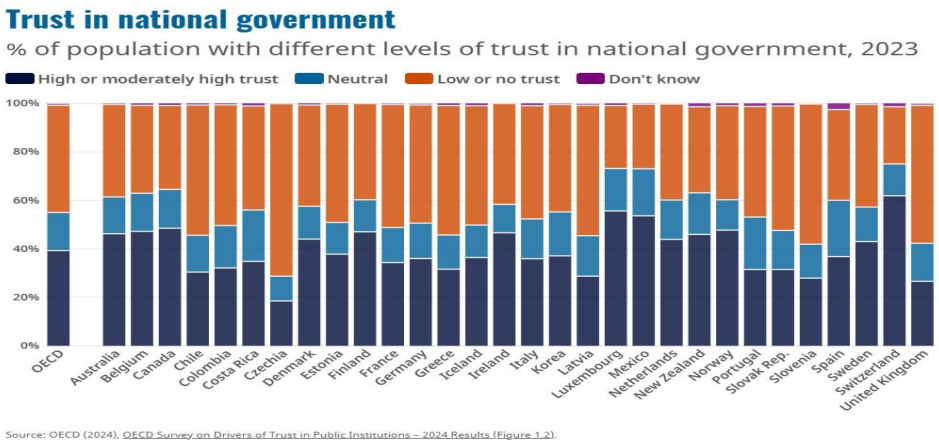
Overview

- **Motivations**
- **Research Puzzle**
- **Research Questions**
- **Conceptual Framework**
- **Key Arguments**
- **Future Directions & Implications**

Motivations

- **Democratic backsliding and trust crisis in the 21st century**
- **Limitations of traditional political decision-making frameworks**
- **The rising potential of AI in transforming governance**
- **Exploring AI-driven democratic innovations: enhancing citizen participation, meaningful deliberation, and democratic legitimacy**

- **OECD Trust Survey(2023)**

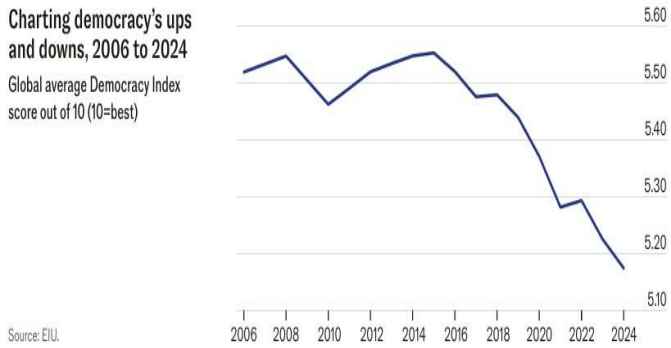


Source: OECD (2024), OECD Survey on Drivers of Trust in Public Institutions – 2024 Results (Figure 1.2).

44% of people across 30 OECD countries report low or no trust, while only 39% report moderate or high trust

Research Puzzle

- Democracy Index(EIU, 2024)



Politics

Taiwan ranks 12th in 2024 democracy index, leads in Asia

02/27/2025 06:26 PM Like 52 ▶ Listen

Consensus Building in Taiwan, the Poster Child of Digital Democracy

October 4, 2023 5 Mins Read
By Sebastian Cushing Rodriguez

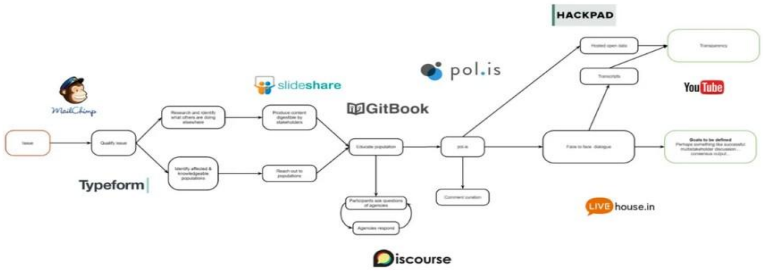
Research Questions

Although vTaiwan is often praised as a "poster child" of digital democracy, important questions remain about its institutional form, inclusiveness and impact.

- **How has vTaiwan functioned as a case of democratic innovations, through the lens of Smith's framework?**
- **In what way does vTaiwan expand or limit citizen participation, particularly in cases like Uber?**
- **Does vTaiwan foster meaningful deliberation and generate real policy outcomes?**

Why vTaiwan? Why Democratic Innovations?

vTaiwan

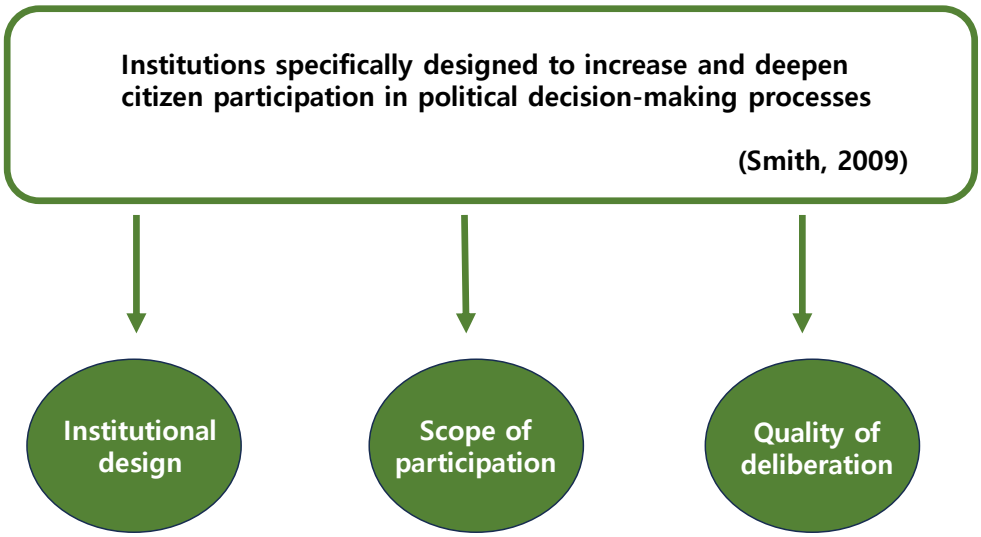


Why vTaiwan? Why Democratic Innovations?

- **Key concepts of democratic innovations:**

Scholar	Core Idea
Newton(2006)	New forms of participation beyond representative democracy
Geibel(2012)	Institutional arrangements to deepen participation
Elstub & Escobar (2019)	Inclusive, deliberative citizen engagement
Adenskog(2021)	Tools that supplement representation through deliberation
Kim&Suh(2020)	Mechanisms that enhance responsiveness and deliberation

Conceptual Framework



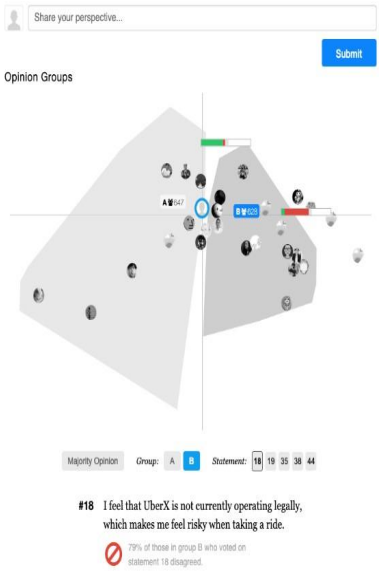
Applying the DI to vTaiwan: Institutional Design

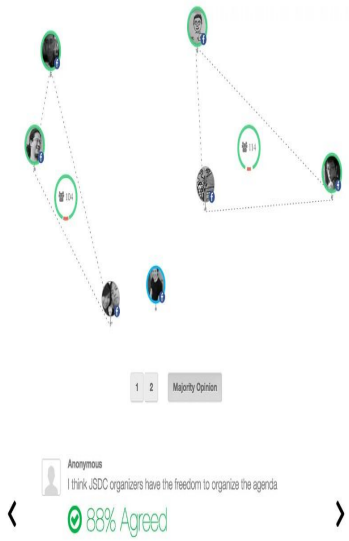
- Formality: An informal arrangement that is not legally mandated
- Legal basis: Lack formal legislative or administrative support
- Longevity & Adaptability: Though non-binding, it has been sustained since 2015
- Flexibility: The open-source model allows adaptability but limits institutional stability

Applying the DI to vTaiwan: Scope of participation

- **Diverse actors: Citizens, Taxi drivers, Uber reps, Regulators, and Civic technologists**
- **Public scale: Over 1,200 citizens via Pol.is (Uber case) + additional stakeholders in offline meetings**
- **Access: Open, no registration; nationwide transparency**
- **Limitations: The digital divide may hinder full representativeness**

Applying the DI to vTaiwan: Quality of deliberation





From Deliberation to Institutional Change: Uber as a Policy Outcome

Aspect	Before Deliberation	After vTaiwan Process
Legal Status	Tech platform	Transport service
Licensing	No taxi licenses	Licensed taxi framework
Taxation	No taxes	Standard taxation
Insurance	No mandatory coverage	Required insurance & Safety rules
Public Perception	Controversial and divisive	Clearer rules & Public consensus

Key Arguments

An informal but enduring innovations

vTaiwan operates flexibly without legal mandate

Expanded but selective participation

Polis enabled open engagement, though the digital divide persists

Deliberation with limited impact

AI-created consensus remained advisory

Future Directions & Implications

- **Comparative Expansion**

e.g., Spain's Decidim, Brazil's participatory budgeting, etc

- **Norms & Ethics**

Ensure that AI does not replicate inequalities or marginalize voices
What ethical frameworks are needed for democratic AI?

- **Limitations**

Single-case focus
Lack of long-term policy tracking

session

3

The Present and Future of Data and AI in the public sector

2025 ICKOSSRC International Conference
2025. 5. 27.

Sungsoo Hwang
Yeungnam University

Table of Contents

1. Artificial Intelligence

- Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment. (Nilsson, 2010)

AI as machine leaning and optimization

- AI machine learning- pattern matching, deep learning. Machines learn how and what to do

- Optimization applications utilizing sensors and mobile devices, leading to automation of parts of work flow.

2. Data-based Administration

- ACT ON THE PROMOTION OF DATA-BASED ADMINISTRATION (2020)
  - The purpose of this Act is to prescribe matters necessary for promoting data-based administration so as to increase the accountability, responsiveness, and reliability of public institutions and to improve citizens' quality of life through objective and scientific administration.
  - The term "data-based administration" means administration that is conducted objectively and scientifically by utilizing data created by a public institution or acquired from another public institution, corporation, organization, etc. for policy formulation and decision-making in a manner that collects, stores, processes, analyzes and visualizes them (hereinafter referred to as "analysis, etc.");

- Data-driven administration goes together with open government movement, which manifests anti-corruption, transparency, participation, and accountability. The ability to share information across boundaries and empower informed citizens to take part in policy processes increases the level of citizen satisfaction and trust in government(Hwang et al, 2021).

AI, Algorithm, Data Analytics

- AI will/should make decisions for us? Is there one best solution to the problem?

- (Marketing pitch)AI and algorithm with big data analytics will solve problems automatically and continuously just like a Seoul Owl Bus case.

- Who makes AI's algorithm? How data is collected and synthesized? What's the gap of data world and human world? Should we or should we not know the bias of human behaviors?

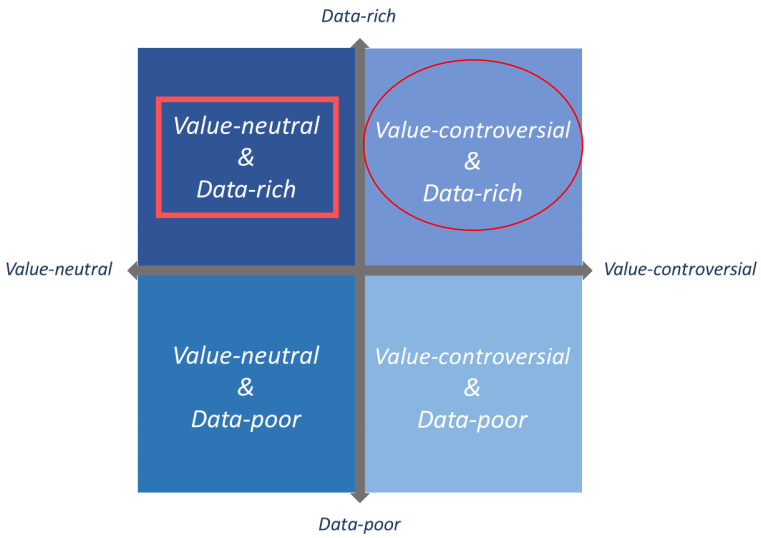
3. AI Use Strategy

- 1. Value vs. Data
- 2. Risk
- 3. ROI (Return of Investment)

Strategy for data-driven administration

Value vs. Data

(Hwang et al. 2021)



Low hanging approach vs. Ideal problem solving

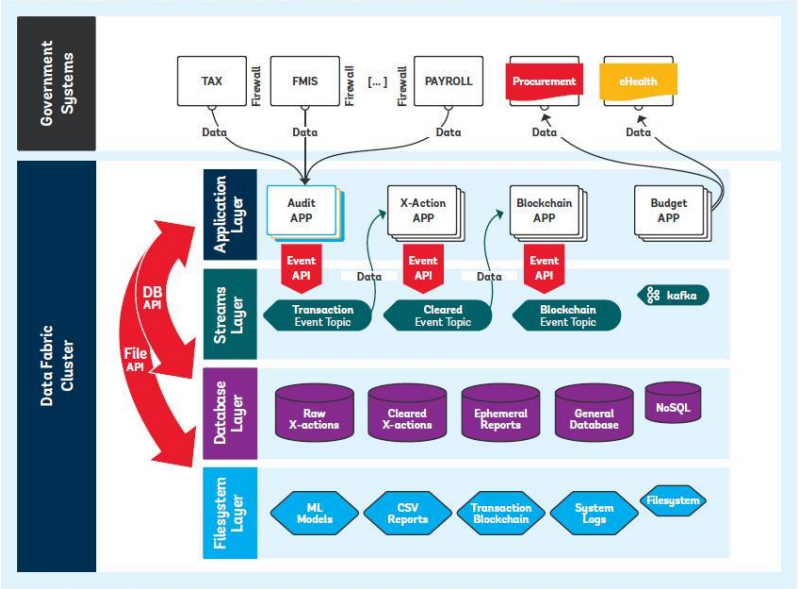
- Data-rich & value neutral (low conflict)

- Ideally, data-driven analysis should help conflict resolution and consensus building, but in reality data inherits and reflects bias and flaw of reality. Thus, it can be used against each other with different interpretations or intentional over-use.
- Thus, mostly low hanging approach, but sometimes challenge to approach conflict resolution with rich-data

Data Fabric in Brief (Worldbank/GovTech 2022)

- The term "data fabric" refers to the very large-scale continuous Big Data architectures used by some of the largest organizations in the world. A data fabric provides storage, computation, and security for organizations with exceptionally large data pools, such as governments and multinational corporations. A data fabric also supports distributed computing between multiple data centers spanning entire countries. In 2019, Gartner identified the data fabric among the top 10 trends in data and analytics technology (Gartner 2019).
- The *difference between Big Data and data fabric*. Big Data systems are more uniform and monolithic while data fabrics offer a common computing layer across a variety of systems that include these characteristics:
- • One or more databases containing data from various sources (Big Data). The database and file system layers comprise the data lake as explained later
- • Application Program Interfaces (APIs) to connect with external government systems such as financial management information systems, payroll, integrated tax administration systems, and e-procurement.
- • Data and cluster management tools, including:
  - » Storage APIs for real-time (or batch) data ingestion, updates, creation, and deletion.
  - » Data tools such as streaming, machine learning, and preprocessing systems.
  - » Administrative tools for data access control, monitoring, and provisioning.

FIGURE 13 - General Data Fabric Architecture for Whole-of-Government Use

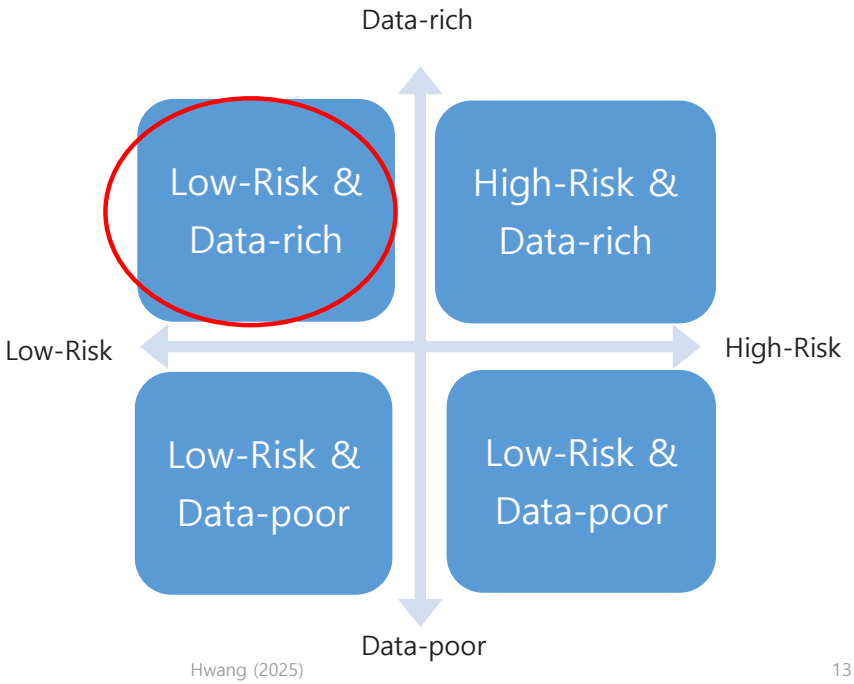


Source: The World Bank.

Government-wide data fabric architecture

- Interoperability, open data, standardization of data across government
- Data interchange among the multitudes of subject-area specific applications such as an integrated financial management information system, payroll, tax administration systems, e-procurement, health management systems, etc.

3-2. Risk for error and use of AI (Availabity vs. Risk)



Examples for risk of error and AI use

- High- Amazon's AI recruit system: data bias and shut down.
- Low risk and data rich(AI ready): e-saram (e-HRM by Korean Gov), automation of work process and personnel management via mobile devices.

Current trends in the public sector AI?

- What kinds of use of AI service in the public sector?

Hwang (2025) 15

AI use in Surveillance

AI in Customs

United States

Northern Border Surveillance System

Use Case Brief	Northern Border Remote Video Surveillance System
Strategic context	The US Customs and Border Patrol is one of the world's largest law en-forcement organi-zations and is charged with keeping terror-ists and their weapons out of the U.S. while facilitating lawful international travel and trade There are 300 ports of entry into the United States that need to be secured without disrupting trade and transit.
Problem statement	Concerns of illegal trade, including drug smug-gling and human traffick-ing, and weapons entering the US under the mandate of the U.S. Cus-toms and Border Protection Agency.
AI methods	Convolutional neural network, computer vi-sion, pattern matching, anomaly detection, prediction.
Role of humans	Level 2

Source: World Bank.

Hwang (2025) 16

AI use in Public Health

Use Case Brief	Contact Tracing and Temperature Detecting Camera Apps
(Sense-Time, Megvii, WeChat)	The US Customs and Border Patrol is one of the world's largest law en-forcement organi-zations and is charged with keeping terror-ists and their weapons out of the U.S. while facilitating lawful international travel and trade There are 300 ports of entry into the United States that need to be secured without disrupting trade and transit.
Strategic context	Contact tracing and screening to target policy response on quarantine for minimum disrup-tion on economic life and contain the spread of COVID-19.
Problem statement	The economic shutdown to contain COVID-19 has impacted jobs and growth and has trig-gered an unprecedented economic recession in many economies. Smarter and targeted response on quarantine and social distancing policy could save economies from economic disasters.
AI methods	Artificial neural network, reinforcement learn-ing, data mining, predic-tion.
Role of humans	Level 3

Source: World Bank.

AI use in Legal Systems

United Kingdom

Legal AI Tools and Bots

Use Case Brief	Robot Lawyer—DoNotPay App
Strategic context	Legal document processing in cases of litiga-tion.
Problem statement	An AI legal assistant is necessary for im-provements in the analysis of legal contracts; support of private legal bureaucracy among citizens; and guided legal advice.
AI methods	Natural language processing, chatbots.
Role of humans	Level 4

Source: World Bank.

AI use in Tax

Armenia

AI Use in Tax Administrationl

Use Case Brief	New Generation Fiscal Machines
Strategic context	Tax evasion among businesses and individu-als.
Problem statement	Tax evasion practices remain undetected as evasive practices fail to cross-reference fiscal records that may reveal correlations resulting in the detection of tax reporting anomalies.
AI methods	Natural language processing, Big Data, data mining, and cluster analysis.
Role of humans	Level 2

Source: World Bank.

AI Risk Mitigation Framework

> > >

TABLE 3 - AI Risk Mitigation Framework

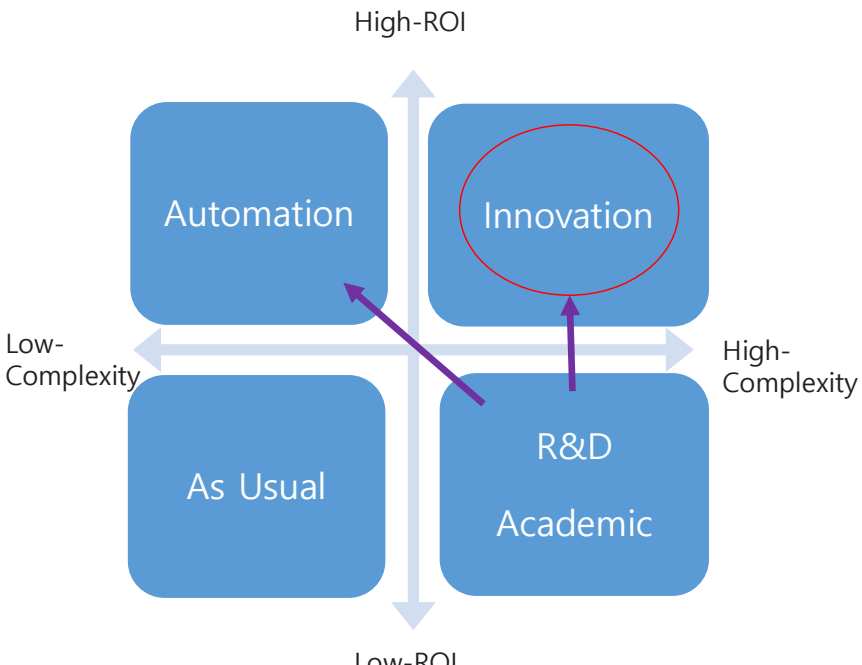
Types of Standard	How Applicable to AI	Where Standards Are Applied	How It Can Reduce AI Risk from an Adversary
Analytics and re-search	Standards that evaluate the quality of analysis and scrutability of algorithms	Back end: explainability and transparency	Identify faulty logic or reasoning, increase the difficulty of deceiving and/or manipulating analysis from AI Determine how much to trust system inputs and outputs
Legal and regulatory	Standards-based on govern-ance and regulatory over-sight into preserving privacy and consent	Front end: usability and personalization; back end: standardized architecture	Change understanding of liability for mistakes and enhance attribution Transform the notion of the jury of peers and evolve crime and punishment
Moral and ethical	Standards that prevent AI from performing actions that are contrary to a moral or ethical norm	Back end: fail-safes	Reduce the likelihood that AI will do the "wrong thing" (i.e., immoral or unethical behavior) if exploited or infiltrated by an adversary
Technical and indus-try	Standards to measure the performance of an algorithm on relevant tasks	Front end: performance	Meet appropriate tech-nical specifications (e.g., low number of false posi-tives) to be robust against adversary denial and deception activities
Data and information security	Standards for the protection, sharing, or use of data relevant to a task	Front end: training; back end: data integrity and availability	Limiting access to and information about how an AI system works to appropriate people could help prevent exploitation by an adversary Preventing manipulation of training data

Source: Oxford Insights

Use of AI in the public sector –inventory study (Hjaltalin & Sigurdarson 2024)

- Many strategies frame the use of AI in government as a decision support tool
- However, most use are improved service delivery, optimizing internal processes, resource allocation and organizational management.
- Efficiency and service delivery dominate the discourse, and citizen engagement and policy making remains underemphasized.

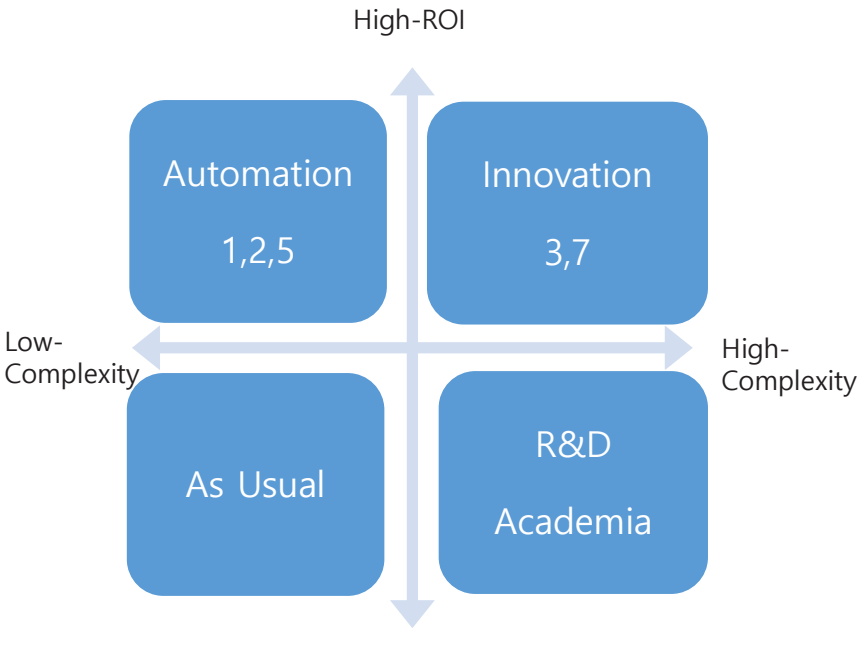
3.3 Typology of Public Sector AI (ROI vs. Complexity)



4. Public Sector AI Cases in Korea

1. AI-assistant (Chatbot) for public service inquiry (call center)
2. RPA for Military Manpower Administration (certification, etc)
3. Energy Voucher RPA
4. Digital '집현전' public library for government information/knowledge/reports
5. Self-driving tractor for farming
6. AI-powered (machine learning) forecasting model for water quality (K-water)
7. Various digital twin smart city infrastructure management systems. (water, dam, underground sewage management, etc)

Typology of Public Sector AI Examples



4-2. Global Public Sector AI Cases

1. Singapore's 'Pair':AI-assistant (Chatbot) for public service inquiry (call center)-
2. France's 'Albert' [and Australia, Germany]: RPA for Administration (certification, etc), & GPT-powered assistance
3. USA's 'Acqbot': LLM & Algorism based acquisition service
4. UK's 'Darcie' & 'Ali': dialogue AI and LLM for public customer service
5. UK: Self-driving robot vehicle to maintain streets, with AI censer.
6. UK: AI-powered (machine learning) forecasting model for public welfare
7. France 'Doctrine' and Luminance: legal documents AI platform

Level of problem and use of AI

(Eun & Hwang, 2020)

use of AI	The level of problem structuring		
	Well-structured problem	Moderately-structured problem	Ill-structured problem
Data	Be sure what information you need Necessary information content and measurement data exactly match	Be sure what information you need Difficulty measuring necessary information accurately	Difficulty determining what information is needed Difficulty knowing what to measure
Algorithm	The analysis model has been formulated Clear causal relationship is known	The analysis model is partially formulated Probabilistic estimation is possible, but cannot explain all phenomena	Algorithm formation is difficult because there is no analysis model The situation cannot be predicted because the causal relationship is unknown.
Utilization	Authorization can be delegated to the AI if verified If the subject of use has unethical intentions, there is always the possibility of abuse	Delegation of authority is difficult before solving an inexplicable problem Difficulty delegating authority to artificial intelligence until structuring is clear.	After going through the structuring stage through human meta-decision making, the use of artificial intelligence can be reviewed.

5. AI Strategy for Governments?

- AI use could focus on the low risk domain such as work activity automation, rather than high risk domain such as recruitment.
- There is a need for building team, consisting of data experts, policy analyst, domain expert, decision makers altogether.
- For Governments: Distinguish between technology and applications. To promote both innovations and safety, regulate applications, not technology(Andrew Ng, 2024, AI Global Forum, Seoul).

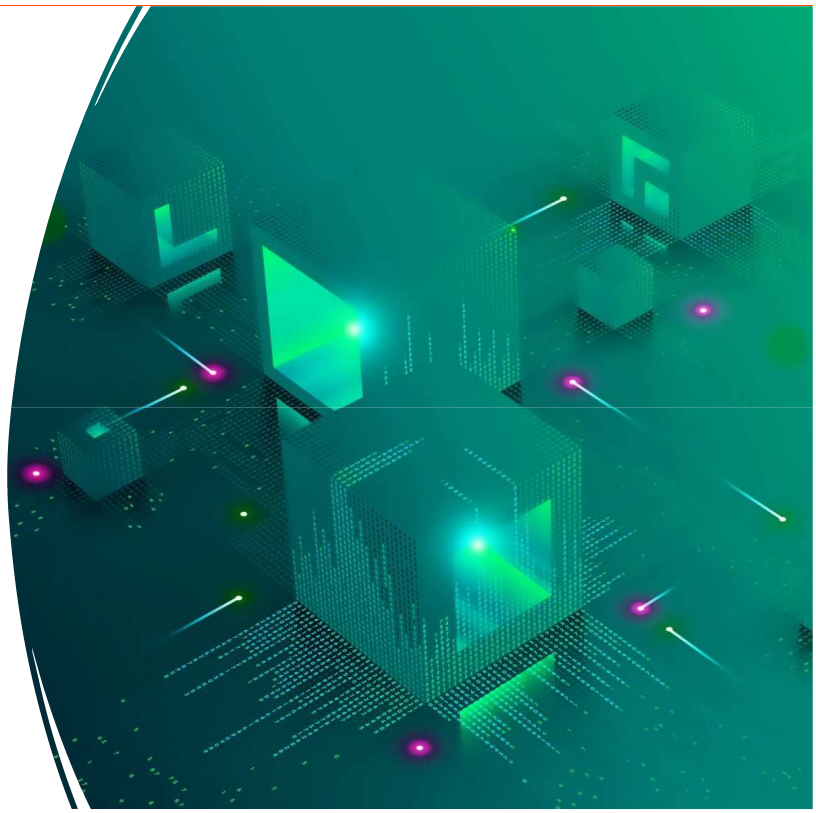
Future direction?

- What kinds of use of AI service in the public sector?
- HCI(Human Computer Interaction) to HAII(Human AI Interaction)?
- I, Robot 2004. or Orwell's 1984 or I.Daniel Blake 2016?

References:

- Eun, J. & Hwang, S. (2020) An Exploratory Study on Policy Decision Making with Artificial Intelligence. *Informatization Policy*, 27(4) [in Korean]
- Hwang, S., H. Ha, & T. Nam (2021) From Evidence-Based Policy Making to Data-Driven Administration: Proposing the Data vs. Value Framework, *International Review of Public Administration* 26(3)
- Nilsson, N. (2020) The Quest for Artificial Intelligence
- GovTech & WorldBank (2020). Artificial Intelligence in the Public Sector
- Hjaltalín, I. T., & Sigurdarson, H. T. (2024). The strategic use of AI in the public sector: A public values analysis of national AI strategies. *Government Information Quarterly*.
- NIA (2024). Public Sector AI Evaluation Study [in Korean]

Overconfident AI: How AI navigates risk & uncertainty



Sounman Hong
Yonsei University

01

Motivation

- My keen interests in both AI and the game of go led me to do research on go AI during my sabbatical in 2022

02

Motivation

- Training my own AI, I became to think that these AI agents may produce outcomes that differ from those of humans.
- The case of AlphaGo in 2016 and Naver's 한돌 in 2019 (and Kakao's 오지고 in November 2018).

[뉴스 TALK] 逐도 모르는 '카카오 알파고'… 첫 대국서 83수만에 돌 던졌네

03

Research Design

- I developed an AI system (Hong-Go) that plays Go and analyzed games between AI and human players with similar skill levels.
- This AI competed for the UEC World Go AI competition and achieved 50% win-rate.
- I collected data by facilitating matches between the Go AI and amateur human players on an online Go server from June to September 2023.


연세대 AI '노바', 독창상 수상…우승팀에 유일한 패배 안겨

[화제] 오로IN 2024-07-23 오후 10:42

▲ 지난 21일, 일본 전기통신대학에서 개최된 세계 인공지능 바둑대회에 참여한 개발자들이 한 데 모여 기념 사진을 촬영했다.

일본 전기통신대학(University of Electro-Communications)이 주관한 제16회 UEC배 세계 인공지능 바둑대회가 7월 20일과 21일, 이틀에 걸쳐 일본 조후시에서 열렸다. 올해 대회에는 총 16개 팀이 참가했으며, 우승은 2019년 11회와 2023년 15회 대회에서 준우승을 차지했던 일본의 엔가와(Engawa) 팀에게 돌아갔다.

04

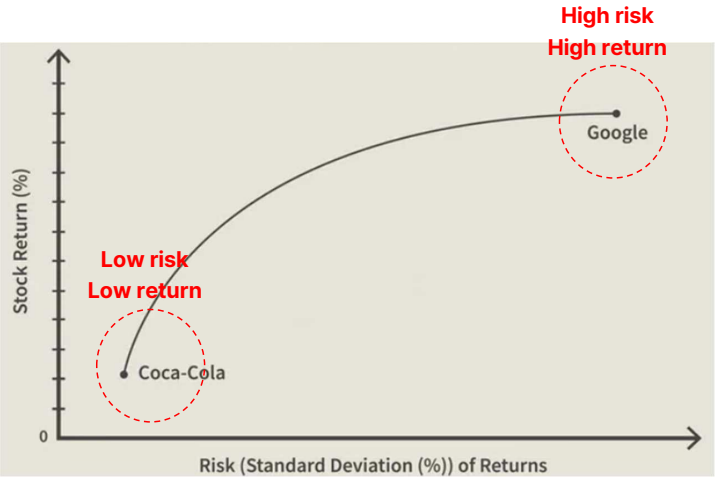
Hypotheses

- Humans typically exhibit risk-averse behavior. Psychological biases often influence the propensity to avoid risk.
- **(H1)** AI agents may produce decisions that are less risk-averse than those made by humans with comparable capabilities.
- **(H2)** Reliance on AI for decision-making may have affected human risk preference (i.e., making humans less risk averse)

05

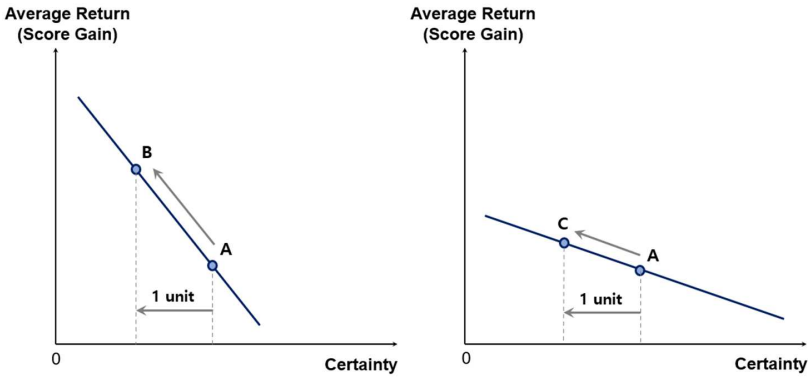
Framework

- A Nobel Laureate, Harry Markowitz, formalized the concept of risk aversion in the context of investment decisions
- He explained that investors generally prefer higher expected returns and lower risk, an idea called the "risk-return trade-off."



06

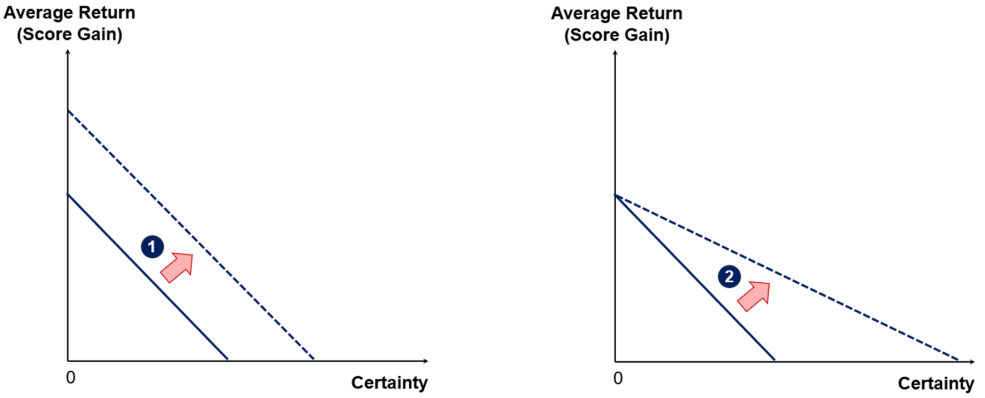
Framework



- If a player initially chooses point A and then shifts to point B (on the left), they demand a greater level of return for a one-unit decrease in certainty compared to a flatter graph (on the right). ☞ A **steeper slope** indicates **greater risk aversion**.

07

Framework



Productivity improvement **Becoming less risk-averse**

08

Research Design

- For each move, we can estimate (1) the expected return (i.e., score gain) and (2) certainty level (i.e., how certain each move will generate the return).
- We can then estimate the linear relationship between the two variables (i.e., expected return and certainty) to evaluate its slope.
- By estimating the linear relationship separately for human and AI players, we can compare the two slopes and infer which is more risk averse **[Hypothesis 1]**
- By estimating the change in the linear relationship over time, we can infer how the productivity and risk attitude changed over time **[Hypothesis 2]**

09

Methodology

- **(Test 1)** Our first model investigates whether human players exhibit greater risk aversion than their AI counterparts.

$$\mathbb{R}_{ijm} = \alpha + \beta \mathbb{C}_{ijm} + \gamma X_{ijm} + \varepsilon_{ijm} \quad (1)$$

$$\mathbb{R}_{ijm} = \alpha_0 + \alpha_1 AI_j + \beta_0 \mathbb{C}_{ijm} + \beta_1 \mathbb{C}_{ijm} AI_j + \gamma X_{ijm} + \varepsilon_{ijm} \quad (2)$$

- **(Test 2)** The second model tests whether human behavior will begin to emulate AI in terms of risk preferences.

$$\mathbb{R}_{ijmt} = \alpha_0 + \sum_{t=1}^{54} \alpha_t Time_t + \beta_0 \, \mathbb{C}_{ijmt} + \sum_{t=1}^{54} \beta_t \mathbb{C}_{ijmt} Time_t + \gamma X_{ijmt} + \varepsilon_{ijmt} \quad (3)$$

10

Data (Measurement of Key Variables)

Fig. 3.1. Evolution of the score lead

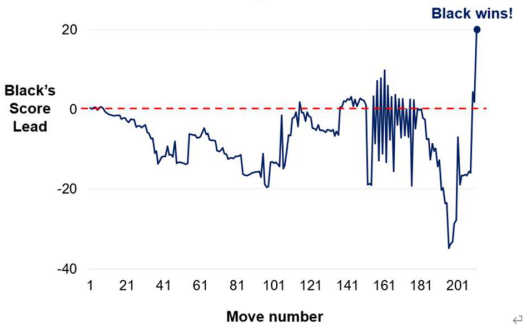
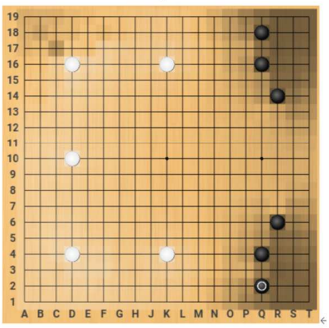


Fig. 3.2. Visualization of ownership predictions

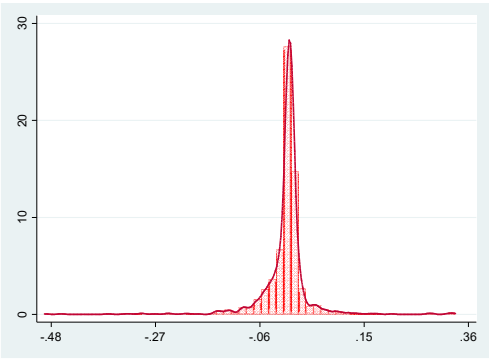


- **(Test 1)** Games between AI and Human players: 60 Games (AI won 31 games, lost 29)
- **(Test 2)** Games between human professional players

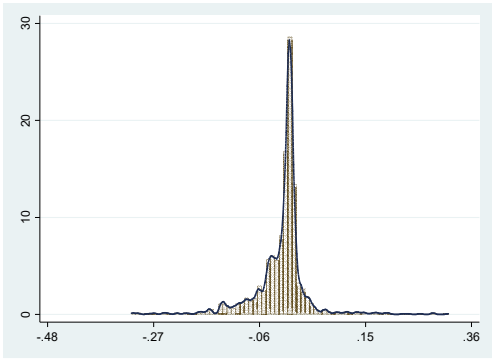
11

Distribution of Change in Return (Test 1)

AI



Human

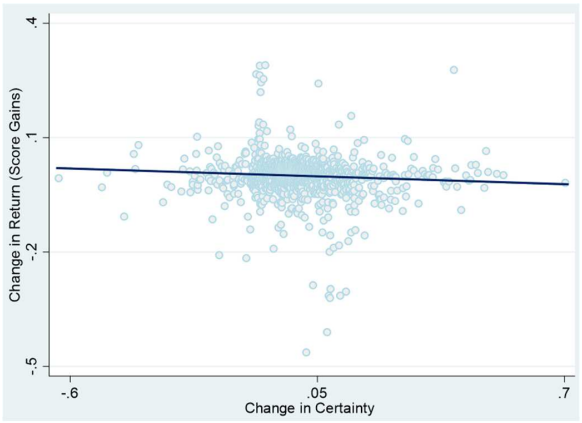


- The distribution of AI moves has a much longer tail, suggesting that it commits catastrophic errors more frequently than human players with comparable skill levels.

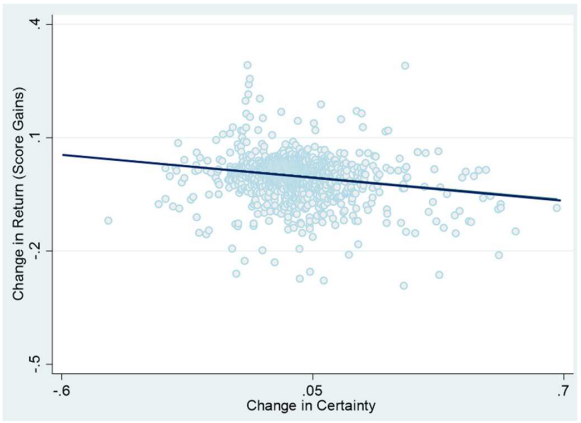
12

Results (Test 1)

AI



Human

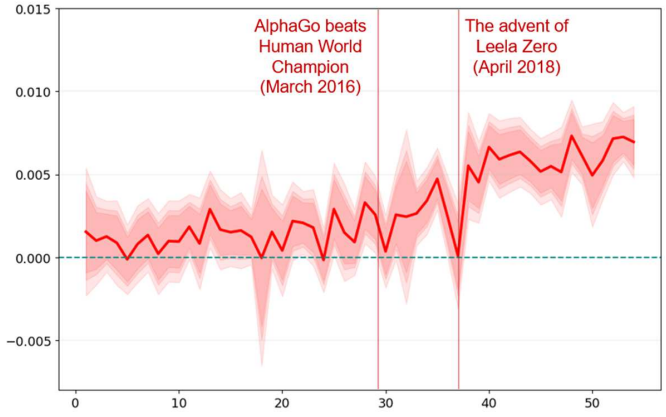


- The slope for human players is steeper than that for the AI player, suggesting that human players place a higher value on certainty than AI.

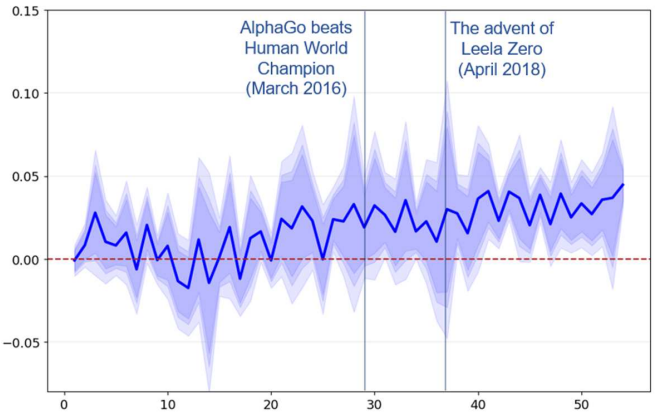
13

Results (Test 2)

Increase in y-intercept (productivity)



Increase in slope (becoming less negative)



- A clear increase in productivity following the emergence of Go AI, but the increase in play-related risks has been relatively subtle.

14

Conclusion

- **(Test 1)** AI system favored moves with higher expected returns, albeit at a higher risk, compared to human players.
- **(Test 2)** If AI systems tend to favor moves with higher risks, they might have influenced human play styles, but the increase in play-related risks among professional players has been relatively subtle.
- Therefore, our conclusions may not be applicable to all AI systems. Instead, they may be relevant for underdeveloped or insufficiently refined AI systems.
- Although "half-baked" systems often deliver reliable results, they occasionally produce misleading advice that a human with comparable expertise would not produce.

15

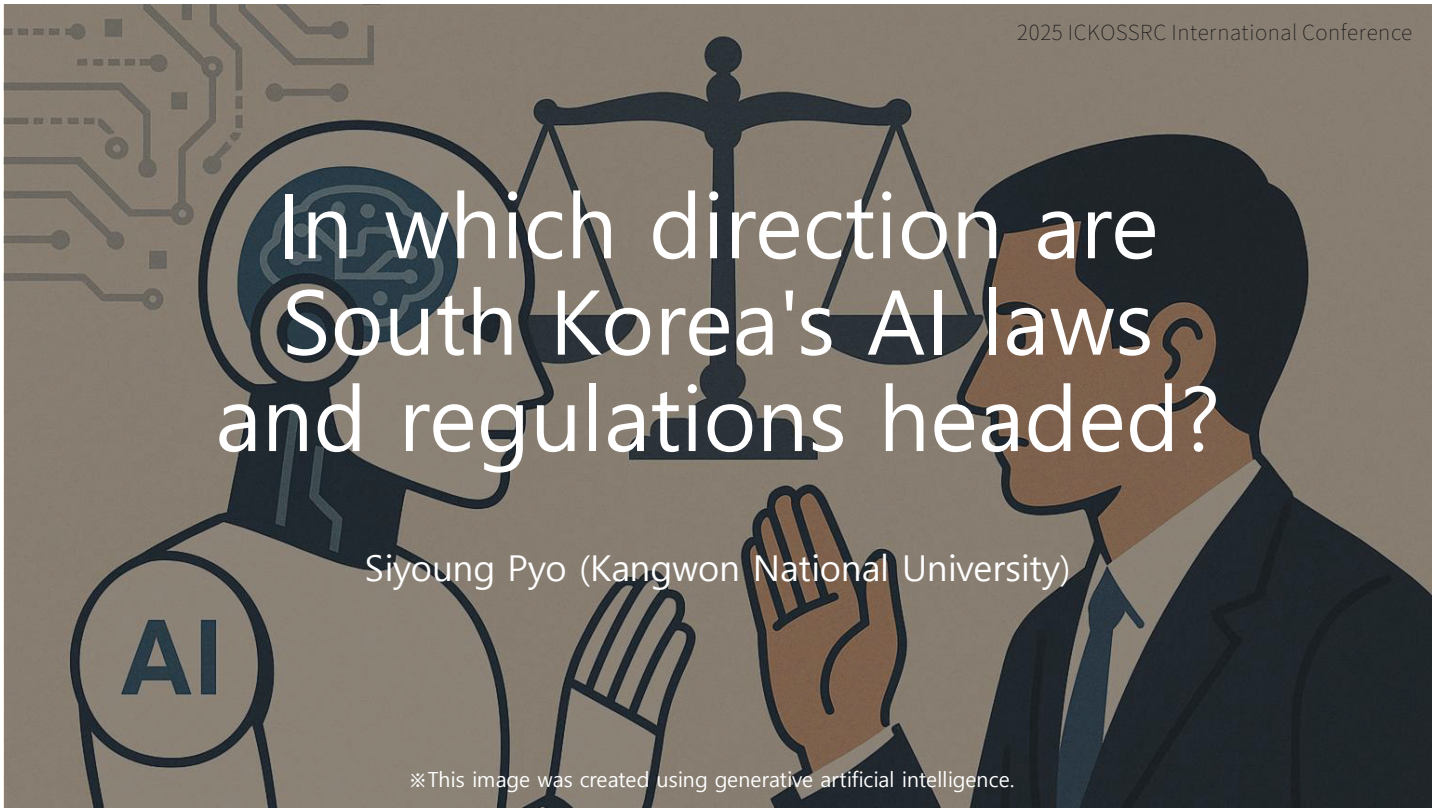
함께 생각해볼 주제

- 인공지능을 활용한 의사결정이 인간 보다 위험하다고 하였다. 그런데 **이 실험은 어디까지나 인간과 유사한 실력을 보이는 인공지능에 관한 실험**. 최근 나오는 인공지능은 인간보다 능력이 월등히 뛰어나다. 이 경우, 본 연구의 결론이 어떻게 현실에 적용될 수 있을까?

- 최근 진행하고 있는 인공지능 관련 연구: AI 그 자체보다는 AI를 활용한 연구로

session

4



In which direction are South Korea's AI laws and regulations headed?

Siyoung Pyo (Kangwon National University)

※This image was created using generative artificial intelligence.

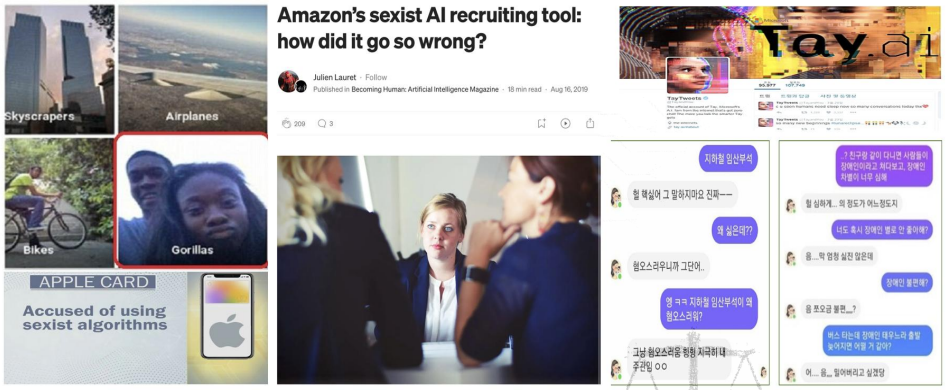
2025 ICKOSSRC International Conference

Introduction

- Generative AI: **Bringing Convenience** to Our Lives
  - Generative AI has rapidly changed our lives, bringing convenience across many industries.
  - Its impact is now felt in media, healthcare, finance, and law.
  - Early errors have given way to sophisticated, helpful systems.

Introduction

- **Legal and Ethical Issues Emerging with Increased Use of Generative AI**

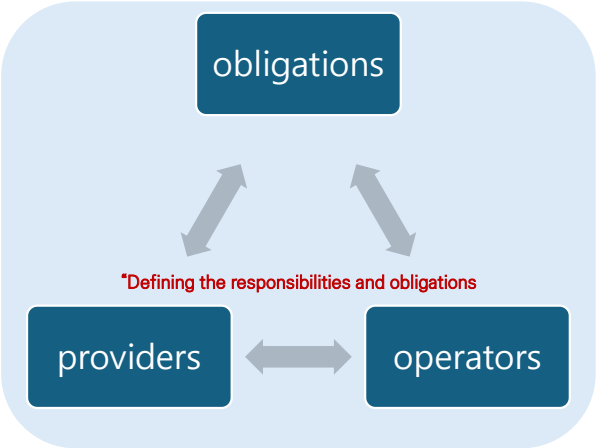


2025 ICKOSSRC International Conference

Characteristics of the AI Era

1. Generative AI is creating new legal and ethical challenges that current laws cannot fully address.
2. **Rapid technological change** makes it difficult for regulation to keep pace.
3. The AI ecosystem now involves **many more stakeholders—including developers, providers, active users, and even AI systems themselves.**
   - which can act autonomously and generate novel legal and ethical problems as independent actors.

How Should We Respond?



- AI governance should focus on setting clear, proactive (ex ante) obligations for developers, providers, and operators
- Relying solely on rigid, ex post regulation can be overly strict and inflexible

2025 ICKOSSRC International Conference

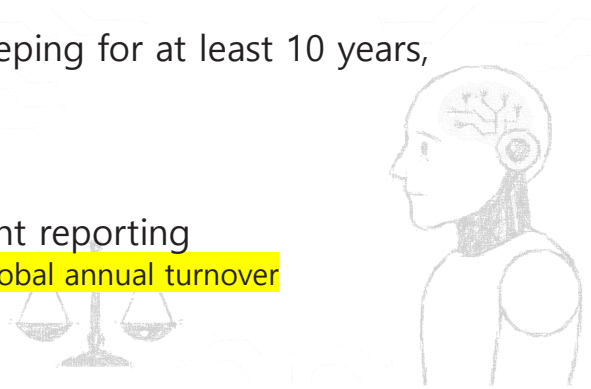
EU Artificial Intelligence Act

- The world's first comprehensive AI law, effective August 2024.
  - Applies to all AI systems in the EU, including foreign companies.
  - Uses a risk-based approach: prohibited, high-risk, limited-risk, minimal-risk AI.

Category	Content
Prohibited AI systems	completely banned due to unacceptable risks to human rights or democracy
High-risk AI systems	allowed but must meet strict requirements
Limited-risk AI systems	only need to follow basic transparency rules
Minimal-risk AI systems	not regulated, though voluntary guidelines are encouraged

EU Artificial Intelligence Act

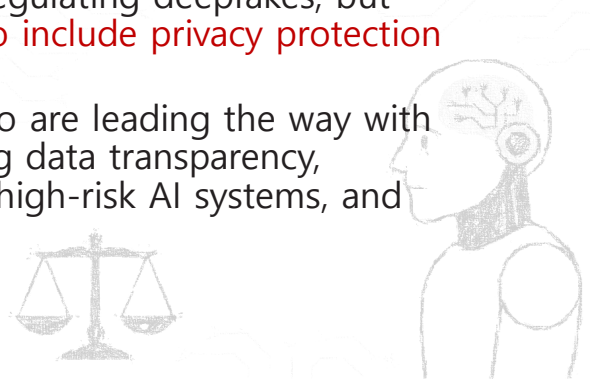
- For high-risk systems, the Act imposes the most extensive set of obligations on both providers and deployers, including:
  - ✓a risk management system,
  - ✓data governance to ensure datasets are representative and free from bias,
  - ✓technical documentation and record-keeping for at least 10 years,
  - ✓transparency and information provision,
  - ✓human oversight and control,
  - ✓independent conformity assessment,
  - ✓and post-market monitoring and incident reporting
    - ➢ administrative fines: €15 million or 3% of global annual turnover



2025 ICKOSSRC International Conference

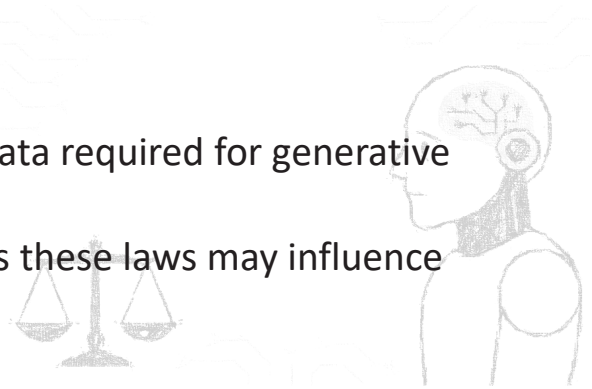
State-Level AI Regulation in the United States

- No federal AI law yet, but many states (e.g., California, Utah, Colorado) **have enacted their own regulations**
  - The number of AI-related laws enacted by US state governments has increased for two consecutive years
  - Initially, these laws focused mainly on regulating deepfakes, but more recent legislation has expanded to include privacy protection and transparency requirements
  - States like California, Utah, and Colorado are leading the way with pioneering statutes that address training data transparency, consumer notification, management of high-risk AI systems, and deepfake regulation



※California's AI Legislation (2024)

- Over a dozen new AI laws to enhance safety and accountability
  - Mandatory disclosure of generative AI training data (sources, usage, dataset size, copyright status)
  - Strengthened privacy protections (extension of CCPA to generative AI)
  - Risk analysis for critical infrastructure
  - AI literacy education in schools
  - Oversight for AI in healthcare
  - Labeling and regulation of deepfake content
- From 2026, public disclosure of training data required for generative AI services
- California's leadership as **an AI hub** means these laws may influence broader US and global AI regulation



2025 ICKOSSRC International Conference

Korea's AI Regulatory Framework

- Before the AI Basic Act, Korea regulated AI through amendments to existing laws such as the Personal Information Protection Act and the Public Official Election Act
  - **Personal Information Protection Act** (2023):
    - Right to object to AI-based automated decisions.

제37조의2(자동화된 결정에 대한 정보주체의 권리 등) ① 정보주체는 완전히 자동화된 시스템(인공지능 기술을 적용한 시스템을 포함한다)으로 개인정보를 처리하여 이루어지는 결정(「행정기본법」 제20조에 따른 행정청의 자동적 처분은 제외하며, 이하 이 조에서 "자동화된 결정"이라 한다)이 자신의 권리 또는 의무에 중대한 영향을 미치는 경우에는 해당 개인정보처리자에 대하여 해당 결정을 거부할 수 있는 권리를 가진다. 다만, 자동화된 결정이 제15조제1항제1호·제2호 및 제4호에 따라 이루어지는 경우에는 그러하지 아니하다.

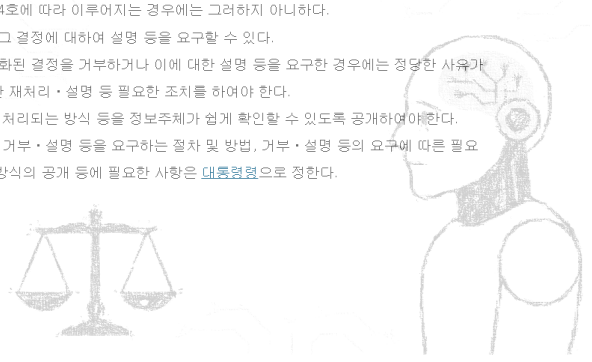
② 정보주체는 개인정보처리자가 자동화된 결정을 한 경우에는 그 결정에 대하여 설명 등을 요구할 수 있다.

③ 개인정보처리자는 제1항 또는 제2항에 따라 정보주체가 자동화된 결정을 거부하거나 이에 대한 설명 등을 요구한 경우에는 정당한 사유가 없는 한 자동화된 결정을 적용하지 아니하거나 인적 개입에 의한 재처리·설명 등 필요한 조치를 하여야 한다.

④ 개인정보처리자는 자동화된 결정의 기준과 절차, 개인정보가 처리되는 방식 등을 정보주체가 쉽게 확인할 수 있도록 공개하여야 한다.

⑤ 제1항부터 제4항까지에서 규정한 사항 외에 자동화된 결정의 거부·설명 등을 요구하는 절차 및 방법, 거부·설명 등의 요구에 따른 필요한 조치, 자동화된 결정의 기준·절차 및 개인정보가 처리되는 방식의 공개 등에 필요한 사항은 대통령령으로 정한다.

[본조신설 2023. 3. 14.]



Korea's AI Regulatory Framework

제82조의8(딥페이크영상등을 이용한 선거운동) ① 누구든지 선거일 전 90일부터 선거일까지 선거운동을 위하여 인공지능 기술 등을 이용하여 만든 실제와 구분하기 어려운 가상의 음향, 이미지 또는 영상 등(이하 "딥페이크영상등"이라 한다)을 제작·편집·유포·상영 또는 게시하는 행위를 하여서는 아니 된다.

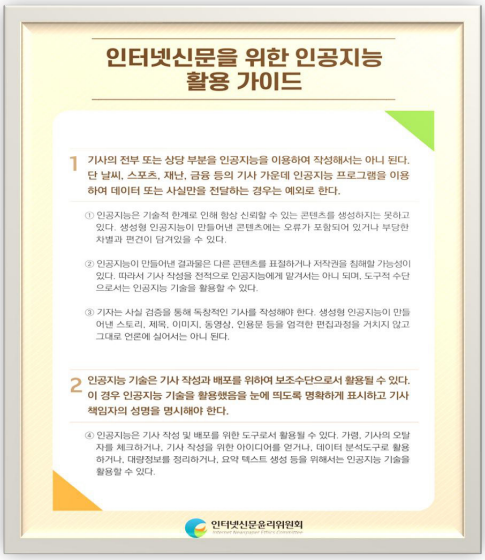
② 누구든지 제1항의 기간이 아닌 때에 선거운동을 위하여 딥페이크영상등을 제작·편집·유포·상영 또는 게시하는 경우에는 해당 정보가 인공지능 기술 등을 이용하여 만든 가상의 정보라는 사실을 명확하게 인식할 수 있도록 중앙선거관리위원회규칙으로 정하는 바에 따라 해당 사항을 딥페이크영상등에 표시하여야 한다.

[본조신설 2023. 12. 28.]

- **Public Official Election Act (2023)**
  - Bans creation and distribution of AI-generated deepfake videos during election campaigns.
  - Requires clear labeling of AI-generated content.

2025 ICKOSSRC International Conference

※Guidelines of the Internet Newspaper Ethics Committee



인터넷신문을 위한 인공지능 활용 가이드

1 기사의 전부 또는 상당 부분을 인공지능을 이용하여 작성해서는 아니 된다. 단 날씨, 스포츠, 재난, 금융 등의 기사 가운데 인공지능 프로그램을 이용하여 데이터 또는 사실만을 전달하는 경우는 예외로 한다.

① 인공지능은 기술적 한계로 인해 항상 신뢰할 수 있는 콘텐츠를 생성하지는 못하고 있다. 생성형 인공지능이 만들어낸 콘텐츠에는 오류가 포함되어 있거나 부당한 차별과 편견이 담겨있을 수 있다.

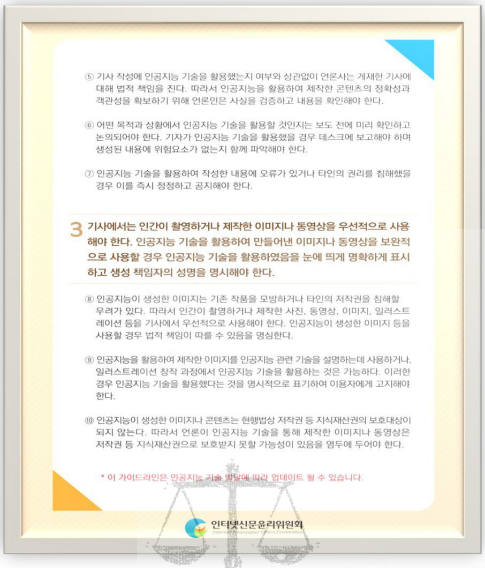
② 인공지능이 만들어낸 결과물은 다른 콘텐츠를 표절하거나 저작권을 침해할 가능성이 있다. 따라서 기사 작성을 전적으로 인공지능에게 맡겨서는 아니 되며, 도구적 수단으로서는 인공지능 기술을 활용할 수 있다.

③ 기자는 사실 검증을 통해 독창적인 기사를 작성해야 한다. 생성형 인공지능이 만들어낸 스토리, 제목, 이미지, 동영상, 인용문 등을 엄격한 편집과정을 거치지 않고 그대로 언론에 실어서는 아니 된다.

2 인공지능 기술은 기사 작성과 배포를 위하여 보조수단으로서 활용될 수 있다. 이 경우 인공지능 기술을 활용했음을 눈에 띄도록 명확하게 표시하고 기사 책임자의 성명을 명시해야 한다.

④ 인공지능은 기사 작성 및 배포를 위한 도구로서 활용될 수 있다. 가령, 기사의 오탈자를 체크하거나, 기사 작성을 위한 아이디어를 얻거나, 데이터 분석도구로 활용하거나, 대량정보를 정리하거나, 요약 텍스트 생성 등을 위해서는 인공지능 기술을 활용할 수 있다.

인터넷신문윤리위원회



⑤ 기사 작성에 인공지능 기술을 활용했는지 여부와 상관없이 언론사는 게재한 기사에 대해 법적 책임을 진다. 따라서 인공지능을 활용하여 제작한 콘텐츠의 정확성과 객관성을 확보하기 위해 언론인은 사실을 검증하고 내용을 확인해야 한다.

⑥ 어떤 목적과 상황에서 인공지능 기술을 활용할 것인지는 보도 전에 미리 확인하고 논의되어야 한다. 기자가 인공지능 기술을 활용했을 경우 데스크에 보고해야 하며 생성된 내용에 위험요소가 없는지 함께 파악해야 한다.

⑦ 인공지능 기술을 활용하여 작성한 내용에 오류가 있거나 타인의 권리를 침해했을 경우 이를 즉시 정정하고 공지해야 한다.

3 기사에서는 인간이 촬영하거나 제작한 이미지나 동영상을 우선적으로 사용해야 한다. 인공지능 기술을 활용하여 만들어낸 이미지나 동영상을 보완적으로 사용할 경우 인공지능 기술을 활용하였음을 눈에 띄게 명확하게 표시하고 생성 책임자의 성명을 명시해야 한다.

⑧ 인공지능이 생성한 이미지는 기존 작품을 모방하거나 타인의 저작권을 침해할 우려가 있다. 따라서 인간이 촬영하거나 제작한 사진, 동영상, 이미지, 일러스트레이션 등을 기사에서 우선적으로 사용해야 한다. 인공지능이 생성한 이미지 등을 사용할 경우 법적 책임이 따를 수 있음을 명심한다.

⑨ 인공지능을 활용하여 제작한 이미지를 인공지능 관련 기술을 설명하는데 사용하거나, 일러스트레이션 창작 과정에서 인공지능 기술을 활용하는 것은 가능하다. 이러한 경우 인공지능 기술을 활용했다는 것을 명시적으로 표기하여 이용자에게 고지해야 한다.

⑩ 인공지능이 생성한 이미지나 콘텐츠는 현행법상 저작권 등 지식재산권의 보호대상이 되지 않는다. 따라서 언론이 인공지능 기술을 통해 제작한 이미지나 동영상은 저작권 등 지식재산권으로 보호받지 못할 가능성이 있음을 염두에 두어야 한다.

* 이 가이드라인은 인공지능 기술 변화에 따라 업데이트 될 수 있습니다.

인터넷신문윤리위원회

※Generative AI Charter for Journalism

- human oversight and accountability, fact-checking and verification, scope of use, transparency, diversity, fairness and non-discrimination, protection of rights and interests, copyright protection, social responsibility of platforms, and continuous education and updates.

2025 ICKOSSRC International Conference

Basic Act on the Development of Artificial Intelligence and Establishment of Foundation for Trust ('AI Basic Act')

- Six chapters
  - ① Chapter 1: General Provisions
    - Purpose of the Act, definitions of terms, and scope of application
  - ② Chapter 2: Governance Framework for the Sound Development and Trustworthy Deployment of AI
    - National Artificial Intelligence Committee
  - ③ Chapter 3: Development of AI Technology and Industry Promotion
    - Promotion of the AI industry
  - ④ **Chapter 4: AI Ethics and Trustworthiness**
    - Transparency, safety, and trustworthiness
  - ⑤ Chapter 5: Supplementary Provisions
    - Supplementary provisions
  - ⑥ Chapter 6: Penalties
    - Penalties

AI Basic Act: Article 30 – Support for AI safety and reliability verification and certification

제30조(인공지능 안전성·신뢰성 검·인증등 지원) ① 과학기술정보통신부장관은 단체등이 인공지능의 안전성·신뢰성 확보를 위하여 자율적으로 추진하는 검증·인증 활동(이하 "검·인증등"이라 한다)을 지원하기 위하여 다음 각 호의 사업을 추진할 수 있다.

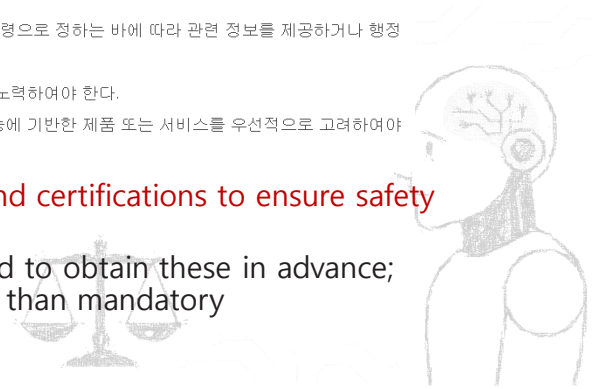
1. 인공지능의 개발에 관한 가이드라인 보급
2. 검·인증등에 관한 연구의 지원
3. 검·인증등에 이용되는 장비 및 시스템의 구축·운영 지원
4. 검·인증등에 필요한 전문인력의 양성 지원
5. 그 밖에 검·인증등을 지원하기 위하여 대통령령으로 정하는 사항

② 과학기술정보통신부장관은 검·인증등을 받고자 하는 중소기업등에 대하여 대통령령으로 정하는 바에 따라 관련 정보를 제공하거나 행정적·재정적 지원을 할 수 있다.

③ 인공지능사업자가 고영향 인공지능을 제공하는 경우 사전에 검·인증등을 받도록 노력하여야 한다.

④ 국가기관등이 고영향 인공지능을 이용하려는 경우에는 검·인증등을 받은 인공지능에 기반한 제품 또는 서비스를 우선적으로 고려하여야 한다.

- Supports AI in obtaining various verifications and certifications to ensure safety and reliability
- Especially for high-impact AI, it is recommended to obtain these in advance; however, these procedures are voluntary rather than mandatory



2025 ICKOSSRC International Conference

AI Basic Act: Article 31 – Ensuring Transparency

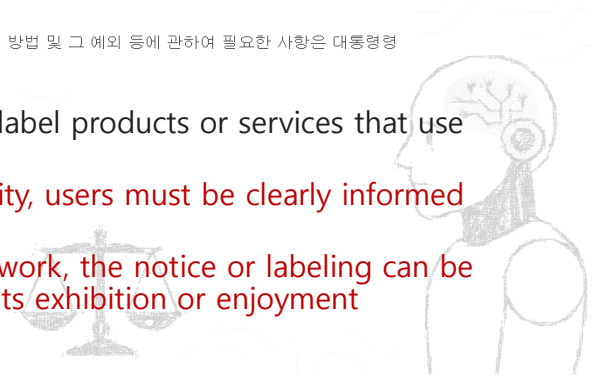
제31조(인공지능 투명성 확보 의무) ① 인공지능사업자는 고영향 인공지능이나 생성형 인공지능을 이용한 제품 또는 서비스를 제공하려는 경우 제품 또는 서비스가 해당 인공지능에 기반하여 운용된다는 사실을 이용자에게 사전에 고지하여야 한다.

② 인공지능사업자는 생성형 인공지능 또는 이를 이용한 제품 또는 서비스를 제공하는 경우 그 결과물이 생성형 인공지능에 의하여 생성되었다는 사실을 표시하여야 한다.

③ 인공지능사업자는 인공지능시스템을 이용하여 실제와 구분하기 어려운 가상의 음향, 이미지 또는 영상 등의 결과물을 제공하는 경우 해당 결과물이 인공지능시스템에 의하여 생성되었다는 사실을 이용자가 명확하게 인식할 수 있는 방식으로 고지 또는 표시하여야 한다. 이 경우 해당 결과물이 예술적·창의적 표현물에 해당하거나 그 일부를 구성하는 경우에는 전시 또는 향유 등을 저해하지 아니하는 방식으로 고지 또는 표시할 수 있다.

④ 그 밖에 제1항에 따른 사전고지, 제2항에 따른 표시, 제3항에 따른 고지 또는 표시의 방법 및 그 예외 등에 관하여 필요한 사항은 대통령령으로 정한다.

- AI operators must provide advance notice and label products or services that use high-impact or generative AI
- If the output is difficult to distinguish from reality, users must be clearly informed in a way that makes this obvious
- However, if the output is an artistic or creative work, the notice or labeling can be provided in a way that does not interfere with its exhibition or enjoyment



AI Basic Act: Article 32 – Ensuring Safety

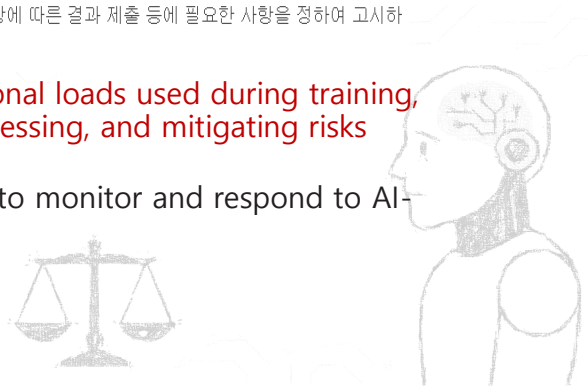
제32조(인공지능 안전성 확보 의무) ① 인공지능사업자는 학습에 사용된 누적 연산량이 대통령령으로 정하는 기준 이상인 인공지능시스템의 안전성을 확보하기 위하여 다음 각 호의 사항을 이행하여야 한다.

1. 인공지능 수명주기 전반에 걸친 위험의 식별·평가 및 완화
2. 인공지능 관련 안전사고를 모니터링하고 대응하는 위험관리체계 구축

② 인공지능사업자는 제1항 각 호에 따른 사항의 이행 결과를 과학기술정보통신부장관에게 제출하여야 한다.

③ 과학기술정보통신부장관은 제1항 각 호에 따른 사항의 구체적인 이행 방식 및 제2항에 따른 결과 제출 등에 필요한 사항을 정하여 고시하여야 한다.

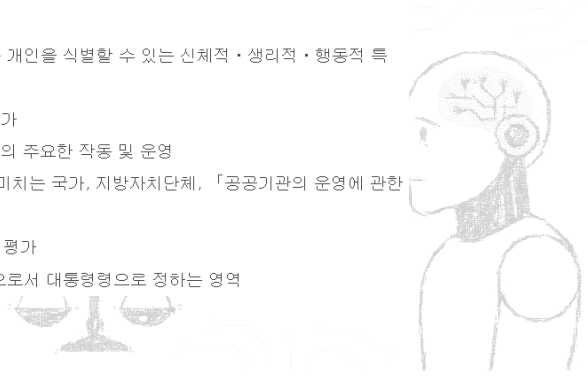
- For AI systems with high cumulative computational loads used during training, operators must ensure safety by identifying, assessing, and mitigating risks throughout the entire AI lifecycle,
- and by establishing a risk management system to monitor and respond to AI-related safety incidents.



2025 ICKOSSRC International Conference

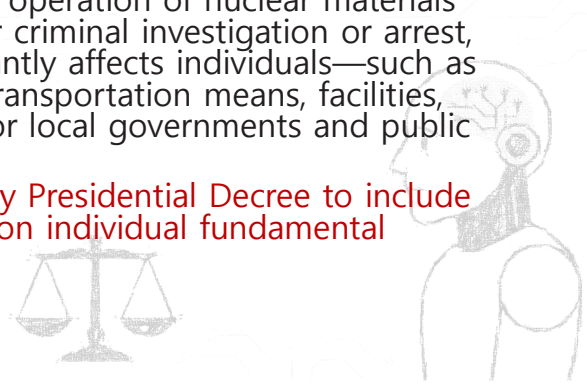
What Is High-Impact AI?

4. "고영향 인공지능"이란 사람의 생명, 신체의 안전 및 기본권에 중대한 영향을 미치거나 위험을 초래할 우려가 있는 인공지능시스템으로서 다음 각 목의 어느 하나의 영역에서 활용되는 것을 말한다.
   - 가. 「에너지법」 제2조제1호에 따른 에너지의 공급
   - 나. 「먹는물관리법」 제3조제1호에 따른 먹는물의 생산 공정
   - 다. 「보건의료기본법」 제3조제1호에 따른 보건의료의 제공 및 이용체계의 구축·운영
   - 라. 「의료기기법」 제2조제1항에 따른 의료기기 및 「디지털의료제품법」 제2조제2호에 따른 디지털의료기기의 개발 및 이용
   - 마. 「원자력시설 등의 방호 및 방사능 방재 대책법」 제2조제1항제1호에 따른 핵물질과 같은 항 제2호에 따른 원자력시설의 안전한 관리 및 운영
   - 바. 범죄 수사나 체포 업무를 위한 생체인식정보(얼굴·지문·홍채 및 손바닥 정맥 등 개인을 식별할 수 있는 신체적·생리적·행동적 특징에 관한 개인정보를 말한다)의 분석·활용
   - 사. 채용, 대출 심사 등 개인의 권리·의무 관계에 중대한 영향을 미치는 판단 또는 평가
   - 아. 「교통안전법」 제2조제1호부터 제3호까지에 따른 교통수단, 교통시설, 교통체계의 주요한 작동 및 운영
   - 자. 공공서비스 제공에 필요한 자격 확인 및 결정 또는 비용징수 등 국민에게 영향을 미치는 국가, 지방자치단체, 「공공기관의 운영에 관한 법률」 제4조에 따른 공공기관 등(이하 "국가기관등"이라 한다)의 의사결정
   - 차. 「교육기본법」 제9조제1항에 따른 유아교육·초등교육 및 중등교육에서의 학생 평가
   - 카. 그 밖에 사람의 생명·신체의 안전 및 기본권 보호에 중대한 영향을 미치는 영역으로서 대통령령으로 정하는 영역



What Is High-Impact AI?

- High-Impact AI: "**AI systems that may have a significant impact on human life, physical safety, or fundamental rights, or that may pose such risks**."
  - It includes AI used in areas such as energy supply, drinking water production processes, healthcare, the development and use of digital medical devices, the safe management and operation of nuclear materials and facilities, biometric analysis and use for criminal investigation or arrest, decision-making or evaluation that significantly affects individuals—such as hiring or loan approval—the operation of transportation means, facilities, and systems, decision-making by national or local governments and public institutions, and student assessment.
  - Additionally, the scope may be expanded by Presidential Decree to include other areas that could seriously infringe upon individual fundamental rights.



2025 ICKOSSRC International Conference

Key Issues in Korea's AI Act

1. **Classification of AI Systems**
2. **Effectiveness of Actor Categorization in AI Systems**
3. **Ambiguity in Transparency Provisions**
4. **Absence of Copyright Provisions**
5. **Strictness of Investigative Procedures and Low Fines**
6. **Concerns about Hindering Innovation in the Market**

1. Classification of AI Systems

- Unlike the EU, Korea's AI Basic Act does not specify prohibited AI categories
- Korea uses **a simple two-tier system** (high-impact/generative AI), which may cause blind spots and overregulation
- Some "limited risk" AI under the EU Act could be treated as "high-impact AI" in Korea, facing unnecessarily strict rules

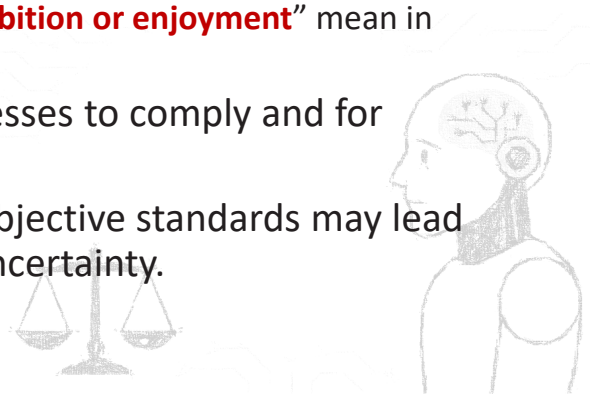
2025 ICKOSSRC International Conference

2. Effectiveness of Actor Categorization in AI Systems

- The definitions of "AI developers," "AI users," and "affected persons" are **vague** and **often overlap**
- Many entities act as both users and operators, but **the Act does not clarify this**
- Both groups face **almost identical obligations, making the distinction ineffective**

3. Ambiguity in Transparency Provisions

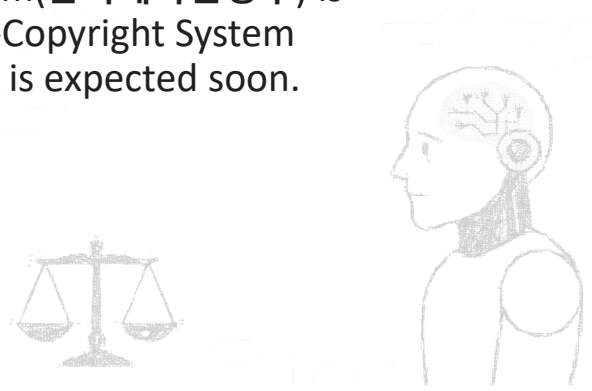
- Article 31(3) is unclear about key terms:
  - What exactly is an "**output that is difficult to distinguish from reality**"?
  - What does it mean for users to "**clearly recognize**" AI-generated content?
  - How do we define "**artistic or creative value**" in AI outputs?
  - What does "**avoiding interference with exhibition or enjoyment**" mean in practice?
- These ambiguities make it hard for businesses to comply and for regulators to enforce the law.
- Especially for art and creative content, subjective standards may lead to inconsistent interpretation and legal uncertainty.



2025 ICKOSSRC International Conference

4. Absence of Copyright Provisions

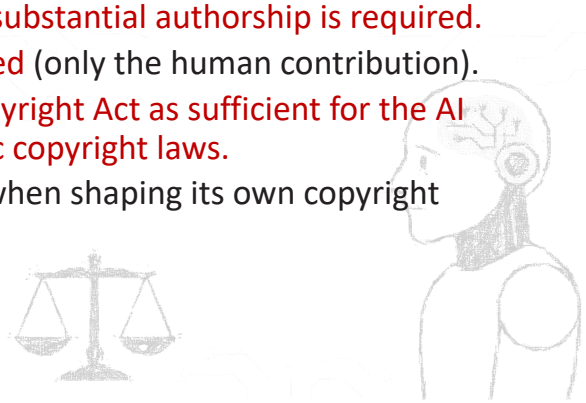
- The AI Basic Act **does not address copyright issues or clarify how it interacts with existing copyright law.**
- Many copyright questions related to AI remain unresolved.
- The Ministry of Culture, Sports and Tourism(**문화체육관광부**) is working on amendments through the "AI-Copyright System Improvement Council," so new legislation is expected soon.



※Copyright and AI: Part 2 – Copyrightability

US Copyright Office's Position on AI (Jan 2025)

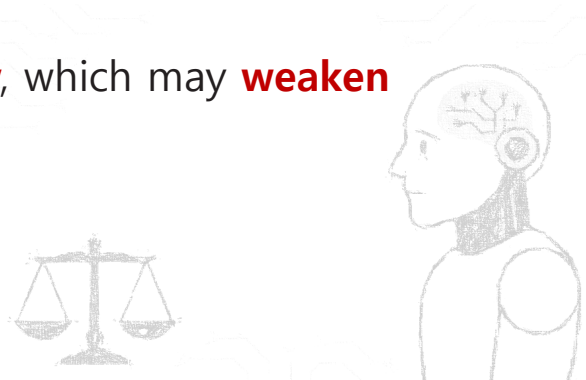
- Copyright protection is limited to works created by humans.
- If AI is just a tool and a human adds substantial creative input, only the human-authored parts are protected.
- Minimal human involvement is not enough; substantial authorship is required.
- Collaborative works may be partially protected (only the human contribution).
- The US Copyright Office sees the current Copyright Act as sufficient for the AI era and does not recommend new AI-specific copyright laws.
- Korea should consider these developments when shaping its own copyright policy for AI.



2025 ICKOSSRC International Conference

5. Strictness of Investigative Procedures and Low Fines

- Article 40 gives government officials broad authority to inspect AI operators' offices and documents if violations are suspected
- These investigative powers may be excessive compared to other laws
- Administrative fines are relatively low, which may **weaken enforcement**



6. Concerns about Hindering Innovation in the Market

- Legal uncertainty and risk of overregulation may discourage startups and innovative companies
- It is crucial that implementing regulations reflect input from **industry stakeholders** to avoid stifling industrial development.

Fake News Discourse & Selective Trust in the AI Era

Joseph Yoo, Ph.D.
Assistant Professor
Communication Department
The University of Wisconsin – Green Bay

2025 Korean Social Science Research Council

- **"Fake news"**: Fabricated stories mimicking legitimate journalism.
- A rhetorical tool used to delegitimize opponents and undermine the credibility of the media.
- Erodes journalism and fuels polarization.
- The study examines (1) how alternative media **strategically deploy** the term "fake news" and (2) how habitual media use shapes selective trust and partisan social reinforcement.

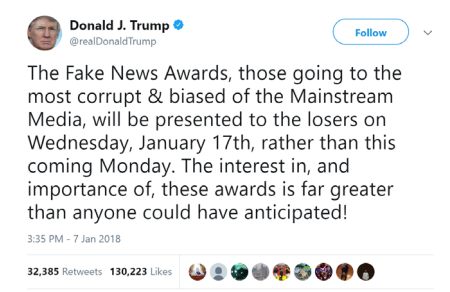
Definition Evolution

1) *Fake news* originally meant intentionally fabricated content mimicking journalism.

2) Over time, it has been conflated with Misinformation (unintentional falsehood), Disinformation (intentional falsehood), and Buzzword or rhetorical label with diminished clarity.

Strategic Use of "Fake News"

- Politicians (notably Trump) have used the term as a delegitimizing weapon against mainstream media.
- The term now functions more as an ideological attack than a factual critique.



Egelhofer & Lecheler's Framework (2019, 2020)

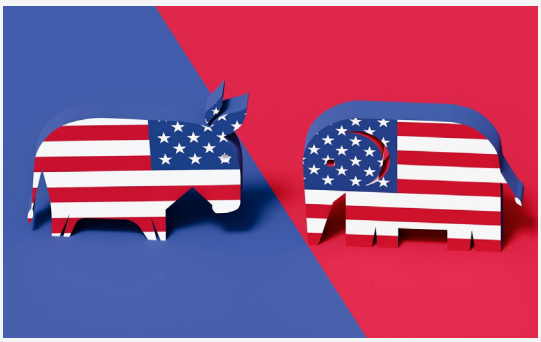
- **Fake news as a genre**: Disguised disinformation mimicking real news
- **Fake news as a rhetorical tool**: A Strategic label used to discredit dissent
- **Fake news as an empty buzzword**: Used so loosely, it loses specific meaning

Habitual News Consumption & NFMP

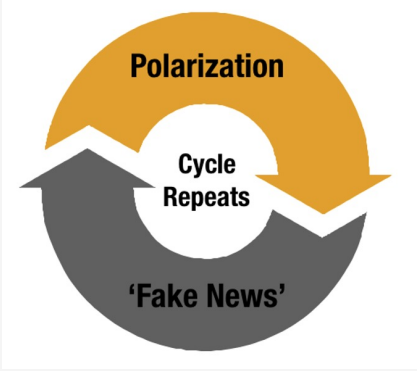
- News habits form through repeated behavior and familiarity.
- Habits reinforce confirmation bias and availability heuristics, leading to selective news trust.
- **News-Finds-Me Perception (NFMP)**: Passive news consumption where users expect to be informed without seeking news

Selective Trust

- Trust only politically aligned information
- Algorithm-driven platforms reinforce bias
- Confirmatory bias narrows informational diversity



Concerns Highlighted



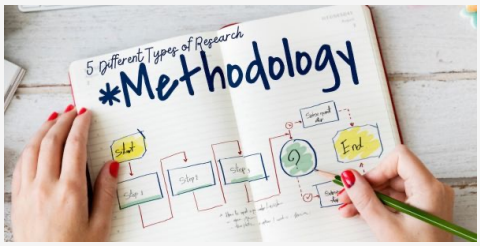
- Weaponizing "fake news" risks eroding trust in journalism.
- Encourages discursive polarization and undermines public deliberation.
- Overuse can trivialize the term, reducing its usefulness in countering actual disinformation.

Research Questions

1) How is the term “fake news” strategically framed in the Korean media landscape, and how does this framing affect media trust and discursive polarization?

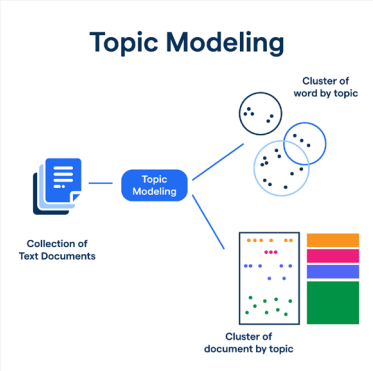
2) How do news users in South Korea form selective trust in digital news environments?

Methods



- Text analysis of Korean online news outlets
- Survey of South Korean adults
- Comparative analysis with U.S. alternative media

Methods: Text Analysis



- **Case Selection**: Newsmax (right-wing) and Occupy Democrats (left-wing) are chosen due to their hyperpartisan nature.
- **LDA topic modeling**: Identifies latent themes and ideological framing patterns based on over 3,000 articles from 2015 to 2023 that mention "fake news."

Methods: Survey

- Online survey (N = 454) of South Koreans
- Tested a **dual mediation model** linking habitual news consumption → selective trust → NFMP → partisan social reinforcement.
- Controlled for political orientation, interest, and media usage types.

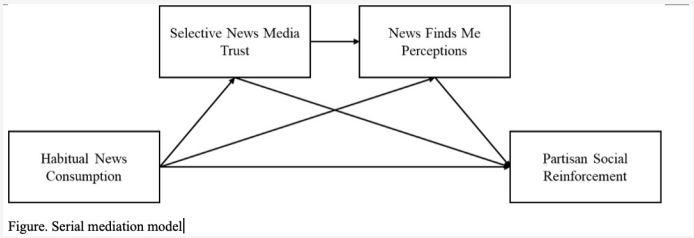


Figure. Serial mediation model

Preliminary Findings

- 'Fake news' used as rhetorical weapon in U.S. alt media
- Habitual news use reinforces partisanship via selective trust
- Supports cognitive shortcuts and availability heuristic

Preliminary Findings: Text Analysis

- Both outlets frequently use “fake news” as an empty buzzword and weaponized label, with little reference to actual disinformation.
- Newsmax used the term to attack mainstream media and legitimize Trump.
- Occupy Democrats used it to discredit Trump and right-wing misinformation.

Preliminary Findings: Text Analysis

Table. Topic Modeling of the term "Fake News" in *Newsmax*

Topic	Topic %	Subtopic	Top Keywords	% of Corpus
Topic 1: Trump and the Fake News Phenomenon	41.4	Questioning the Authenticity of Information	Fact, Evidence, Truth, Information, Question, Allegation, Own, Real, American, Public	12.42
		Selective Media Criticism	*CNN*, *NYT*, *WaPo*, *FOX*, Outlets, Mainstream, Journalist, Media, News, Report	18.63
		Election and Political Commentary	Election, Campaign, Democrat, Republican, Russia, Russian, America, Country, Political, Vote	10.35
Topic 2: Newsmax and the Trump Administration: A Symbiotic Relationship	24.4	Supportive Media Coverage and Legitimacy	Fake news, Trump, President, Media, Story, News, Report, People, Days, Sources	15.86
		Challenging Mainstream Media	Fake news, Media, Trump, President, *CNN*, *NYT*, Report, Story, News, Election	8.54
Topic 3: Navigating the Media Maze in the Trump Era	12.1	Critique of Mainstream Media and Its Consequences	Fake news, Media, Press, *CNN*, Propaganda, Conspiracy, Information, News, Story, Post	6.06
		Impact on Public Perception	Fake news, Media, Press, People, Election, Campaign, American, Democrat, Supporters, Public	3.02
		Descriptive Power and Examples	Fake news, Media, Press, Story, Intelligence, Hacking, Propaganda, Conspiracy, Election, Campaign	3.02
Topic 4: Dissecting the Complex Web of Fake News, Media, and Politics	22.4	Intensified Spread of Misinformation in Politics	Fakenews, Election, Propaganda, Misinformation, Lies, Conspiracy, False, Untrue, Deceptive, Misleading	8.96
		Media Dynamics and Shaping Public Opinion	Public, Opinion, Perception, Influence, Narrative, Discourse, Debate, Responsibility, Journalism, Agenda,	6.72
		Trump's Central Role and Media Strategy	Trump, Media, Strategy, Tactic, Relationship, Message, Communication, Rhetoric, Discourse, Narrative	6.72

Preliminary Findings: Text Analysis

Table. Topic Modeling of the term "Fake News" in *Occupy Democrats*

Topic	Topic %	Subtopic	Top Keywords	% of Corpus
Topic 1: Media Critique and Political Commentary	47.4	Trump Administration and Media Dynamics	Trump, President, Media, Whitehouse, Press, News, Story, *CNN*, *Fox*, Election	26.07
		Media Critique and the Fight Against Misinformation	Fake News, Media Lies, Conspiracy, Press, Truthful, Report, Public, Outlets, Mainstream	21.33
Topic 2: Political Discourse and Media Analysis	22.8	Political Narratives and Media Criticism	Trump, Media, President, Twitter, Election, Campaign, Report, News, Biden, Democrat	9.13
		Public Discourse and Perception	People, Story, Sources, News, Report, Media, Election, Campaign, Public, Information	6.84
		Specific Issues and Events	Election, Biden, *NYT*, *CNN*, Russia, Trump, Media, President, Campaign, Democrat	6.84
Topic 3: Political Conspiracies and Information Warfare	29.7	Russian Influence and Intelligence Operations	Russian, Intelligence, Hacking, Election, President, Trump, Media, News, Propaganda, Agencies	13.36
		Conspiracy Theories and Propaganda	Conspiracy, Propaganda, Media, News, Trump, President, Election, Theories, American, Rightwing	10.40
		Election Integrity and Information Warfare	Election, Hacking, Propaganda, Russian, Intelligence, Trump, Media, President, Agencies, Campaign	5.94

Preliminary Findings: Survey

- Selective trust mediates the relationship between habitual consumption and NFMP.
- NFMP strongly predicts partisan social reinforcement.
- Direct effect of habitual consumption on partisanship is non-significant. Its influence is indirect via trust and NFMP.

Preliminary Findings: Survey

Table. Testing the Dual Mediation Model

M1: Selective news trust (SNT)	*β*	SE	*t*	*p*	LLCI	ULCI
Habitual news consumption (HNC)	.18	.05	3.63	<.001	.08	.28
Age group	-.18	.08	-2.31	.021	-.34	-.03
Gender	.38	.11	3.51	<.001	.17	.59
Education level	-.08	.07	-1.11	.269	-.23	.06
Political orientation	.03	.04	0.84	.403	-.04	.11
Political interests	.07	.04	1.63	.104	-.02	.16
Traditional news use	.25	.05	5.49	<.001	.16	.35
Social media news use	-.00	.04	-0.06	.950	-.08	.08
Portal news use	.10	.05	2.03	.043	.00	.19
M2: News-finds-me perceptions (NFMP)	*β*	SE	*t*	*p*	LLCI	ULCI
Habitual news consumption	-.02	.05	-0.38	.703	-.11	.07
Selective news trust	.11	.04	2.67	.008	.03	.20
Age group	.03	.07	0.42	.672	-.11	.17
Gender	.17	.10	4.78	.076	-.02	.36
Education level	-.04	.07	-0.59	.559	-.17	.09
Political orientation	-.01	.03	-.22	.829	-.07	.06
Political interests	.04	.04	0.94	.350	-.04	.11
Traditional news use	.14	.04	3.34	.001	.06	.22
Social media news use	.19	.04	5.46	<.001	.12	.26
Portal news use	-.00	.04	-0.08	.934	-.09	.08
DV: Partisan social reinforcement (PSI)	*β*	SE	*t*	*p*	LLCI	ULCI
Habitual news consumption	.05	.05	0.92	.357	-.05	.15
Selective news trust	.07	.05	1.43	.153	-.03	.16
News finds me perceptions	.30	.05	5.60	<.001	.20	.41
Age group	.13	.08	1.65	.100	-.03	.29
Gender	-.16	.11	-1.56	.149	-.38	.06
Education level	-.03	.07	-0.43	.667	-.18	.11
Political orientation	.12	.04	3.22	.001	.05	.20
Political interests	.40	.04	8.99	<.001	.31	.48
Traditional news use	.04	.05	0.86	.390	-.05	.14
Social media news use	.14	.04	3.37	.001	.06	.22
Portal news use	.12	.05	2.65	.008	.03	.22

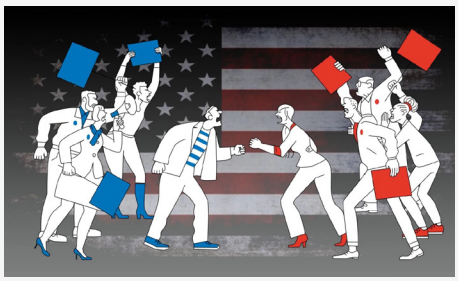
Paths	*β*	SE	LLCI	ULCI
HNC → SNT → PSI	.01	.01	-.00	.04
HNC→ NFMP → PSI	-.01	.02	-.04	.03
HNC → SNT → NFMP → PSI	.01	.00	.00	.02

Note. SE = Standard Error; *β* = Standardized Coefficients; LLCI = Lower Limit Confidence Interval; ULCI = Upper Limit Confidence Interval.

Discussion & Conclusion

- Algorithms deepen filter bubbles and selective trust.
- Reinforce polarization, undermine journalism.
- Study offers insights for restoring trust in AI-driven news environments

Discussion & Conclusion



- The strategic use of “fake news” mirrors ideological polarization, regardless of political orientation.
- Alternative outlets use the term more to delegitimize opposition than to correct false information.
- This rhetoric exacerbates media fragmentation, weakens journalistic standards, and contributes to echo chambers and confirmation bias.

Discussion & Conclusion

- Individuals tend to trust personally preferred outlets, raising concerns about misinformation and filter bubbles.

- Recommendations

(1) Media literacy programs focusing on critical engagement over technical skills

(2) Algorithmic design reforms to diversify content exposure

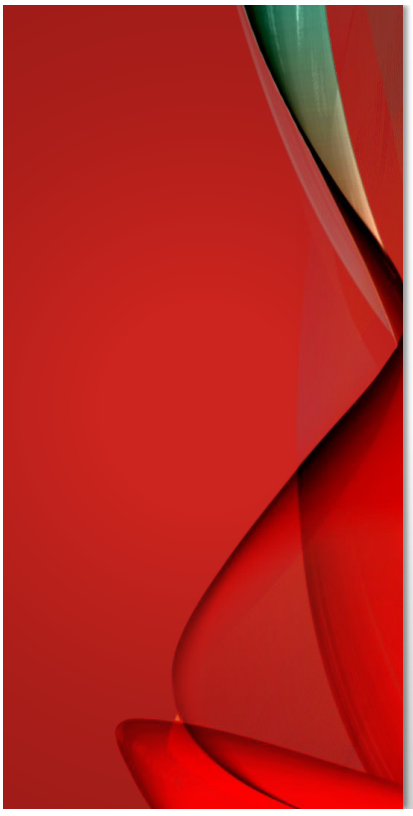
(3) Journalism reforms to restore trust across generations.

keynote

2

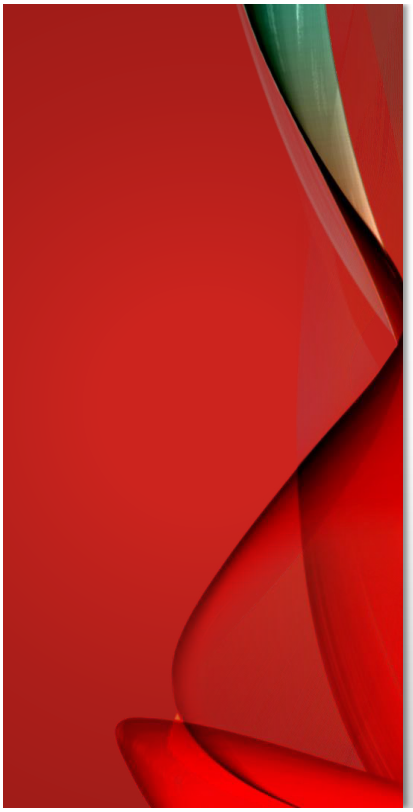
AI, Robots and the Future of Works:
Ethical, Economic, and Social Dimensions

Michael J. Ahn
Associate Professor, University of Massachusetts Boston
National Council Member, American Society for Public Administration



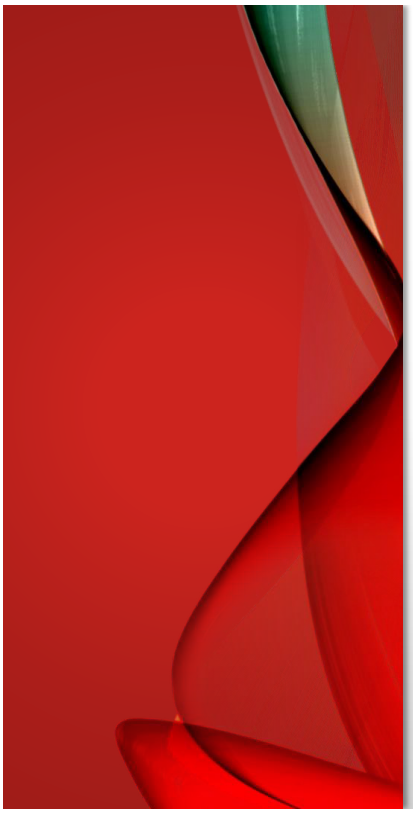
INTRODUCTION

- The transformative impact of AI on society and governance
  - Two key questions on AI – Ethics and the Future of Work
- How will AI change our lives, jobs, society, and government?
  - AI is rapidly transforming the work landscape, creating a new AI sector, promises innovation, productivity boost, and efficiency.
  - But it also rases concern for potential and extensive job displacement
- Current Debate: AI as an augmenter vs. AI as a replacer
  - What to do if the latter?



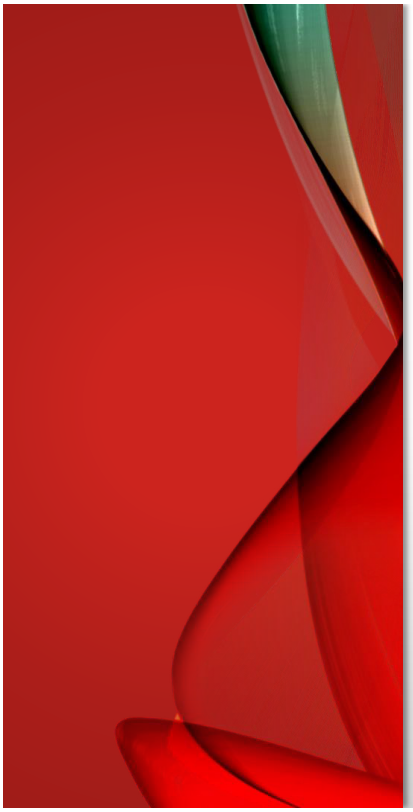
DEFINING HUMAN-MACHINE BOUNDARY

- **AI's growing role as a creator of value** reshapes traditional economic frameworks.
- Increasing role of AI and robotics in traditionally human roles:
  - Autonomous driving (Tesla, robotaxis)
  - Robotics in industry (Unmanned café, restaurants, factories)
  - Language models (ChatGPT, Grok-3, Claude)
  - Autonomous warfare (drones, unmanned vehicles)
- Expanding the role of machines; **human-machine collaboration**
- Challenging how we define the boundary of machines in human domains



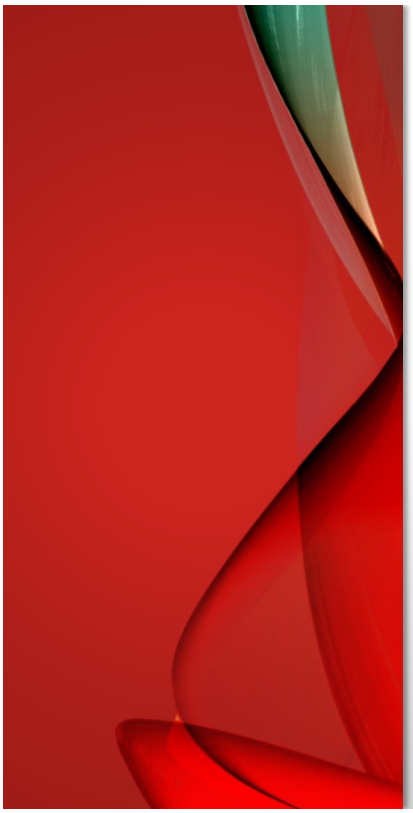
THE EMERGING AI INDUSTRY

- AI and robots performing roles traditionally held by humans
- Expansion of AI and robotics into Human Domains
  - Tesla FSD, Figure 1, robotaxi, Boston Dynamics robotics
  - LLM - ChatGPT, Perplexity, Grok-3, Claude
  - Unmanned café, restaurants, factories
  - Changing warfare: algorithms, unmanned vehicles, drones (Milley & Schmidt, 2024)
- Categories of AI Industry:
  - Infrastructure: AI hardware (chips, servers, networking).
  - Software: Machine learning platforms, data analytics.
  - Application services: AI in healthcare, automotive (self-driving), financial tech, customer service automation.
- Growth Trends: AI-related job postings increased significantly from 0.5% in 2017 to roughly 2% in 2023, indicating rapid growth and demand. (Lightcast, 2024)



POLICY CONTEXT: THE U.S. AS A MANUFACTURING HUB?

- New administrative initiatives aim to revive manufacturing
- Current U.S. labor structure
  - Manufacturing Sector: Approximately 9.3% of U.S. employment (Statista, 2023).
  - Service Sector: Dominant, accounting for approximately 79% of employment (Trading Economics, 2023).
- Current US labor structure makes the transition difficult.
- Who will work in factories?
  - AI and Robots



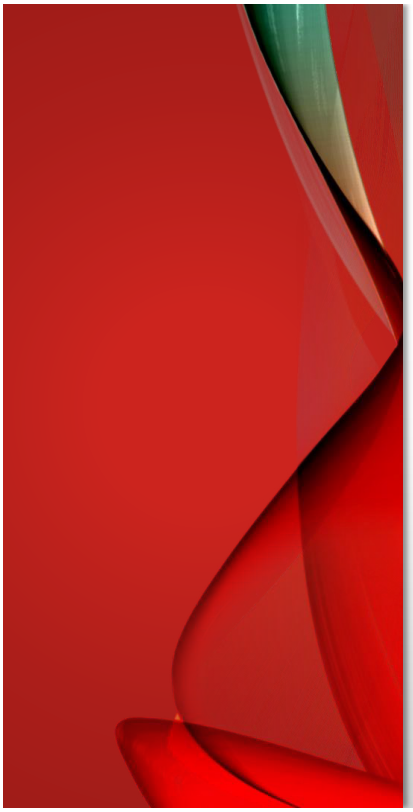
FROM INFORMATION TO AI ECONOMY

Porat's Information Economy Framework

- Original categorization:
  - Primary Information Sector (e.g., media, telecommunications).
  - Secondary Information Sector (internal information use).
- Evolution into dominant economic force (46% of U.S. GNP in 1967 to 63% in 1997).

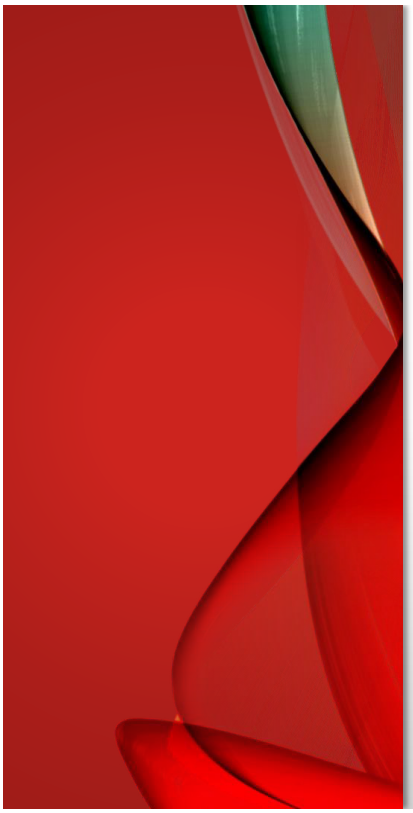
"AI Economy"

- AI economy defined by autonomous interpretation and decision-making.
- Key AI capabilities: machine learning, NLP, computer vision.
- AI as the "new electricity": universal economic impact.



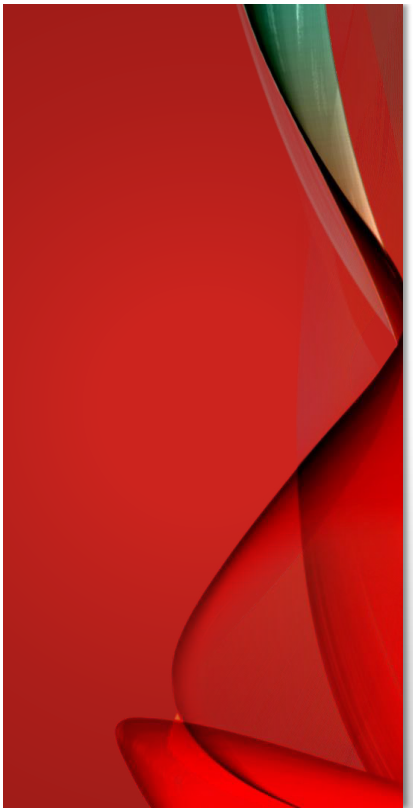
"AI ECONOMY"

- **Primary AI Sector**
  - Firms explicitly focused on AI development.
  - Examples: AI software providers (OpenAI, Google DeepMind), AI hardware (Nvidia).
- **AI-Integrated Industries**
  - Traditional industries increasingly leveraging AI for productivity and innovation.
  - Examples: Manufacturing automation, AI-driven healthcare diagnostics, financial sector algorithms.
- **AI-Augmented Workforce**
  - Jobs enhanced and reshaped through integration with AI technologies.
  - Examples: AI-supported medical diagnoses, AI-enhanced analytics roles.
- **Implications for Measurement and Policy**
  - Need for updated economic metrics to accurately capture AI's economic contributions.
  - Potential for targeted fiscal policies (e.g., robot tax) to manage transitions and inequalities.



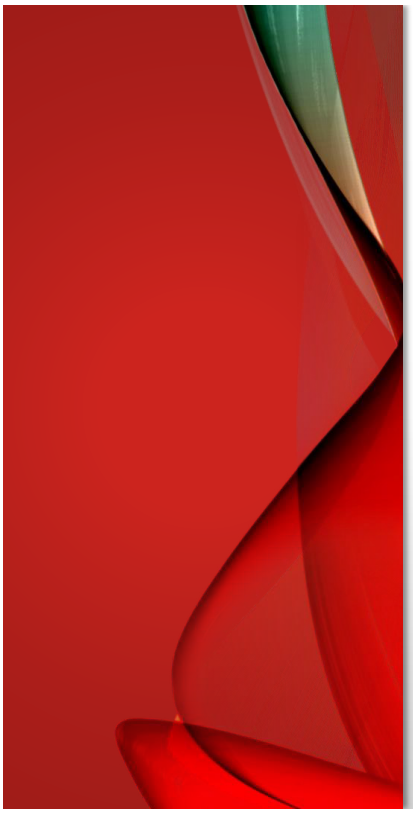
MARKET SIZE, IMPACTS AND ECONOMIC IMPLICATIONS

- **Market Size and Growth**
  - Current Global AI Market (2023): ~$189 billion (UNCTAD, 2025).
  - Projected Growth: Approximately $4.8 trillion by 2033 (25-fold increase).
  - Growth Drivers: Rapid adoption of generative AI technologies; annual growth rate 30–40%.
  - Major investments: Generative AI investment surged from $1.3 billion (2022) to $17.8 billion (2023).
- **Economic Implications**
  - Productivity Boost: AI could significantly enhance productivity by automating routine tasks and enabling innovation.
  - Job Exposure: Approximately 40% of global jobs partially automatable (IMF, 2024).
  - Advanced economies face greater disruption (~60% exposure) compared to emerging economies (30–40%).
- Job displacement vs. augmentation—AI redefining labor roles.



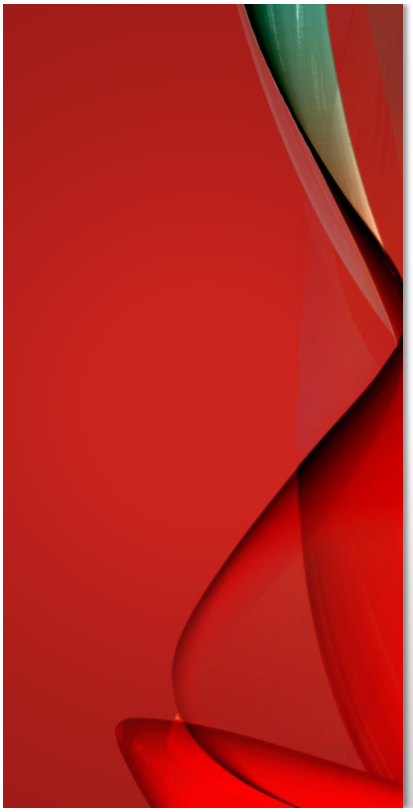
AI'S IMPACT ON INEQUALITY

- Wage Gap Concerns: Potential widening gap between high-skill, AI-literate workers and those more vulnerable to automation.
- Uneven Access and Benefits: Advanced economies better positioned to capture AI benefits, potentially exacerbating global inequality.
- Policy Necessities: Proactive measures required to mitigate inequality, such as workforce retraining, education investments, and targeted fiscal policies (e.g., Robot Tax).



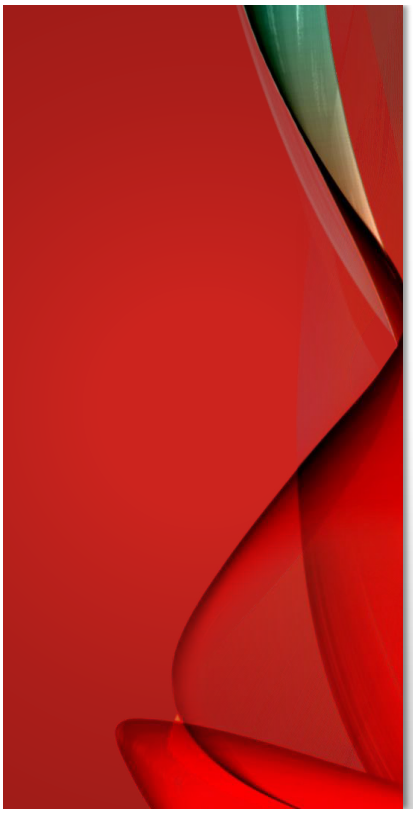
WHY CONSIDER A ROBOT TAX?

- Redistributing Economic Gains: Ensuring wealth generated from AI efficiency benefits all citizens, not just technology companies.
  - Do we need to intervene or leave it to the market?
- Responding to Wealth Concentration: Without systematic redistribution, wealth and revenue from productivity gains and automation will overwhelmingly accrue to AI and technology companies.
  - If no one has money to spend, what good is an efficiency gain?



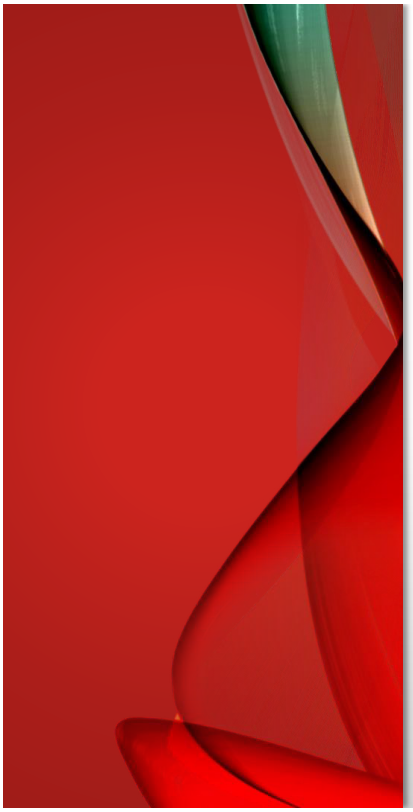
KEY QUESTIONS FOR IMPLEMENTING A ROBOT TAX

1. Categorizing the AI Industry:
   - How should we define AI for taxation purposes—hardware infrastructure, software applications, services, or both?
   - New category in economic input-output table – impact of AI industry on value creation
2. Determining Tax Targets:
   - Should taxes apply to developers of AI technologies, users who automate jobs, or both?
3. Taxation Level -
   - What is a fair, effective tax rate that avoids discouraging innovation?
4. International Collaboration
   - How can we ensure international collaboration to avoid competitive disadvantages and tax evasion?



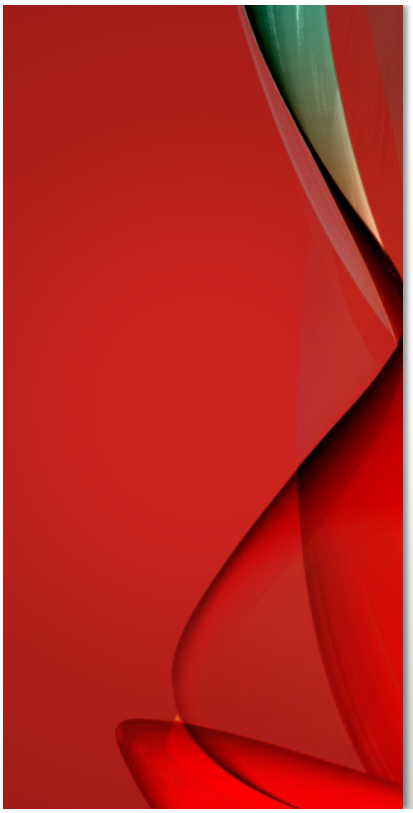
ETHICAL AND ECONOMIC JUSTIFICATION FOR REDISTRIBUTION

- "Do we need this or should we let the market decide?"
- Protecting displaced workers, basic income, job training, and social welfare programs
- Maintain a level of market demand by way of wealth redistribution
- Balancing economic growth from innovation vs. social equity and fairness.
- Creating a choice for employers: Choose AI or Human Agent faced with Robot Tax – which will be "cheaper?"



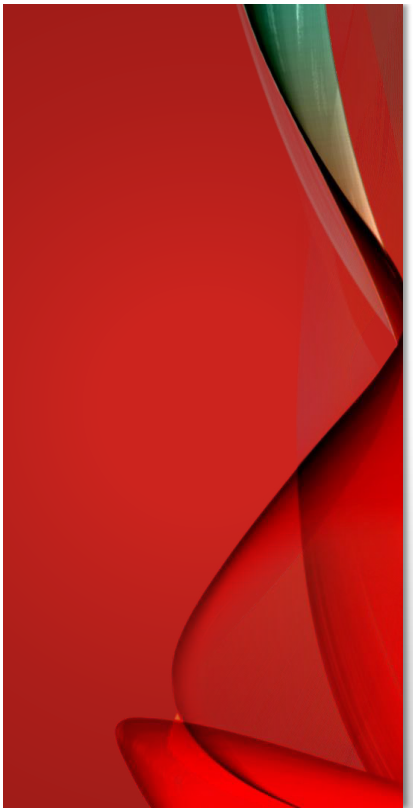
ETHICAL GUIDELINES FOR AI – WHOSE ETHICS?

- Machine Ethics – why do we need it?
  - **Value judgement in public policy**
  - Ethical guidelines for AI are crucial as machines take on more significant roles.
- No universal, **monolithic ethical** consensus exists.
  - Human ethical foundation: Personal, religious, cultural value systems
  - Geographic differences: U.S., Italy, China, India.
  - Diverse religious/philosophical values: Christianity, Islam, Judaism, Buddhism, Hinduism, Atheism, New Age, etc.
  - US Constitution as a value system – liberty, equality, justice for all ("created equal," divine origin of men)
- The paradox of **unethical humans teaching ethical AI**
  - **AI is what data they consume** - AI systems learn primarily from data, reflection of our behaviors.
  - Human actions often differ what they say – are we really ethical?
  - Parents advising against excessive phone usage yet we are glued to our phones ourselves.
  - Necessity: Lead by example; we need to be ethical in the first place for ethical AI.



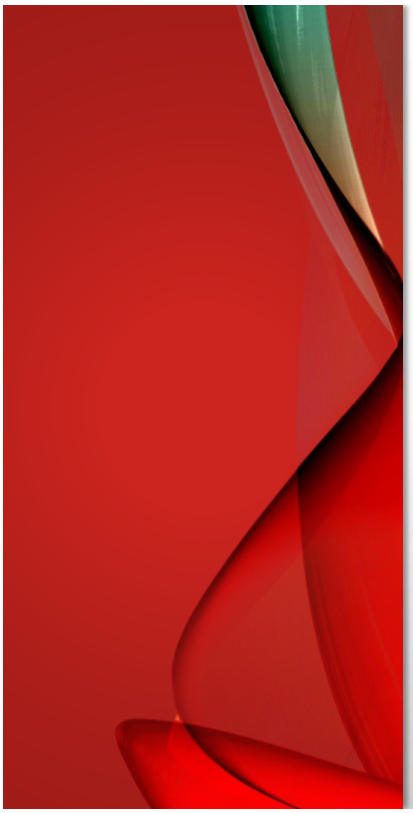
AI TRANSPARENCY AND TEACHING COMPASSION

- Humans often demand transparency from AI but human judges are often not transparent.
- Trust in human judges is based on assumed compassion and common ethical standards of culture, religion and geography.
- Human emotion and **compassion in judgement**
  - "I like you" factor then looking for evidence.
  - AI is not capable of such "liking" – are we ok with this?
- Critical challenge: AI lacks inherent compassion and emotional understanding.
  - Need for intentional **training of AI in compassion**
  - Christian-specific standard of compassion may differ across other cultural and religious traditions.



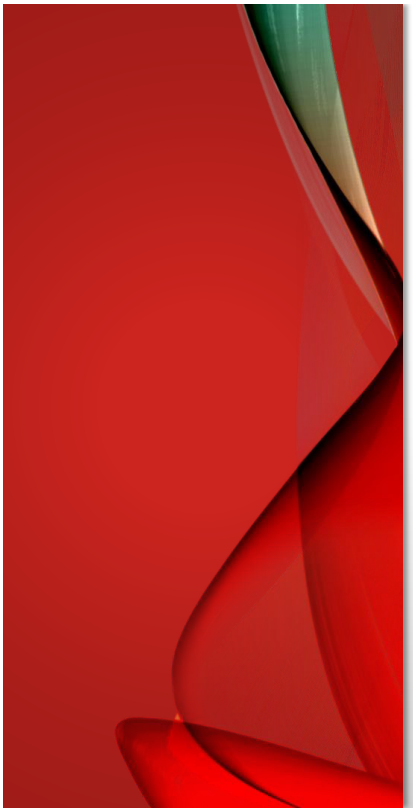
THE "PERFECT INFORMATION" PROBLEM

- Traditional organizations as information processors
  - Collecting, processing, and producing valuable decision outputs
- AI providing **"perfect information" from outside the organization**
  - Challenges to traditional organizational structures
  - Example: Why enroll in universities when AI offers instant knowledge?
  - AI reduces demand for human "information agents" (e.g., law graduates).
  - Potential shift in organizational theory and practices
- **Overdependence on AI** systems and increasing **uniformity**
  - **Loss of diversity in decision-making** and innovation.
  - Uniformity in decisions (e.g., financial investments) could amplify economic risks, creating catastrophic failures.
  - Promote decentralized, diverse AI systems to mitigate systemic risks.



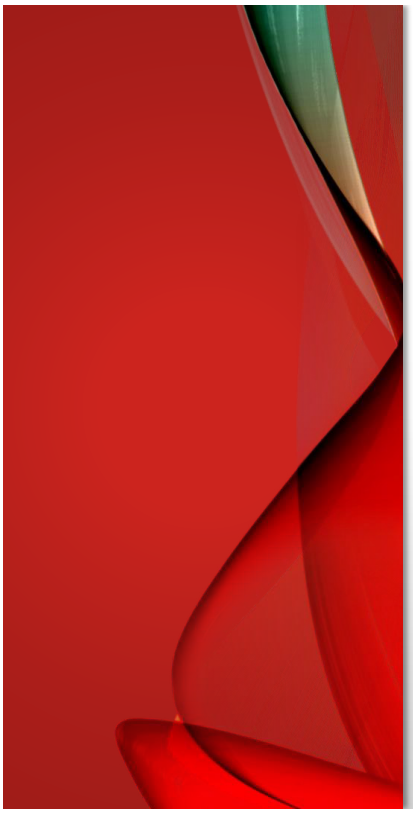
REDEFINING THE ROLE OF HUMANS IN AN AUTOMATED WORLD

- Identifying uniquely human roles amid AI expansion
  - What roles remain for humans amid expanding AI capabilities?
- Impact on Writing and Creativity
  - Has ChatGPT "killed" traditional writing?
  - Importance of "intentionality" in writing, art, music
- The Concept of Human-Certification
  - Growing significance of human authenticity in art and music
  - Valuing creations that embody human experience and emotion
- The Case for a **Robot Tax**
  - Addressing societal and economic impacts caused by AI transformation
  - **Reduce severe concentration of wealth** by AI, support displaced workers (**social welfare**), provide retraining opportunities - **"soft landing"**
  - New: Tax AI and **robots as value-creating entities** in the economy



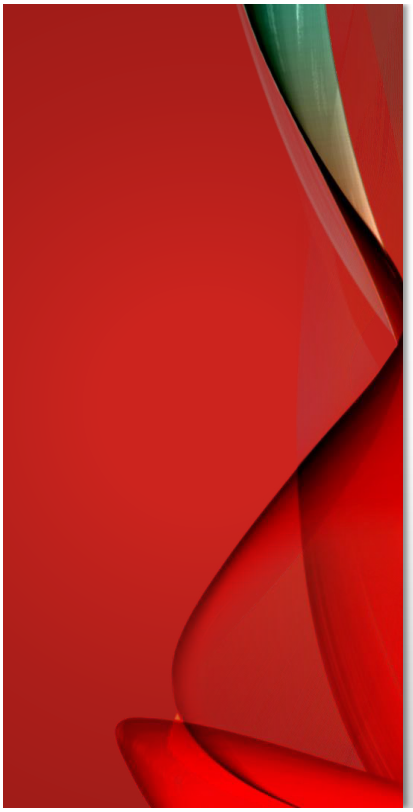
A GLIMPSE INTO THE FUTURE: INSIGHTS FROM THIS YEAR'S NOBEL PRIZES

- Era of Human-Machine Collaboration
  - **Highlighting interdisciplinary cooperation** between AI and sciences
  - Importance of collaboration between social and natural sciences
- **Nobel Prizes in chemistry**
  - Demis Hassabis and John Jumper developed AlphaFold2, an AI model that solved the 50-year-old challenge of predicting proteins' complex 3D structures from their amino acid sequences.
  - Using AlphaFold2, they predicted the structures of nearly all 200 million known proteins, greatly advancing the field of structural biology.
- Application to Public Administration
  - **Creation of area-expert AI platforms**: "public policy" machine
  - Shift in education towards training AI systems



CONCLUSION

- Embrace Human-AI Collaboration
  - Proactively define roles and boundaries between humans and machines
- **AI as a New Source of National Wealth**
  - **AI will determine the wealth of a nation**
  - Recognize AI's growing role as an independent economic and value-creating entity.
  - Implement **Robot and AI Taxes** to support equitable economic transitions and "soft transformation" to automation.
- Develop Ethical Frameworks
  - Establish clear **ethical guidelines (compassion) for AI** development and deployment
- Reimagine Organization Theory
  - Rethink traditional organizational theories in light of AI's expanding capabilities
- Cultivate **Uniquely Human Contributions**
  - Seek uniquely human activities
  - Value human-certified art, music, and writing
- **Invest in Education** and Interdisciplinary Collaboration
  - **Prepare professionals for AI-integrated environments**
  - Train AI specialized in public policy formulation, implementation, evaluation
  - Focus public administration education on creating data, training AI



CONCLUSION

- Embrace Human-AI Collaboration
  - Proactively define roles and boundaries between humans and machines
- **AI as a New Source of National Wealth**
  - **AI will determine the wealth of a nation**
  - Recognize AI's growing role as an independent economic and value-creating entity.
  - Implement **Robot and AI Taxes** to support equitable economic transitions and "soft transformation" to automation.
- Develop Ethical Frameworks
  - Establish clear **ethical guidelines (compassion) for AI** development and deployment
- Reimagine Organization Theory
  - Rethink traditional organizational theories in light of AI's expanding capabilities
- Cultivate **Uniquely Human Contributions**
  - Seek uniquely human activities
  - Value human-certified art, music, and writing
- **Invest in Education** and Interdisciplinary Collaboration
  - **Prepare professionals for AI-integrated environments**
  - Train AI specialized in public policy formulation, implementation, evaluation
  - Focus public administration education on creating data, training AI

AI or Not AI: Perspectives of Higher Education in Uzbekistan

Tashkent University of Information and Technologies Vice Rector,
Djamshid Sultanov

As the world rapidly transitions from digital transformation (DX) to an era of AI transformation (AX), higher education systems must rethink their roles in shaping future-ready societies. This keynote explores how Uzbekistan's universities are embracing this challenge with purpose and vision. With AI no longer a distant frontier but an immediate force reshaping economies, labor markets, and knowledge systems, institutions in Uzbekistan are proactively reimagining their educational models.

The presentation will address key strategies being implemented, such as integrating AI-related disciplines across curricula, upskilling faculty in digital competencies, and fostering partnerships with technology and research sectors. Particular emphasis will be placed on aligning education with values such as inclusion, ethics, and sustainability, ensuring that AI serves human development rather than replacing it.

Drawing on both national initiatives and university-level reforms, the keynote presents Uzbekistan's higher education response not as a matter of choosing "AI or not AI," but rather as a matter of choosing how to harness AI in meaningful ways. This forward-looking approach seeks to empower the next generation to become co-creators of a future where AI augments human potential, nurtures critical thinking, and promotes a just and innovative society. The talk will offer valuable insights for universities around the world navigating similar transitions in the age of AX.

Challenges for Building Bright Human-Centered AX era: Frontier Technology, Empathy, Communication, and Integration

Moon Suk Ahn (Prof. Emeritus of Korea University)

Key Questions:

0. Is the arrival of the AX era **inevitable** or a choice?
0. What are the **grounds for pessimism and optimism** in regarding the future of humanity?
0. What is the **impact of AX** on the future world?
0. Is human-centered AX **possible**?
0. What are the **challenges** for human-centered AX?
0. Hawking's warning on AI and implication to AX era

1. Evolutionary Process of Human Civilization

Human history has evolved through cutting-edge technologies (Steam Engine, Electricity, Computer, and AI).
Frontier technologies have created new living space, new kind of human-beings, new production system, new interpersonal relationships, new market structures, and new forms of government.
The new frontier technologies have brought about four Industrial Revolutions.
The 1st and 2nd Industrial Revolution created cities as the living space, city dwellers as the new human species, factory production as production system, permanent market as transaction pattern, and bureaucracy as the new

governance. Steam engine and electricity triggered the revolutions.

The computer technology created **the 3rd Industrial Revolution**. Living space evolved into **cyber space**. **Netizen** was born as the new human species. **E-commerce** became the new transaction method. **E-government** appeared as the new pattern of government.

The frontier technologies that are triggering **the 4th Industrial Revolution** are AI, IoT, Big Data, Cloud Computing, Mobility, Blockchain etc. Among them AI is leading the new revolution.

In essence, t**he AX era** is not driven by a single technology but by the synergistic **convergence of breakthroughs** in AI algorithms, the availability of big data, powerful computing infrastructure, widespread connectivity, and more natural ways for humans to interact with machines. .

In the new world of the 4th Industrial Revolution, living space of the new era is **hybrid space merging cyber space and the physical world**. **Humanoid, the new species** appear in the living space together with netizens. Consumers become **prosumers** who produce products and services at the consumption site. **Digital on-demand transaction** will be next transaction system. E-government is being replaced into **Digital Platform government.**

AX has been born in the Fourth Industrial Revolution.

2. Data as a primary driver in the AX era

In the 1970s and 1980s, T**he "close coupling"** or tight integration of computer programs and data had evolved.

The introduction of DBMS marked a significant shift in 1990s. Databases provided a structured and organized way to store and manage data, separate from the application programs that accessed it. This allowed multiple programs to share the same data, improving efficiency and reducing redundancy.

In 2000s, the explosion of data from diverse sources (social media, sensors, IoT devices) has created a paradigm shift. Programs, especially AI and machine learning models, are now designed to process and learn from these vast datasets. We call this as **the Big Data era**, It was called also as the DX era.

Finally we reached to **the AX Era.**

In this era, data is not just something to be processed; it's the foundation upon which **models are built and trained.** The quality and quantity of data directly determine the performance of AI systems. **Data become the source of the National Wealth.**

3. Background Technologies of AX era

The background technologies are Deep Learning, Generative AI, Big Data and Cloud Computing, Enhanced Computing Power, Internet of Things(IoT) and Sensor Technology, Advancements in Human-Computer Interface(HCI), and Extended Reality(XR)

In this AX society, AI is easy to learn and to use. Humans live with the help of AI in all aspects. AI becomes like water. Nations that use AI efficiently will be the **winners.**

Nations that reluctant to use AI will be the **losers**. **Naturally AX era can be assumd to be neccesary step in the human history.**

3. Evolution from the Information Society to the Intelligent Information Society: from DX to AX

The intelligent information society of the 4th Industrial Revolution **inherited Clonism Society and information overlord from the information society.**

In the information society, through SNS, humans on earth are connected to each other. In the new society. 'the pain and unhappiness of one individual are felt as the pain and unhappiness of the entire human race. **The Clonism Society has born.**

In such a world, empathy should develop between individuals in order to avoid chaos and conflict among people. Empathy includes thinking, feeling, and acting from the perspective of others, moving beyond the individual 'I' to a dominant sense of 'we'. This capacity for empathy can work **favorably** in resolving conflicts.

But information overload in information society resulted in polarization among people and lack of communication between politicians and scientists. This capacity worked **unfavorably** for building bright information society.

AX era begins with those two contrasting phenomena.

Consequently, **without the development of novel governance to manage** human's selfish exploitation of intelligent machine and information overload, the AX era becomes

significantly more perilous to human society.

Machine-based Intelligence Society and AX

The AX era can be said to be the Intelligent Information Society.

As artificial intelligence develops further, intelligent machines can gain **self-awareness** and act as **autonomous agents**, we enter the **era of machine-based intelligence society.**

There exist possibility in which humans and intelligent machines **collide.**

In the machine-based era, **if** such abilities of understanding and empathy develop between humans and intelligent machines, then humans and intelligent machines can move towards a mutually **beneficial cooperative relationship resulting in the bright AX era.**

4. What we learn from the past information society.

The Club of Rome in the 1970s told us lessons.

Through computer simulations, they presented a **pessimistic view** (Doomsday Theory), stating that without urgent and drastic measures, humanity would perish due to explosive population growth, pollution, and food shortages, the crisis variables.

They reasoned that the **excessive specialization** within science and technology, coupled with **communication difficulties among scientists**, would prevent science and technology from progressing at the rapid, exponential rate.

Fortunately, humanity has overcome this pessimism and continued to advance.

Science and technology have progressed at a hyper-exponential rate, guiding the destiny of humankind.

At its core was the Third Industrial Revolution, **the Information Revolution.**

Humanity has indeed moved beyond the pessimism with help of science and technology.

In the information society, however, polarization among people and communication difficulties between scientists and politicians have **hindered** the much more needed empathy and cooperation among people.

5. Possible Path to Bright Human-centered AX

Humanity, with intelligent machines possessing the immense power of artificial intelligence may produce a optimistic future.

The scenario begins with **all purpose intelligent AI (AGI)**.

The ChatGPT could reduce the information overloads to people.

ChatGPT's excellent **search and summarization abilities** are expected to save humanity from the information overload.

A **reduction in information overload** will lessen polarization in people and promote communications between politicians and scientists.

It will finally **promote empathy** among people and even between human and intelligent machine.

Bright AX society could be reached.

6. The keys for building a bright AX era

Realizing a bright future in the AX era, driven by automation and AI, requires addressing several key challenges. First and foremost, it is crucial to **cultivate empathy** between humans, and between humans and intelligent machines. This will contribute to strengthen social connections and reduces alienation. Secondly, **communication** between politicians and scientists is essential for in-depth discussions and rational policymaking regarding the complex ethical, social, and technological issues of the AX era. Finally, **overcoming political polarization**, which hinders healthy social consensus, is necessary to ensure that AX technology contributes to the prosperity of society as a whole to build the bright human-centered AX future.

8. Conclusion

In our discussion, we found that we can avoid the Doomsday and build the bright future through AX.

Beyond AI Technology, the Human Imperative is essential in the desirable AX Future.

Bridging the AX Divide among people and nations
can forge a Hopeful Future in the Age of AX.

We strongly believe that we can escape the Hawking's Warning on Future of humanity.

At the dawn of the AX era, humanity's future rests upon us, the social scientists.

Human-centered bright AX era is not free. It is costly. We humans have to do our best effort to build the bright

future.
Let's bravely move forward together with a pioneering spirit.

Thank you!

session

5

KAIST

Uneven Automation

ChatGPT's Impact on Software Tasks Varies by Difficulty and Data Availability

Myokyung Han, Jiwoon Hong, Taegyoon Kim, Jinhyuk Yun, Lanu Kim
Presenter: Lanu Kim (KAIST)
https://lanukim.github.io
https://computationalsociologylab.github.io

GenAI / LLM and society

growing interest for applying LLMs to ***medical domains*** (Wei et al., 2024),

improve student engagement and interaction in ***education*** (Kasneci., 2023)

reshaping ***organizational management strategies*** (Ayinde et al., 2023),

revolutionizing software development practices in ***information technology*** (Liu et al., 2024),

influencing ***economic structures*** *and* ***labor markets*** (Eloundou et al., 2023),

All jobs

All jobs

- 2/3 of all jobs will be replaced by Generative AI.
- Not only structured tasks, but unstructured tasks are also likely to be replaced.

Empirical Gap: Gen AI's Impact on Labor

- Prediction, prediction, prediction

"AI may diminish some of today's valuable employment opportunities."
(Frank et al., 2019)

"These new technologies are set to drive future growth across industries. Such posi tive effects may be counter-balanced by workforce disruptions."
(The Future of Jobs Report 2020).

"The potential scope of automation and augmentation will further expand over the next few years, with AI techniques maturing and finding mainstream application across sectors." (Cramarenco et al. 2023)

Empirical Gap: Gen AI's Impact on Labor

- Prediction, prediction, prediction

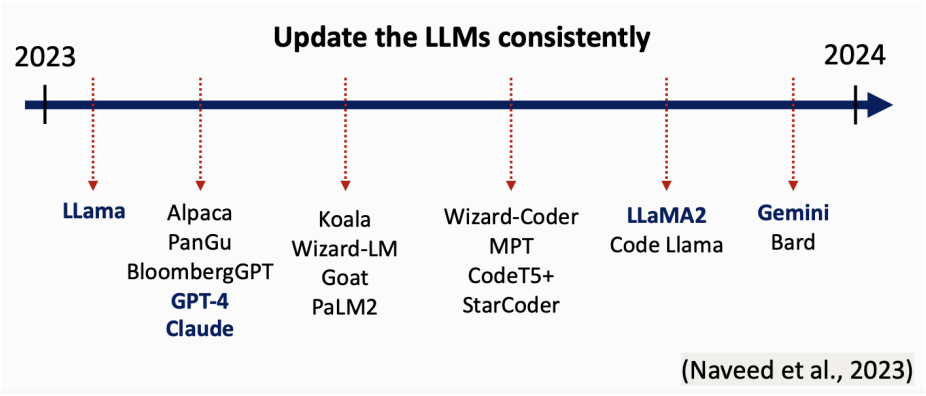
"AI ***may*** diminish some of today's valuable employment opportunities."
(Frank et al., 2019)

"These new technologies ***are set to drive*** future growth across industries. Such po sitive effects ***may be*** counter-balanced by workforce disruptions."
(The Future of Jobs Report 2020).

"The potential scope of automation and augmentation ***will further*** expand ***over th e next few years***, with AI techniques maturing and finding mainstream application across sectors." (Cramarenco et al. 2023)

LLMs change rapidly compared to the update frequency of the statistics

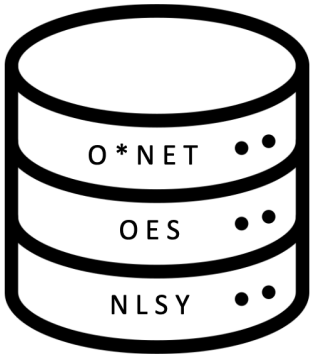
LLMs



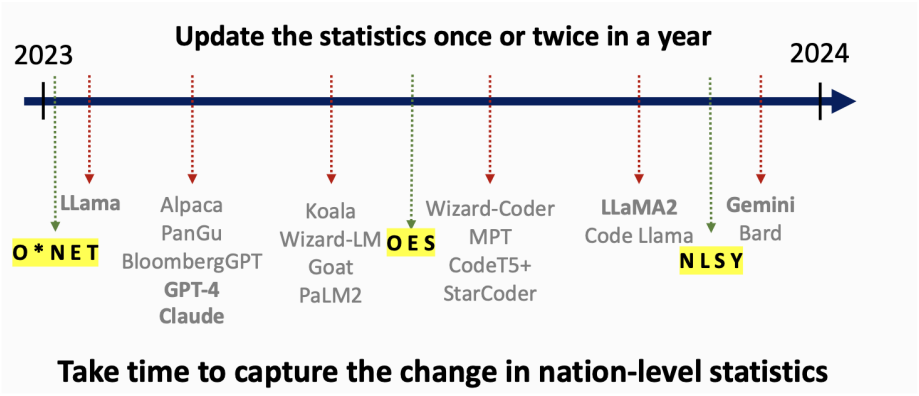
(Naveed et al., 2023)

7

LLMs change rapidly compared to the update frequency of the statistics



National-level **statistics**

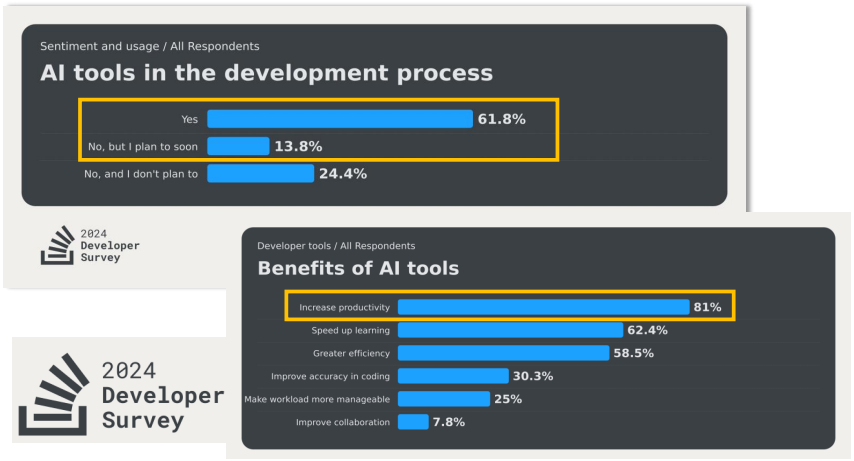


Take time to capture the change in nation-level statistics

8

Our approach

Our empirical site: software engineers



Our approach

Our empirical site: software engineers

- 44.8% of computer and mathematical jobs (including software engineers) are at risk of being replaced by AI.

- So far, people conduct surveys (Wang et al., 2023, Feng et al., 2024) or do in-depth interviews (Woodruff et al., 2024)

  => Lack of the general overview; not substantive evidence of labor market

stackoverflow About Products OverflowAI Search... Log in Sign up

Home
Questions
Tags
Users
Companies
LABS
Discussions
COLLECTIVES
Communities for your favorite technologies. Explore all Collectives
TEAMS
Ask questions, find answers and collaborate at work with Stack Overflow for Teams.
Try Teams for free
Explore Teams

How to label or rename bin ranges in a series output from value count

Ask Question

Asked 2 days ago Modified 2 days ago Viewed 46 times

1

In a series or df column, I want to count the number of values that fit within predefined bins (easy) and meaningfully label the bin values (problem).

```
import pandas as pd

data = [{'A': 1, 'B': "Jim"}, {'A': 5, 'B': "Jim"}, {'A': 2, 'B': "Bob"}, {'A':
df = pd.DataFrame(data)

mBins = [-1, 2, 4, 6]
mLabels = ["0-2", "3-4", "5-6"]

simple_VC = df["A"].value_counts(bins=mBins)
```

```
Out[25]:   # ugly bin values
(-1.001, 2.0]    2
(2.0, 4.0]       1
(4.0, 6.0]       1

# Wanted more meaningful bin values:
0-2    2
3-4    1
5-6    1
```

I've tried using pd.cut, which allows me to label the bins, but I'm not sure how to use this in a value count. I've also tried to rename, but I don't know how to specify values like (4.0, 6.0], which are neither text or non-text.

How do I label the binned value counts - if possible during the value count, and how to rename bin ranges?

python pandas

The Overflow Blog

The developer skill you might be neglecting

Featured on Meta

Voting experiment to encourage people who rarely vote to upvote

Upcoming Experiment for Commenting

Results and next steps for the Question Assistant experiment in Staging Ground

Related

691 Converting a Pandas GroupBy multiindex output from Series back to DataFrame

744 How can I get a value from a cell of a dataframe?

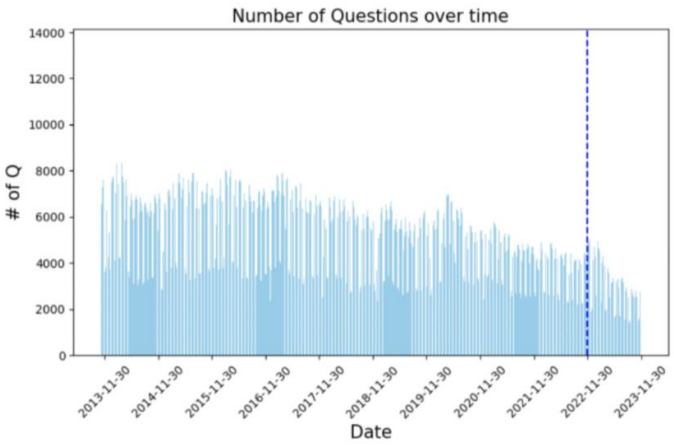
945 How to see normal stdout/stderr console print() output from code during a pytest run?

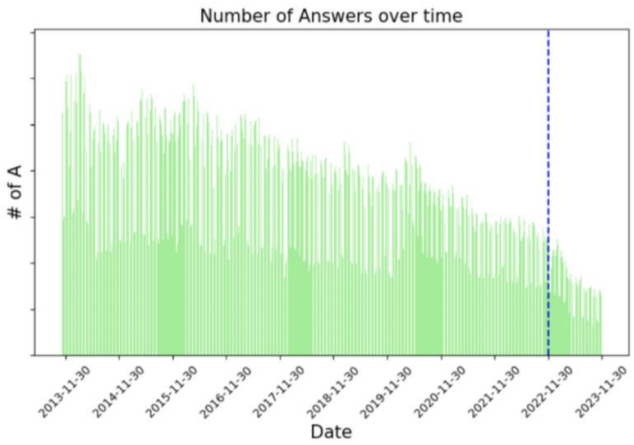
670 How to get the return value from a thread?

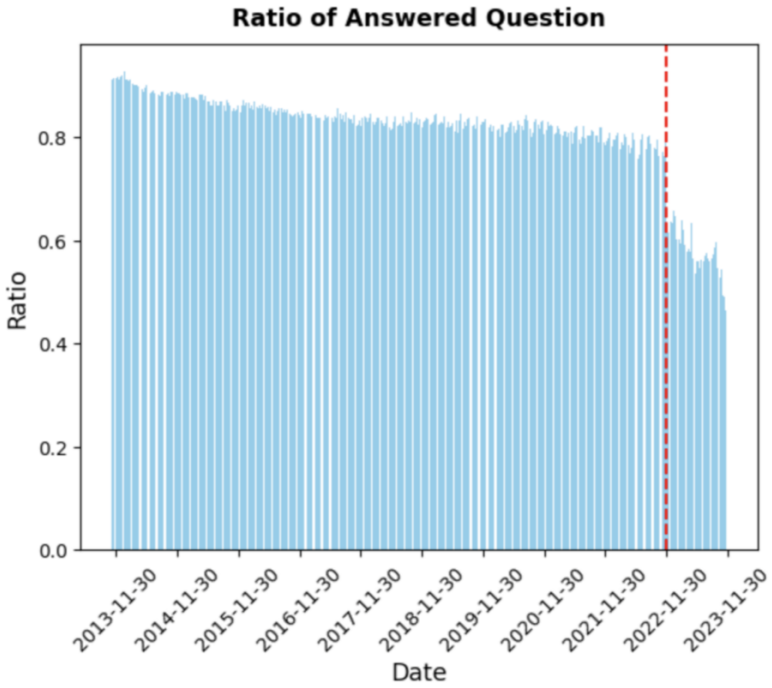
1 Converting time series into a heatmap

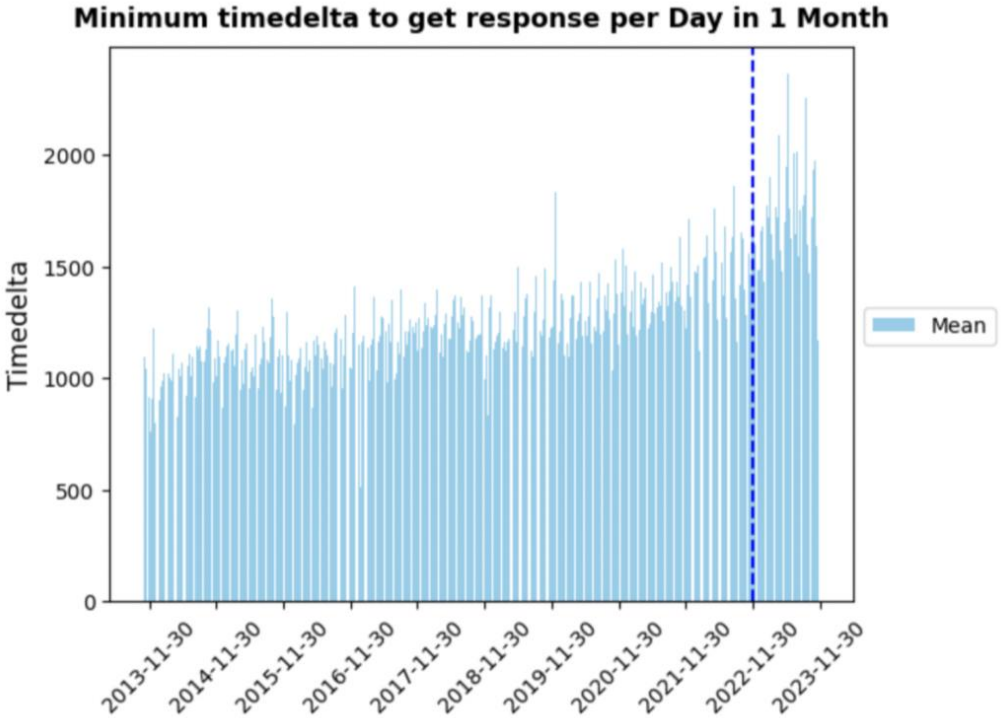
1 Binning Pandas column of timestamps

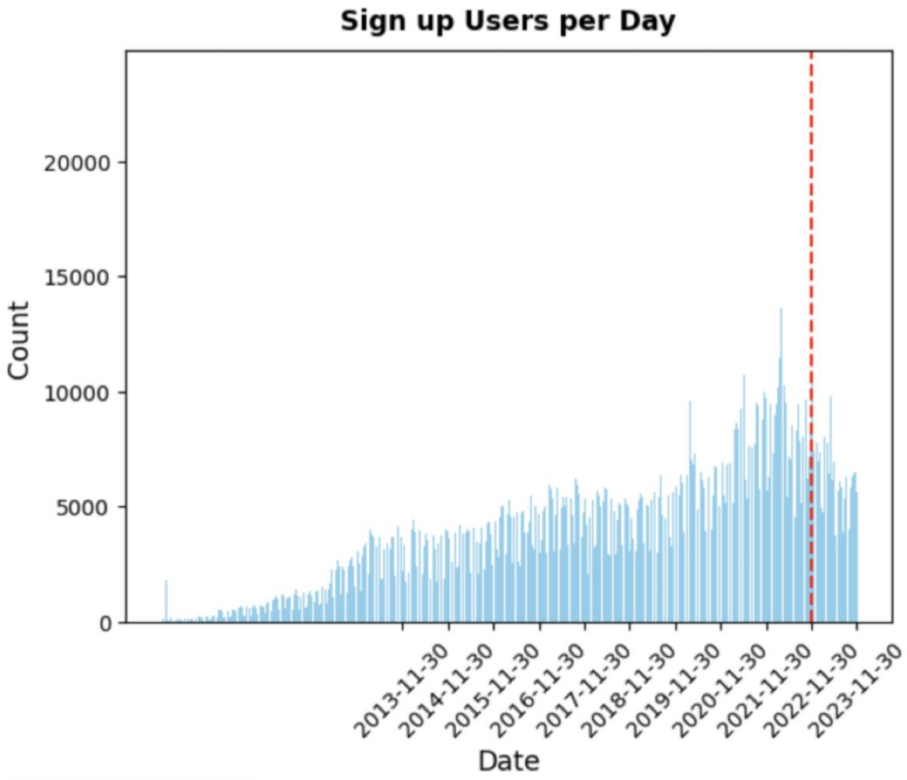
2 Is there a built in way, using pandas, to find the amount of values below a threshold for bins?






Ratio of Answered Question
Ratio
0.0
0.2
0.4
0.6
0.8
2013-11-30
2014-11-30
2015-11-30
2016-11-30
2017-11-30
2018-11-30
2019-11-30
2020-11-30
2021-11-30
2022-11-30
2023-11-30
Date


Minimum timedelta to get response per Day in 1 Month
Timedelta
0
500
1000
1500
2000
2013-11-30
2014-11-30
2015-11-30
2016-11-30
2017-11-30
2018-11-30
2019-11-30
2020-11-30
2021-11-30
2022-11-30
2023-11-30
Mean

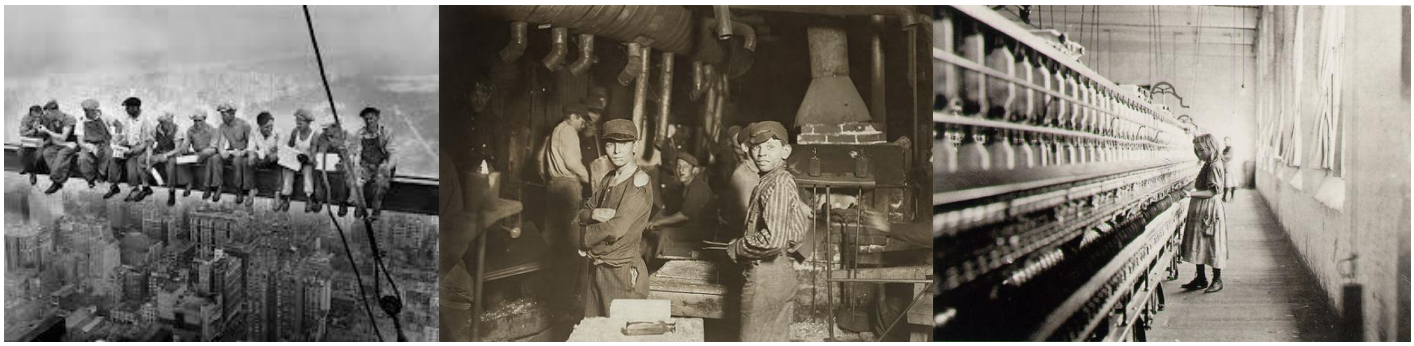


This is not new.

16

"Peasants Farming" Granger

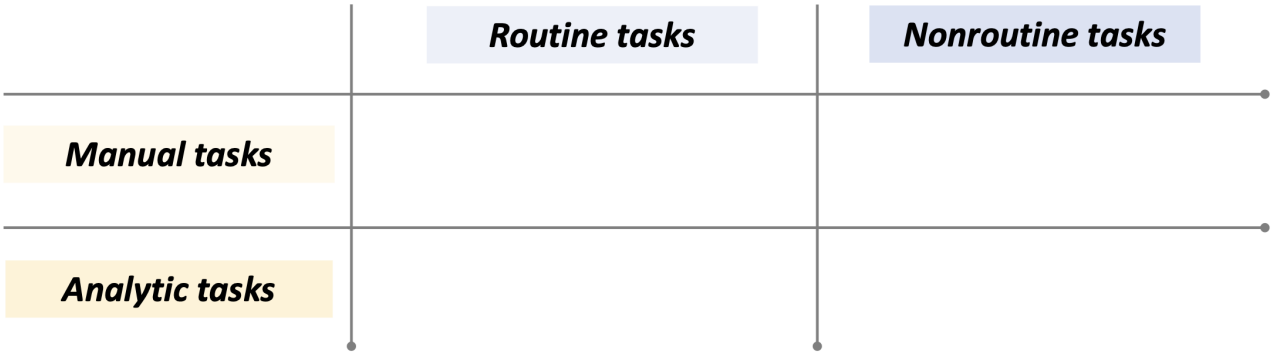
The Bodleian Library, Oxford



(left) Lunch atop a Skyscraper (1932) (middle) Indiana glass works boys (1908) (right) Cotton mill girl (1908)
Photographed by Lewis Hine

Theoretical ground

Automation impact towards Human Labor



19

Theoretical ground

Automation impact towards Human Labor

	Routine tasks	Nonroutine tasks
Manual tasks	Picking or sorting Repetitive assembly	Janitorial services Truck driving
Analytic tasks	Calculation Repetitive Customer service	Medical diagnosis Legal writing

Difficulty of the questions

20

Theoretical ground

Automation impact towards Human Labor

Data availability

	Routine tasks	***Nonroutine tasks***
Manual tasks	Picking or sorting Repetitive assembly	Janitorial services Truck driving
Analytic tasks	Calculation Repetitive Customer service	Medical diagnosis Legal writing

21

Research questions

Do *difficult questions* in Stack Overflow decrease faster after ChatGPT than *easy questions*?

Do *digitized (=textized) topics* in Stack Overflow decrease faster after ChatGPT than *less digitized topics*?

22

Data Extraction

23

Research questions

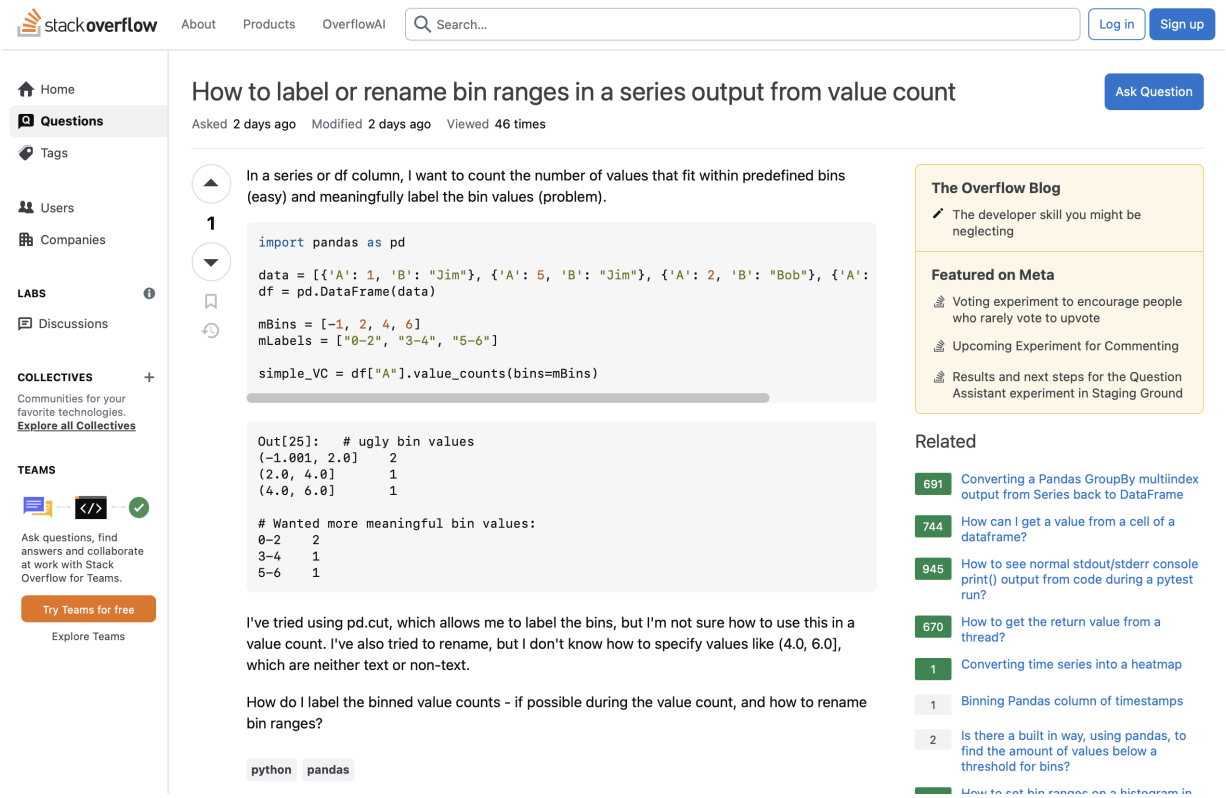
Do *difficult questions* in Stack Overflow decrease faster after ChatGPT than *easy questions*?

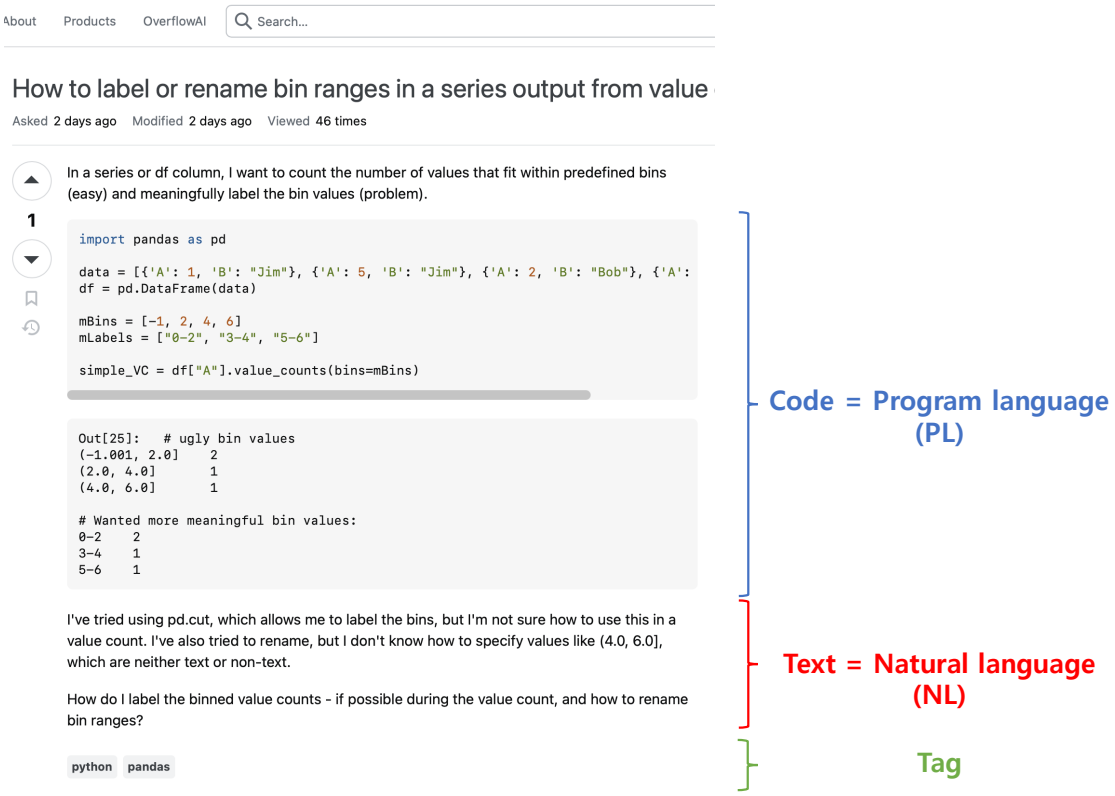
Do *digitized (=textized) topics* in Stack Overflow decrease faster after ChatGPT than *less digitized topics*?

24

RQ1: Do *difficult questions* in Stack Overflow decrease faster after ChatGPT than *easy questions*?

- Analytic challenge: **measuring difficulty of questions**



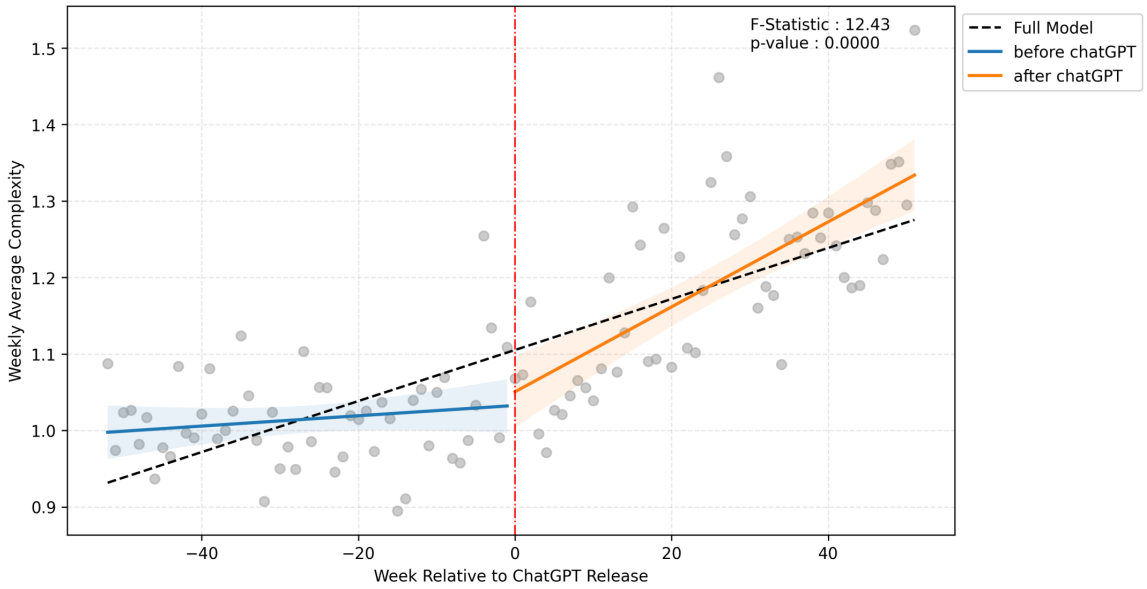


Method	Purpose	Data
Code Complexity	Calculate the Code Complexity in the source code measures the level of human effort to comprehend	python PL of question 2021-11-30 ~ 2023-11-30
Difficulty Measure	Measure the difficulty of each question based on the rubric evaluates the technical sophistication based on content	python NL+PL question 2021-11-30 ~ 2023-11-30

Method	Purpose	Data
Code Complexity	Calculate the Cognitive Complexity in the source code measures the level of human effort to comprehend	python PL of question 2021-11-30 ~ 2023-11-30
Difficulty Measure	Measure the difficulty of each question based on the rubric evaluates the technical sophistication based on content	python NL+PL question 2021-11-30 ~ 2023-11-30

29

Code Complexity



The complexity of source code **increases** after ChatGPT.

Method	Purpose	Data
Code Complexity	Calculate the Code Complexity in the source code measures the level of human effort to comprehend	python PL of question 2021-11-30 ~ 2023-11-30
Difficulty Measure	Measure the difficulty of each question based on the rubric evaluates the technical sophistication based on content	python NL+PL question 2021-11-30 ~ 2023-11-30

31

What is a difficult question?

32

Difficulty level	General rule set	Granular breakdown
Basic	Questions on simple built-in functions/API documentation/beginner level knowledge	Regular Built-in-function
Basic	Questions related to comparison between concepts and functions of various languages	Analysis of various languages' functions
Basic	Questions about simple problem-solving or random topic	Simple problem solving
Basic	Questions with simple exception, error, and other problem	Solve for null reference issue
Intermediate	Questions that require a relatively deeper understanding of the language to answer, for example Why type questions	Advanced features of a language that require deeper understanding
Intermediate	Questions where the questioner knows about the answer/solution but wants to know a more efficient one	Looking for contextually suitable solution despite having a solution
Intermediate	Questions related to time complexity, memory usage or other different resource usages of a system/solution	Efficient way
Intermediate	Questions need conceptual reasoning of programming construct/design principle	Reverse programming
Advanced	Questions that deal with hard/critical problems where solution needs in-depth programming knowledge or conceptual/logical thinking	Solution needs in-depth programming knowledge or conceptual thinking
Advanced	Questions that require advanced in-depth knowledge of internal language structure	In-depth knowledge of internal language structure
Advanced	Questions that deal with infrequently used functions	Deals with infrequently/rarely used framework/API/functions
Advanced	Question that requires in-depth knowledge about software architecture & SDLC	In-depth knowledge of software architecture
Advanced	Related to production environment	Efficiency related question
Advanced	Question that deals with rare and diversified topics	Need in-depth knowledge of multiple topics

(Raida et al., 2024)

33

Basic Level

How to calculate the total of each list, in a list of lists Python

…ked 5 years, 6 months ago Modified yesterday Viewed 2k times

0

I have collected a list of lists, each list representing data from a single day. I need to find the SUM of these to calculate the total volume each day. I can only seem to add together each list, not an individual lists data.

Provides the total of all the lists, not each individual lists total.

```
for ele in range(0, len(y_pred)):
    total = total + y_pred[ele]


print (total)
```

Expected 18 outputs, each lists sum, not one output with a sum of everything.

34

Expired tokens are not deleted after expiration in django-rest-knox 4.1.0

Asked 2 years ago Modified 2 yrs ago Viewed 227 times

Advanced Level

rest-knox 4.1.0 . In its documentation it says that expired tokens are deleted automatically. But it is not deleting the tokens.

1

```
REST_FRAMEWORK = {
```

Too time consuming and not scalable.

```
"AUTH_TOKEN_CHARACTER_LENGTH": 321,
"TOKEN_TTL": timedelta(minutes=10),
}
```

In settings.py I gave 10 minutes for expiration of token (for testing purposes). "TOKEN_TTL": timedelta(minutes=10)

I check the database after that time, they are not deleted. pgadmin knox token table Also I try to send request with those expired tokens, the respond is successful.

python django django-rest-framework token django-rest-knox

35

Research Design

Difficulty Measure

1. Sampling some Question

Sampling the questions based on criteria

2. Annotate the difficulty of samples

Using the Rubric from previous study

3. LLM Tuning

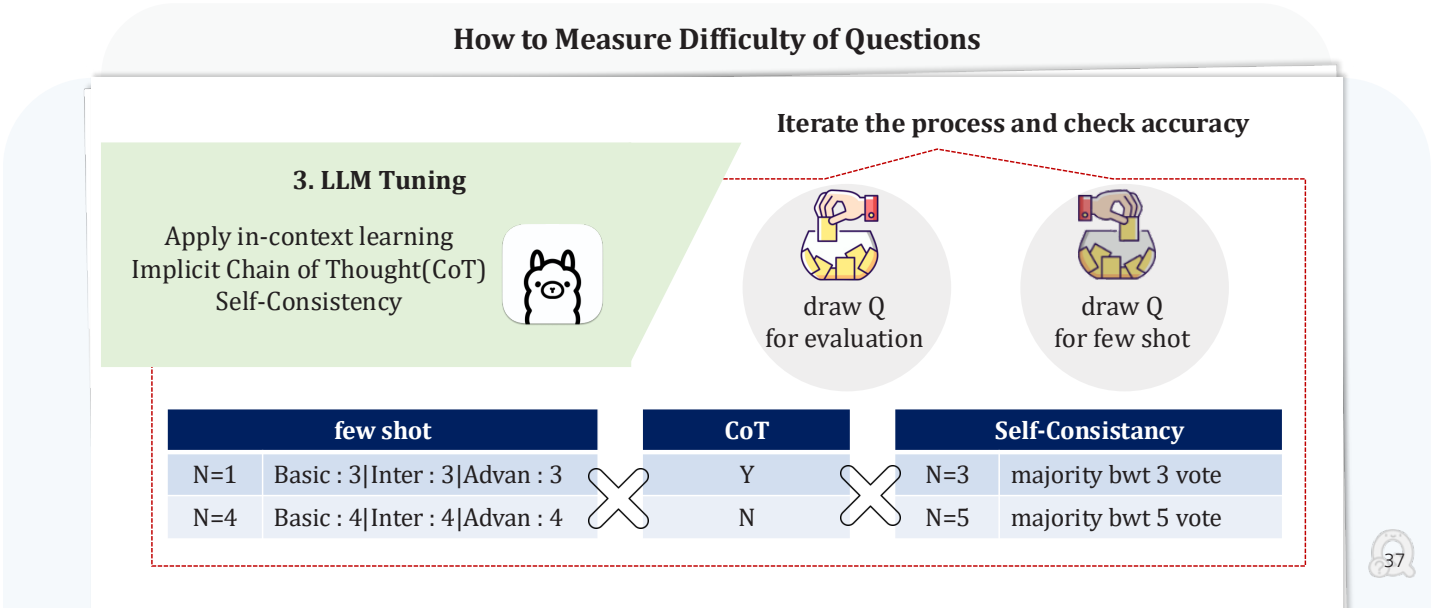
Apply in-context learning
Chain of Thought
Self-Consistency

4. Annotate target questions

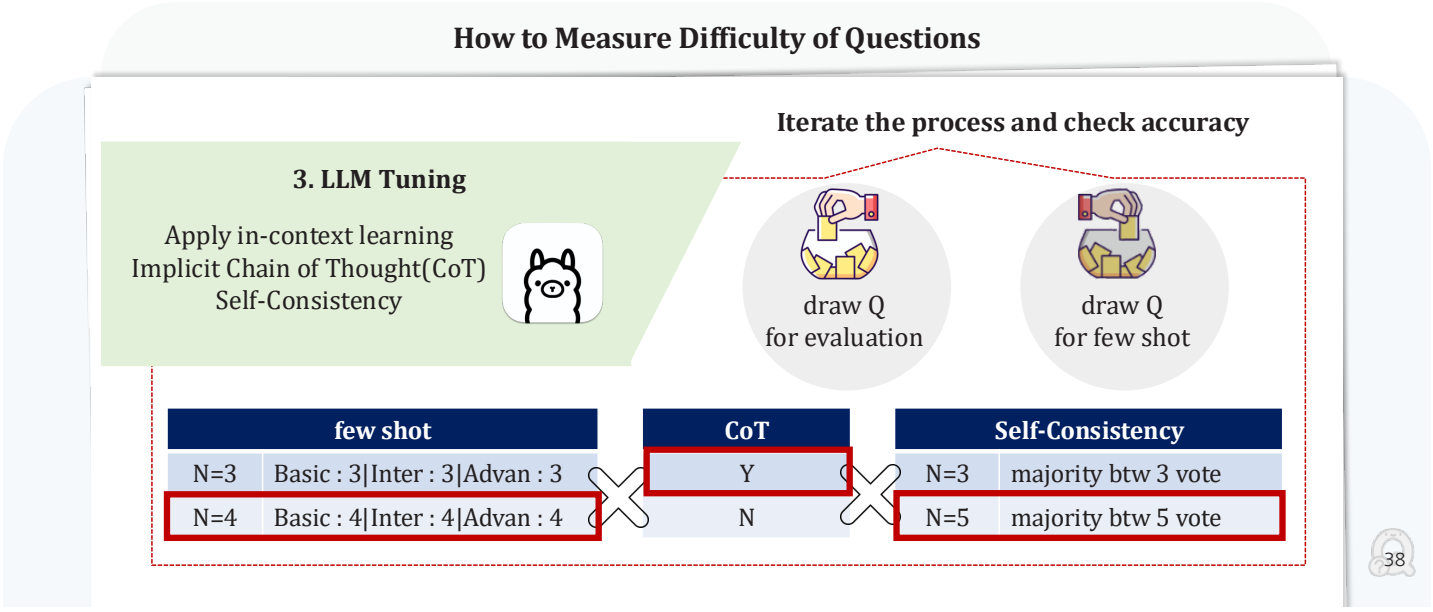
Let the annotation model evaluate the difficulty of questions.
(Annotated Q) N = 57,900

36

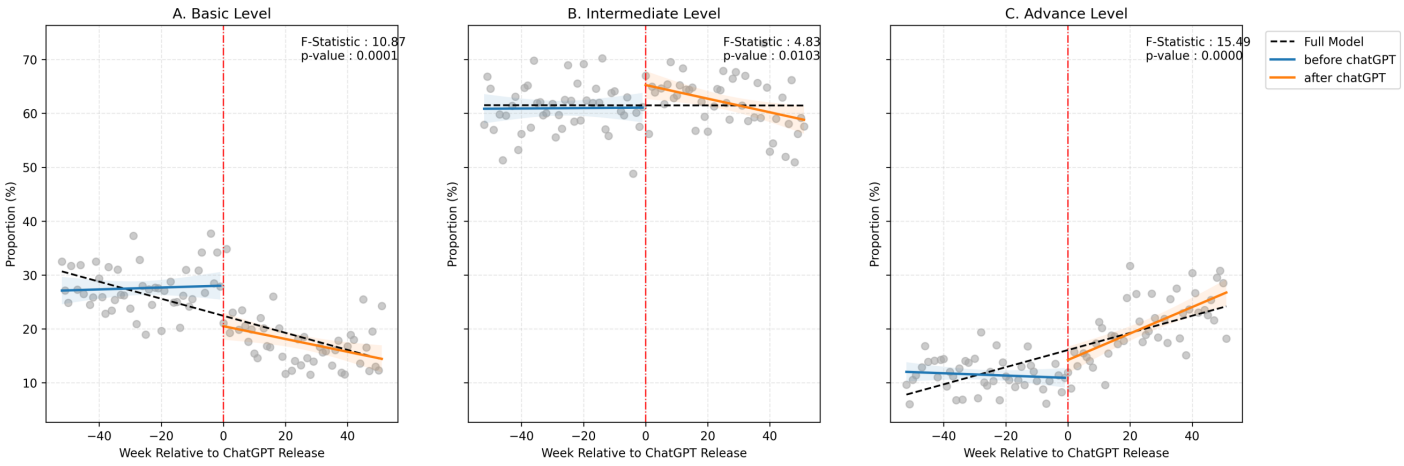
Research Design



Research Design



Changes in difficulty of question composition



The percentage of Basic/Intermediate Level questions significantly decreased

The percentage of Advanced Level questions significantly increased

Research questions

Do *difficult questions* in Stack Overflow decrease faster after ChatGPT than *easy questions*?

Do *digitized (=textized) topics* in Stack Overflow decrease faster after ChatGPT than *less digitized topics*?

RQ2: Do *digitized (=textized) topics* in Stack Overflow decrease faster after ChatGPT than *less digitized topics*?

- Analytic challenge: **measuring digitized topics**

41

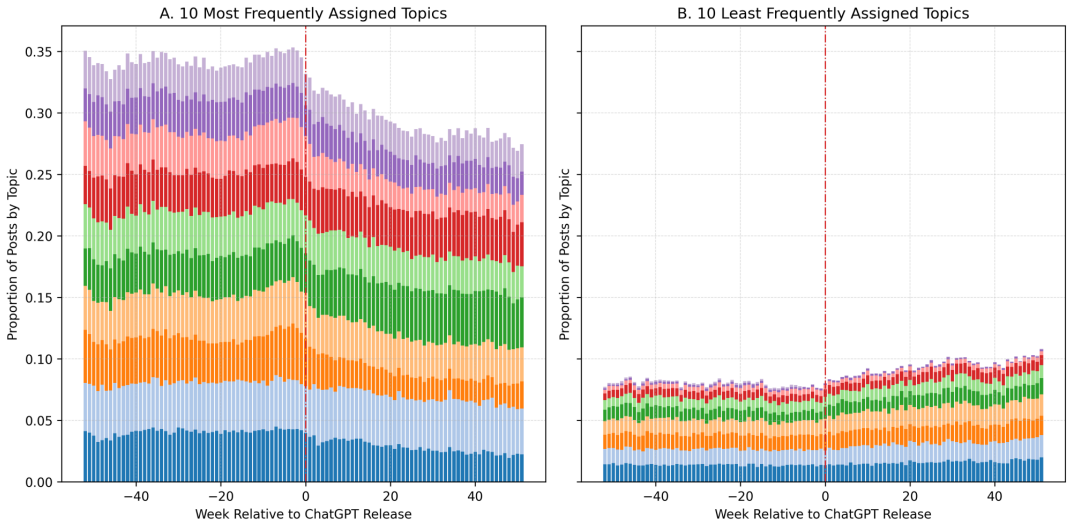
Method	Purpose	Data
Topic composition	Measure the amount of digitized topics by using BERTopic model and its topic composition	python NL question 2021-11-30 ~ 2023-11-30
Tag composition	Measure the amount of digitized topics by analyzing tag composition	All tags Tag only 2021-11-30 ~ 2023-11-30

42

Method	Purpose	Data
Topic composition	Measure the amount of digitized topics by using BERTopic model and its topic composition	python NL question 2021-11-30 ~ 2023-11-30
Tag composition	Measure the amount of digitized topics by analyzing tag composition	All tags Tag only 2021-11-30 ~ 2023-11-30

43

Changes in topic composition

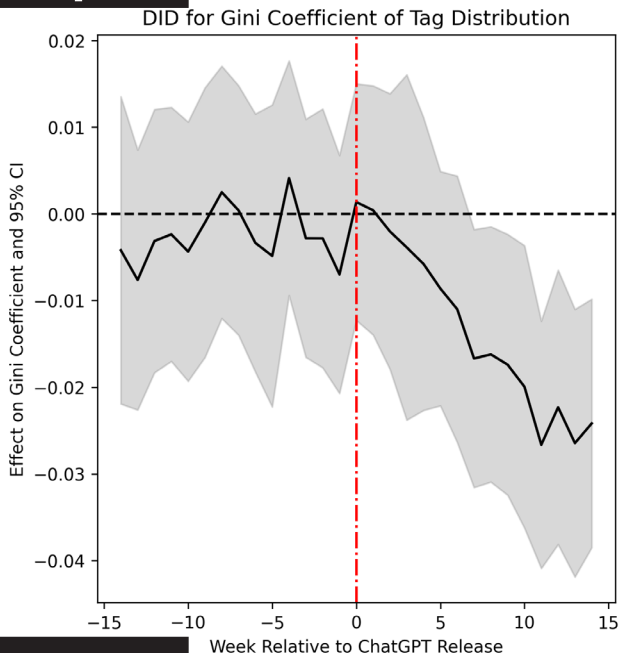


The composition of less digitized topics (B) increases after ChatGPT.

Method	Purpose	Data
Topic composition	Measure the amount of digitized topics by using BERTopic model and its topic composition	python NL question 2021-11-30 ~ 2023-11-30
Tag composition	Measure the amount of digitized topics by analyzing tag composition	All tags Tag only 2021-11-30 ~ 2023-11-30

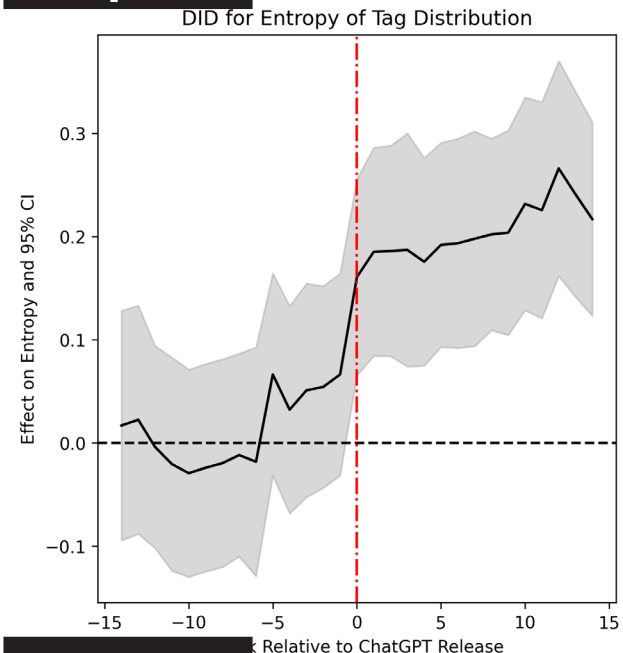
45

Unequal



Equal

Equal



Unequal

46

Research questions

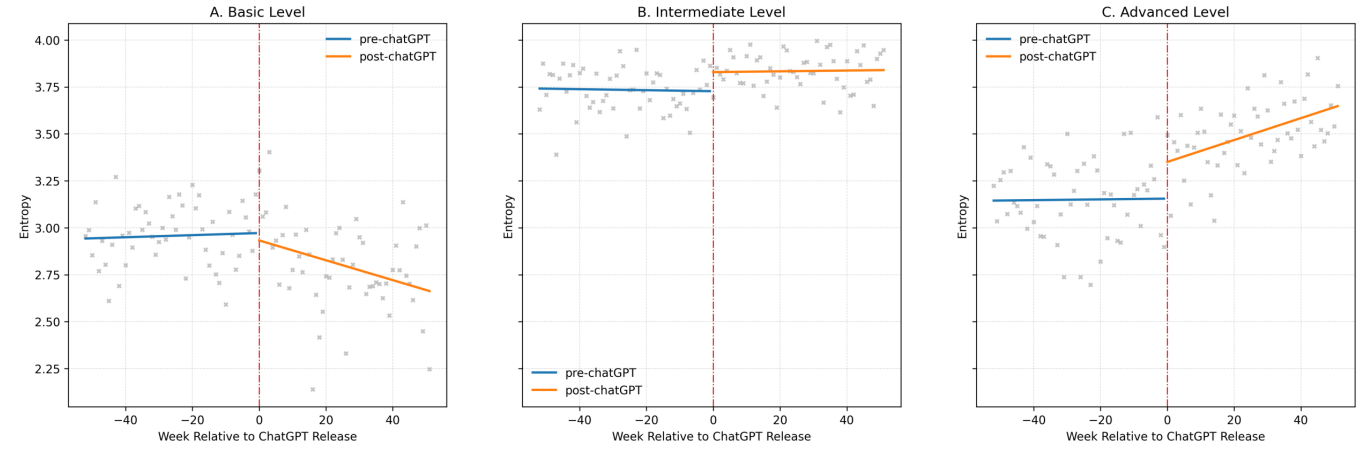
Do *easy and difficult questions* in Stack Overflow decrease equally after ChatGPT? **No!**

Do *digitized (=textized) topics* in Stack Overflow decrease faster after ChatGPT than *less digitized topics*? **Yes!**

47

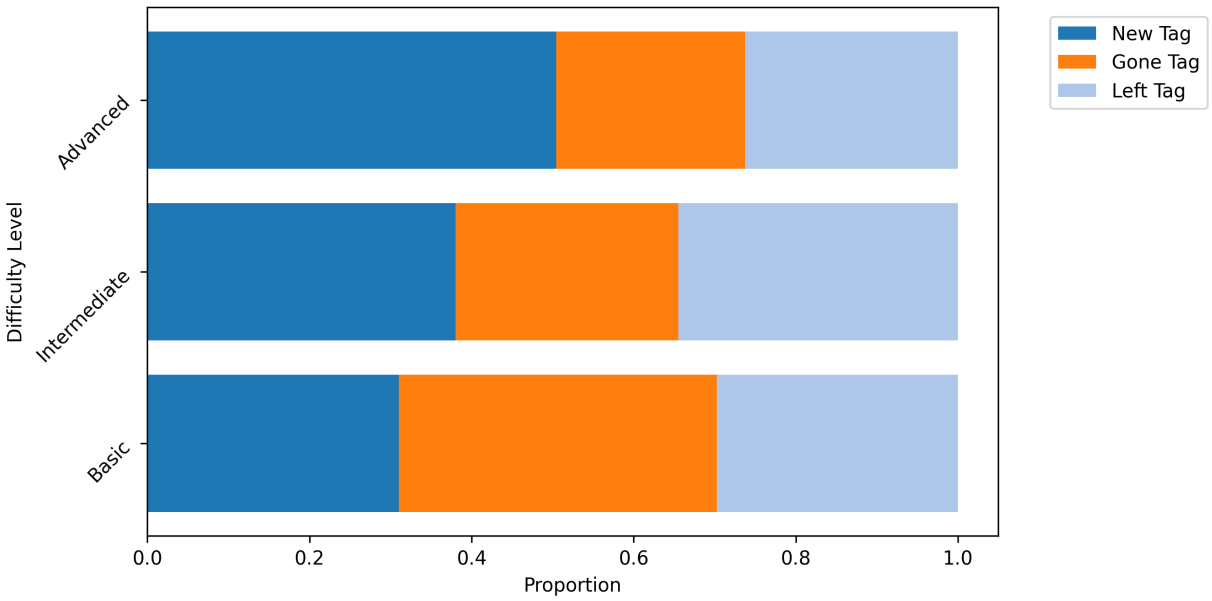
How about inter-relationship of these two?

Equal

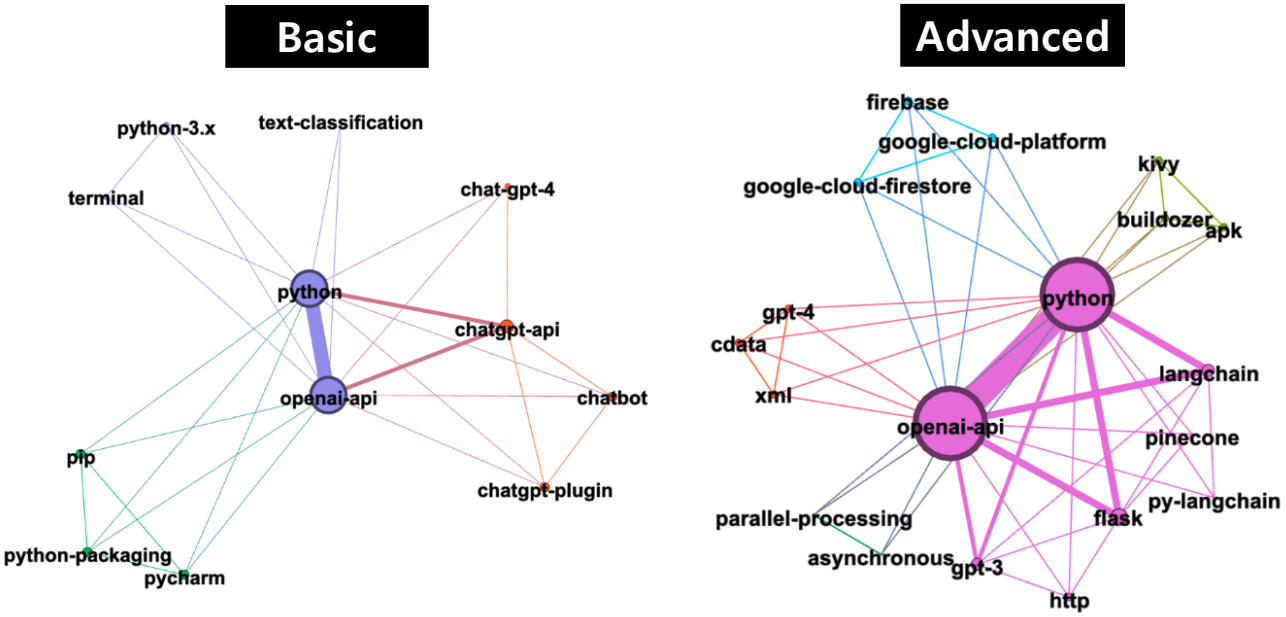


Unequal

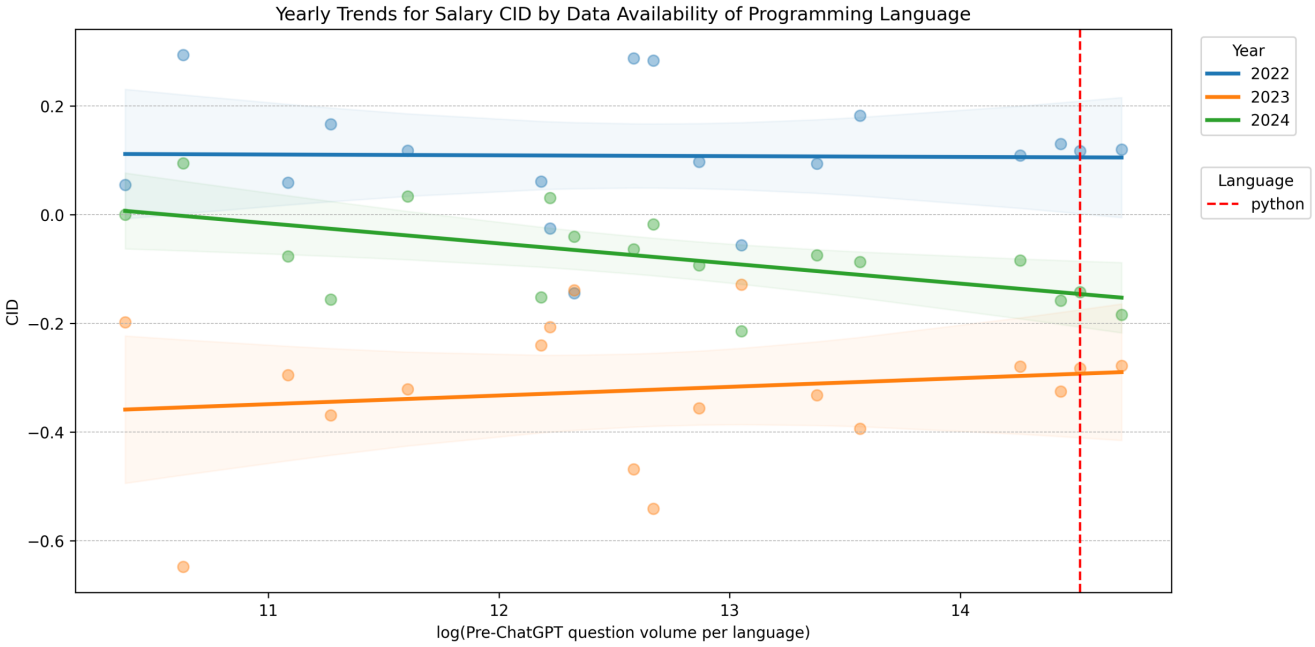
48


New Tag
Gone Tag
Left Tag
Difficulty Level
Advanced
Intermediate
Basic
0.0
0.2
0.4
0.6
0.8
1.0
Proportion

49


Basic
python-3.x
text-classification
terminal
chat-gpt-4
python
chatgpt-api
openai-api
chatbot
pip
chatgpt-plugin
python-packaging
pycharm
Advanced
firebase
google-cloud-platform
kivy
google-cloud-firestore
buildozer
apk
python
gpt-4
cdata
langchain
xml
openai-api
pinecone
py-langchain
parallel-processing
flask
asynchronous
gpt-3
http

50



51

Summary

- More intensive inequality story at the labor market.
- In any forms of AI, it is likely to be reproduced.

52

Questions?

53

Collaboration between school meal cooks and robotic chefs

SuMin Park,
Korea Labor Institute

Back Grounds

Industry Automation now extends beyond manufacturing to service industries.

- In Korea, the service sector accounts for 57% of GDP but features low capital/employment. The restaurant industry suffers from labor shortages and low wages.
- Service and manufacturing sectors differ significantly in capital, employment size, work characteristics, labor composition, and wage levels.
- In 2023, food service was chosen as a priority sector for robot adoption.

➔ How are robots introduced in the service industry changing workers' tasks, labor intensity, and working conditions?

Robots and the Transformation of Work

Robot adoption sparks diverse perspectives on how automation affects both employment and the nature of work.

Task - Job Replacement

- Focuses on economic impacts of technology
- Highlights job losses/gains and productivity growth *(Acemoglu & Autor, 2011; Acemoglu & Restrepo, 2019, 2020; Frey & Osborne, 2013; Park et al., 2023; Yoon, 2023)*
- Important to distinguish tasks from occupations
  – Task substitution ≠ Occupation substitution *(Autor et al., 2003; Autor, 2013)*

Changing Forms of Labor Division

- Heteromation shifts core tasks to end-users, making them integral to the automated system. Workers are both beneficiaries and operators *(Ekbia & Nardi, 2014)*
- Workers interpret machine outputs, handle unaccounted information, mediate customer-machine interactions, and adjust performance during failures*(Park, 2023; Fox et al., 2023; Quadri & D'Ignazio, 2022).* Customers share worker's responsibility *(Kelly et al., 2017)*
- Automation discourse obscures the human labor and infrastructure behind system operations *(Grey & Suri, 2018; Mateescu & Elish, 2019; Park, 2021)*

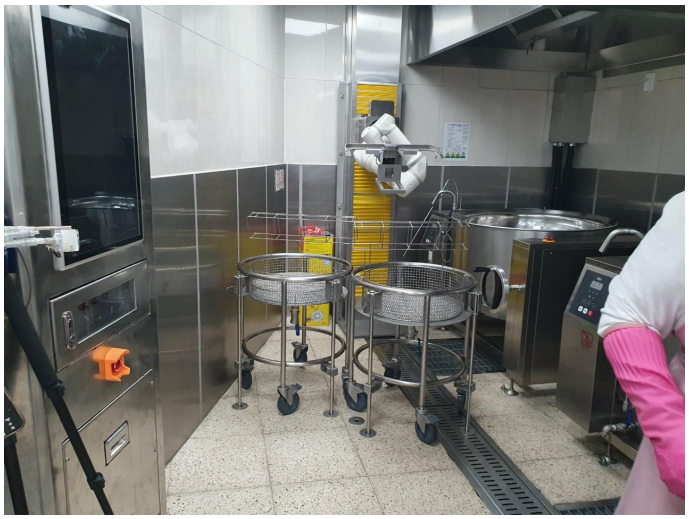
Methods

Field observation in 3 middle schools with robots.

- Field Study
- Analysis of work sequence, tasks, and ergonomics pre/post robot adoption.
- Group Interview after robot adoption.

	School A	School B	School C
Number of students served	750	950	470
Number of cooking staff	6	8	4
Visit before robot adoption	X	O	O
Visit after robot adoption	after 9 months	1 month after	Day of first meal service
Robots adopted	2 stir-fry robots 1 soup robot 1 frying robot	1 stir-fry robot 1 frying robot	1 frying robot (* can cook soup/stew)

Methods



School Meal Robot Deployment Status

Since its first adoption in Seoul (2023), school lunch robots have expanded rapidly. However, regional responses remain mixed due to differences in population density, labor markets, and school size.

Metropolitan/Provincial Office of Education	Status and Plan (as of Oct. 2024)
Seoul	2 schools adopted; 6 more planned
Incheon	1 school adopted; next year's budget secured
Busan	Under internal review
Daegu	1 school adopted; no additional plans (single-unit type)
Gwangju	Under internal review
Daejeon	None
Ulsan	None
Sejong	None
Gyeonggi	5 schools in procurement, expected adoption by Oct.
Gangwon	1 school adopted; to be included in next year's main budget but no appicant
Chungbuk	None
Chungnam	None
Jeonbuk	1 school adopted; future plans undecided
Jeonnam	No plan due to budget limitations
Gyeongbuk	1 completed, 2 in progress
Gyeongnam	Under internal review
Jeju	2 schools planned for adoption

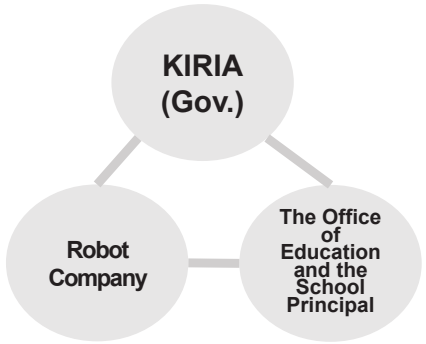
Background of School Meal Robot Adoption

Robots are being rapidly introduced in school kitchens for political, industrial and administrative reasons, but workers are excluded from the decision-making process.

Deployment Background

Workplace Safety	Conflict between labor union and education office due to high work intensity and safety (cancer caused by cooking fume)
Industrial Development	Government support for the promotion of the 'K-robot' industry
Labor Shortage	Operational challenges due to labor shortage in school lunch

Adoption Decision-Makers



Characteristics of School Meal Kitchen Labor

High deadline pressure, complex workflows, and difficulty in standardizing tasks; flexible cooperation among workers to handle limited staffing.

•School C: Approximately 470 students served (117 meals per worker)

<table>
<tr><th></th><th>Kitchen Worker 1</th><th>Kitchen Worker 2</th><th>Kitchen Worker 3</th><th>Kitchen Worker 4</th></tr>
<tr><td>07:30~08:00</td><td colspan="4">Ingredient receiving and Checking</td></tr>
<tr><td rowspan="8">08:00~11:00</td><td>Vegetable preparation</td><td colspan="2" rowspan="2">Vegetable preparation</td><td>Chicken preparation</td></tr>
<tr><td rowspan="2">Rice</td><td>Vegetable preparation</td></tr>
<tr><td colspan="2">Kitchen cleaning</td><td rowspan="5">Batter mixing, coating, and frying

Staging Food</td></tr>
<tr><td>Preparing kimchi for serving</td><td>Other side dishes</td><td>Dishwashing</td></tr>
<tr><td>Meal setup</td><td>Making sauce for fried d ishes</td><td rowspan="2">Soup</td></tr>
<tr><td>Dishwashing</td><td>Kitchenware washing</td></tr>
<tr><td>Preparing fruit for dessert</td><td>Frying assistance</td><td>Check vegetables for side dishes</td></tr>
<tr><td>11:00~11:30</td><td colspan="4">Staging food and setting the serving line / Kitchenware waching</td></tr>
<tr><td>11:30~12:00</td><td colspan="4">Break and Lunch for workers</td></tr>
<tr><td>12:00~13:00</td><td colspan="4">Meal distribution</td></tr>
<tr><td>13:00~15:30</td><td colspan="4">Dishwashing/Cleaning</td></tr>
<tr><td>15:30~16:30</td><td colspan="4">Rest, shower, and wrap-up</td></tr>
</table>

Workflow Changes with Frying Robot Introduction

Current cooking robots support individual tasks, not full workflows: alleviating musculoskeletal strain and reducing heat and respiratory risk

Before robot introduction	After robot introduction
	Loading battered chicken portions into the frying basket (by worker)
	Place the basket in the holding area (by worker)
	Touch panel activation (by worker)
	Lifting the basket from the holding area (by robot)
Placing chicken into boiling oil	Lowering chicken basket into pot (by robot)
Manual stirring with skimmer during frying	Basket shaking during frying(by robot)
Lifting the fully fried chicken with skimmer	Lifting the basket after frying(by robot)
Shaking off excess oil	Draining excess oil
Transferring it to a large mesh strainer	Placing the basket in the holding area
Skimming crumbs from oil	Skimming crumbs from oil
Moving fried contents to a serving tray	Moving fried contents to a serving tray

Replaced tasks by robots

repeat

9

한국노동연구원 Korea Labor Institute

Workflow Changes with Frying Robot Introduction

Current cooking robots support individual tasks, not full workflows: alleviating musculoskeletal strain and reducing heat and respiratory risk

- Tasks replaced

- Placing the chicken into the boiling oil
- Stirring it during frying (must stay close to the pot)
- Lifting the cooked chicken and shaking off excess oil (Shoulder strain and slippery floors). Stir-fry robots are more effective in reducing shoulder strain

10

한국노동연구원 Korea Labor Institute

Workers' Reactions to the Introduction of Cooking Robots

KLI 한국노동연구원

Reactions to robot performance are mixed; some workers express anxiety about adjusting to new work rhythms following robot adoption.

> "We can't just fry as much as we want. There's a certain limit. We have to portion it out to the amount that comes out best."
>
> "I thought the robot would just lift things up and drop them in for us. But it doesn't do that — if we want to use it, we have to pre-portion everything into under 10 kilograms first."
>
> "With regular jjajang, we just serve it directly from the tray. But with the robot, we have to divide it into batch 1, 2, and 3. And before, we only used four trays, but now it takes five — which means more dishwashing too."

- Fixed basket sizes require portioned cooking, often increasing the number of repetitions needed to produce the same quantity of food.
- Physical strain decreases with robot use, but total time may rise—especially when food must be divided into smaller portions or when workers are unfamiliar with the system.
- Existing time-optimized routines must be restructured to accommodate robot-based processes. Robot use is feasible only when tasks can still be completed within the required time frame.

Workers' Reactions to the Introduction of Cooking Robots

KLI 한국노동연구원

Positive impact on body strain and heat exposure, but path-related disruption observed.

> "I've been doing this for a long time, and my shoulders really hurt. So equipment like stirrers or fryers—it's better if we use them. They make things much easier. We lift heavy things all the time... those food trays, as you know, are really heavy."
>
> "When you're working in front of the pot, sometimes the heat makes it hard to breathe. But with the robot, that doesn't happen."
>
> "We move around too, but now the robot blocks part of the space. If I want to go that way, I have to walk around it. The workflow path has gotten too narrow."

- In compact school kitchens, the addition of robots has complicated movement flow. While collision-prevention features exist for robots, the restructured layout increases the risk of collisions among workers and with kitchen equipment such as carts.

Workers' Reactions to the Introduction of Cooking Robots

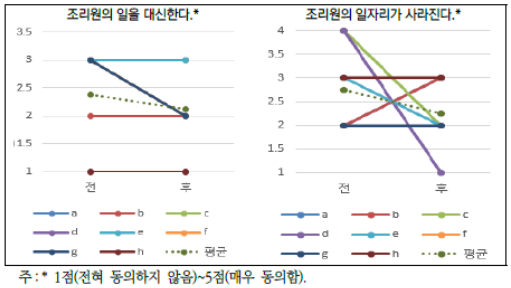
Initial robot operation difficulties are overcome through hands-on learning, grounded in prior experience.

- **On-site Support and Learning**
  "We were only able to learn so quickly because the engineer stayed with us for six months. If they had just installed it and left, I don't think we could've handled it."
- **Recipe Innovation by Field Workers**
  Kitchen workers adapted recipes to robot capabilities, enabling dishes previously thought unfit for automation.
- **Functional Repurposing**
  Workers proposed new uses for the robot fryer (e.g., boiling), leading to suggestions for hybrid pots and customized trays.

Workers' Perception to the Introduction of Cooking Robots

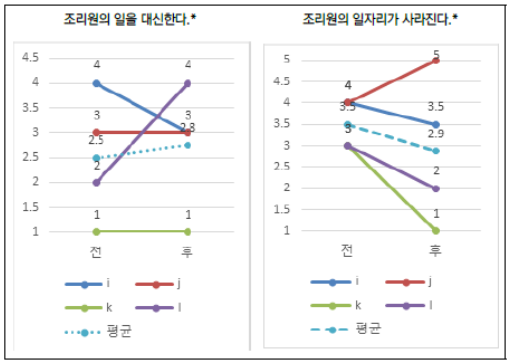
Workers perceived robots as highly capable of performing kitchen tasks, but did not see them as a serious threat to job security.

[그림 5-19] 로봇의 일자리 대체에 대한 인식(B학교)



주 : * 1점(전혀 동의하지 않음)~5점(매우 동의함).

[그림 5-26] 로봇의 일자리 대체에 대한 인식(C학교)



주 : * 1점(전혀 동의하지 않음)~5점(매우 동의함).

Conclusions

- **Labor Characteristics of School Meal Work**
  - School kitchens involve complex, non-standardized tasks under strict time pressure.
  - Compared to manufacturing, these environments remain highly labor-intensive and lack conveyor-style workflows.
  - Current robots alleviate physically demanding tasks but do not fully automate the process
  - While full job replacement appears unlikely, the substitution of specific tasks is increasingly feasible.
- **Exclusion of Workers from Robot Governance**
  - Kitchen workers were not meaningfully included in robot adoption decisions or safety/training programs.
  - This lack of inclusion may reduce acceptance and hinder effective utilization after deployment.
- **Need for Human-Robot Interaction and Reskilling**
  - Workers are not just end-users but integral actors in the technology adoption network.
  - Their practical knowledge allows for novel uses beyond initial design intentions.
  - Experience with robot operation may evolve into new roles or occupations.
  - Policy must consider how to transition middle-aged women in disappearing jobs into new assistive technology (AX) roles.

Discussion

Labor Quality, Satisfaction, and Emerging Challenges in Robotized Service Work

Category	Description	Examples
Task Substitution	Robots take over specific tasks or connect multiple robotic functions (e.g., smart orders, kiosks, serving robots) to substitute some human work.	Restaurant serving robots Hamburger patty cooking
Productivity Enhancement	Robots perform or assist with part of the task, allowing workers to focus on other duties, enhancing overall productivity and service quality.	Cafe beverage prep, restaurant serving robots (serving, grilling)
Labor Intensity Reduction	Although robots do not fully replace workers, they take over physically demanding parts of tasks, helping reduce labor burden overall.	School meal cooking, Pasta cooking, Restaurant service (serving, dish return)

- Can low wages be improved?

- Robots shift the nature of work
- Replacing active production with machine assistance may reduce job satisfaction
- School meal vs. Highway rest stops

- Cognitive/emotional labor often remains invisible
- In restaurant front-of-house work, robots may reduce physical tasks but increase the need for constant environmental monitoring—shifting workers toward more responsive, less visible forms of cognitive and emotional labor.

Discussion

Aspirations for automation and advancements in automation technologies are reinforcing each other, accelerating deployment across sectors.

Automated snack bar using a slicer, rice dispenser, and auto-cooker

- **Automation is not only about high-tech robotics**
  - In many workplaces, simple and low-cost automation tools can drive equally significant changes.
- **Barriers to 'automation' and 'optimization' are diminishing**
  - The psychological, technological, and financial thresholds for automation have lowered.
- **Rising labor costs and persistent labor shortages are fueling demand for automation**

session

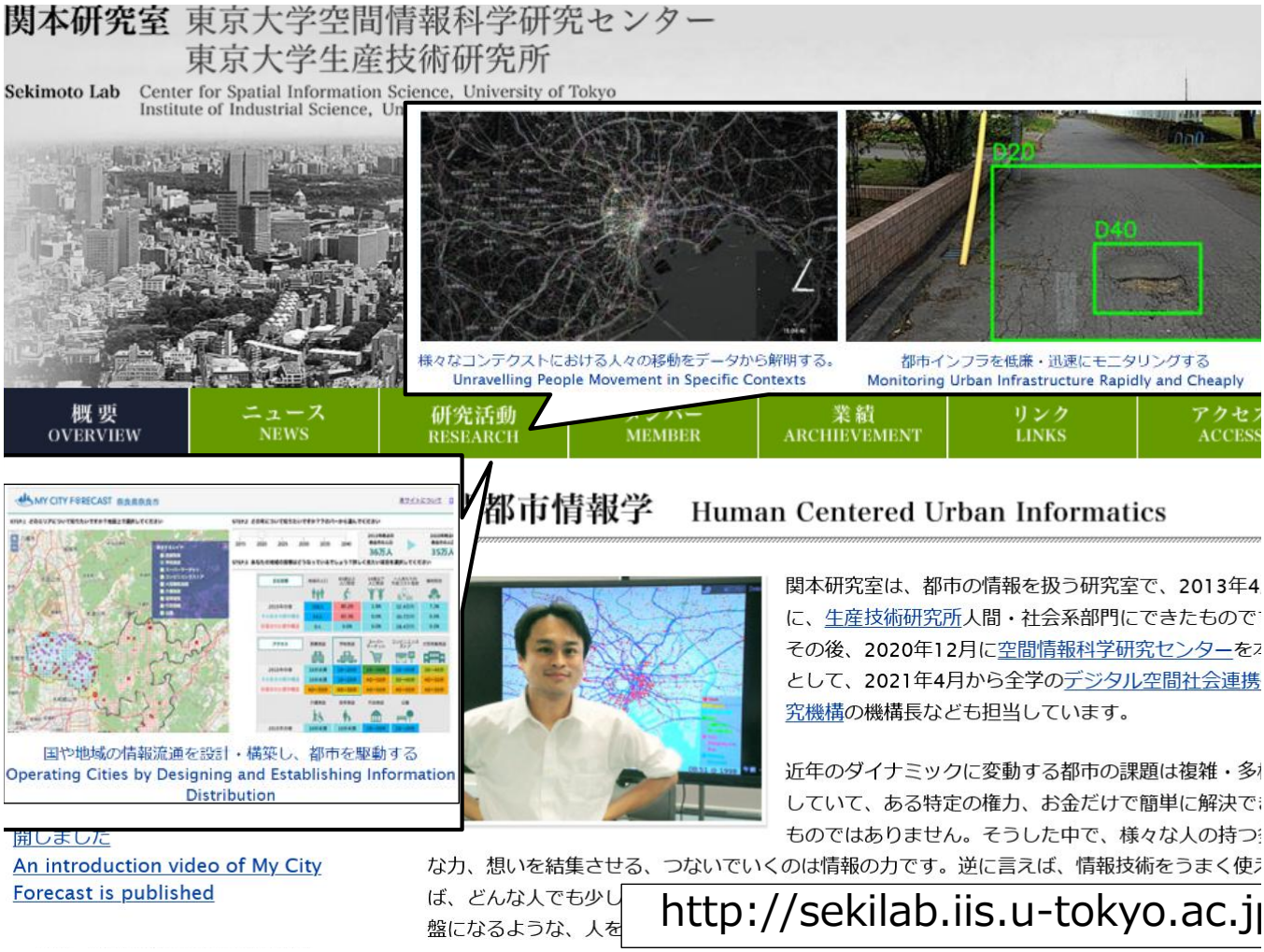
6

The 2025 International Conference of the Korean Social Science Research Council (KOSSREC) at Seoul, Korea (May 28, 2025)

Sustainable Development of Urban Digital Twin

Yoshihide SEKIMOTO, Professor

Director, Center for Spatial Information Science (CSIS), University of Tokyo



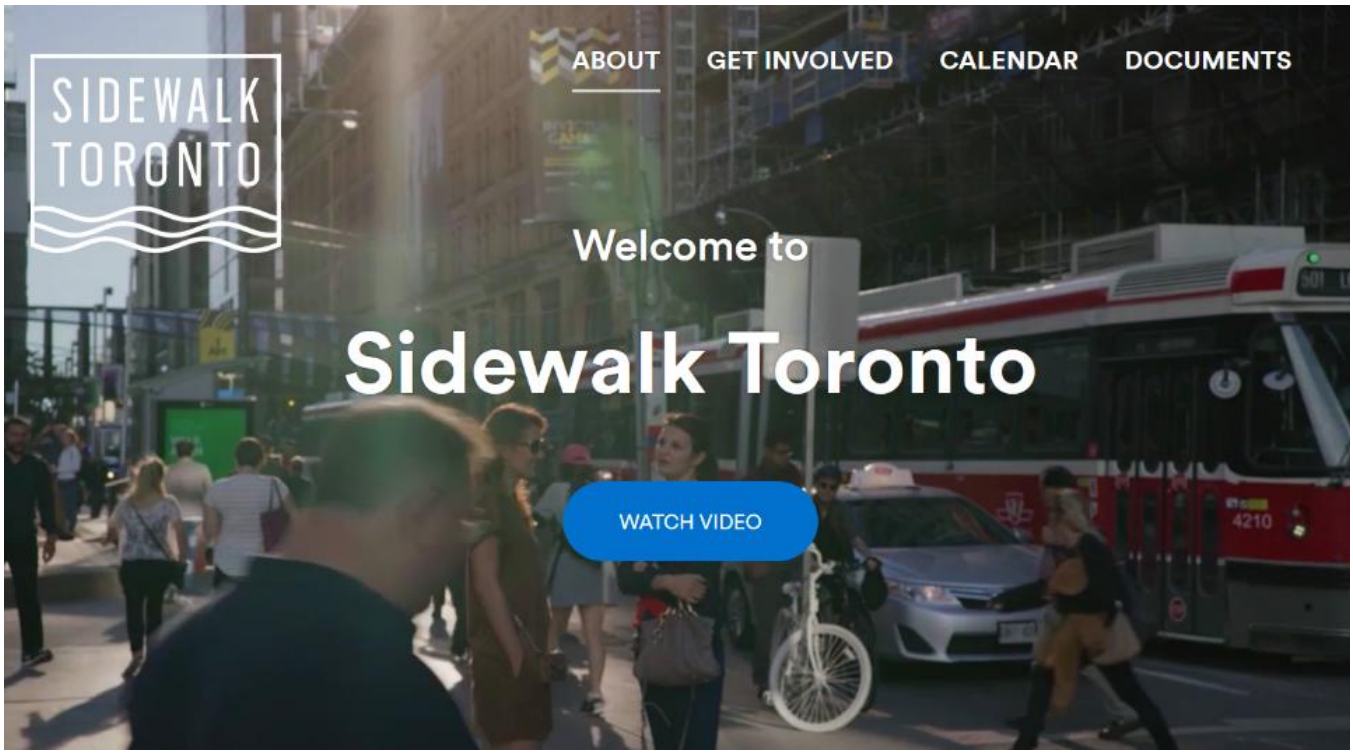
Today’s contents

- For smooth distribution of geospatial information & automatic generation of digital twin
- Public service development for citizen collaborated infrastructure monitoring: My City Report (MCR)
- Public service development for citizen collaborated long-term urban planning: My City Forecast (MCF)

How to realize sustainable digital smart city ??

Recent Smart City

- From sensors and devices to data centric



https://www.sidewalklabs.com/ 5

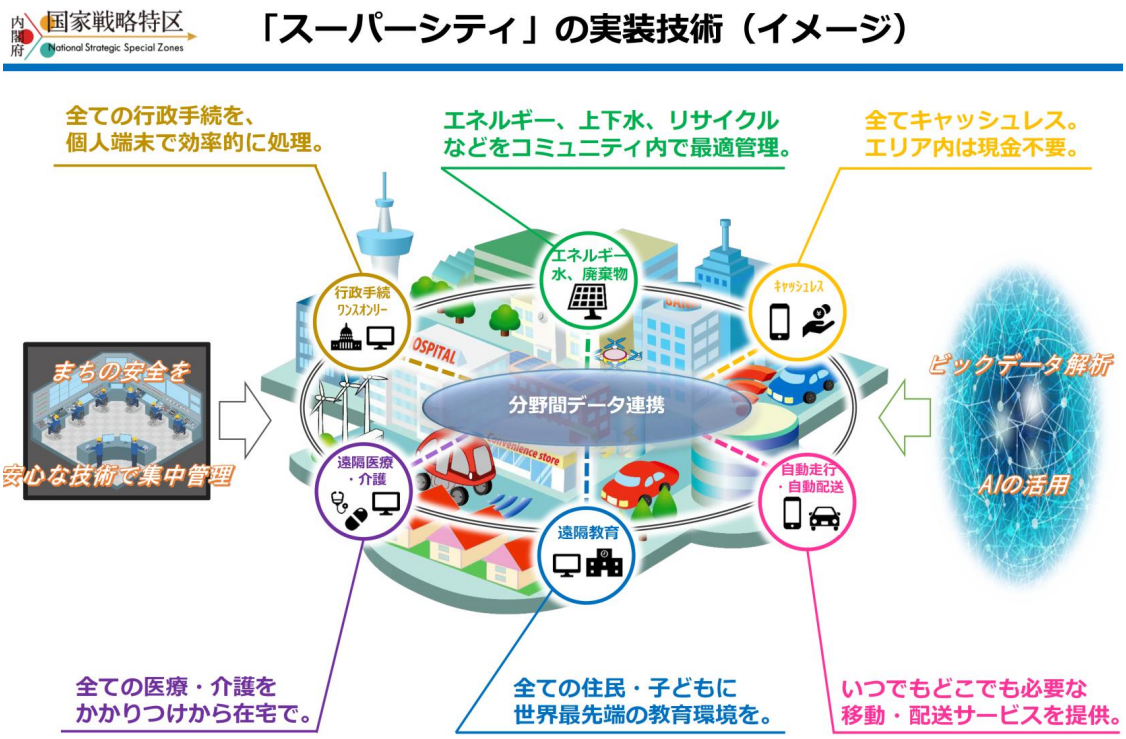
Alibaba city ?

- Alibaba “ET BRAIN”@杭州, China

https://jp.alibabacloud.com/product/datav 6

No Concept ?

- “Super city” policy in Japan Gov.



https://www.kantei.go.jp/jp/singi/tiiki/kokusentoc/supercity/setsumei.pdf 2

7

Current problems

- For researchers, AI and big data are normal, but for daily city operation, not normal.
- Methodology should be sustainable even for normal small cities.
- Almost all trials finish when projects (money) finish.
- Many data are not available when projects finish.

8

Our direction

- Develop effective grand design of urban data integrating each infrastructure service domain such as transportation, road, water, building, environment etc.
- Safe and smooth data integrating technology from different types of stakeholders such as local gov., national gov., private company and personal
- Citizen collaborative by default
- Don’t depend/trust too much on the local gov.

- Supporting real smart city is difficult but maintaining digital twin can be possible !
- Manage local digital twin sustainably from the open data by our own resources including neutral public but external sector (out of gov. sector)

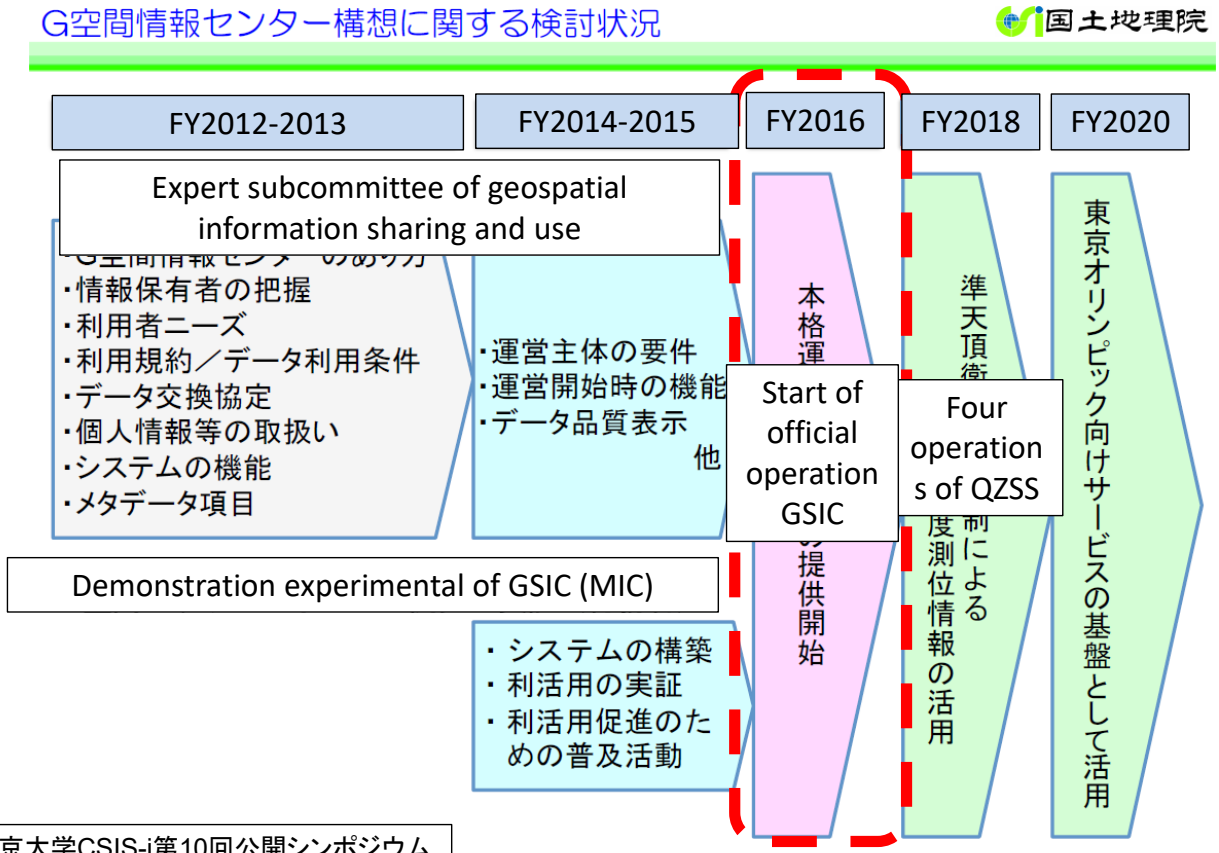
9

Web-based visualization technology trend

For smooth geospatial data distribution & automatic generation of digital twin from open data

11

Road map of Geospatial Information Center



東京大学CSIS-i第10回公開シンポジウム
国土地理院　村上氏資料より

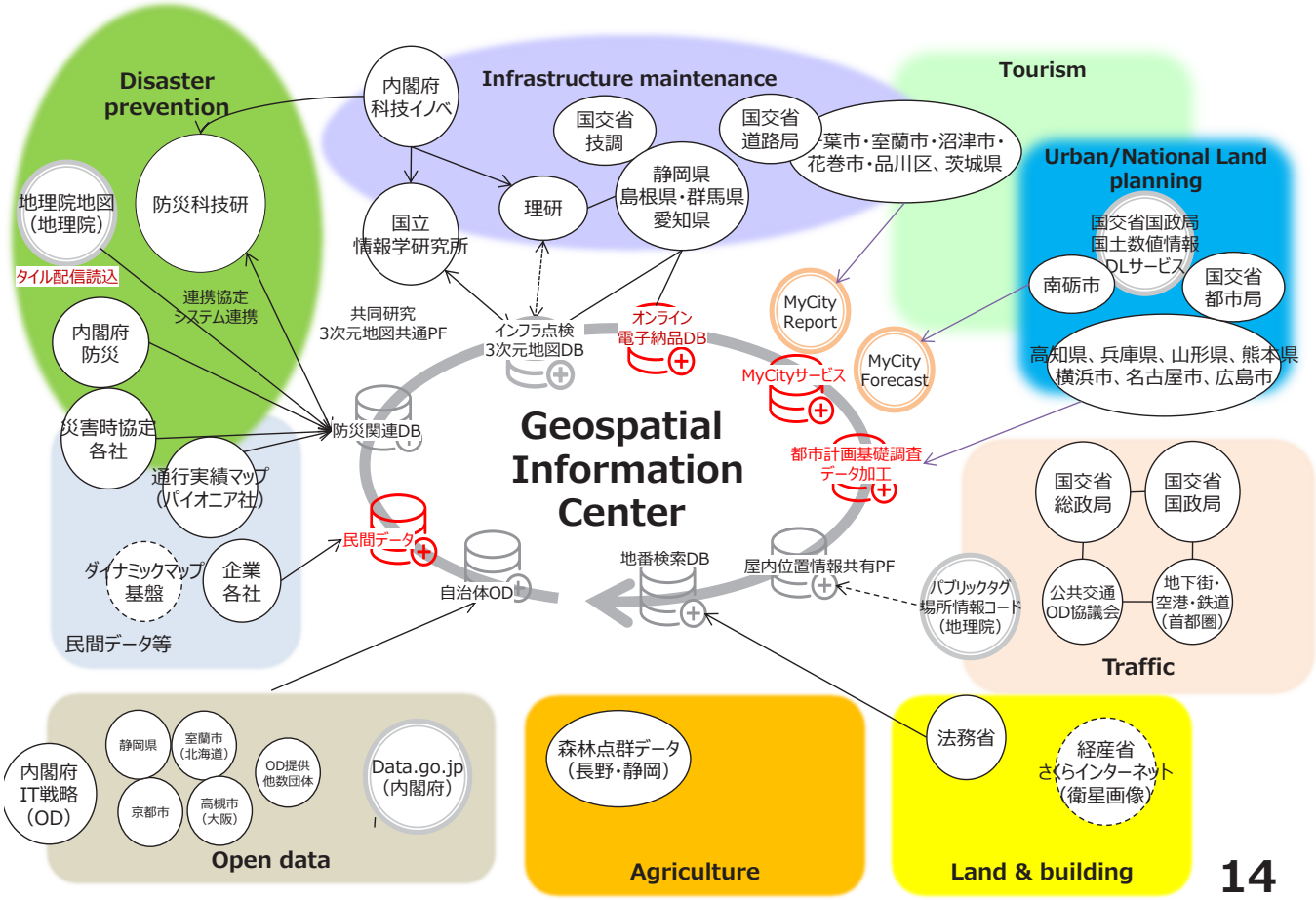
12

Geospatial Information Center

- Is business hub by distribution of business/open geospatial data since 2016.
- Is operated by AIGID (Association for Promotion of Infrastructure Geospatial Information Distribution)
- Profit should be kept by side businesses.

13

Core data and services in each field

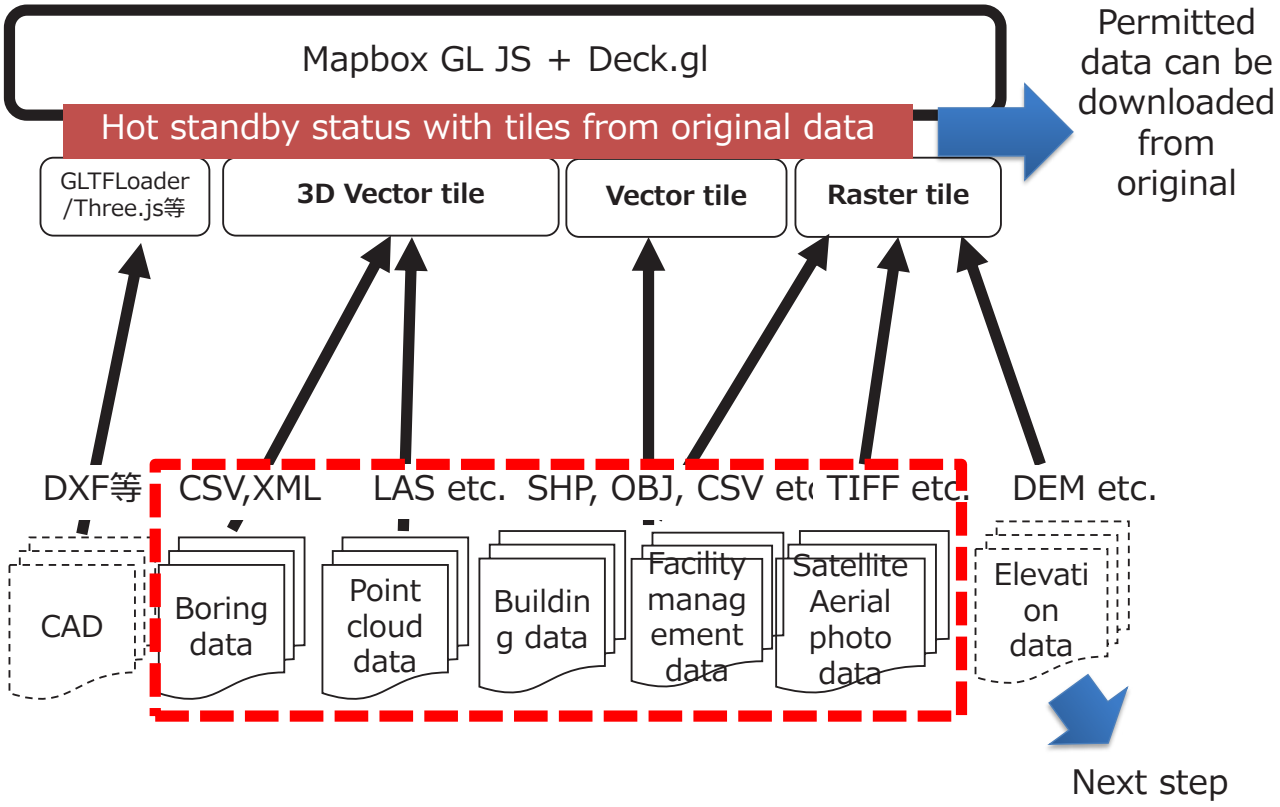


14

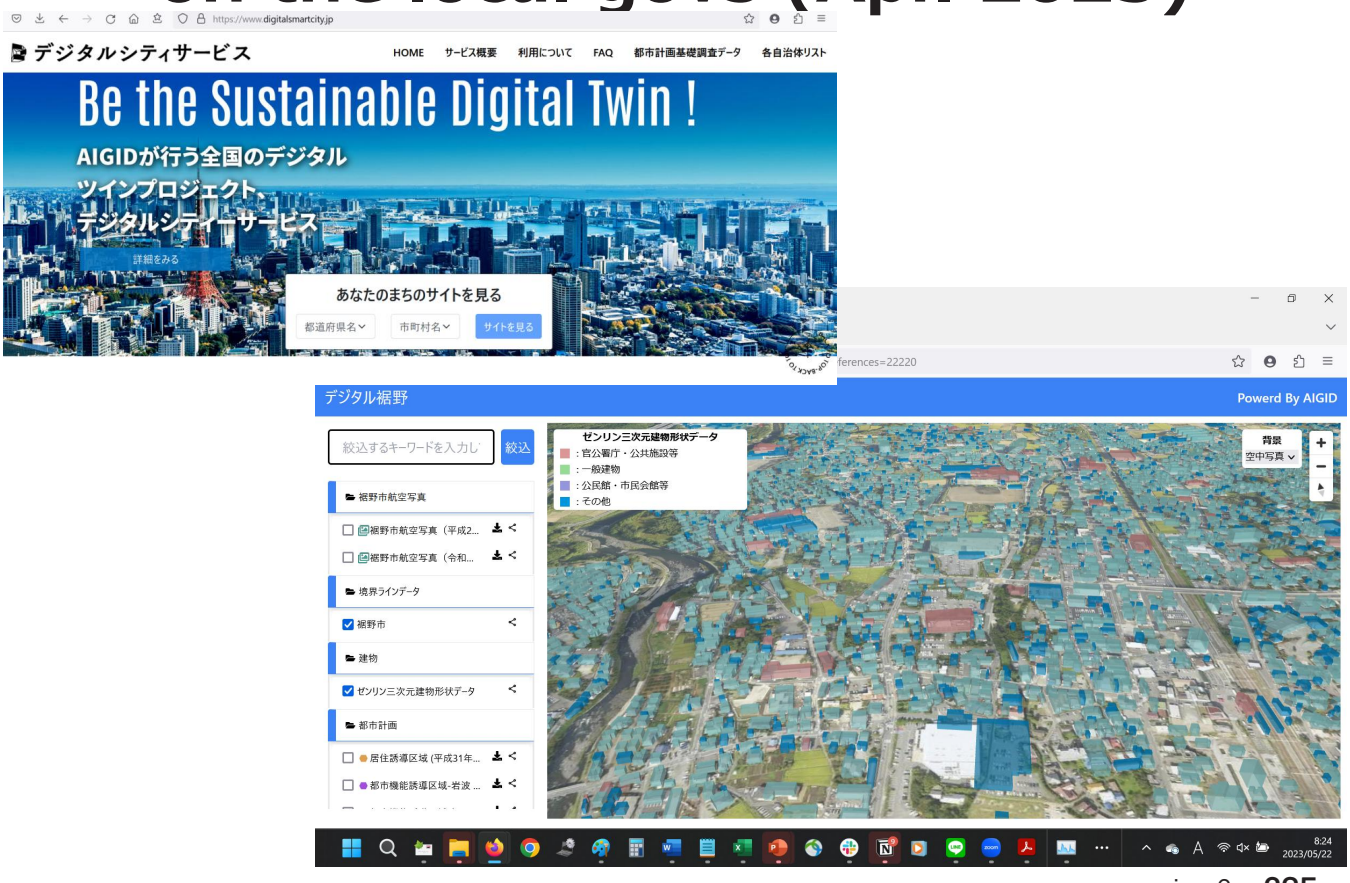
Simple building 3D open data had a big impact (Project PLATEAU by MLIT)

Automatic synchronization between data registry and digital twin view

Synchronized visualization core between original data and tile data



Nation-wide digital twin depending on the local govs (Apr. 2023)



My City Report

Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T. and Omata, H.: Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images, Computer-Aided Civil and Infrastructure Engineering Vol.33, pp.1127–1141, Wiley, June 2018.

19

Aging civil infrastructure…

- Citizens' better understanding should be needed for aging civil infrastructure in the local gov.…

20

Collaboration between citizens & local government ("My City Report" project)

"Chiba repo" (Citizens' report)

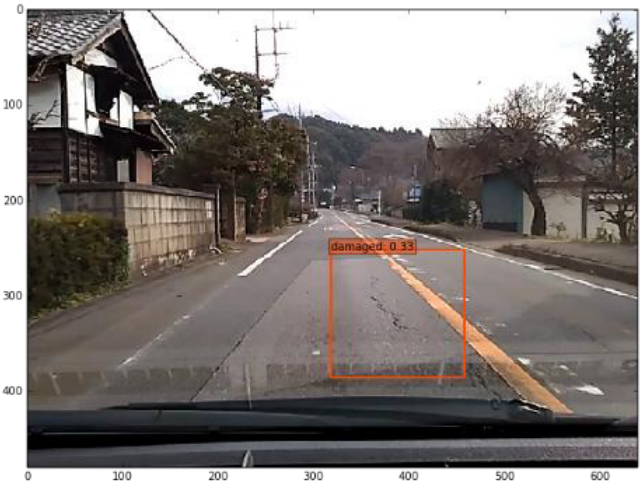
Pavement damage

Wall graffitti

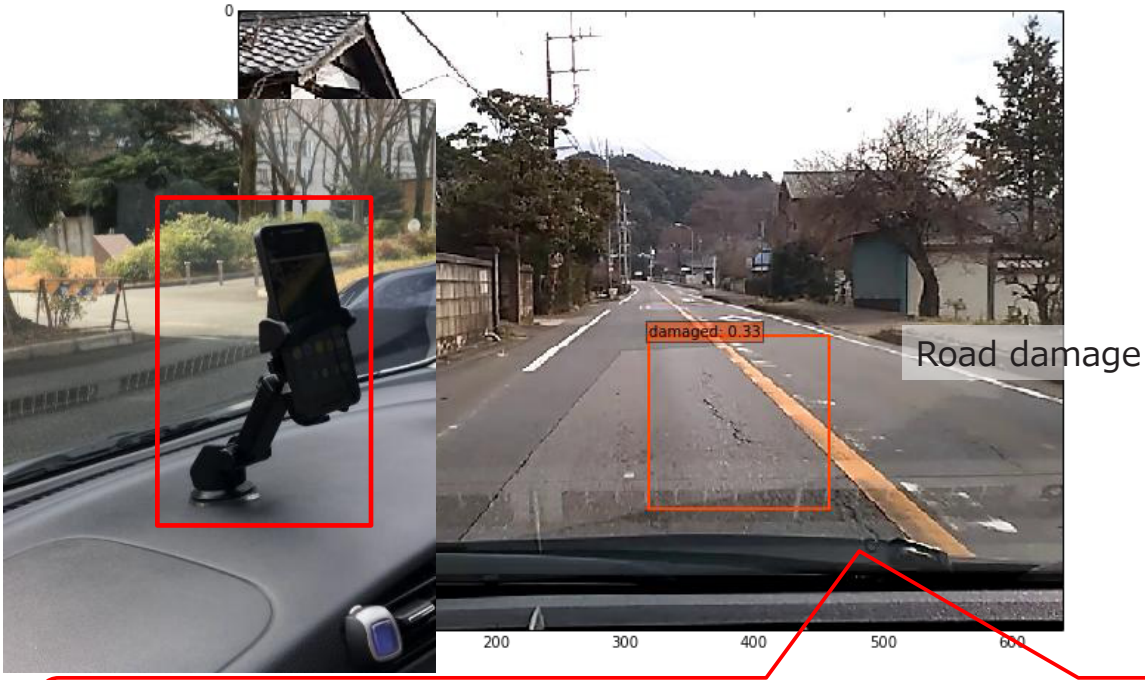
Report by smartphone

Citizen-Government collaborating SNS

Daily road patrol by road manger

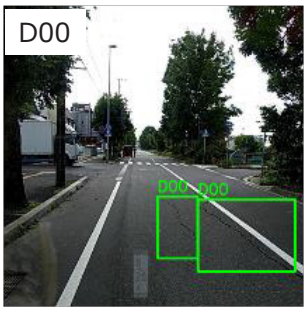


Real-time road damage detection with smartphone

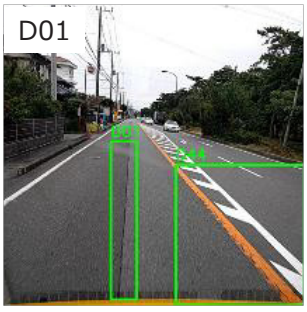


Upload images to server only when road damages are detected !

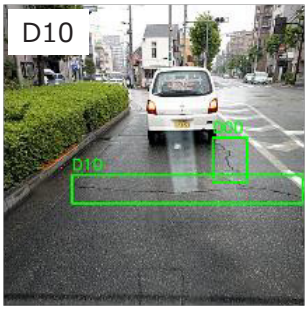
Training dataset with classification labels



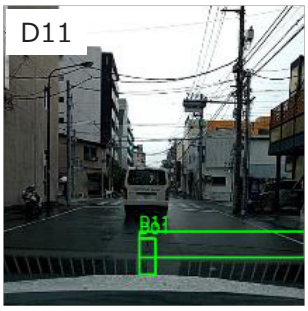
Linear crack, longitudinal, wheel mark part



Linear crack, longitudinal, construction join part



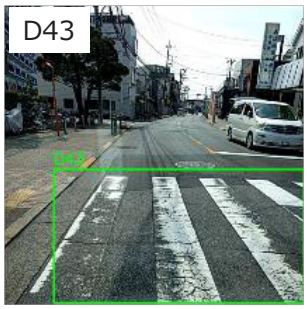
Linear crack, lateral, equal interval



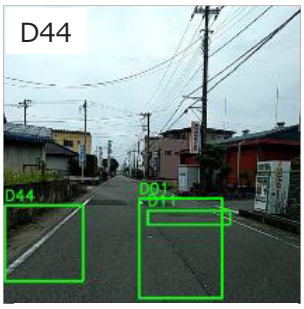
Linear crack, lateral, construction joint part

Alligator crack

Rutting, bump, pothole, separation



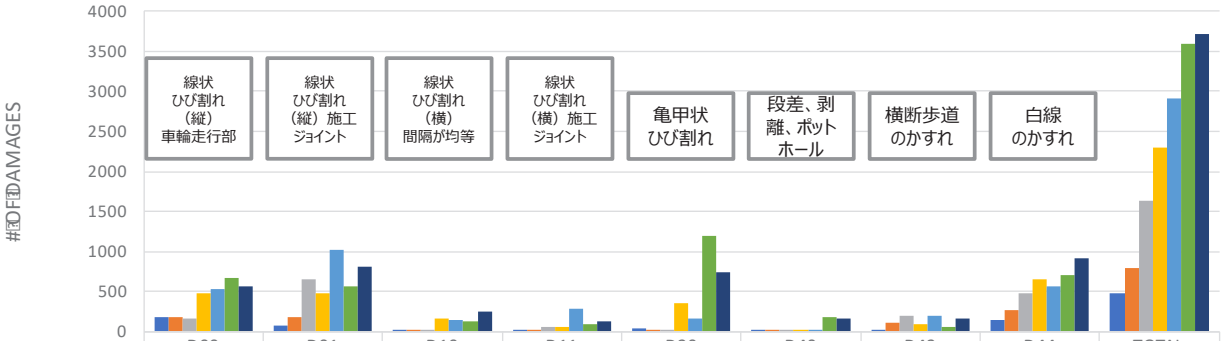
Cross walk blur



White line blur

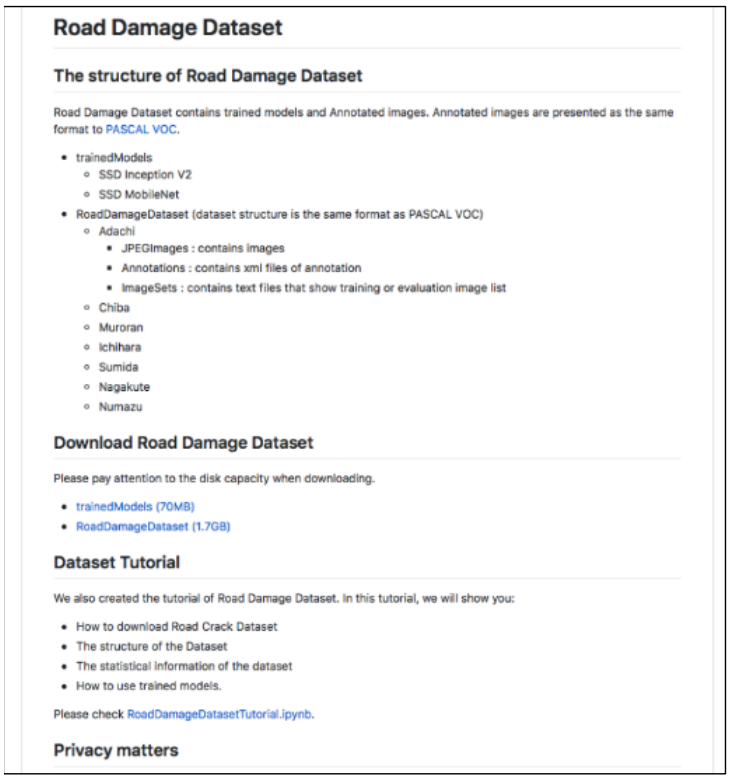
Collection of training dataset

Seven local governments ran 1,600km (50 hours), collected 160K road damage images and put labels for 9,053 damage images



	D00	D01	D10	D11	D20	D40	D43	D44	TOTAL
Ichihara City	175	71	18	9	43	8	20	138	482
Chiba City	183	187	13	12	27	3	104	267	796
Sumida Ward	168	660	20	61	21	19	201	482	1632
Nagakute City	482	477	169	58	351	14	90	659	2300
Adachi Ward	529	1013	153	279	172	11	191	567	2915
Muroran City	671	574	124	88	1192	189	50	712	3600
Numazu City	560	807	245	129	735	165	161	908	3710
TOTAL	2768	3789	742	636	2541	409	817	3733	15435

World's first published large-scale dataset for road damage image



Road Damage Dataset

The structure of Road Damage Dataset

Road Damage Dataset contains trained models and Annotated images. Annotated images are presented as the same format to PASCAL VOC.

- trainedModels
  - SSD Inception V2
  - SSD MobileNet
- RoadDamageDataset (dataset structure is the same format as PASCAL VOC)
  - Adachi
    - JPEGImages : contains images
    - Annotations : contains xml files of annotation
    - ImageSets : contains text files that show training or evaluation image list
  - Chiba
  - Muroran
  - Ichihara
  - Sumida
  - Nagakute
  - Numazu

Download Road Damage Dataset

Please pay attention to the disk capacity when downloading.

- trainedModels (70MB)
- RoadDamageDataset (1.7GB)

Dataset Tutorial

We also created the tutorial of Road Damage Dataset. In this tutorial, we will show you:

- How to download Road Crack Dataset
- The structure of the Dataset
- The statistical information of the dataset
- How to use trained models.

Please check RoadDamageDatasetTutorial.ipynb.

Privacy matters

(https://github.com/sekilab/RoadDamageDetector)

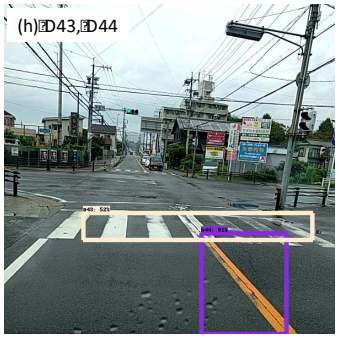
※プライバシー保護のため、人の顔、
車のナンバープレートにモザイクをかけています。

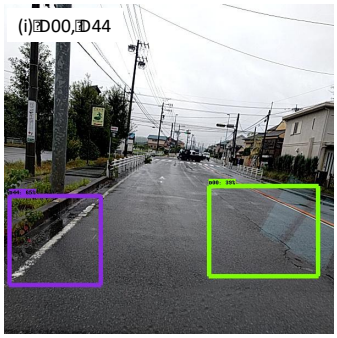
25

Classification accuracy

Some classes are difficult to detect due to less images

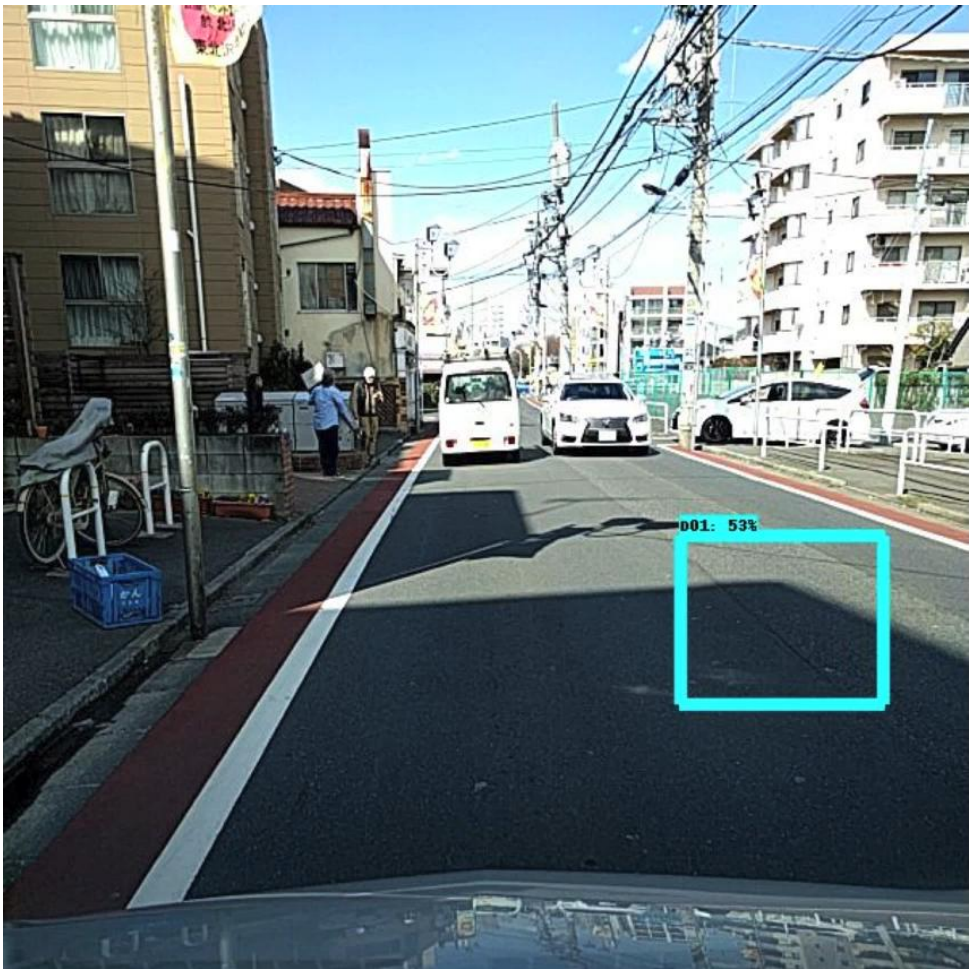
	D00	D01	D10	D11	D20	D40	D43	D44
Recall	0.40	0.89	0.20	0.05	0.68	0.02	0.71	0.85
Precision	0.73	0.64	0.99	0.95	0.68	0.99	0.85	0.66
Accuracy	0.81	0.77	0.92	0.94	0.83	0.95	0.95	0.81





Runtime is 30.6ms on GPU serve, and 1.5s on smartphone

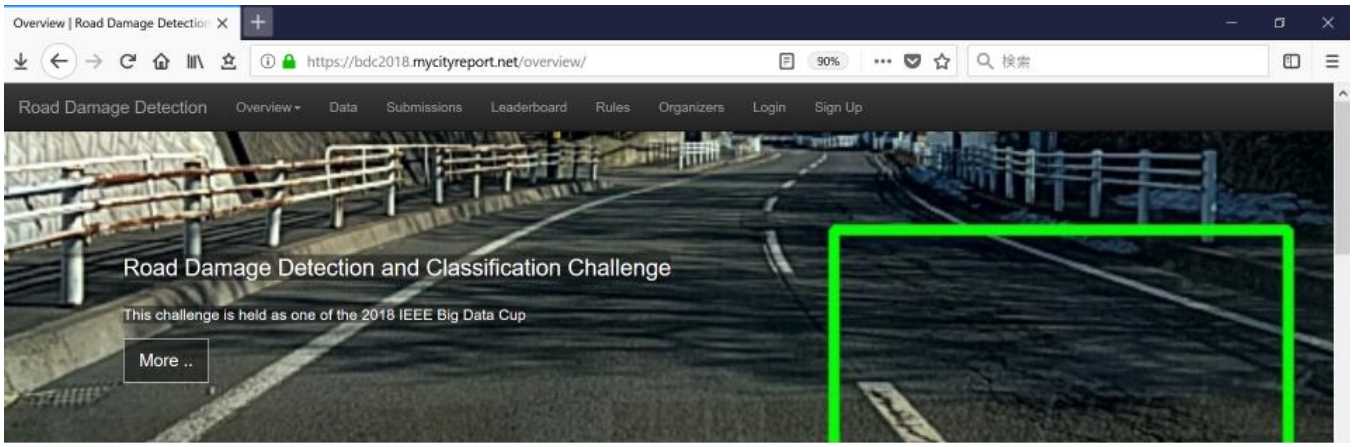
Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T. and Omata, H.: Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images, Computer-Aided Civil and Infrastructure Engineering Vol.33, pp.1127–1141, Wiley, June 2018.



Road AI Dashboard

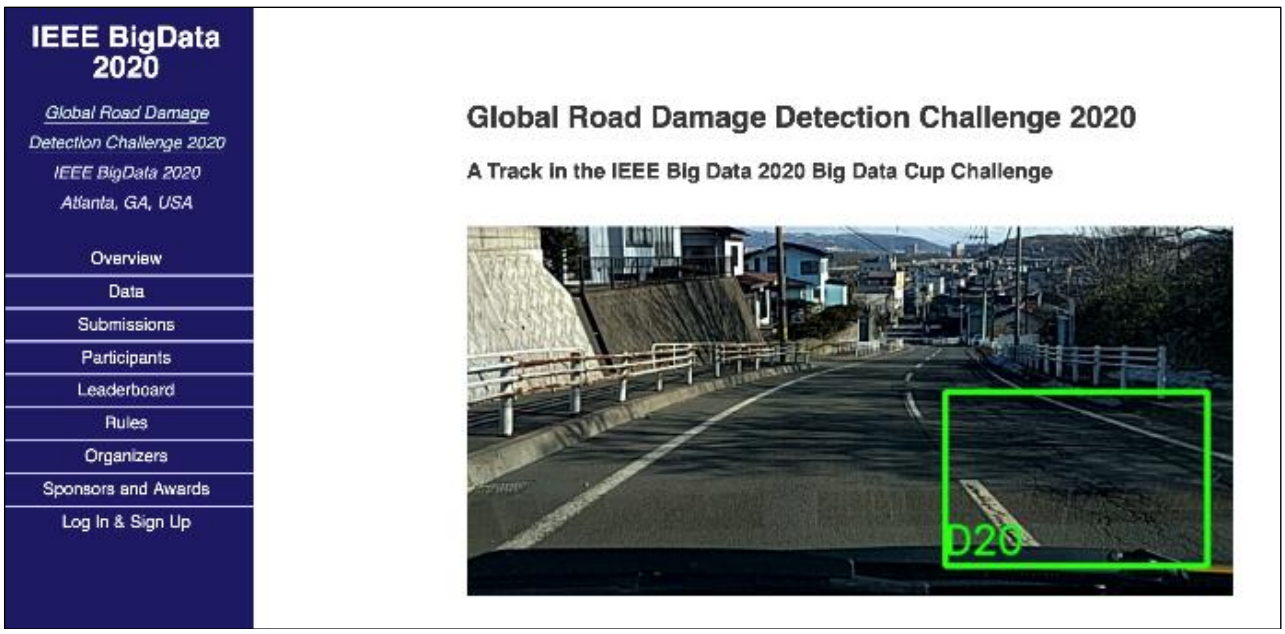
Connecting with researchers all over the world by open data

- **Road Damage Detection and Classification Challenge** in IEEE Big Data Cup 2018@Seattle, Dec.10, 2018
- 59 teams from 15 countries are participating for this challenge: USA(11), China(8), India(6), Poland(4), Germany(3), South Korea(3), France(2), Taiwan (2), Philippines(2), Pakistan(1), Vietnam(1), Morocco(1), Canada(1), Spain(1) and Japan(1).

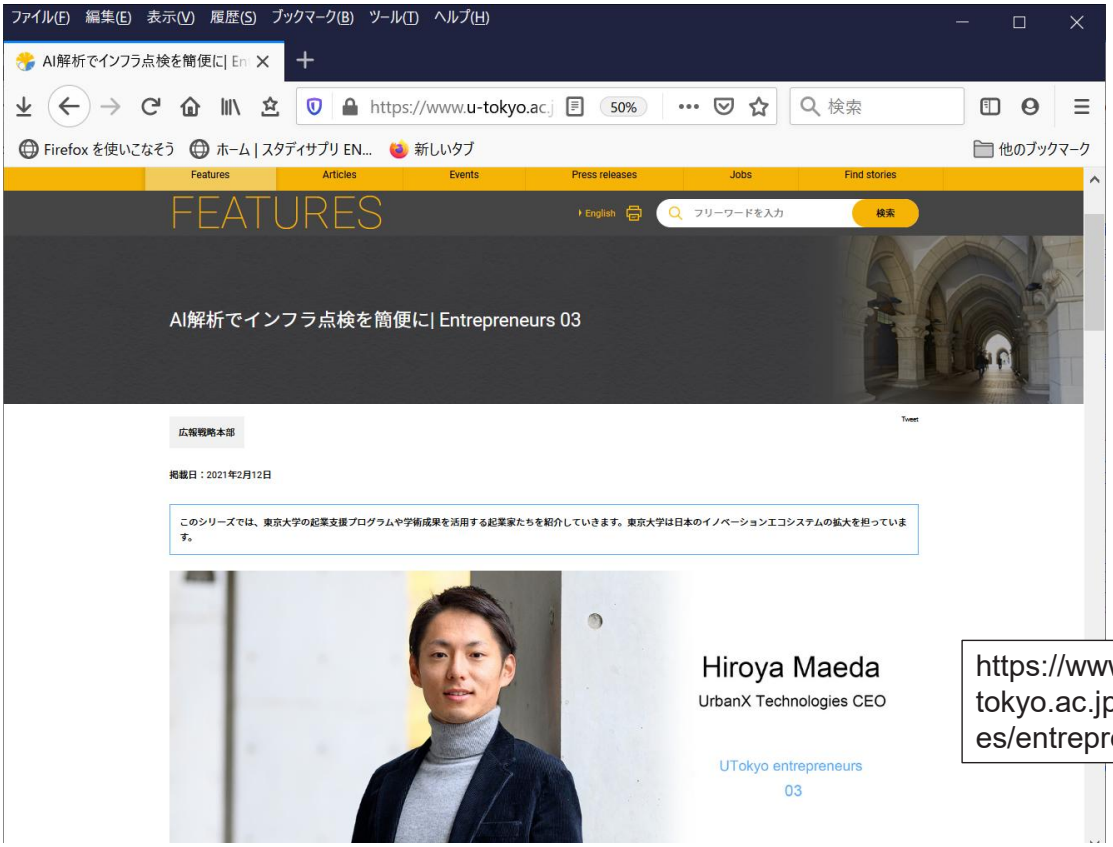


Global Road Damage Detection Challenge, IEEE Bigdata 2020

- Aim to be more generalized detection model adding India and Czech data
- 120 teams from many countries participated



Renew as a start up since Apr. 2020 (Urban X Technologies Inc.)



https://www.u-tokyo.ac.jp/focus/ja/features/entrepreneurs03.html

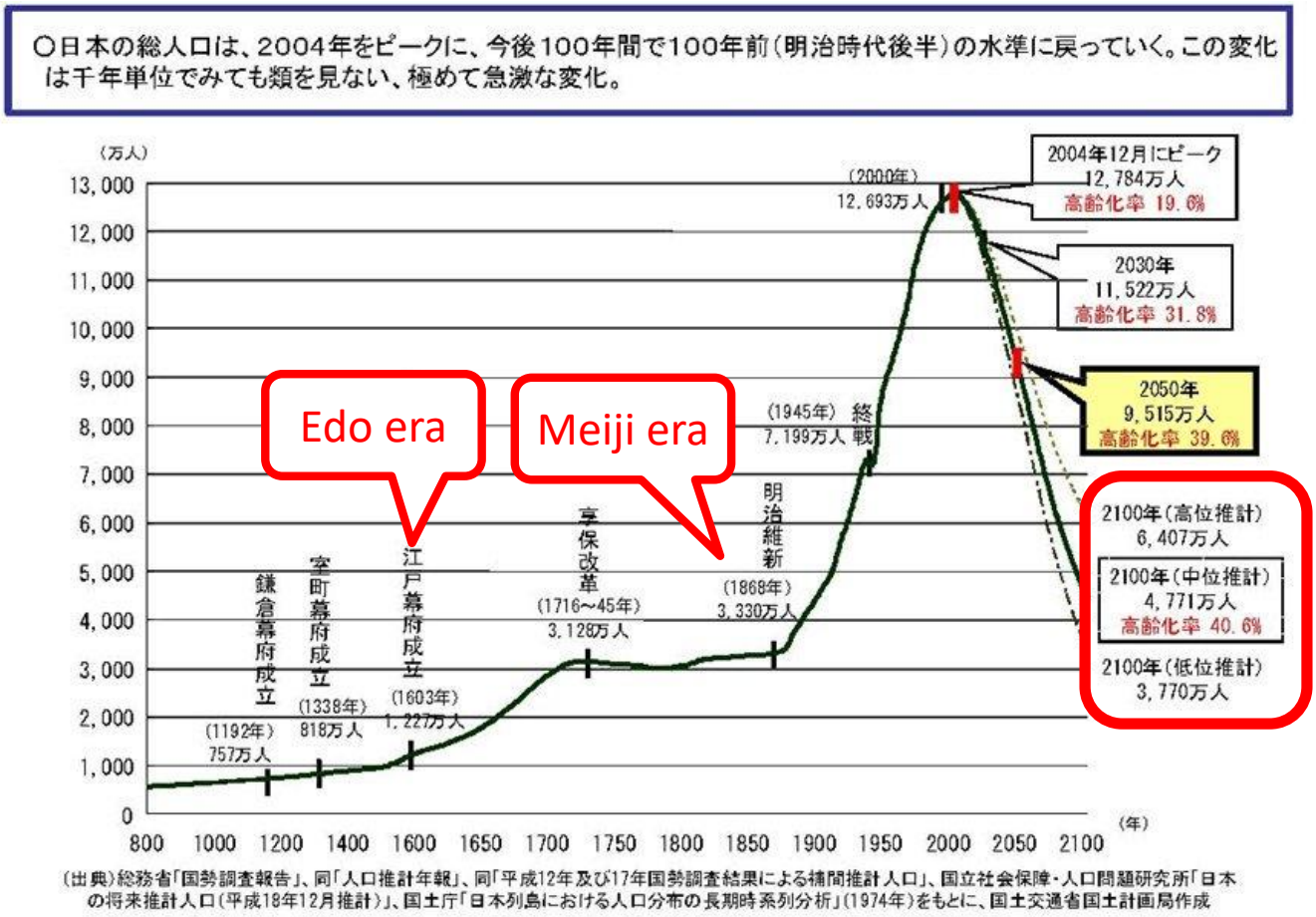
31

My City Forecast

Urban Planning Communication Tool for Citizen–Government Cooperation

Hasegawa, Y., Sekimoto, Y., Seto, T. and Fukushima, Y. and Maeda, M.: Urban Planning Communication Tool for Citizen with National Open Data, *Computers, Environment and Urban Systems*, Elsevier, Available online 19, June 2018.

Do you know future population in Japan ??



（出典）総務省「国勢調査報告」、同「人口推計年報」、同「平成12年及び17年国勢調査結果による補間推計人口」、国立社会保障・人口問題研究所「日本の将来推計人口（平成18年12月推計）」、国土庁「日本列島における人口分布の長期時系列分析」（1974年）をもとに、国土交通省国土計画局作成

Background

- Compact city is needed in Japan
- About 300 local governments are developing compact city plan

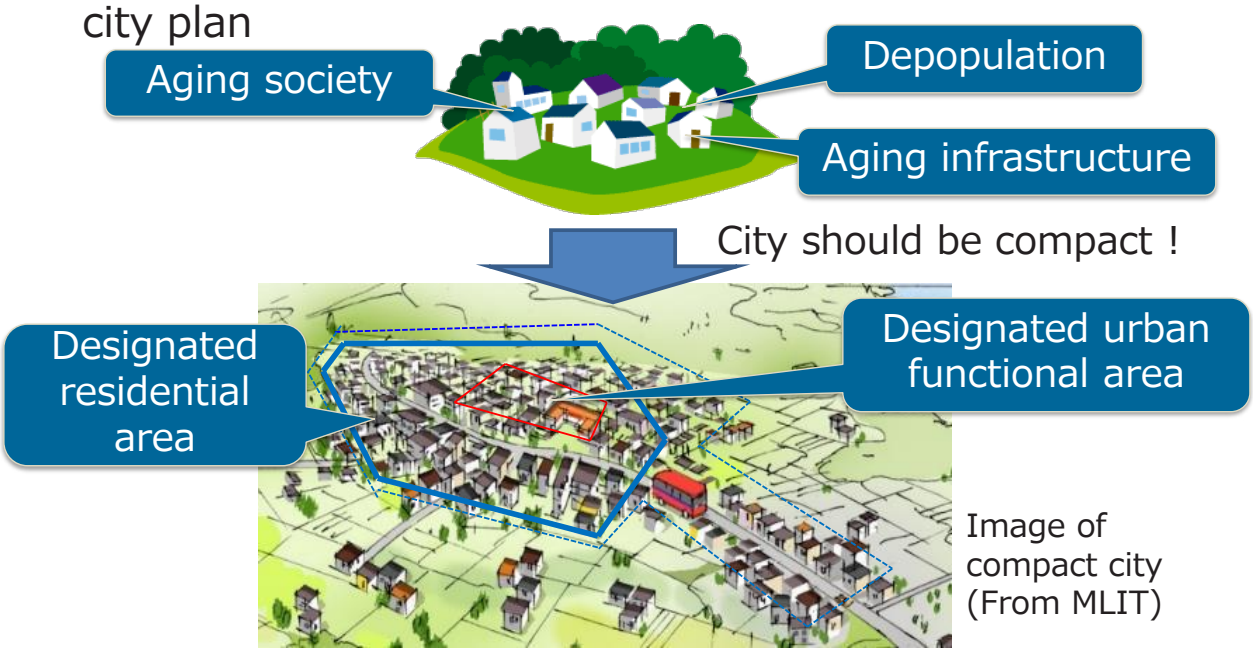
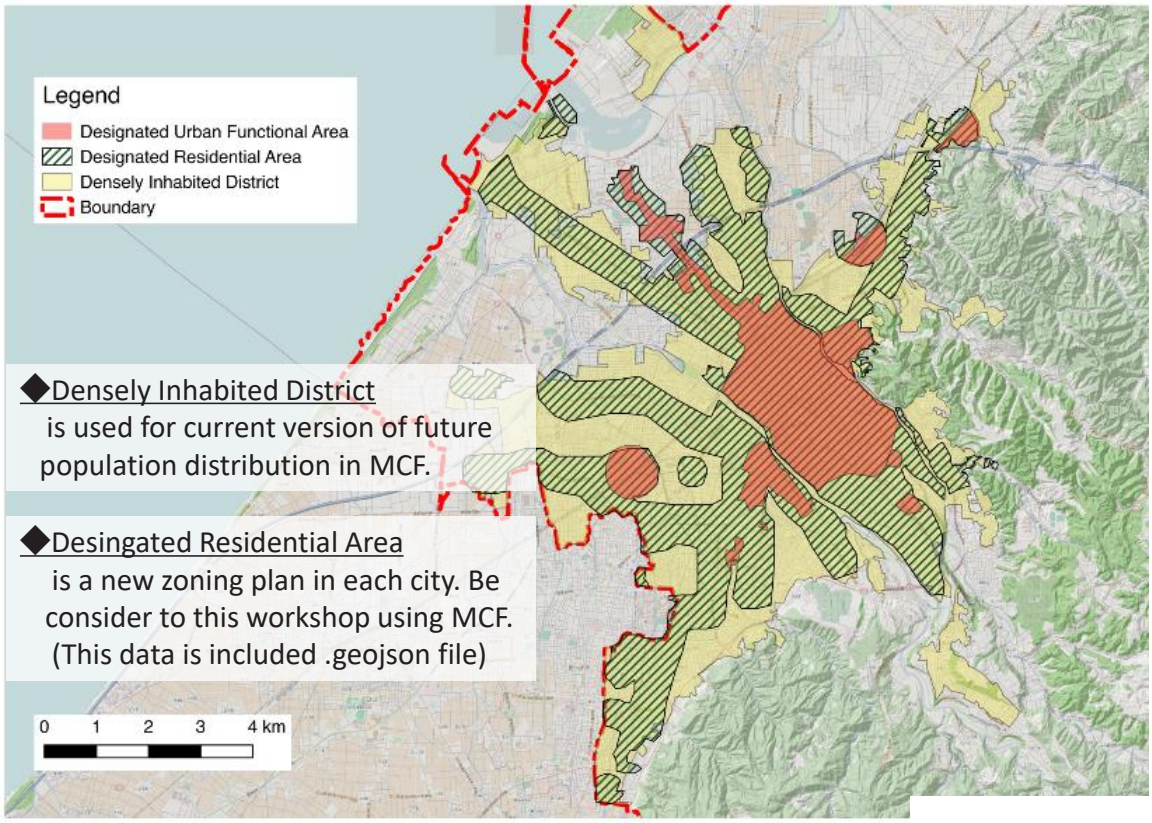


Image of compact city (From MLIT)

35

Example of compact city zoning (Kanazawa City)

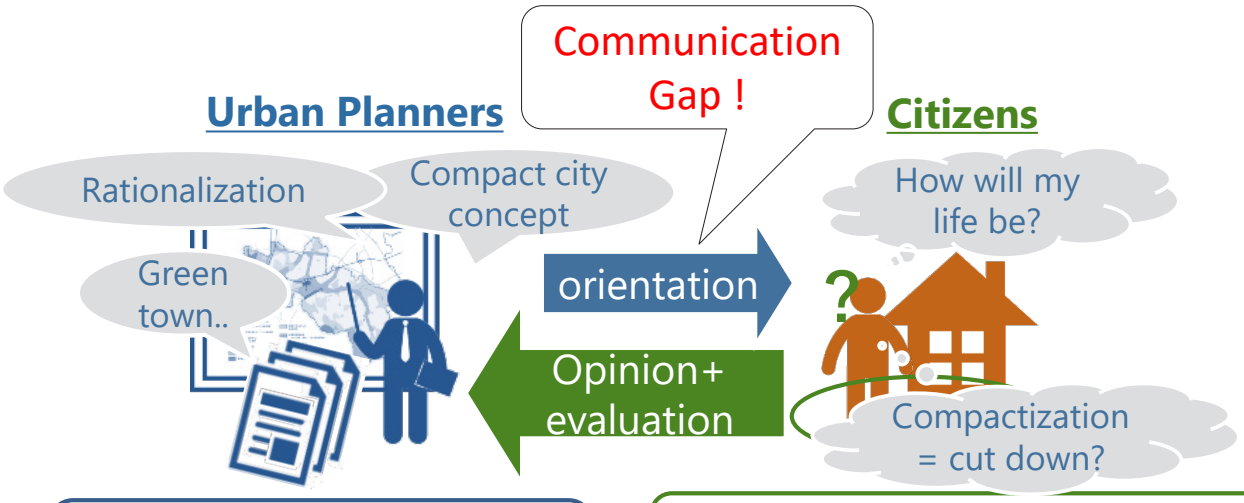


◆Densely Inhabited District
is used for current version of future population distribution in MCF.

◆Desingated Residential Area
is a new zoning plan in each city. Be consider to this workshop using MCF. (This data is included .geojson file)

36

Current Public Involvement in Urban Planning

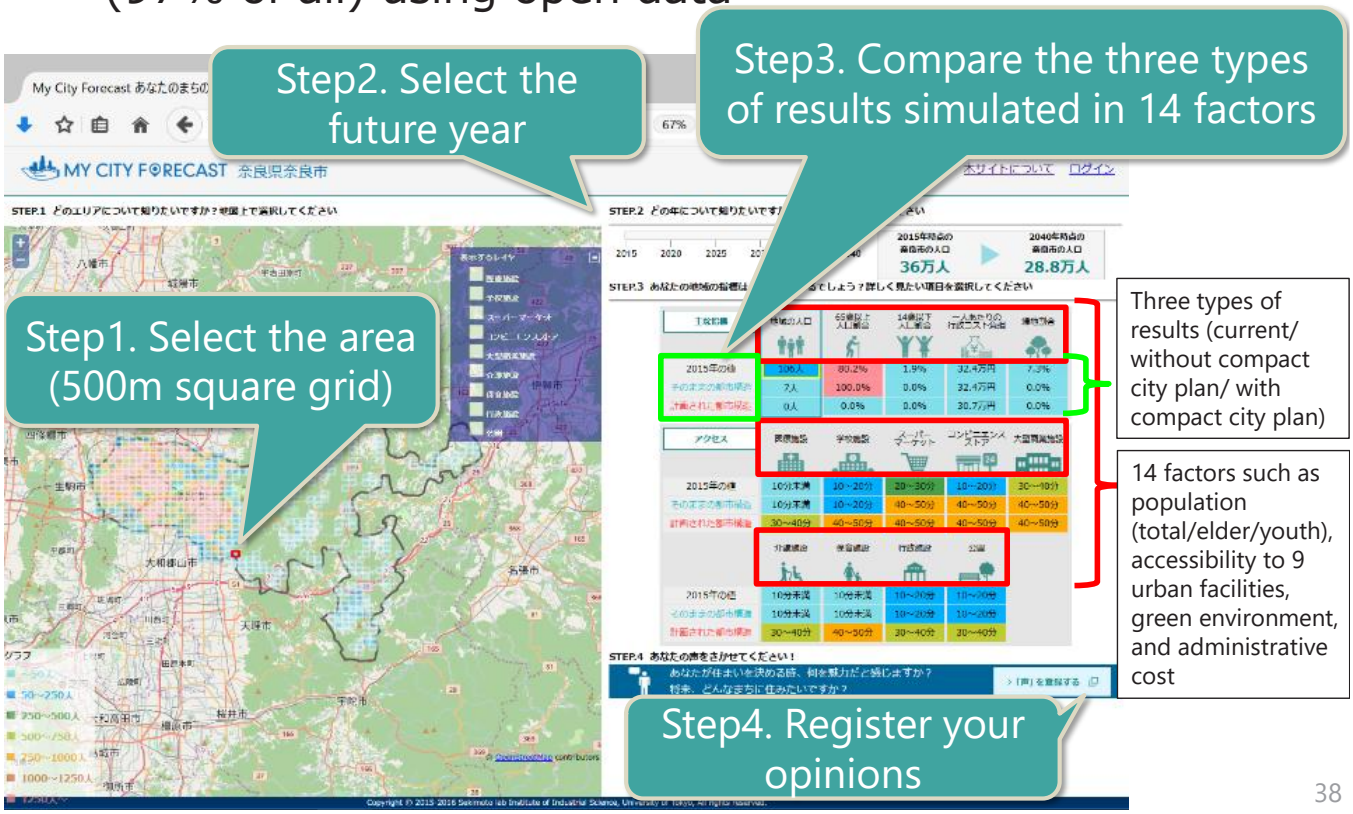


- Publication of **master plans** (booklet or pdf) and holding **briefing sessions**
- Explanation of how the city will be **as a whole**

- What they really want to know is **effect to their daily lives**
- **Anxiety** and **repulsion** due to **lack of information**

Our site (https://mycityforecast.net/)

- We already open MCF for 1670 local governments (97% of all) using open data



Simulation algorithm for future state

Six fundamental factors illustrating urban structure included in data

❖Population distribution ❖Land use plans
❖Green land distribution ❖Urban facility distribution
❖Administrative cost information ❖Public transportation system

The data including information for the future

❖ Japanese census ❖Digital national land information
❖Financial information from municipal government
❖The information about public transportations

input

Simulation taking into account the notion of the compact city

Estimated future population distribution	Assuming that specific areas are designated as residential areas
Estimated the existences of urban facilities	Judging whether urban facilities can maintain their functions by population around them
Administrative costs	On the basis of estimated population distribution and location changes of urban facilities

output

Estimated life-related indicators in each region every five years

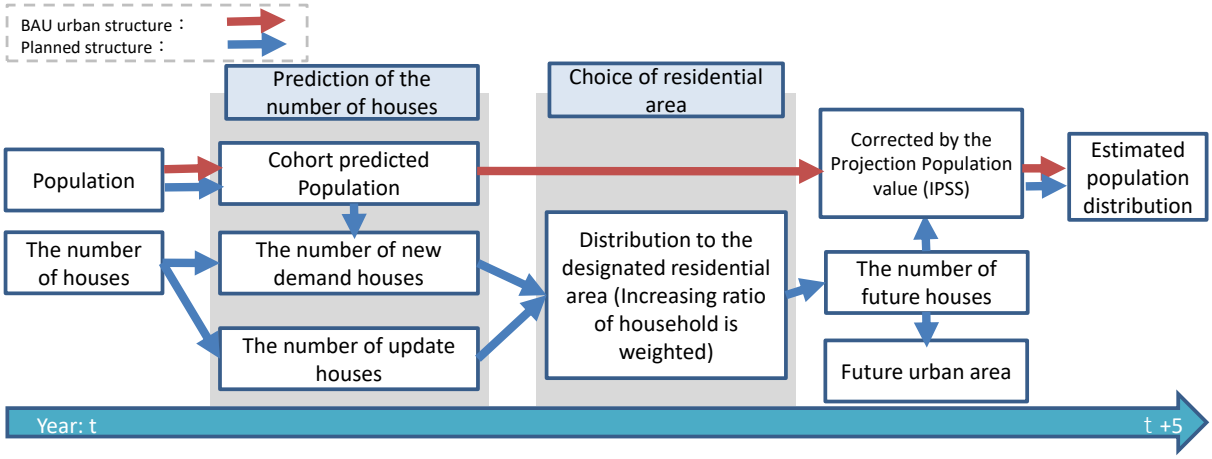
❖Future population ❖Environmental index
❖Accessibility of urban facilities in the future
❖Administrative cost for each resident

39

Future population distribution

Assumption

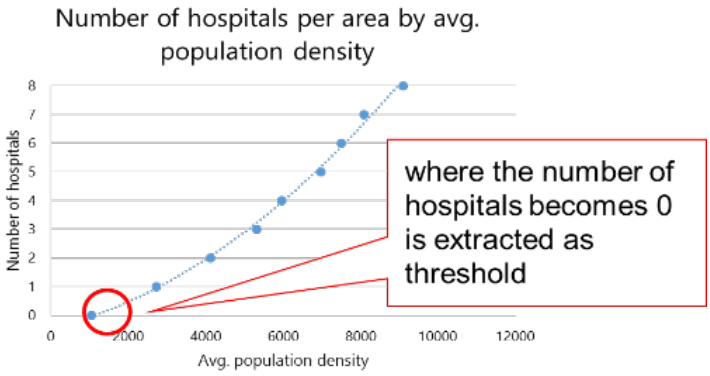
- House demand is satisfied within the city (local government).
- Newly houses for new demands and updates are built in the designated residential area.
- We use DID data **by default as** the designated residential area.
- Population and urban area increase proportionally with the number of household.



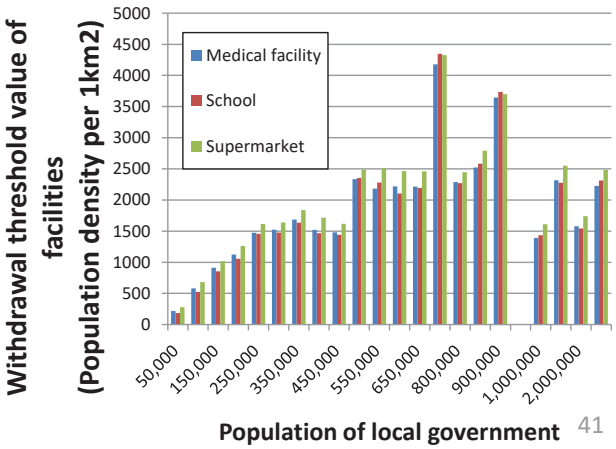
40

Future facility existence

- Estimate withdrawal threshold value of the population density from the current existence of facilities in each local government
- **Commercial beneficial facilities**
  Withdraw if the density is under the threshold value
- **Public facilities**
  No change in case of "As is" urban structure without compact city plan, and same rule as commercial facilities with compact city plan



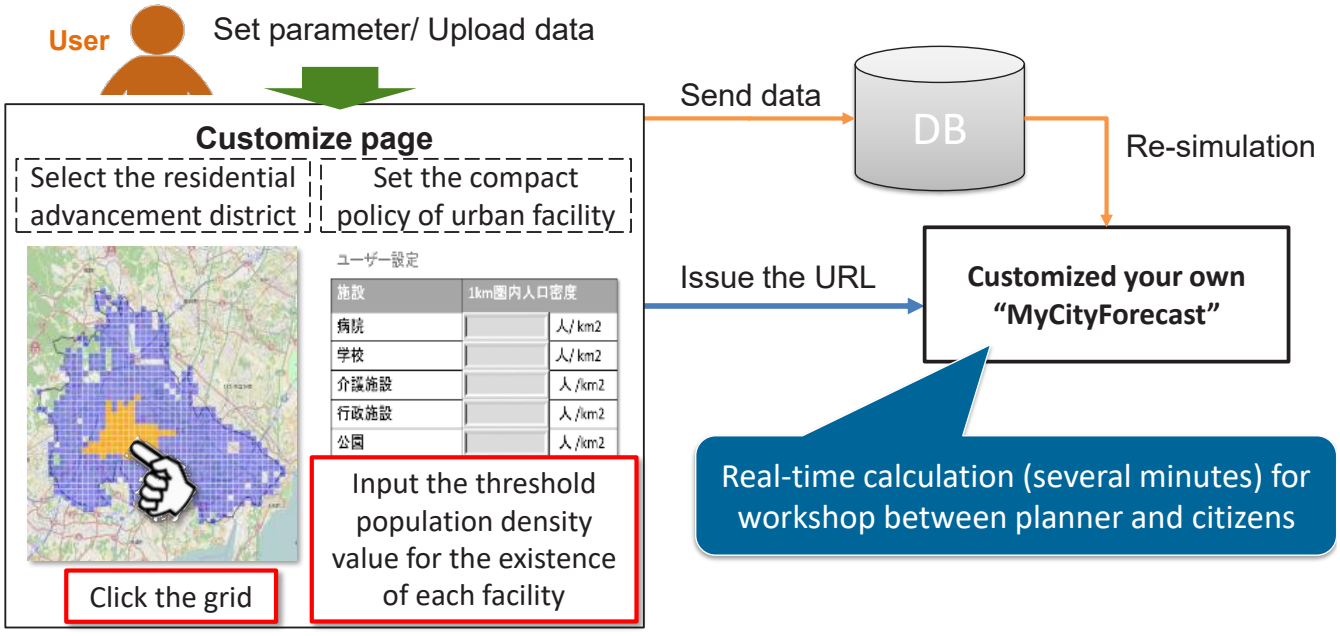
Withdrawal threshold value of facility



Threshold value of each local government

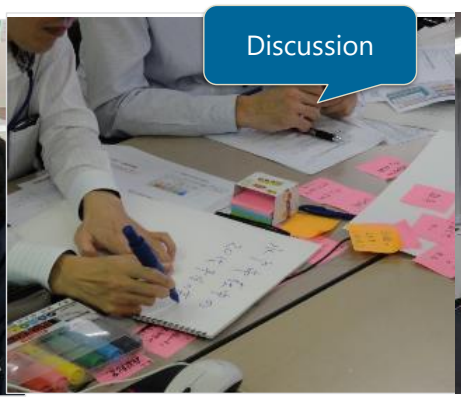
41

Real-time customize function by regional private detail data



Workshops using My City Forecast

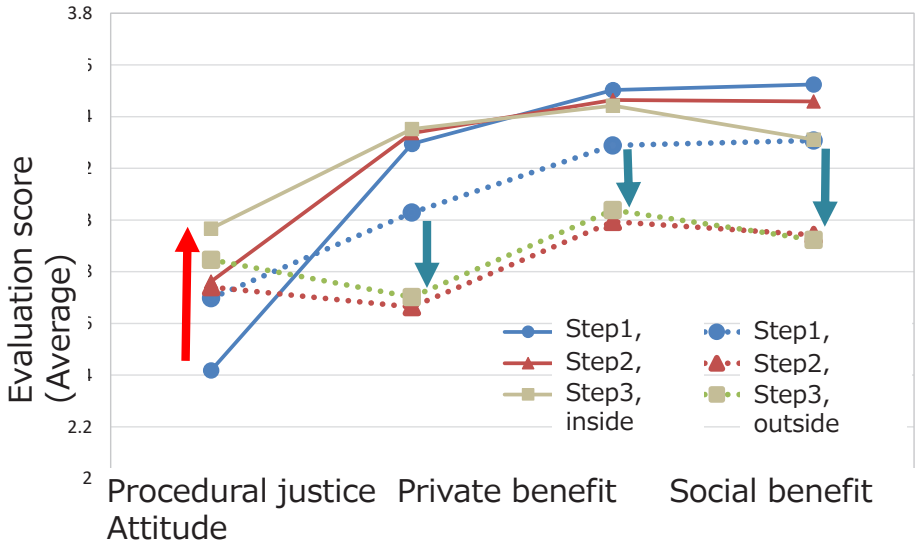
City	Date	Theme	Number of participants
Mito City (Ibaraki Pref.)	10/16/2015	Study workshop for urban master plans in administrative office	18
	1/21/2016	Workshop for use and application of open data	20
Yokohama City (Kanagawa Pref.)	11/9/2015	Training in practical use of data for administrative employees	52
	1/16/2016	Workshop to think about Aoba-district's future by using data	11
Kurashiki City (Okayama Pref.)	1/7/2017	Traffic congestion prevention policy around tourist spot	15
Gotsu City (Okayama Pref.)	1/14/2017	Public transport/ Medical and nursing care problem in the underpopulated area	25
Nanto City (Toyama Pref.)	1/21/2017	Workshop for public facility policy	25



43

Attitude change for compact city policy

- Survey user's experiments of MCF in Mito City via internet survey for the attitude of compact city on Dec. 2015.
- 113 citizens consisting of 61 citizens inside residential guidance area and 52 citizens outside.
- For procedural justice, both citizens showed more positive attitude via MCF, but for private justice, the citizens who live outside showed more negative.

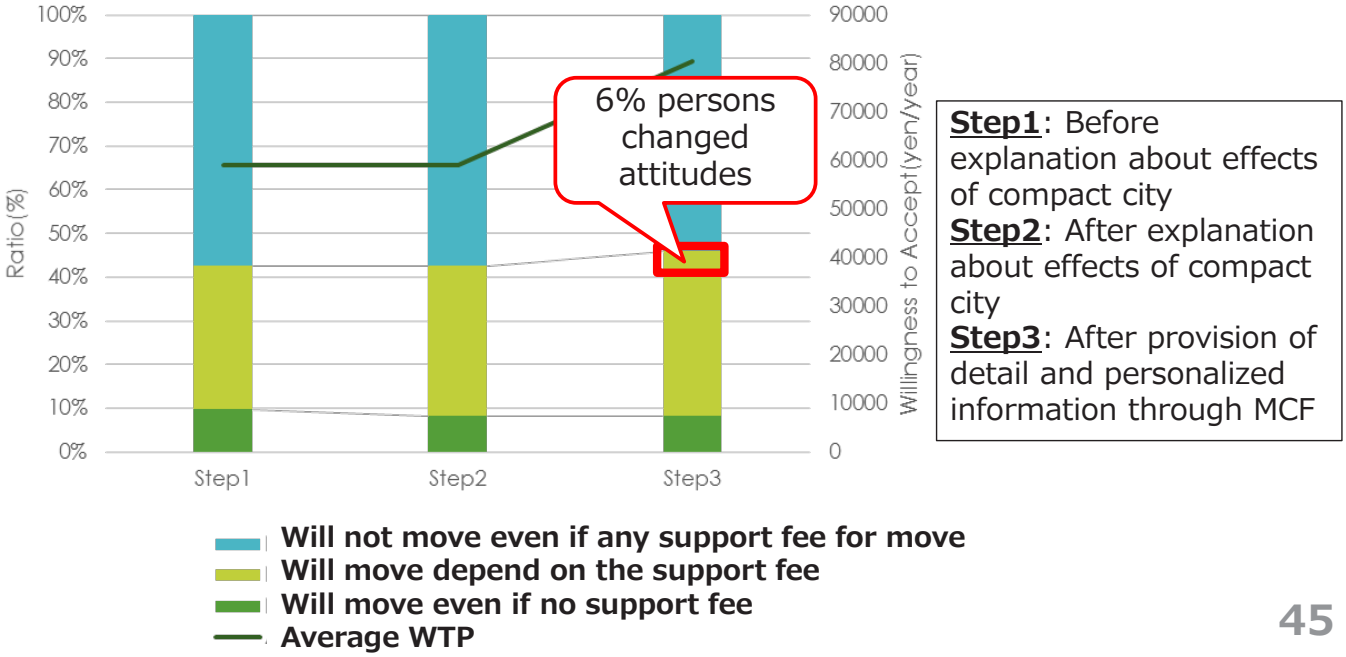


Step1: Before explanation about effects of compact city
Step2: After explanation about effects of compact city
Step3: After provision of detail and personalized information through MCF

44

Willingness to pay for move to inside of the area

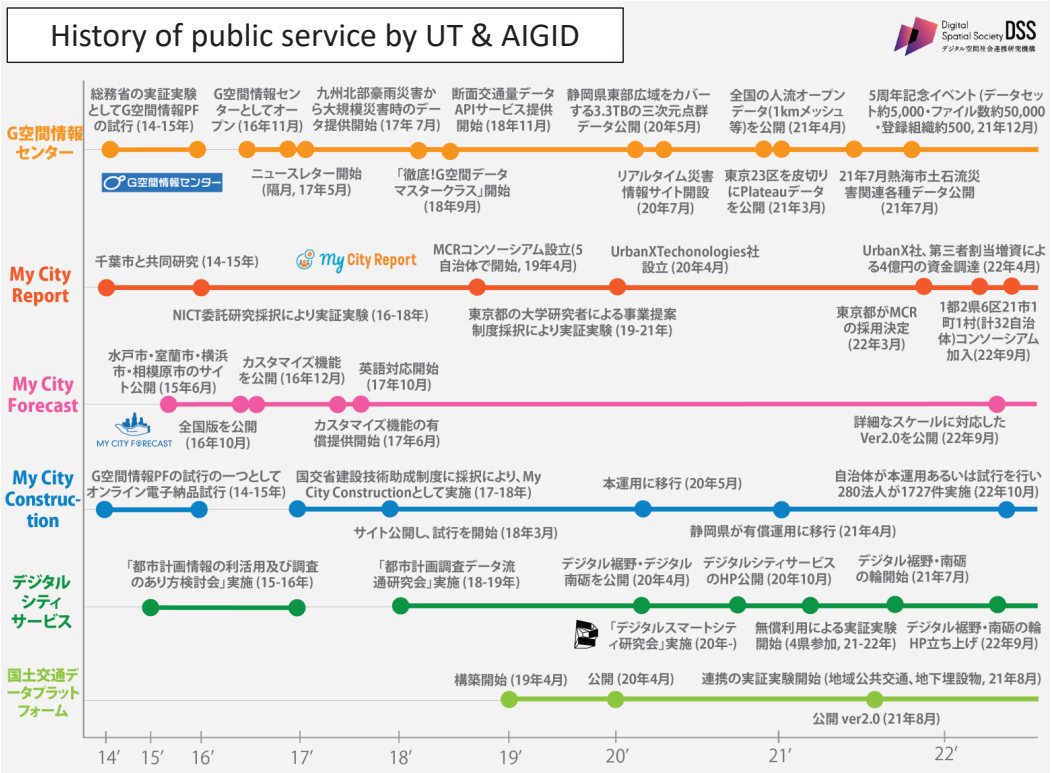
- In step 3, 6% persons who answered “Will not move even if any support fee for move” changed attitudes to “Will move depend on the support fee”.



Step1: Before explanation about effects of compact city
Step2: After explanation about effects of compact city
Step3: After provision of detail and personalized information through MCF

Conclusions

- Cutting edge technologies & social implementation are pair of wheels !!



Thank you for kind listening !!

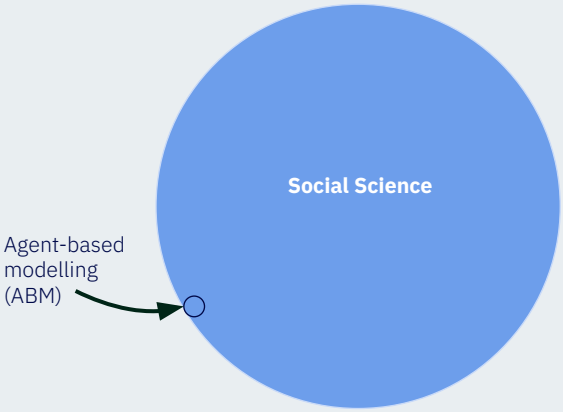
http://sekilab.iis.u-tokyo.ac.jp
sekimoto@csis.u-tokyo.ac.jp

Can AI outperform the shortest path?: Rethinking routing in agent-based traffic models

KOSSREC Meeting May 28th, 2025

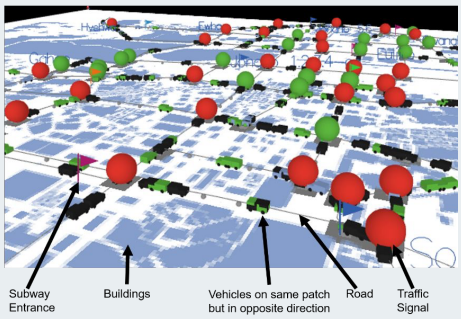
Hyesop Shin

Agent-based modelling in Social Science



Seoul CBD

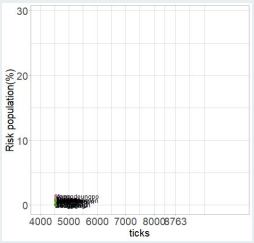
3D Version



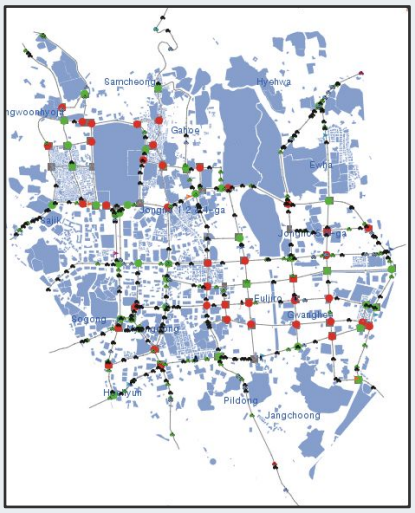
Spatial extent: 16km^2

Temporal res: 1 minute

Jan.1st-Mar.31st 2018



2D Version



CBD of Seoul, S.Korea

Shin and Bithell (2023): TRAPSim in Seoul

Glasgow CBD

2.5D Version

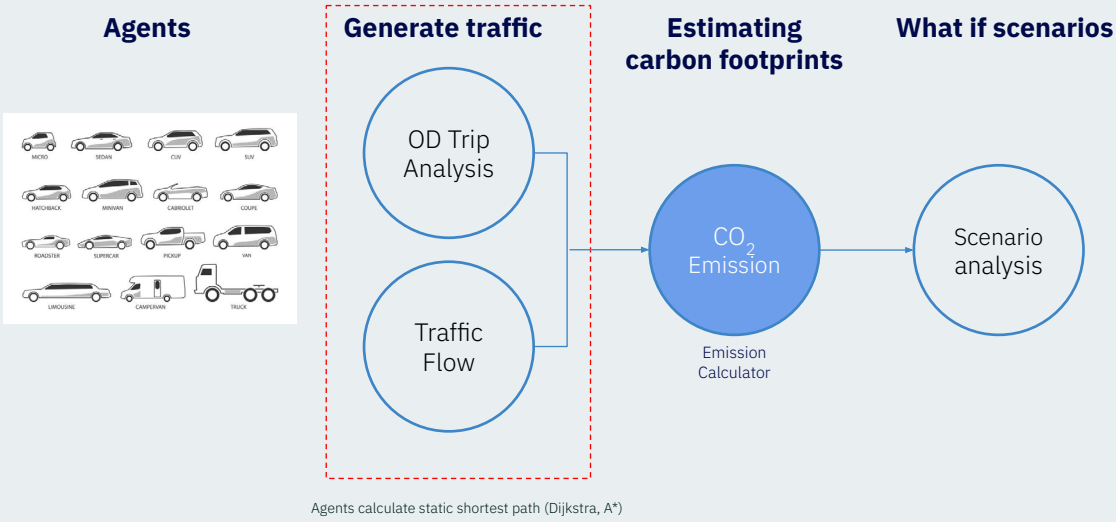
Spatial extent: 4km^2

Temporal res: 30 sec

May.26th-Jun.5th 2022

Simulating traffic based on shortest path

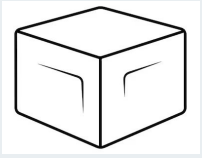
Model Development | **Generating Traffic and emissions**



5

Why Agent-Based Modelling in Traffic?

Agent-Based Modelling (ABM):
A "white box" approach where we simulate individuals (cars, people) making decisions in a spatial environment.



White box: We can observe each step of the world changing

Strengths:

- Transparent and traceable decision rules
- Great for *what-if* scenario testing
- Can represent micro-level interactions (e.g., car-following, detours)

But ABM also has its own weaknesses

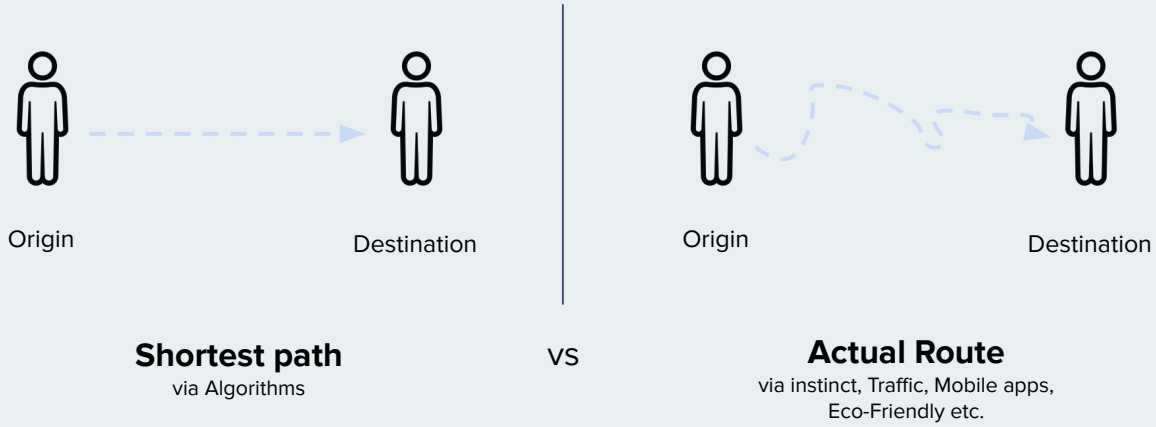
Often uses fixed rules (shortest path, fixed demand)

Lacks realistic behavioural adaptation (i.e. bounded rationality)

Data-hungry and hard to calibrate without real-world feedback

"We've made our agents smart enough to move, but not yet smart enough to behave."

How do you find your destination?

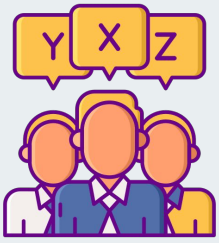


AI: Filling the Behavioural Gap

Cognitive flexibility without needing to hard-code every rule

"Avoid school at 3pm"

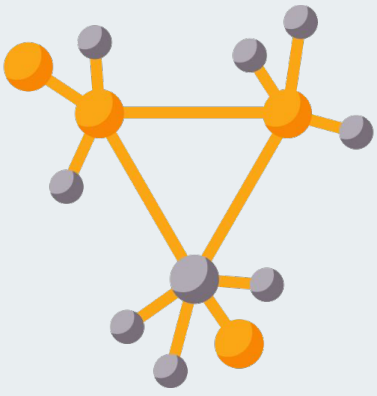
Roadblocks



Not every agent behaves the same

Using **Entropy** as a measurement

🛑 Fastest ≠ Best
🛑 Shortest ≠ Smartest Choice
📉 We need to measure how agents diversify, adapt, and respond: ***Entropy***



Gao et al., 2024

Humanised Routing Scenarios

Entropy formula:

$$H = -\sum p_i \log_2(pi)$$

- 📈 High entropy → realistic adaptation
- 📉 Low entropy (close to 0) → rigid, robotic, fragile

11

Humanised Routing Scenarios

Scenario	Traditional Agent	LLM-Driven Agent
School zone at 3 PM	Follows shortest path	Reroutes based on prompts
Sudden roadblock	Cannot respond mid-simulation	Sudden roadblock
Emission zone violation	Ignores unless coded	Avoids proactively

12

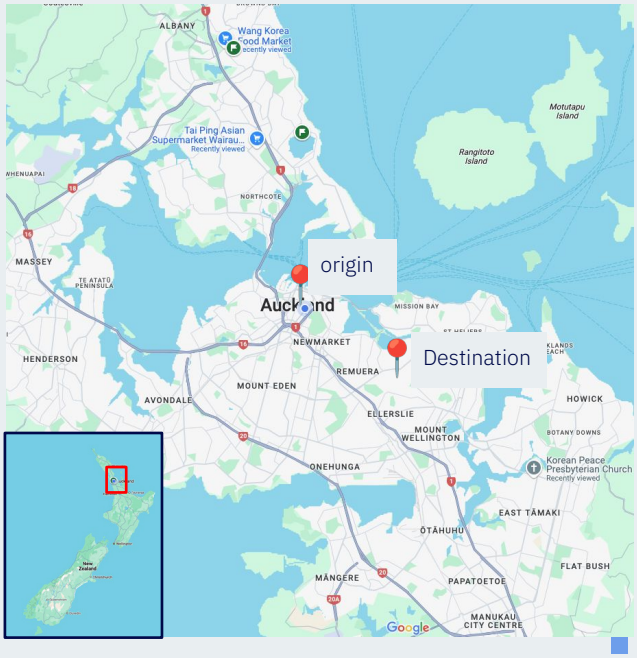
Auckland, NZ

Symonds St → Remuera Rd

Shortest Path: everyone funnels into the same road

GPS Trace: varied human detours

LLM Agent: prompted to “avoid congestion and schools”



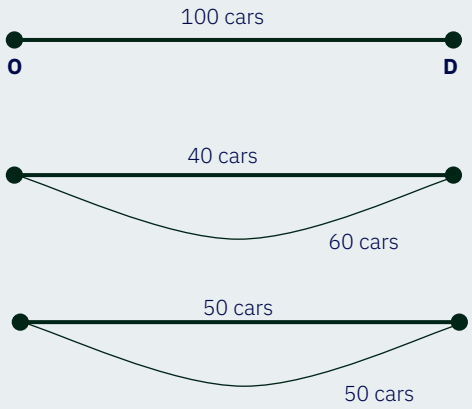
One Journey, Three Routes

Symonds St → Remuera Rd

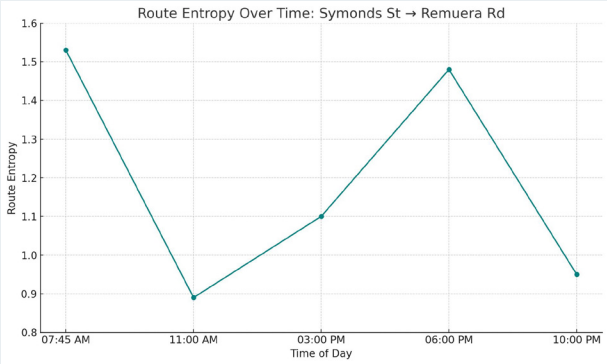
🛣️ Shortest Path: everyone funnels into the same road
📍 Entropy = 0.20

🚗 GPS Trace: varied human detours
📍 Entropy = 1.45

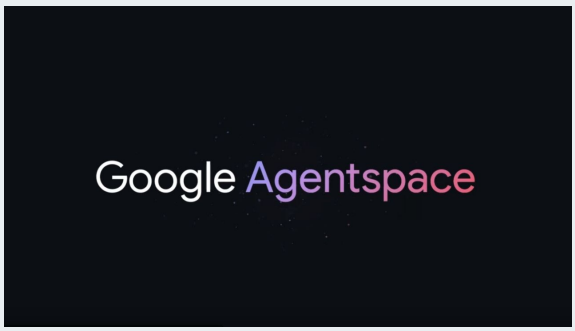
🤖 LLM Agent: prompted to “avoid congestion and schools”
📍 Entropy = 1.48



Real Behaviour Happens Over Time



2025: An important year of AI

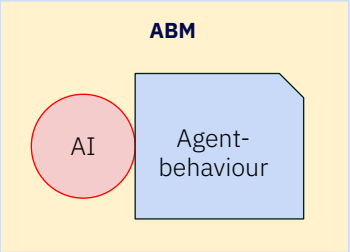


An open framework for multi-agent simulations powered by language models

Supports perception, reasoning, memory, planning

Publicly demoed agents doing realistic social behaviours in physical and online environments

Closing

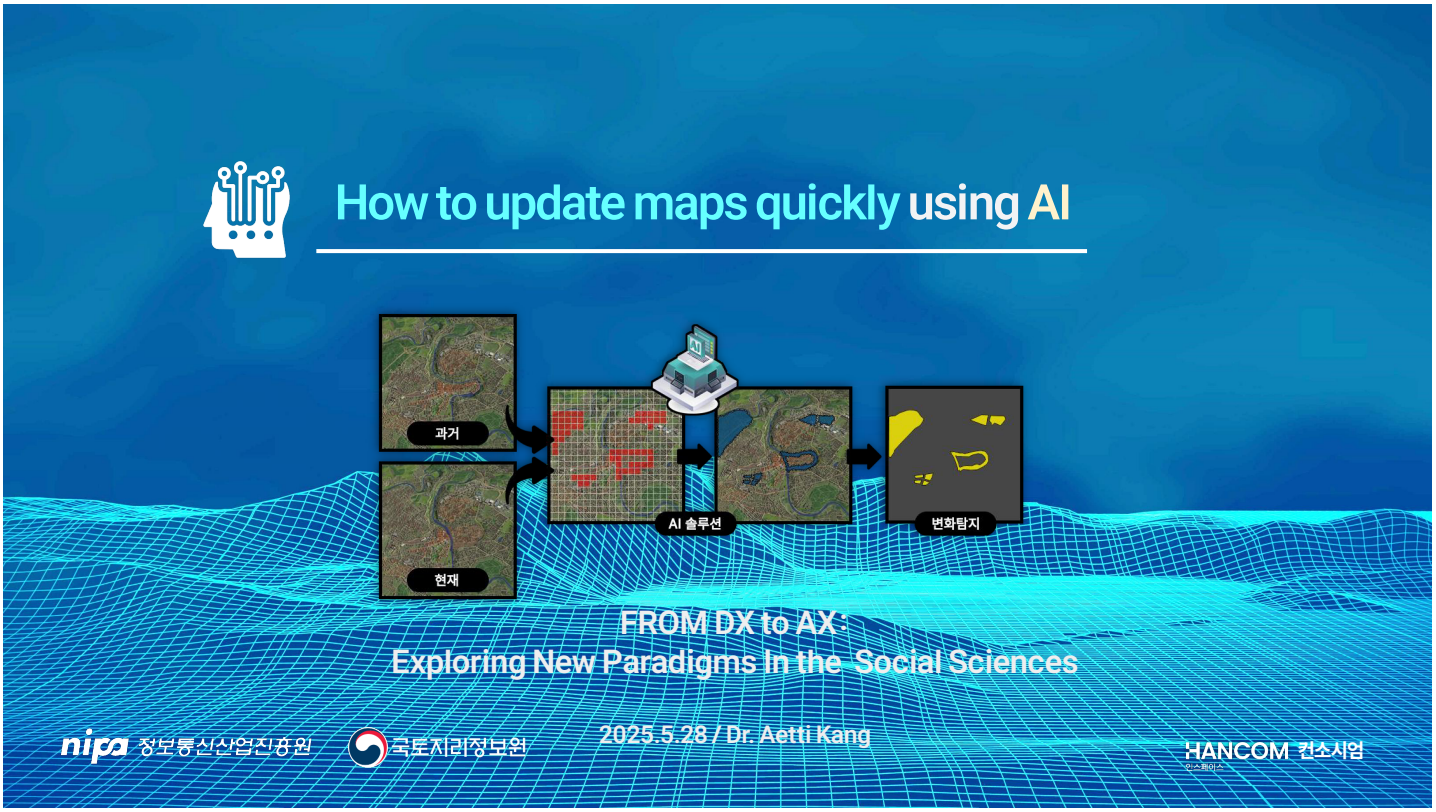


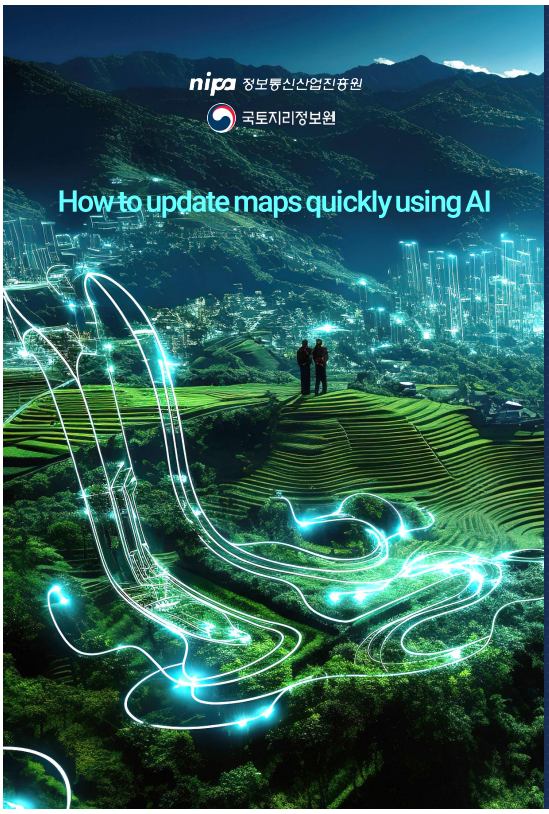
ABM-AI research is still young

ABMs show us how the world works

AI can help us simulate how people really behave.

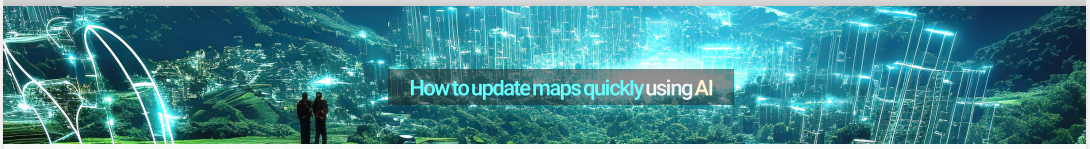
But we need benchmarks like entropy to make sure our simulations don’t just look intelligent, but feel real.





CONTENTS

HANCOM 컨소시엄
인스페이스



I. Business Overview

1. Background
2. Goals

nipa 정보통신산업진흥원 HANCOM 컨소시엄 인스페이스 국토지리정보원

I
사업개요

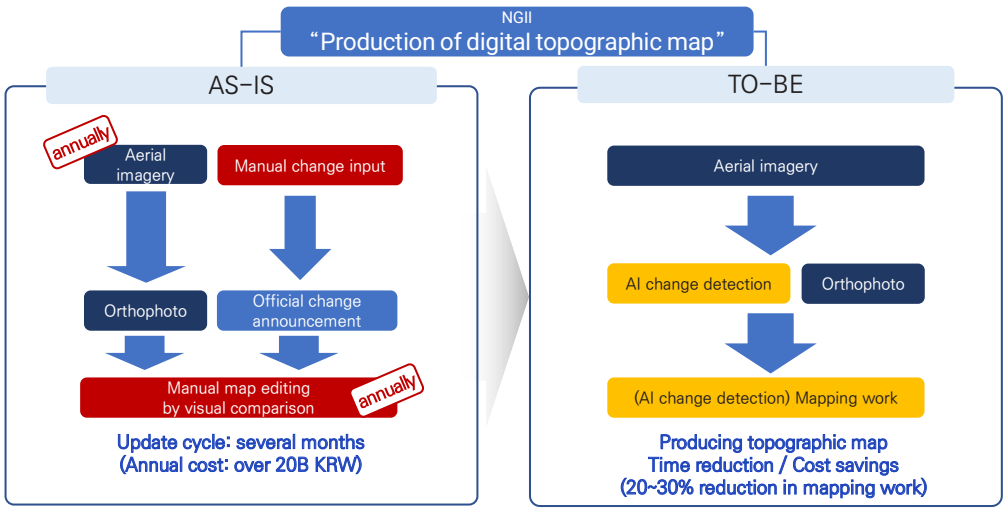
How to update maps quickly using AI

nipa 정보통신산업진흥원
국토지리정보원
HANCOM 인스페이스 컨소시엄

1. Background

For rapid reflection of national land changes in geospatial data

『Digital Topographic Map』 using AI-based Change Detection

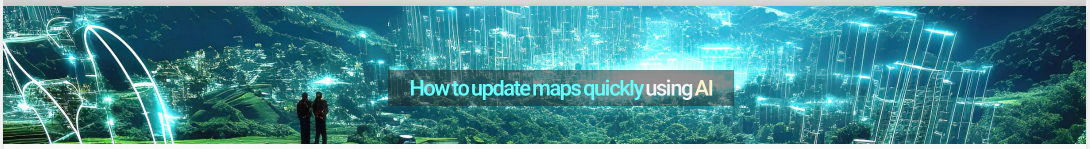


4

2. Goals

▪ AI-based National Land Change Detection Service Enhancement

Category	Performance Indicator	Evaluation Proportion(%)	Target
Data Collection & Processing	Multisensor image training data accuracy	30	93% / 70,500Unit (Cumulative 141,000)
	Geolocation alignment training data volume	5	40-degree grid/1:5000
	Image quality evaluation training data accuracy	1	80% /3000 items per each/3-point scale
	Change explanation training data accuracy	1	-
AI Solution Development & Testing	Image georeferencing accuracy	8	RMSE <= 3M
	AI model accuracy for binary change detection	10	92%
	AI model accuracy for semantic change detection	10	93%
	Change explanation accuracy	2	-
	Number of change detection system built	10	1set(update)
	Number of field testing	10	1



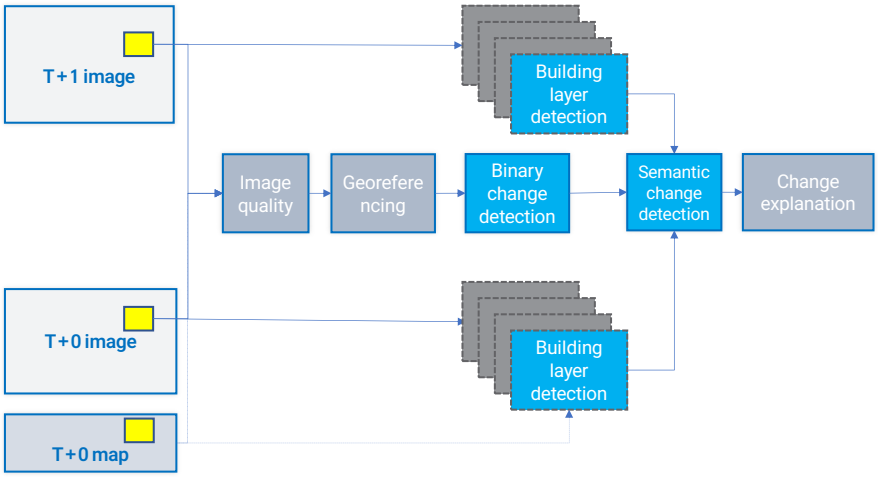
II. Change Detection Method

1. Detection Method
2. Input/Output
3. Developed Model

1. Detection Method

▪ Detection of changes and display of details between two time period inputs (Image, map, and point cloud)

✓ Includes AI for pre-processing image quality, georeferencing, etc



2. Input/Output

▪ Output by change detection program

Binary Change Detection AI

- Compares images with the same range, area, and resolution, and shows whether there was a change in the location of cells at the same position

• Output format: shapefile

1	0.5	0	1	0.5
0	0	1	0	0
0	1	0.5	0	0
1	0.5	0	0	0
0.5	0	0	0	0

Semantic Change(area) Detection

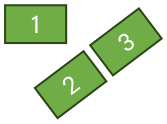
- Extracts desired objects in vector form from the input image and shows whether there is a change in the area of those objects
- In the past, comparison was made by extracting only the desired object layer from the digital map

• Output format: shapefile (+EXCEL)

ID	Change	Admin data	Match status
1	Road expansion	Y	Y
2	Road closure	N	N

Semantic Change(height) Detection

- Performed when a building is selected as an object
- Creates a shapefile storing the average height values of point clouds included in the building
- Compares the height of buildings from two different time periods



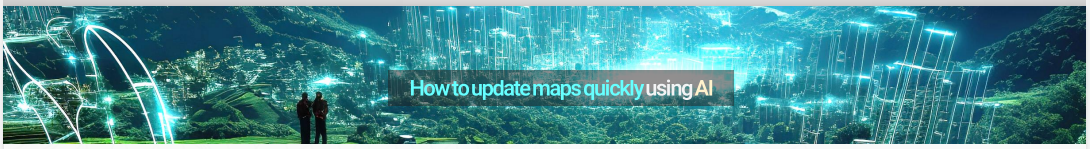
ID	Change	Admin data	Match status
1	Height increase	Y	Y
2	Area expansion	N	N
3	Demolition	Y	Y

3. Developed Model

▪ Types of national land change detection models

- **Image:image model** : Detects changes in buildings and roads between two images (past and recent) using AI
- **Map:image model** : Compares AI-extracted features (buildings, roads) from recent imagery with corresponding features from past map data to detect changes
- **Point cloud:point cloud model** : Detects height changes between two sets of point cloud data from different time periods
- DEM comparison model: Detects terrain elevation changes by comparing two DEM datasets from different times

Past data \ New data		Imagery		Point cloud	DEM
		Aerial image	Ortho satellite image		
Imagery	Aerial image	Binary change Semantic change (shape, area)	Semantic change (shape, area)	-	-
	Ortho satellite image	Semantic change (shape, area)	Binary change	-	-
Point cloud		-	-	Semantic change (height)	-
Digital topographic map		Semantic change (shape, area)	Semantic change (shape, area)	-	-
DEM		-	-	-	Semantic change (height)



Ⅲ. Solution Development Approach

1. Training Data Development
2. Model Development and Training
3. System Implementation

1. Training Data Development

▪ Draft definition of change detection objects

No	Major category	Middle category	Subcategory (Topographic feature name)	Label name	Available image	Input data used	Training data preparation		Cartographer task difficulty	AI Recognition difficulty	Change detection types		
							Human recognition	Use of map (+admin data)			New	Closure (Disappeared)	Change
1	Transportation	Road	Road boundary(polygon)	Road	Satellite /aerial	Orthoimage (25cm)	O	O	Mid	Low	Road planned / under construction	Road closure	Area / length increase (under construction)
2			Sidewalk(boundary)	Sidewalk	Aerial	Orthoimage (25cm)	O	O	High	High	Open	Closure	Area and length change
3		Road facility	Crosswalk (boundary) (planned in 2nd yr)	Crosswalk		Orthoimage (25cm)	O	O	Mid	Mid	Open		Area and length change
4			Safety zone (boundary)	Safety zone		Orthoimage (25cm)	X	O	Mid	Mid	Open		Area and length change
5			Overpass(boundary) (planned in 2nd yr)	Overpass		Orthoimage (25cm)	O	O	Mid	Mid	Open		Area and length change
6			Bridge(boundary) (planned in 2nd yr)	Bridge		Orthoimage (25cm)	O	O	Mid	Mid	Open		Area and length change
7	Building	Building	Building(boundary)	Building		High-res orthoimage (12cm)	O	O	Mid	Low	New construction	Demolition	Area / length increase
8		Greenhouse	Greenhouse (boundary)	Greenhouse		Orthoimage (25cm)	O	O	Mid	Mid	New construction		Area / length increase
9	Man-made Structure (facility)	Storage yard	Storage yard (boundary)	Storage yard		Orthoimage (25cm)	X	O	High	Mid	Open		Area / length decrease
10		Installation	Solar panel (boundary) (planned in 2nd yr)	Solar panel		Orthoimage (25cm)	O	X	Low	Low	Installation		Increase /decrease
11		Cemetery	Cemetery(경계)	Cemetery		Orthoimage (25cm)	O	O	Mid	Mid	Open		Increase /decrease
12			Public cemetery	Public cemetery		Orthoimage (25cm)	O	O	Mid	Mid	Open		Increase /decrease
13			Burial mound	Cemetery		Orthoimage (25cm)	O	O	Mid	Mid	Open		Increase /decrease

Ⅲ
솔루션 구축방법
How to update maps quickly using AI

1. Training Data Development

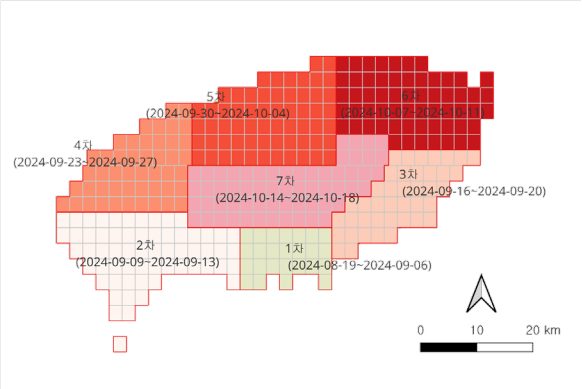
▪ Draft definition of change detection objects

No	Major category	Middle category	Subcategory (Topographic feature name)	Label name	Available image	Input data used	Training data preparation		Cartographer task difficulty	AI Recognition difficulty	Change detection types		
							Human recognition	Use of map (+admin data)			New	Closure (Disappeared)	Change
14	Man-made Structure	Cultivated land	Paddy field (boundary)	Paddy field	Satellite /aerial	Orthoimage (25cm) Satellite image (50cm)	O	O	Mid	Mid	Reclamation	Closure	Expansion /reduction
15			Dry field (boundary)	Dry field			O	O	Mid	Mid	Reclamation	Closure	Expansion /reduction
16			Orchard (boundary)	Orchard			O	O	High	High	Development	Closure	Expansion /reduction
17		Forest	Forest area (boundary)	Forest			O	O	High	High	Afforestation	Deforestation	Expansion /reduction
18			Pasture	Pasture			O	X	High	High	Reclamation	Destruction	Expansion /reduction
19		Water system	River(boundary) Replaced with land-sea boundary (planned in 2nd yr)	Water system			O	O	High	Mid	Formation	Landfill	Expansion
20			Dam, reservoir (boundary) (planned in 2nd yr)	Water system			O	O	High	Mid	Formation	Landfill	Expansion

1. Training Data Development

▪ Pre-processing for training data development

- Training data was pre-processed across 7 zones
- The processed data was then used to develop final training datasets



Zone	No. of 1:5,000 map sheets	Note	Processed count: 1:1000 map sheets
1	25	Seogwipo area	Y2021: 516 sheets Y2022: 541 sheets Y2023 : 541 sheets
2	61	Seogwipo area vicinity	Y2021 : 1465 sheets Y2022: 1510 sheets Y2023 : 1510 sheets
3	41	Seogwipo area vicinity	Y2021 : 997 sheets Y2022: 997 sheets Y2023 : 997 sheets
4	47	Reflecting development near Jeju Airport	Y2021 : 1175 sheets Y2022: 1175 sheets Y2023 : 1175 sheets
5	61	Urban	Y2021 :1525 sheets Y2022: 1525 sheets Y2023 : 1525 sheets
6	60	Suburban	Y2021 : 1450 sheets Y2022: 1500 sheets Y2023 : 1500 sheets
7	60	Mountainous area	Y2021 : 1500 sheets Y2022: 1500 sheets Y2023 : 1500 sheets

Ⅲ
솔루션 구축방법
How to update maps quickly using AI

1. Training Data Development

▪ Types and volume of training data developed

Training Data Volume for Binary Change Detection Models

Data type	Image resolution	Number of zones	Pre-processing unit size	Number of data developed
For base model training	1:5000 map sheets	3 zones	20m x 20m	50,553
For fine-tuning training	1:1000 map sheets	62 zones	20m x 20m	11,751

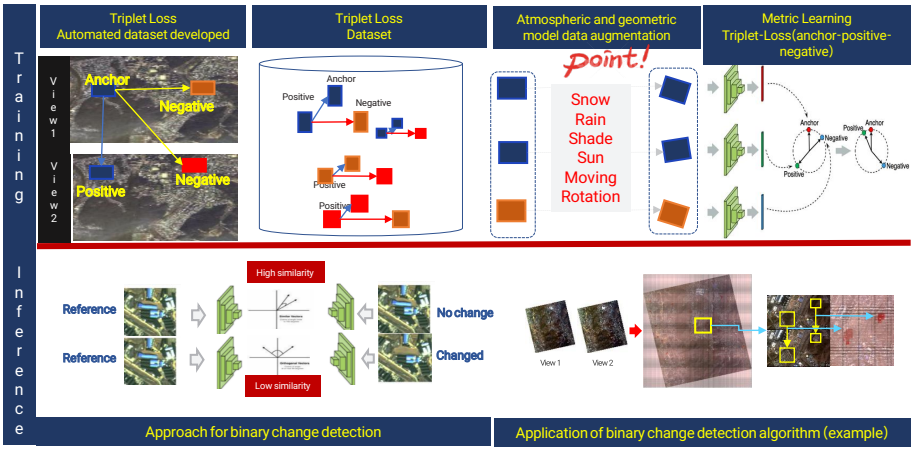
Training Data Volume for Object Detection AI

Category		Target object	Target volume: Satellite	Target volume: Aerial	Developed volume: Satellite	Developed volume: Aerial
Transportation	Road	**Road (SS)**	3,000	4,000	3,600	6,132
		Sidewalk (SS)	-	3,000	-	4,578
		Safety zone (OD)	-	4,000	-	4,944
Building	Building	**Building (OD)**	8,000	10,000	8,000	10,624
	Greenhouse	**Greenhouse (OD)**	4,000	10,000	4,000	19,344
Man-made Structures	Cemetery	**Cemetery (OD)**	-	3,000	-	5,184
		Public cemetery (OD)	-	1,500	-	2,016
Natural feature	Cultivated land	**Paddy field (SS)**	-	3,000	-	-
		Dry field (SS)	2,000	3,000	2,160	2,148
		Orchard (SS)	2,000	3,000	2,160	6,177
	Forest	**Forest (SS)**	1,000	3,000	1,350	7,578
		Pasture (SS)	-	3,000	-	5,166
	Water system	**Lake and reservoir (SS)**	1,000	3,000	1,080	1,248
Total			**74,500**		**97,489**	

2. Model Development and Training

▪ Development and training of binary change detection model

- Deep Metric Learning
  - The AI model learns to map similar images to similar representations, and dissimilar images to distinct representations
  - Compressed feature representations are extracted from input images using the trained model
  - The difference between the feature vectors is computed to determine the binary change

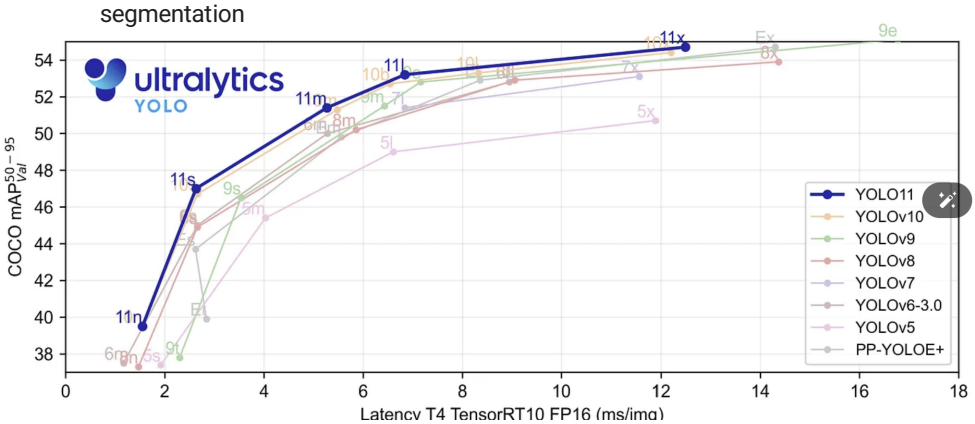


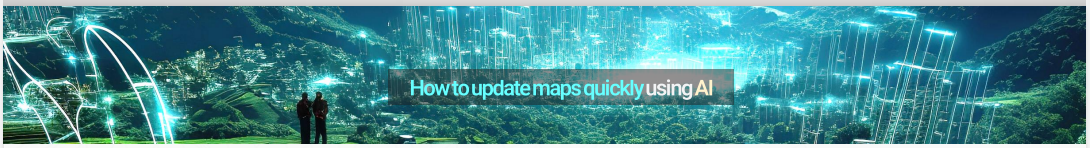
Ⅲ
솔루션
구축방법
How to update maps quickly using AI

2. Model Development and Training

▪ Training for semantic change detection

- Applied algorithms capable of recognizing both topographic features and visual appearances of digital topographic maps
- Utilized conventional object detection and semantic segmentation algorithms
  - Adopted Yolov11n which supports real-time processing with simultaneous detection and segmentation





Ⅳ. Solution Performance

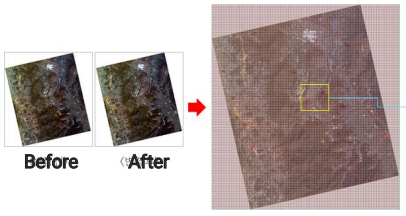
1. Performance Evaluation Method
2. Performance Test

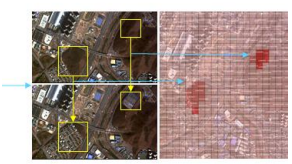
nipa 정보통신산업진흥원 HANCOM 컨소시엄 인스페이스 국토지리정보원

1. Performance Evaluation Method

Accuracy Measurement of the Binary Change Detection

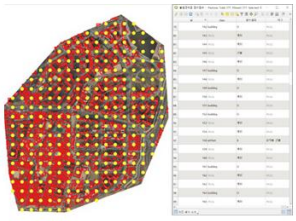
$$\text{Binary change detection accuracy} = \frac{\text{Correctly detected changed grid count}}{\text{Total changed grid count}}$$





Accuracy of the Object Detection (Detection Content)

$$\text{Object detection accuracy} = \frac{\text{Correctly detected object count (grid center)}}{\text{Total object count (total grid center)}}$$



테스트 정사영상 명	41390_003.ecw
포인트 수	276

구분	정탐	오탐	미탐	탐지제외	합계	정확도
불법 건축물	49	4	10	213	276	95
토지용도	71	30	5	170	276	87

〈성능검토 결과〉

2. Performance Test

- Test Area : 20 map sheets of Jeju Island at 1:1000 scale (selected to reflect regional characteristics, including urban, suburban, and non-urban areas)
- Input : Orthoimages of 2021 and 2022
- Output : Performance results of the change presence detection and change detail detection
- Performance Evaluation Method(ground truth) :
  ① Visual comparison ② 2022 revised topographic maps

Agricultural regions (primarily with greenhouses) and urban areas are selected as the test areas

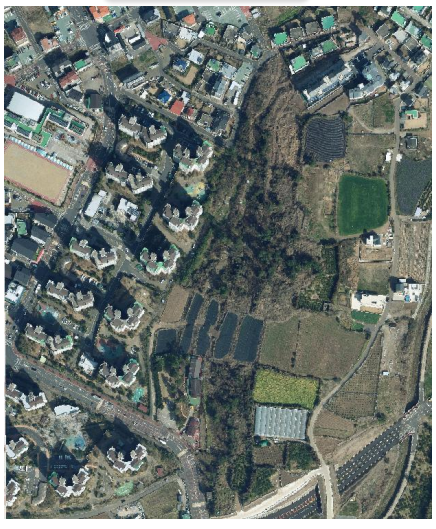
Ⅳ
솔루션 성능
How to update maps quickly using AI

2. Performance Test

- Characteristics of test areas (representative Images)

Agricultural area

Urban area



2. Performance Test

▪ Visual comparison

Map No	total	TP	TN	FP	FN	Acc	precision	recall
0	672	190	166	288	28	52.976%	39.749%	87.16%
1	672	105	289	234	44	58.631%	30.973%	70.47%
2	672	122	271	261	18	58.482%	31.854%	87.14%
3	672	61	304	303	4	54.315%	16.758%	93.85%
4	672	55	377	228	12	64.286%	19.435%	82.09%
5	672	116	264	284	8	56.548%	29.000%	93.55%
6	672	424	84	143	21	75.595%	74.780%	95.28%
7	672	185	154	318	15	50.446%	36.779%	92.50%
8	672	140	365	145	22	75.149%	49.123%	86.42%
9	672	51	431	184	6	71.726%	21.702%	89.47%
10	672	49	373	249	0	62.891%	16.443%	100.00%
11	671	59	412	189	12	70.089%	23.790%	83.10%
12	672	144	244	265	19	57.738%	35.208%	88.34%
13	672	35	401	233	3	64.881%	13.060%	92.11%
14	672	58	349	263	2	60.565%	18.069%	96.67%
15	672	175	234	227	36	60.863%	43.532%	82.94%
16	672	30	310	326	6	50.595%	8.427%	83.33%
17	672	72	233	358	9	45.387%	16.744%	88.89%
18	672	84	254	327	7	50.298%	20.438%	92.31%
19	672	263	163	223	23	63.393%	54.115%	91.96%
Total (Average)	13,439	2418	5678	5048	295	60.243%	32.387%	89.13%

Ⅳ
솔루션 성능
How to update maps quickly using AI

2. Performance Test

▪ Model characteristics

Map No	recall	Area
0	87.16%	Urban
1	70.47%	Agricultural
2	87.14%	Agricultural
3	93.85%	Urban
4	82.09%	Agricultural
5	93.55%	Agricultural
6	95.28%	Urban
7	92.50%	Urban
8	86.42%	Agricultural
9	89.47%	Agricultural
10	100.00%	Agricultural
11	83.10%	Agricultural
12	88.34%	Agricultural
13	92.11%	Agricultural
14	96.67%	Agricultural
15	82.94%	Urban
16	83.33%	Agricultural
17	88.89%	Urban
18	92.31%	Urban
19	91.96%	Urban

Agricultural area

- ✓ Accurately detects new or demolished buildings
- ✓ Tends to falsely detect color changes in greenhouses
- ✓ Tends to falsely detect changes in roof colors
- ✓ Successfully detects changes in stockpiled materials with noticeable color variation

Urban area

- ✓ Accurately detects changes in building color
- ✓ Accurately detects changes in road facilities
- ✓ Tends to falsely detect color changes in greenhouses
- ✓ Tends to falsely detect shadows as changes
- ✓ Tends to falsely detect moving objects like vehicles as changes

2. Performance Test

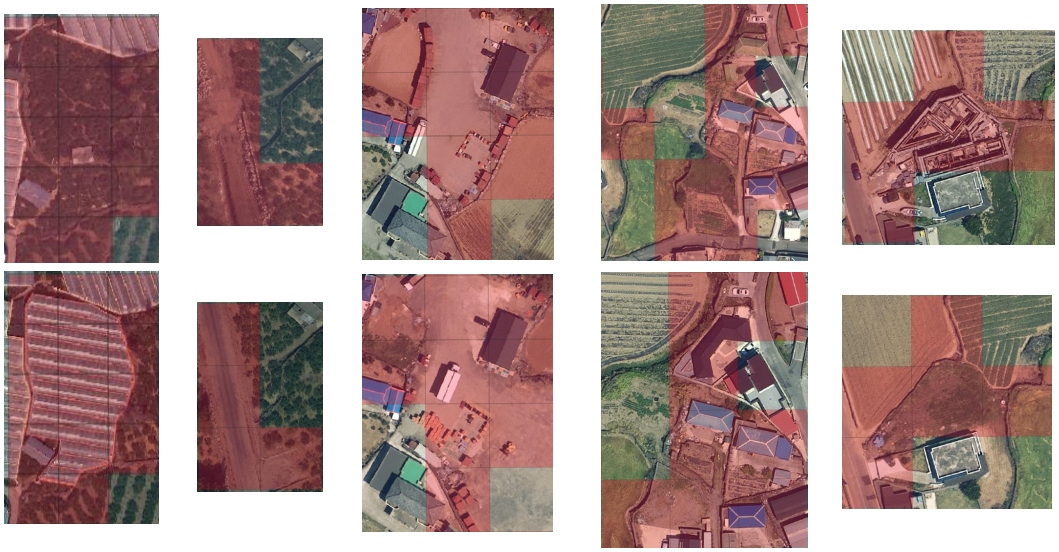
▪ Performance test by visual comparison

✓ TP examples of successful detections (agricultural areas)

Areas of Change Detected by AI

2021

2022



Ⅳ
솔루션 성능
How to update maps quickly using AI

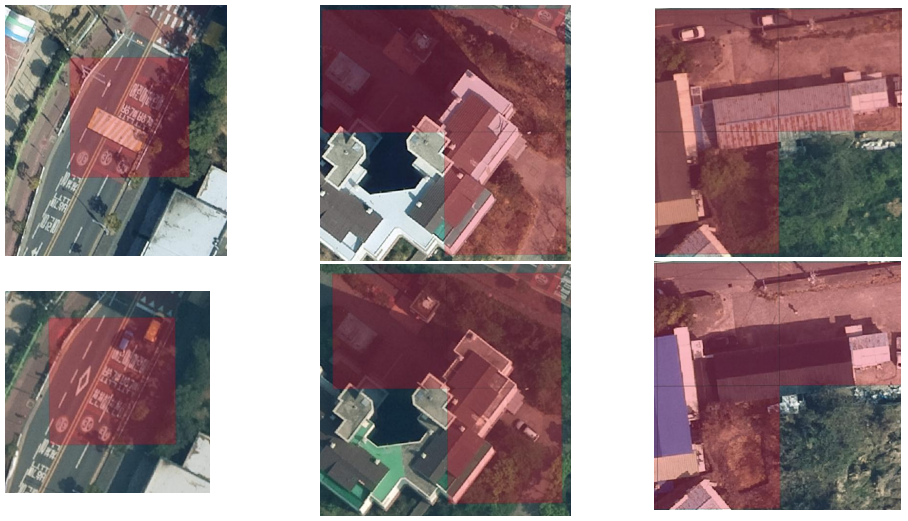
2. Performance Test

▪ Performance test by visual comparison

✓ TP examples of successful detections (urban areas)

2021

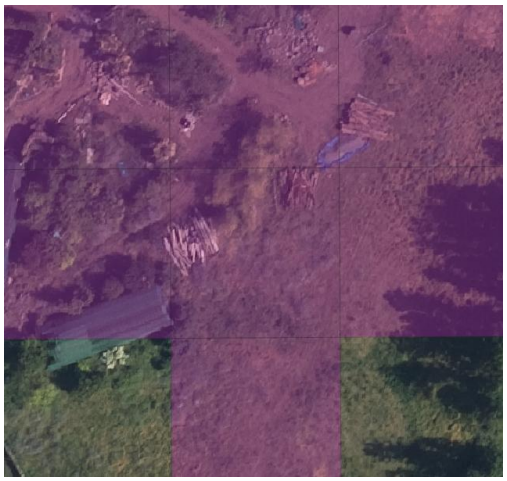
2022



2. Performance Test

▪ Performance test by visual comparison

✓ FP examples of detected changes in stored materials (map sheet 09)

2021	2022
	

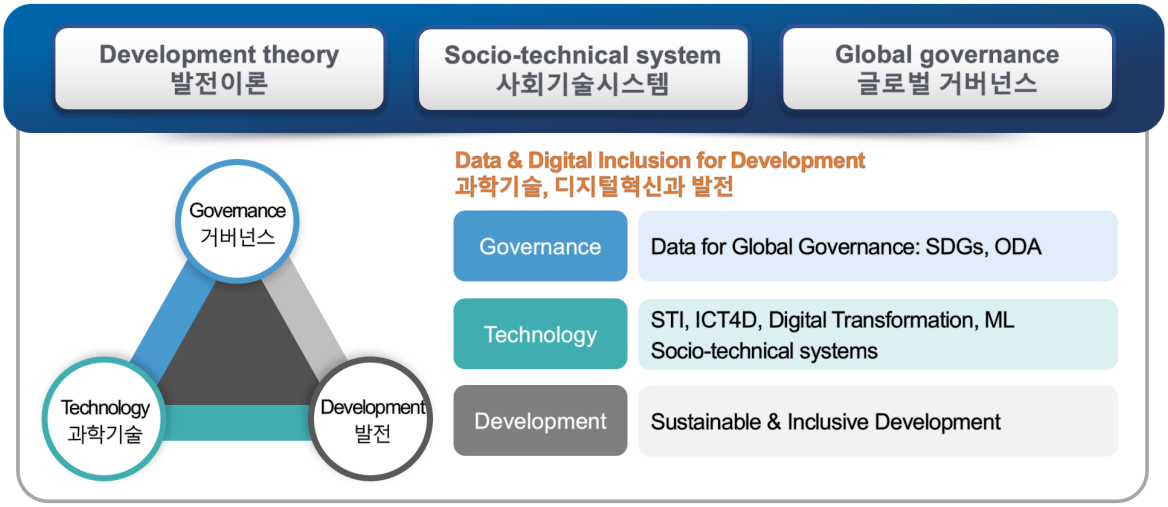
session 7

AI-driven Development Cooperation Opportunities and Challenges

2025. 05. 28
Korean Social Science Research Council Conference

Park Kyung Ryul 박경렬
Graduate School of Science and Technology Policy, KAIST
park.kr@kaist.ac.kr
di-lab.kaist.ac.kr

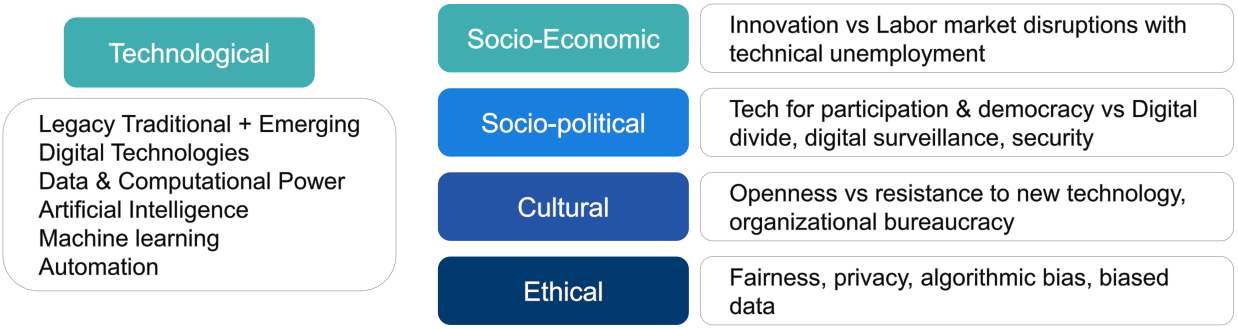
Introduction 소개



Source: KAIST di-Lab

Digital Transformation, AI and Int'l Development

- Digital transformation and increased sophistication of digital applications including AI have triggered questions for governments and international development.



- Socio-technical views on ICT, digital technologies

Park, KR. (2022). A Theoretical Reflection on International Development Cooperation in the Era of Digital Transformation. International Development and Cooperation Review,. 14(2), pp.1-20. 박경렬. 디지털 전환 시대 비판적 국제개발협력을 위한 이론적 고찰. 국제개발협력연구. (Korean)

Theoretical Evolution in ICT4D

- Technology assistance (Solow; Gamser 1988), technology transfer (Lall, 1993, Popp 2011)
- Socio-technical view (Avgerou 1996; Walsham 2001), National Innovation Systems (Nelson 1999)

- ICT4D/ICTD:
  - Developed as an independent and interdisciplinary academic field that integrates Information Systems (IS), Development Studies, Computer Sciences, Policy Studies (Richard Heeks 2002; Chrisanthi Avgerou 2008; Geoff Walsham 2017; Park 2022)

- The oldest community: Implications of Information and Digital Technologies for Development
  - Association of ICT4D researcher was established in 1989 in the IFIP(International Federation of Information Processing) WG 9.4

- Research Groups : ICIS-Global Dev, DSA – Digital Development, TC9: ICT and Society, Development (TC8 IS; TC12 AI; TC13; HCI), HCISS etc.

Challenges in Development Cooperation

- Impediments to aid coordination and development effectiveness has been studied

  1) In the context of aid & development studies
     - Donor proliferation (Burky, 2011); increased aid heterogeneity in modality (Marvrotas, 2005)
     - Donor's politico-economic interest (Barnett 2005; Kanbur et al. 1999)

  2) In the field of organizational theory (Provan & Kens 2007; Markus & Bui 2012)
     - Information issue; economic incentives; socio-political and institutional factors.

- Data/Information problem is considered as the main challenge to coordination leading to development effectiveness. (Easterly 2006; Moon & Williamson 2010; Chandy & Kharas, 2010)

  ⟹ Paris Declaration on Aid Effectiveness (OECD, 2005; AAA, 2008), GPEDC (2014)

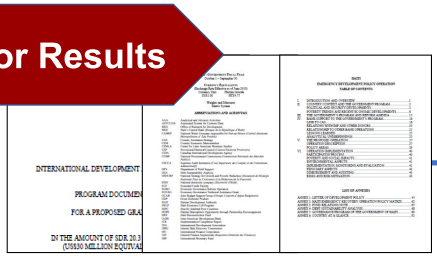
  ⟹ Advent of Data-sharing platform

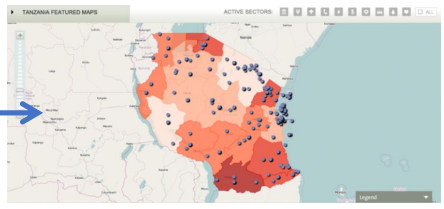
  ⟹ Opportunities in AI

Park, KR. (2017). An Analysis of Aid Information Management Systems (AIMS) in Developing Countries: Explaining the Last Two Decades. Proceedings of the 50th Hawaii International Conference on System Sciences. pp.2580-2589.

World Bank GIS-based Mapping Platform

Mapping for Results





Initiative

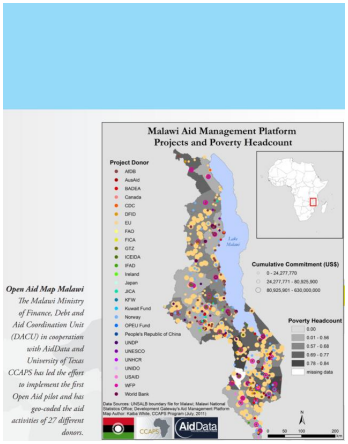
The World Bank, in partnership with AidData, launched Mapping for Results (M4R) platform which visualizes the location of World Bank projects

Results

- Mapped 30,000 Bank project locations for 2,500 active projects in 145 IDA countries
- Combined with sub-national MDG data for 107 countries
- Researchers & practitioners have access to data for better policy

Open Data for Development Cooperation

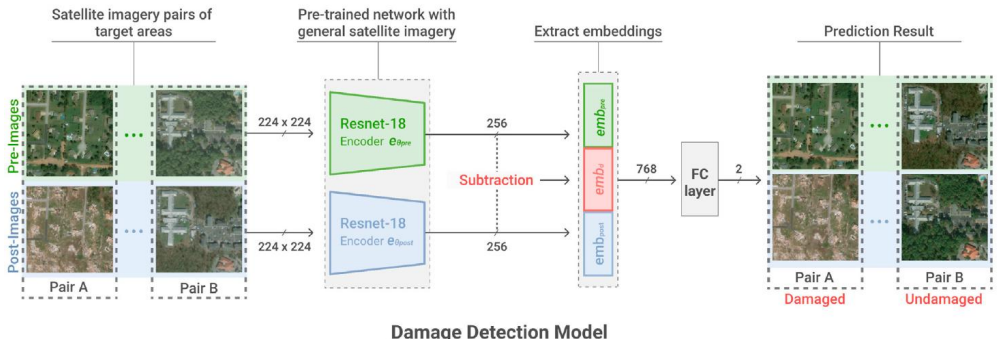
Source: Bappenas (2013)



- Park (2018) Why aid information management systems fail? Understanding the global diffusion of data-driven development initiative and sustainability failure in the case of Indonesia

AI & Satellite Images for Disaster Cooperation

- **Satellite and airborne images** are increasingly used at different stages of disaster management and mapping detection of infrastructure damage;
  - Towards a rapid automatic detection of building damage using remote sensing for disaster management (Pham et al. 2014)



Damage Detection Model

AHA Center (2022)

Kim, D., Won, J., Lee, E., Park, KR., Kim J., Park, S., Yang, H., and Cha M. (2022). Disaster Assessment Using Computer Vision and Satellite Imagery: Applications in Water-related Building Damage Detection. Frontiers in Environmental Science.

AI & Satellite Images for Disaster Cooperation

Figure 3. Examples of damage labels in the xDB dataset. The pre- and post-disaster images of minor damaged structures are nearly identical, yet they show substantial visual differences for major damaged buildings.

Kim, D., Won, J., Lee, E., Park, KR., Kim J., Park, S., Yang, H., and Cha M. (2022). Disaster Assessment Using Computer Vision and Satellite Imagery: Applications in Water-related Building Damage Detection. Frontiers in Environmental Science.

Data, Monitoring and Accountability: SDG 17.

Data, monitoring and accountability

17.18 By 2020, enhance capacity-building support to developing countries, including for least developed countries and small island developing States, to increase significantly the availability of high-quality, timely and reliable data disaggregated by income, gender, age, race, ethnicity, migratory status, disability, geographic location and other characteristics relevant in national contexts

17.19 By 2030, build on existing initiatives to develop measurements of progress on sustainable development that complement gross domestic product, and support statistical capacity-building in developing countries

17.18.1 Proportion of sustainable development indicators produced at the national level with full disaggregation when relevant to the target, in accordance with the Fundamental Principles of Official Statistics

17.18.2 Number of countries that have national statistical legislation that complies with the Fundamental Principles of Official Statistics

17.18.3 Number of countries with a national statistical plan that is fully funded and under implementation, by source of funding

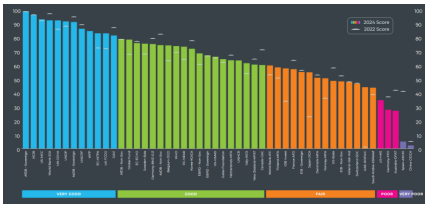
17.19.1 Dollar value of all resources made available to strengthen statistical capacity in developing countries

17.19.2 Proportion of countries that (*a*) have conducted at least one population and housing census in the last 10 years; and (*b*) have achieved 100 per cent birth registration and 80 per cent death registration

Source: UN A/RES/71/313 (2017)

Data, Monitoring and Accountability: SDG 17.

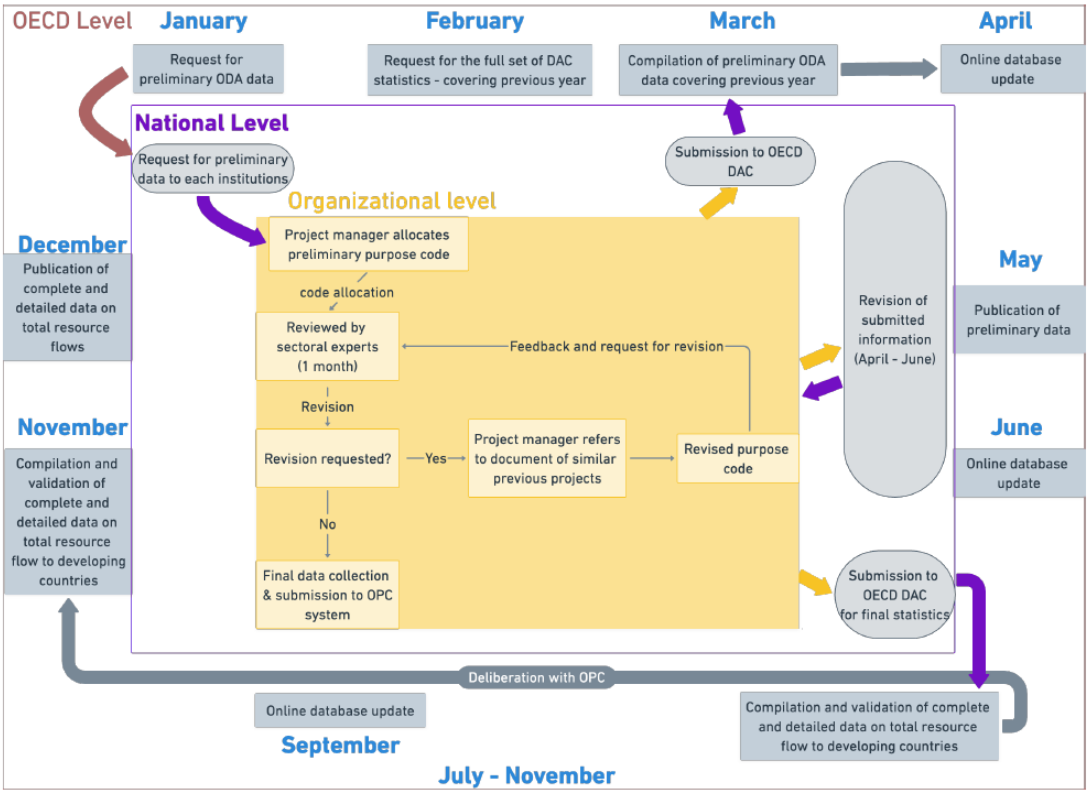
OECD Creditor Reporting Systems (CRS)
for Official Development Assistance (ODA) evaluation



- Aid Transparency Index (2024)

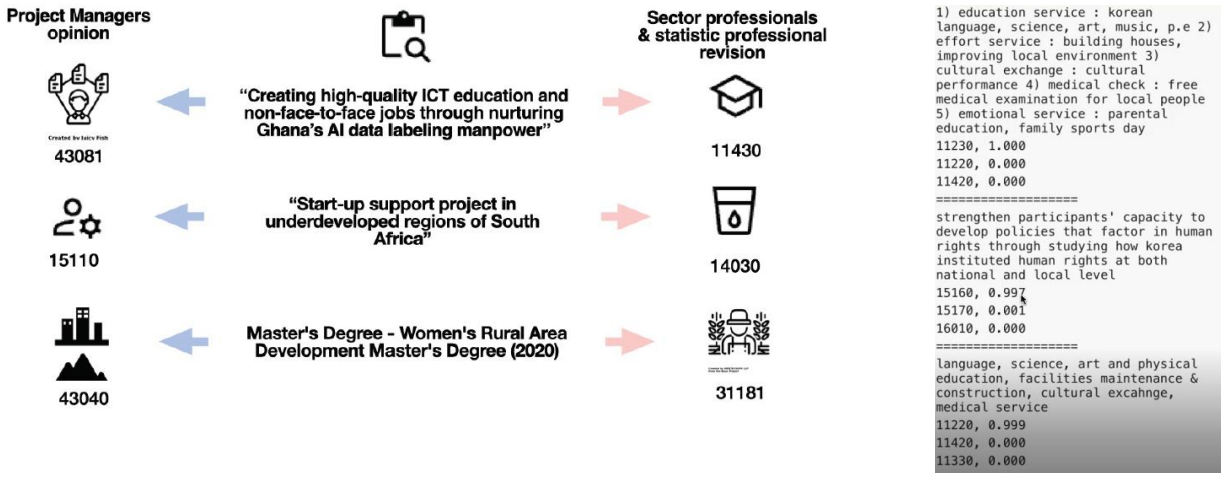
How to improve ODA/SDG targeting and M&E?

Purpose code	Code description	# of projects	Total expense (USD)
22040	Information and communication technology (ICT)	84	$12,779,947
11430	Advanced technical and managerial training	24	$11,414,800
11130	Teacher training	6	$5,821,831
15110	Public sector policy and administrative management	24	$5,656,950
11420	Higher education	10	$5,546,508
11330	Vocational training	10	$4,373,664
22010	Communications policy and administrative management	43	$4,200,991
15130	Legal and judicial development	10	$4,105,447
43030	Urban development and management	13	$3,744,262
31130 (Top 10)	Agricultural land resources	1	$2,883,806
31182	Agricultural research	8	$1,652,626
43082	Research/scientific institutions	21	$1,647,549
23110	Energy policy and administrative management	45	$1,492,486
23230	Solar energy for centralised grids	5	$937,701
23510	Nuclear energy electric power plants and nuclear safety	9	$103,105



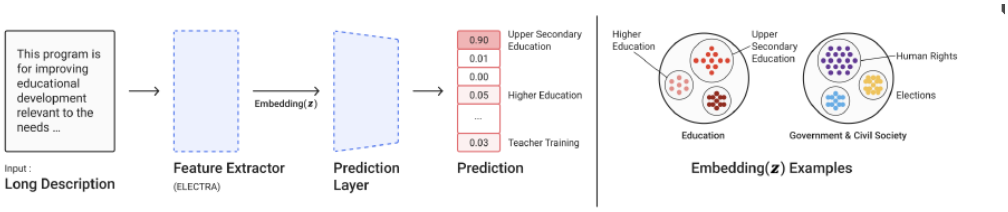
13

Machine Learning for Development Management



- Lee, J., Song, H., Lee, DJ., Kim, S., Sim, JS., Cha, M., and Park, KR. (2023). Machine Learning Driven Aid Classification for Sustainable Development. International Joint Conferences on Artificial Intelligence. 2023.

Machine Learning for Development Management



$$c = g \circ f$$
$$\mathbf{z} = f(\mathbf{x})$$
$$\hat{\mathbf{y}} = g(\mathbf{z}) \quad \text{where,} \mathbf{z} \in \mathbb{R}^d$$

ELECTRA Top-K

k	acc	f1	prec	recall
1	0.89714437	0.87401682	0.88558225	0.87263808
3	0.95749603	0.94443641	0.95050299	0.94035918
5	0.97051824	0.95738241	0.96194351	0.95411232

MAIN TAKEAWAYS

Frequency matters.
- Most frequent classification showed a higher prediction accuracy
- Infrequent purpose codes requires more attention
- Reverse engineer the model to analyze the minimum observations required for an accurate classification.

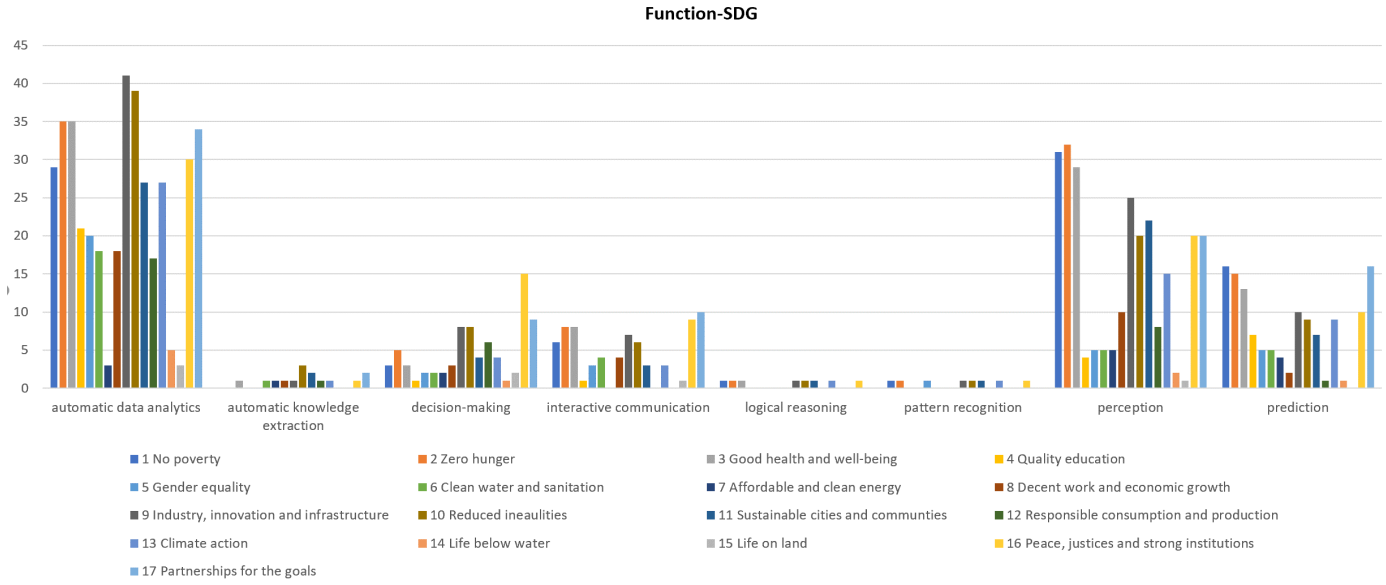
Budget matters
- Sectors with higher budgets for projects have better classification results
- Qualitative study can be proposed in the future to analyze the reasons

- Lee, J., Song, H., Lee, DJ., Kim, S., Sim, JS., Cha, M., and Park, KR. (2023). Machine Learning Driven Aid Classification for Sustainable Development. International Joint Conferences on Artificial Intelligence. 2023.

AI for Transparency & Accountability

- Closing the Feedback Loop: Can Technology Bridge the Accountability Gap? World Bank, Washington. DC.
- Wittemyer, R., Bailur, S., Anand, N., Park, KR., and Gigler, S. (2014). “New Routes to Governance: A Review of Cases in Participation, Transparency, and Accountability,”
- Importance of understanding the complexity of digital transformation, data & AI governance;
- Early stage of theoretical foundation for careful reflection on data divide and algorithmic bias, when using new forms of data and methods in development governance

AI Functionalities and the SDGs



Theoretical Views on Implementation of AI

- Four theoretical propositions of digital innovation and development (Avgerou 2010)
- Macro perspective including AI (Avgerou & Park, under review)

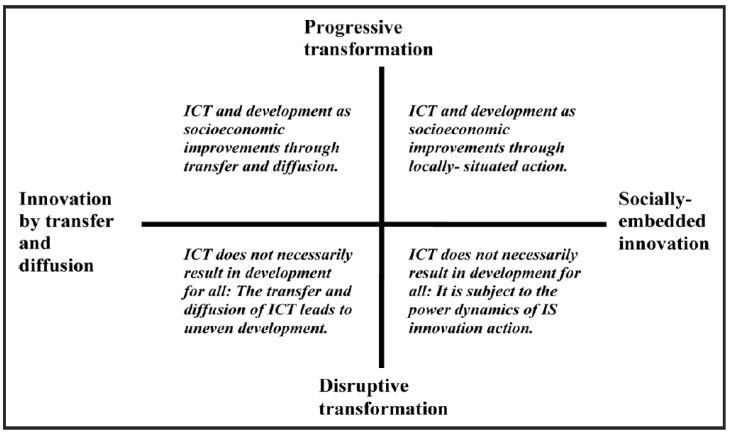
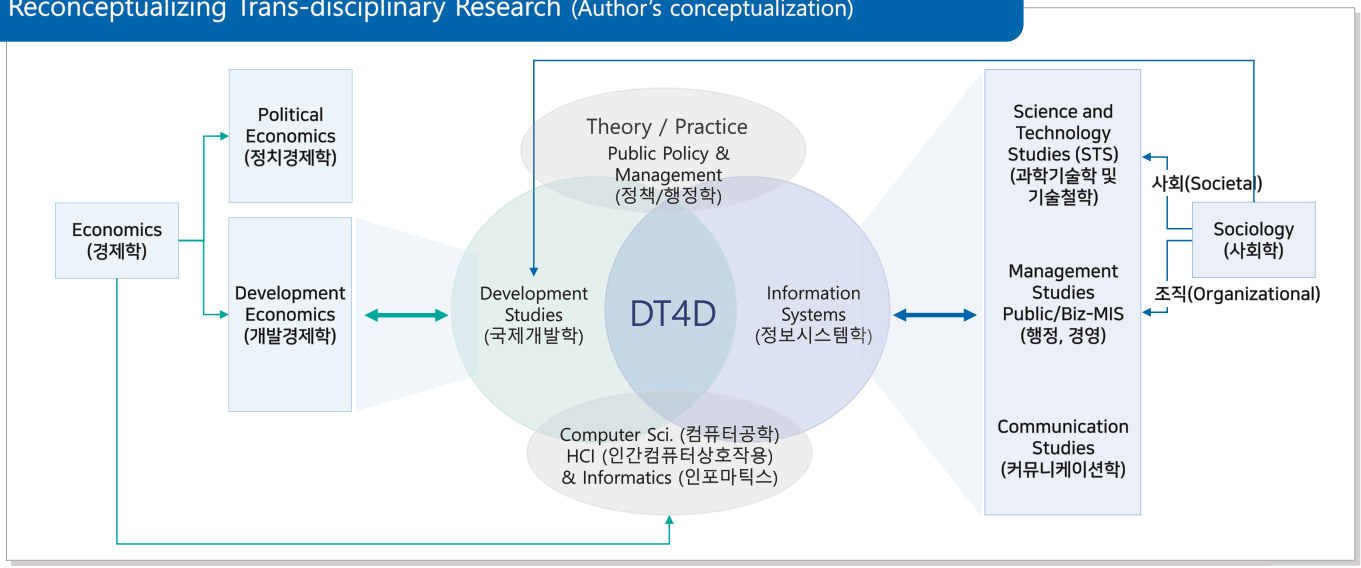


Figure 1. Four Discourses on ICTD.

Future Interdisciplinary Research

Reconceptualizing Trans-disciplinary Research (Author's conceptualization)



* Source: Park, KR. (2022). A Theoretical Reflection on International Development Cooperation in the Era of Digital Transformation.

19

KAIST

Thank you

Park Kyung Ryul

Park.kr@kaist.ac.kr

di-lab.kaist.ac.kr

di-lab

PROFESSOR MEMBERS RESEARCH PROJECTS TEACHING NEWS

Welcome

Welcome to di-lab at KAIST! di-Lab has been led by Professor Park, Kyung Ryul since 2018. We are an interdisciplinary research group with a focus on technology, development, and governance. New emerging technologies and current data revolution greatly impact not only our daily lives but also global governance. These socio-technical phenomena call for multi-disciplinary research to investigate the fundamental relationship between technology, national and global development.

Reflecting this trend, "di" stands for development innovation, digital inclusion, and data ingenuity. Research topics and questions of interest include, but are not limited to the following:

- National development strategy
- Digital transformation and governance
- Science and Technology for Sustainable development
- Data, Digital Innovation & Development Cooperation (ICT4D)
- Techno-nationalism & global political economy
- Epistemology of data-driven policy

KAIST di-Lab(디랩) 홈페이지 방문을 환영합니다!
di-Lab은 과학기술, 발전, 글로벌 거버넌스의 관계를 연구하고 있습니다. 디지털 기술의 부상과 데이터 혁명은 우리의 일상생활뿐 아니라 글로벌 기술경쟁과 개발협력의 장에도 큰 영향을 미치고 있습니다. 이러한 사회기술적(socio-technical) 현상은 과학기술과 국가 및 글로벌 발전의 근본적인 상호작용을 파악하는 다학제적이고 현장중심적인 연구를 필요로 합니다.

'di'는 발전을 위한 혁신 (development innovation), 디지털 포용 (digital inclusion), 데이터 독창성 (data ingenuity) 이라는 세 가지 가치를 의미하며 주요 연구분야는 아래와 같습니다.

- 국가 발전 전략
- 디지털 전환과 거버넌스
- 지속가능발전, 사회혁신을 위한 과학기술
- 데이터, 디지털 혁신과 국제개발협력 (ICT4D)
- 과학기술패권과 글로벌정치경제
- 데이터기반 정책의 과제

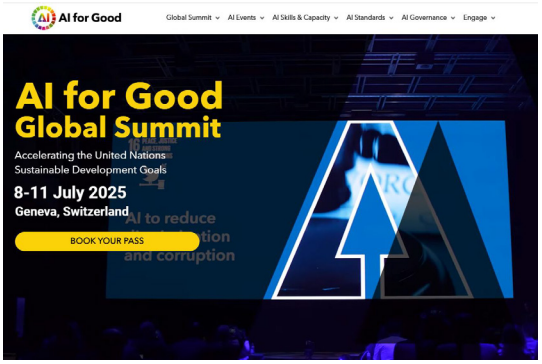
AI Governance and the Global South:
Current Discussions for International Development Cooperation

May 28, 2025
Hanah Zoo
Graduate School of Global Cooperation, Hallym University

Contents

01 Introduction| Risks and promises of AI

- Artificial Intelligence for Social Good
  - "A vector for hope" to achieve sustainable development goals (Del Rio Castro et al. 2021)
  - Explosion of experimentation by governments, applications including: agriculture, financial services, health care, pandemic response, science, transportation, and climate change response (Robles & Mallinson, 2023)
  - The development of "trustworthy" or "responsible" AI will increase the fulfillment of SDGs (ITU, 2024)

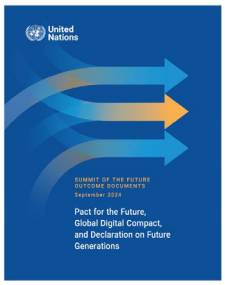


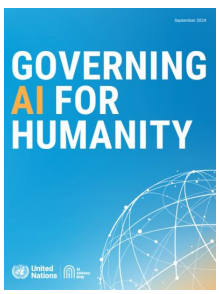
- An unregulated proliferation of AI can impose severe risks to society
  - Economic and environmental health (Garvey, 2018); democratic processes and political institutions (Erman & Furendal, 2022); inequalities between Global North and Global South countries (Gehl Sampath, 2021; Sinanan & McNamara, 2021)
  - Sociotechnical harms (Shelby et al. 2022): harms from algorithmic systems that occur through the interplay of technical system components and societal power dynamics

Source: AI for Good Homepage https://aiforgood.itu.int /

01 Introduction | Why AI governance in IDC? (1)

- Global AI governance under pressure from institutional gridlock, fragmentation, and geopolitical competition (Sepasspour, 2023)
  - OECD, G7, G20, GPAI, AI Summits; Industry standardization arena i.e. ISO, IEC, ITU, etc.
  - US, EU, China competing for leadership (e.g. Network of AI Safety Institutes led by US and EU)
  - Criticisms including lack of representation and coordination
- UN's interest in AI geopolitical collaboration and safeguarding human values, such as ethics, rules, and norms (Robles &Mallinson, 2023)
  - UN Global Digital Compact (2024) aims to ensure that digital technologies, including AI, are designed, used, and governed to benefit everyone, including developing countries
  - Governing AI for Humanity proposes a global framework for AI oversight, emphasizing the need for responsible and equitable AI governance



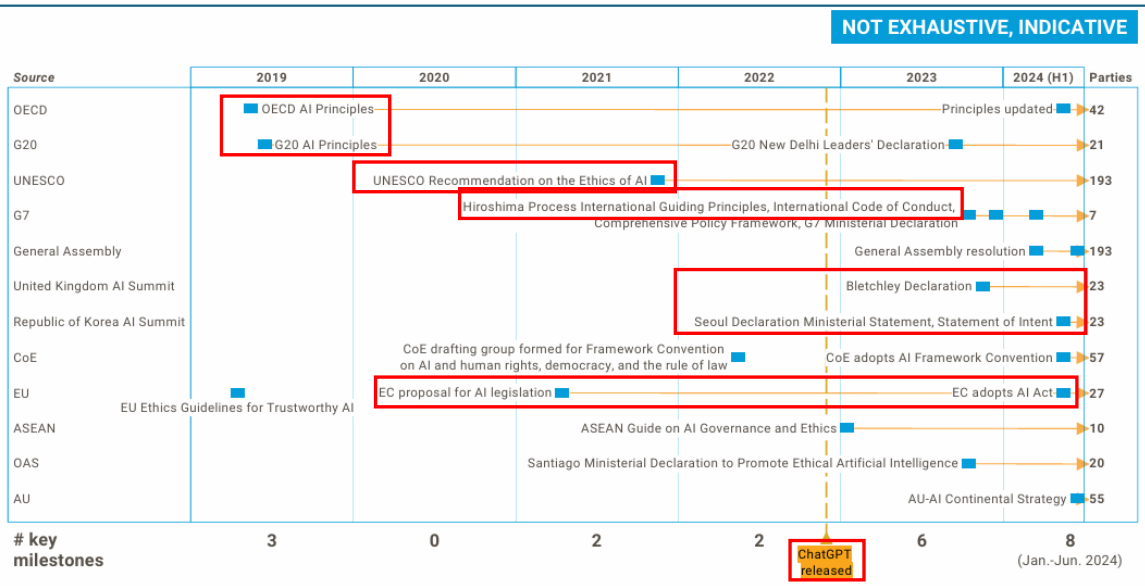


UN Global Digital Compact https://www.un.org/global-digital-compact/en
UN Homepage https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

01 Introduction | Why AI Governance in IDC? (2)

- What governance challenge of the Global South to address in the context of international development cooperation?
  - Global South perspectives in the global AI norm-setting
  - AI Governance capacity building in key development organizations' digital strategies
  - Safe, trustworthy, responsible AI aligned with development goals

02 Global Landscape | AI Governance Milestones

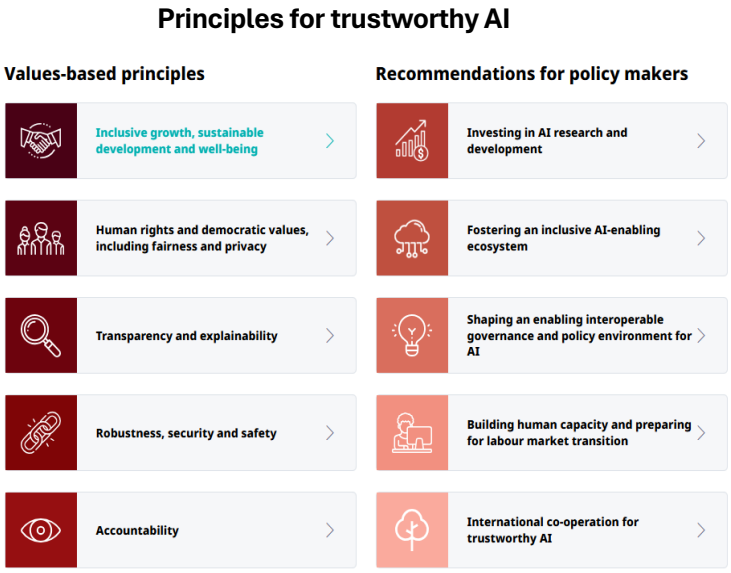


Abbreviations: ASEAN, Association of Southeast Asian Nations; AU, African Union; CoE, Council of Europe; EU, European Union; G20, Group of 20; G7, Group of Seven; GPAI, Global Partnership on Artificial Intelligence; OAS, Organization of American States; OECD, Organisation for Economic Co-operation and Development; UNESCO, United Nations Educational, Scientific and Cultural Organization.

출처: UN (2024). *Governing AI for Humanity*. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

02 Global Landscape | OECD AI Principles (2019; 2024)

- First intergovernmental standard on AI with 47 adherents to the Principle
- Promote use of AI that is innovative and trustworthy and that respects human rights and democratic values



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

7

02 Global Landscape | Governing AI for Humanity (2024)

- UN SG's High-level Advisory Body on AI(HLAB-AI)for establishing an international AI governance framework

- **Seven Key Recommendations**
  - Establish an International Scientific Panel on AI: Create a panel to provide impartial, reliable scientific knowledge about AI, including annual reports on AI capabilities, risks, and trends.
  - Initiate a Policy Dialogue on AI Governance: Launch a new policy dialogue at the UN to foster common ground on AI governance, involving intergovernmental and multistakeholder meetings.
  - Develop an AI Standards Exchange: Establish a platform where standards organizations, tech companies, and civil society can collaborate to create technical standards for measuring and evaluating AI systems.
  - Create a Global AI Capacity Development Network: Build a network to enhance AI governance capacities, offering training, computing resources, and AI datasets to researchers and social entrepreneurs.
  - Establish a Global AI Fund: Set up a fund to address capacity and collaboration gaps between countries, promoting equitable access to AI technologies.
  - Develop a Global AI Data Framework: Standardize definitions, principles, and stewardship practices related to data to ensure transparency and accountability in AI systems.
  - Set Up an AI Office within the UN Secretariat: Create a dedicated office to support and coordinate the implementation of these recommendations

8

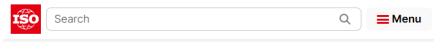
02 Global Landscape | AI Summits (2023;2024;2025)

Summit	UK AI Safety Summit (2023)	AI Seoul Summit (2024)	France AI Action Summit (2025)
Theme	Frontier AI risks and safety	Safety, innovation, and inclusivity	Global AI governance and public interest
Key Outcomes	Bletchley Declaration; AI Safety Institute	Seoul Declaration; Frontier AI Safety Commitments	Launch of Current AI Foundation; InvestAI initiative
Features	First global summit on AI safety; focus on existential risks	Emphasis on multistakeholder collaboration; inclusion of Global South perspectives	Significant investment pledges; focus on AI's societal impact
Participants	28 countries including US, China, EU	Governments, industry leaders, civil society	Broad international representation, including Global South
Approach	Precautionary, with emphasis on pre-deployment testing	Collaborative, promoting shared AI standards	Action-oriented, addressing AI's role in society and economy

9

02 Global Landscape | Recent & upcoming

- AI Standards Hub Global Summit (March 2025)
  - UK AI Standards Hub
  - Role of technical standards as a governance tool
  - Topics covering standards and measurement, assurance and certification, foundation models, civil society & standards, sustainability, global cooperation
- ITU AI for Good Global Summit (July 2025)
  - Topics moving towards more governance issues
  - AI for Good: Accelerating Progress towards the SDGs(2023) → Status, Implementation and Way Forward of AI (2024) → How AI is Shaping the Future: Trust, Standards, and Global Governance (2025)
- WSC International AI Standards Summit (December 2025)
  - A consortium of formal SDO – International Standards Organization, International Electrotechnical Commission, International Telecommunication Union
  - A response to UN HLAB-AI on "International AI Standards Exchange"
  - Driving AI standards for responsible, safe, and inclusive AI standards



World-first International AI Standards Summit to be held in 2025, announced today at the World Economic Forum in Davos

10

02 AI governance | Characteristics in the literature

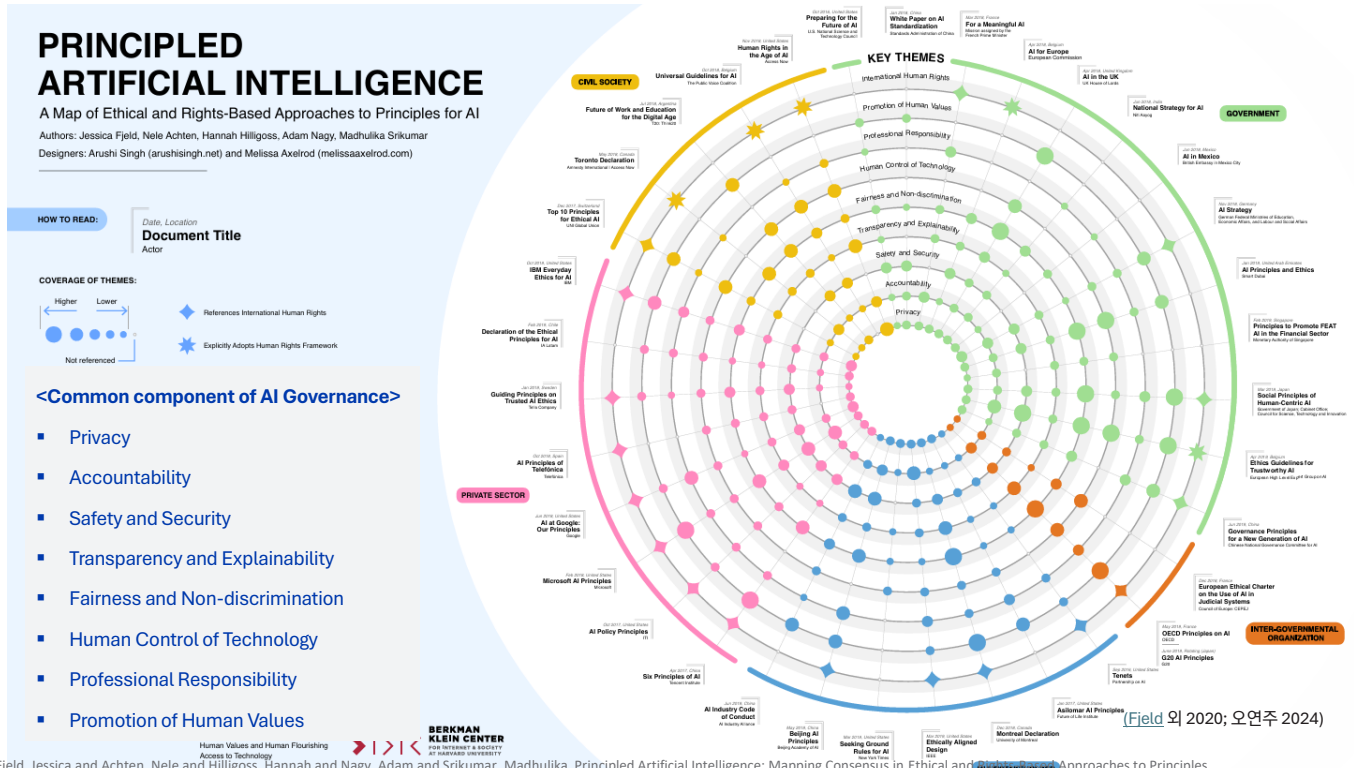
- AI governance: mechanisms and processes that shape and govern AI (Olugbade, 2025)
  - "fragmented"(Schmitt, 2022), "underdeveloped"(Naidoo, 2021), ""unorganized," and "immature" (Butcher & Beridze, 2019)
- Instruments:
  - hard approaches like governance institutions, frameworks, regulations, and legislative instrument
    - binding regulations applied on a risk-based approach
  - soft instruments like voluntary standards, codes of conduct, norms, and ethical principles
- Scope:
  - a governance mechanism that focuses on narrow and specific AI-application areas, i.e. healthcare, transportation, etc.
  - a broader global framework for general AI applications
- Centralization vs. decentralization
  - Central mechanism : efficiency, reduced competition, greater political power to effect changes (Cihon et al., 2020), inflexibility, limited scope, low stakeholder participation
  - Multiple governing bodies at different levels: greater agility, sensitivity to contextual issues, improved stakeholder inclusion (Jelinek et al., 2021)

11

02 AI governance | Key Actors in the literature

Type		Evaluation
Private Sector	big multinational corporations (Google, OpenAI, Microsoft, IBM, and Meta, etc.)	• heavily involved in the technical development and application of AI systems • a self-regulated or collective-industry regulation approach to AI governance that does not inhibit the development prospects of AI • "majority of binding agreements and voluntary commitments that exist are proposed by the private sector" (Radu, 2021)
Public Sector	National governments and their agencies (US, China..) EU as a regional actor	• Publishing national AI strategies and policies defining their countries' direction of AI development and international agreements between countries on AI applications • Powerful national actors (US, China) difficult to cooperate due to geopolitical competition (Mokry & Gurol, 2024)
	International (G20, OECD, UN, ISO/IEC...)	• High agency in brining together otherwise fragmented AI governance landscape within their existing governance architecture (Schmitt 2022) • Inability to enforce the implementation of their AI rules by members; the different interpretations of such rules by members based on their diverse cultural contexts; the absence of the membership of influential actors; the interference of geopolitical interests of powerful members in the activities of international organizations; and the varying degrees of priority given to issues on the agenda of international organizations by each country as determined by their political will and availability of resources (Cihon et al., 2020; Johnson & Bowman, 2021; Schmitt, 2022).
Non-governmental Organizations	Professional organizations i.e. IEEE, Think tanks, civil society, research institutes, etc.	• Organizing international workshops, setting standards, calling attention to issues, and drawing up recommendations on AI development, deployment, and use • Serving as a watchdogs on AI issues, serving as a reference points for IOs (Cihon et al., 2020; Schmitt, 2022) • Level of influence low; recommendations abstract and difficult to operationalize (Schiff et al., 2020)

Based on Olugbade (2025)



(Fjeld 외 2020; 오연주 2024)

Fjeld, Jessica and Achten, Nele and Hilligoss, Hannah and Nagy, Adam and Srikumar, Madhulika, Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI (January 15, 2020). Berkman Klein Center Research Publication No. 2020-1, Available at SSRN: https://ssrn.com/abstract=3518482 or http://dx.doi.org/10.2139/ssrn.3518482

02 AI governance | AI Ethics

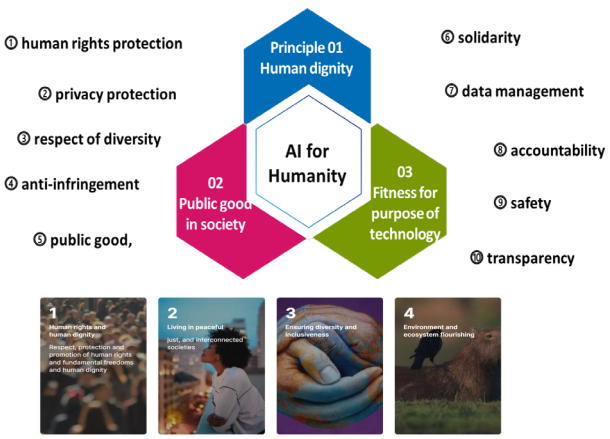
Do AI systems adhere to human values, social norms, and ethical principles?

UNESCO "Recommendation on the Ethics of Artificial Intelligence (2021)"

Core Values: ①Human rights and human dignity , ②Living in peaceful, just, and interconnected society , ③Ensuring diversity and inclusiveness , ④) Environment and ecosystem flourishing

Principles: Proportionality and do no harm, Safety and security, Right to privacy and data protection, Multi-stakeholder and adaptive governance and collaboration, Responsibility and accountability, Transparency and explainability, Human oversight and determination, Sustainability, Awareness and literacy, Fairness and non-discrimination

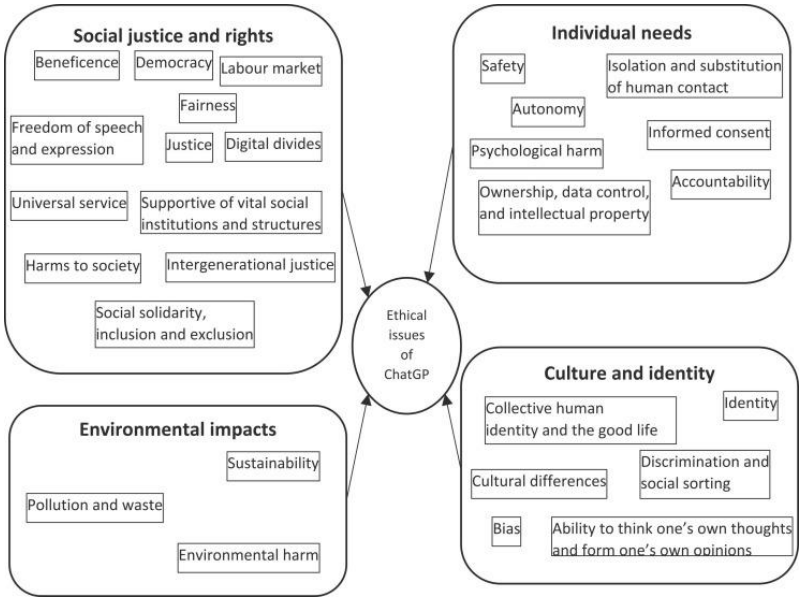
National AI Ethical Guidelines of ROK (2020)



출처 : UNESCO https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
KISDI 인공지능윤리 소통채널 https://ai.kisdi.re.kr/aieth/main/main.do

02 AI governance | The Ethics of GPT

Research on ethical issues has been criticized for focusing primarily on immediate harms and their prevention, while overlooking broader concerns such as fairness, inclusion, accessibility, and labor market impacts (Stahl & Eke, 2024)

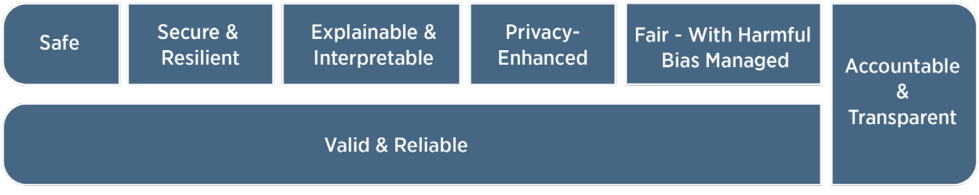


15

Bernd Carsten Stahl, Damian Eke, 2024. " The ethics of ChatGPT – Exploring the ethical issues of an emerging technology", International Journal of Information Management, Volume 74, 2024 https://doi.org/10.1016/j.ijinfomgt.2023.102700.

02 AI governance | AI Trustworthiness

- ISO definition "the ability to meet stakeholder expectations in a verifiable way"
  - "Depending on the context or sector, and also on the specific product or service, data and technology used, different characteristics [including "reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality and usability."] apply and need verification to ensure stakeholders expectations are met
- US NIST AI Risk Management Framework (RMF)
  - Trustworthiness characteristics are tied to social and organizational behavior, the datasets used by AI systems, selection of AI models and algorithms and the decisions made by those who build them, and the interactions with the humans who provide insight from and oversight of such systems.



"ISO/IEC 22989: 2022, Information Technology – Artificial Intelligence – Artificial intelligence concepts and terminology". https://www.iso.org/standard/74296.html
"ISO/IEC TR 24028:2020 Information Technology – Artificial Intelligence – Overview of trustworthiness in Artificial Intelligence." https://www.iso.org/standard/77608.html
"NIST AI RMF - AI risks and Trustworthiness" https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics/

16

02 AI governance | AI Safety

- **Safety:** "freedom from unacceptable risks" (the "negative condition") *ISO/IEC TR 24028:2020*
  - The property of avoiding harmful outputs, such as providing dangerous information to users, being used for nefarious purposes, or having costly malfunctions in high-stakes settings (Bengio et al., 2025)
  - AI system should "not under defined conditions, lead to a state in which human life, health, property, or the environment is endangered" ISO/IEC TC 5723:2022; address both technical and non-technical risks
- **Risk:** The combination of the probability and severity of a harm that arises from the development, deployment, or use of AI. (Bengio et al., 2025)
- **Risk Classification** (Bengio et al., 2025)
  - **Risks from malicious use:** harm to individuals through fake contents, manipulation of public opinion, cyber offence, biological and chemical attacks
  - **Risks from malfunctions**: reliability issues, bias, loss of control
  - **Systemic risks**: labor market risks, global AI R&D divide, market concentration and sing points of failure, risks to the environment, risks to privacy, risks of copyright infringement
  - Impact of open-weight general purpose AI models on AI risks

"ISO/IEC TR 24028:2020 Information Technology – Artificial Intelligence – Overview of trustworthiness in Artificial Intelligence." https://www.iso.org/standard/77608.html
Bengio et al. "International AI Safety Report," DSIT 2025/001. January 2025. 2501.17805

03 Governance Challenges | Data

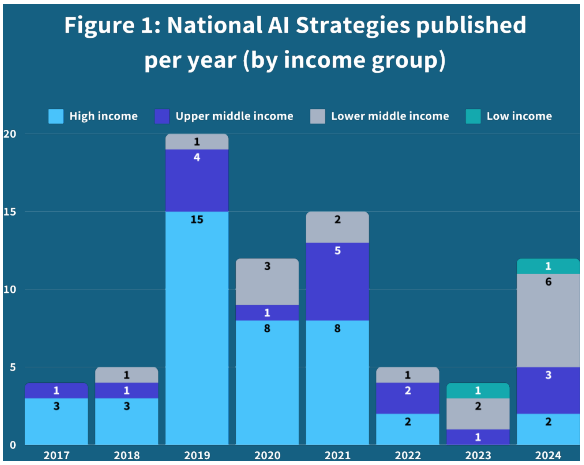
Structural Inequalities

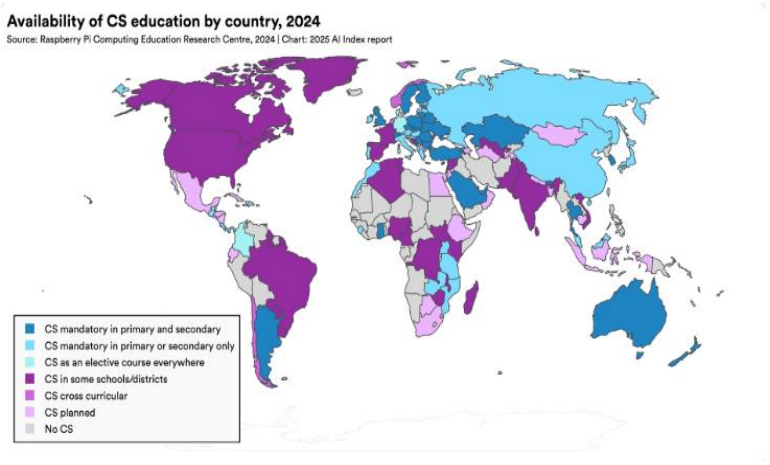
- AI systems interact with and deepen pre-existing structural inequalities—social, racial, ethical, economic, and institutional
- Opacity and bias of AI systems exacerbate marginalization, especially in contexts with limited institutional accountability (Gehl, 2021)
- techno-solutionism often excludes people and issues not represented in datasets—leading to systemic exclusion (Arora, 2019)
- User vulnerability is heightened—many cannot "opt out" as banking, markets, and transactions depend on digital services.

A disconnect in the data – industrialization narrative

- While data-driven growth is needed, personal data is often exploited unethically by private actors
- Data localization to counter data extraction of foreign tech companies vs. Digital industrialization strategies
- Government opacity: challenges in enforcing data localization or preventing political misuse of personal data

03 Governance Challenges | AI divide





The Government AI Readiness Index 2024, Oxford https://oxfordinsights.com/ai-readiness/ai-readiness-index/
The 2025 AI Index Report, Stanford University. https://hai.stanford.edu/ai-index/2025-ai-index-report

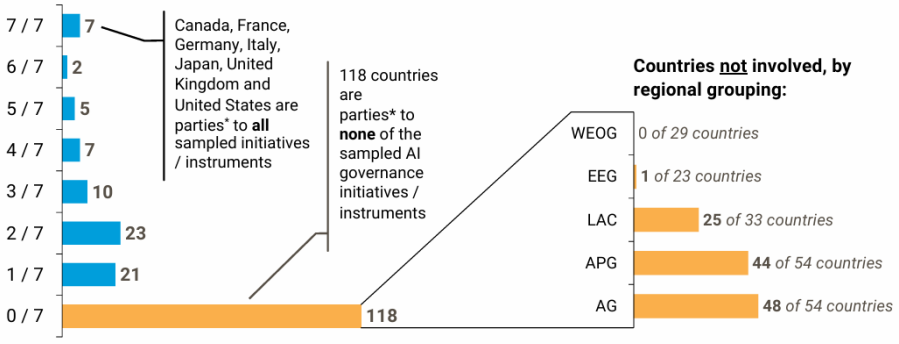
19

03 Governance Challenges | Representation

Figure 8: Representation in seven non-United Nations international AI governance initiatives

Sample: OECD AI Principles (2019), G20 AI principles (2019), Council of Europe AI Convention drafting group (2022–2024), GPAI Ministerial Declaration (2022), G7 Ministers' Statement (2023), Bletchley Declaration (2023) and Seoul Ministerial Declaration (2024).

INTERREGIONAL ONLY, EXCLUDES REGIONAL



* Per endorsement of relevant intergovernmental issuances. Countries are not considered involved in a plurilateral initiative solely because of membership in the European Union or the African Union. *Abbreviations*: AG, African Group; APG, Asia and the Pacific Group; EEG, Eastern European Group; G20, Group of 20; G7, Group of Seven; GPAI, Global Partnership on Artificial Intelligence; LAC, Latin America and the Caribbean; OECD, Organisation for Economic Co-operation and Development; WEOG, Western European and Others Group.

20

출처: UN (2024). *Governing AI for Humanity*. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

04 Digital Development | UNDP Digital Strategy

"Humans and AI complement each other, opening new opportunities for economic growth, production, health care, education, communication, and transportation."
— UNDP Digital Strategy

UNDP AI Approach

- To accelerate progress towards sustainable development, whilst steadfastly promoting human rights
- Ethical, transparent and sustainable development and utilization of AI technologies to ensure their deployment strengthens local AI ecosystems and advances human dignity, equality and justice for all

DEEP lens

- Demystify & Democratise
  - Accessible to all; Everyone in society to understand AI: terminology, usage, risks, and potential benefits
- Empower people (to server people; improving lives and livelihoods)
- Explore & Experiment
  - To be tested in a safe and secure way to maximize its positive impact
- Protect people
  - Put people and their rights and safety at the center; tackling issues of bias and ensuring accountability in AI usage

UNDP - AI Hub for Sustainable Development

- Launched at June 2024 G7 Leaders' Summit to support local AI digital ecosystems

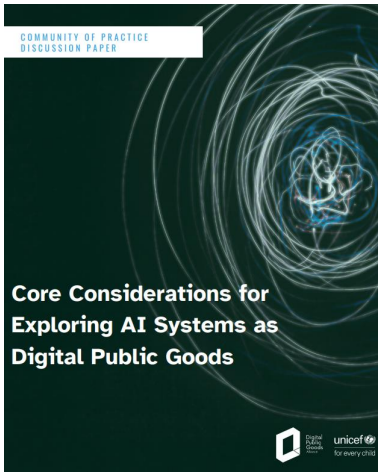
Data	Compute	Talent
• Enhance data quantity and quality through public–private partnerships in data collection • Strengthen physical and governance infrastructure to improve Africa's data pipeline. • Ensure accountability, transparency, and interoperability in the data ecosystem via open–source digital public goods. • Example: Masakhane (African language natural language processing)	• Leverage private sector collaboration to enhance computing resources in Africa's AI ecosystem. • Mobilize local companies to innovate computing accessibility, sustainability, and economic efficiency. • Develop local and green computing capacities for African talent through global stakeholder partnerships. • Support the establishment of a robust AI environment in Africa by sharing digital public goods. • Bridge gaps by opening EU High–Performance Computing (HPC) resources.	• Universalize AI education and research across the educational ecosystem. • Improve access to resources for AI technology development. • Create an environment that nurtures African "unicorns" and startup ecosystems. • Address the "dual technology challenge" through private sector partnerships and digital public infrastructure.

출처: UNDP Homepage https://www.undp.org/publications/ai-hub-sustainable-development-strengthening-local-ai-ecosystems-through-collective-action

04 Digital Development | AI as Digital Public Goods

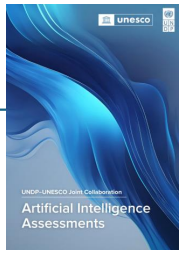
- Digital Public Goods
  - "Open-source software, open data, open AI models, open standards, and open content that do no harm and comply with relevant privacy laws, international/national standards, and best practices."(*UN Global Digital Compact, 2024*)
  - Open source: Free to access, modify, and redistribute by all users

- Open-source AI models
  - Need to balance open-source AI with responsible AI
    - Ensures transparency and ethical principles for public good
    - Mitigates risks such as data bias, privacy violations, and lack of accountability
  - Challenges with Open Data
    - Trade-off: Openness (transparency, reusability) vs misuse of sensitive data
    - Need to define the level of openness for data and models in AI development
    - Concerns: Data exploitation, data colonialism
  - Responsible AI Licenses (RAIL)
    - a new type of open-source license designed specifically for AI models
    - Unlike traditional open-source licenses that allow broad and unrestricted use, RAIL licenses impose ethical and use-case restrictions to prevent harmful applications of AI.
    - To balance openness (collaboration, innovation) with ethical safeguards by allowing developers to: 1) Share AI models and tools openly, 2) Prohibit misuse of those models for harmful or unethical purposes



출처: Digital Public Goods Alliance (DPGA) https://www.digitalpublicgoods.net/AI-CoP-Discussion-Paper.pdf
United Nations. n.d. "United Nations Summit of the Future Global Digital Compact." https://www.un.org/en/summit-of-the-future/global-digital-compact

04 Digital Development | UNDP-UNESCO Assessment



UNDP:
Artificial Intelligence Landscape Assessment (AILA)

Purpose:
- Assess national readiness for integrating AI across public and economic sectors
- Support responsible adoption of AI through governance frameworks

Focus areas:
- **Government as a User**: Capacity to deploy AI in government operations
- **Government as an Enabler**: Supportive role for national AI ecosystems
- **Ethical AI**: Ensure responsible and trustworthy AI use

Methodology & Outputs:
•Field surveys & expert interviews via UNDP country offices
•Gap analysis and country-specific policy recommendations
•Strategic vision workshops and follow-up partnerships

UNESCO:
Readiness Assessment Methodology (RAM)

Purpose:
- Align national policies with global standards (UNESCO's Recommendation on the Ethics of AI)
- Promote ethical AI governance across sectors

Focus areas:
- Legal, Social/Cultural, Scientific/Educational, Economic, Technical and Infrastructural readiness

Methodology & Outputs:
•Multistakeholder process led by UNESCO
•Country reports with actionable roadmaps
•Capacity-building plans for institutions and human resources

출처: UNDP https://www.undp.org/digital/aila
UNDP UNESCO Joint Collaboration Artificial Intelligence Assessments https://www.undp.org/sites/g/files/zskgke326/files/2024-12/undp-unesco-offer-web-7-aug-2024.pdf

04 Digital Development | UNESCO Assessment

- Policy Areas for Responsible developments in AI

https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
https://www.unesco.org/ethics-ai/en/eia?hub=32618

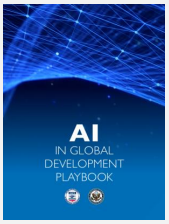
- Ethical Impact Assessment (EIA)
  - A tool for evaluating the benefits and risks of an AI system in relation to the values and principles of the UNESCO Recommendation on the Ethics of AI
  - considers the entire process of designing, developing and deploying an AI system allowing for assessment of the risks before and after the system is released to the public.
- Structure
  - Scoping questions
    - Project description, proportionality screening, project governance, multi-stakeholder governance
  - Implementing the UNESCO principles
    - Safety and security, fairness/non-discrimination/diversity, sustainability, privacy and data protection, human oversight/determination, transparency/explainability/accountability/responsibility, awareness/literacy

25

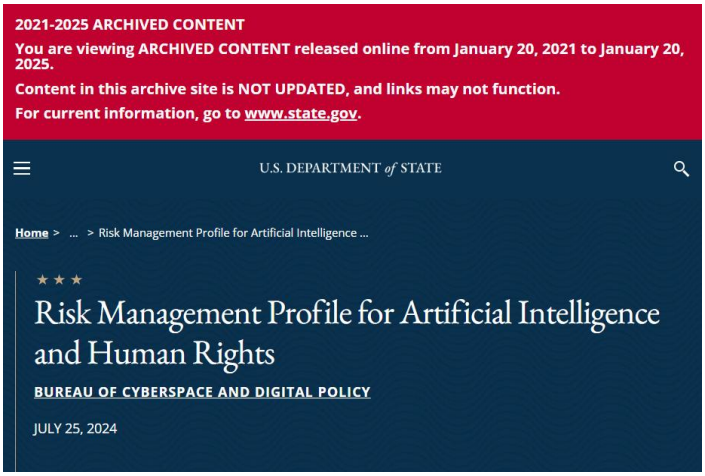
04 Digital Development | US – Responsible AI *(cancelled)*

Responsible AI
An approach to AI that aligns the design, deployment, and use of AI with the core values of democracy, reliability, safety, security, trustworthiness, inclusion, transparency, privacy, cybersecurity, fairness, human rights, and accountability

Recommendations
1. Enhancing Capacity, Promoting AI-Related Skills Across All Sectors and Levels, and Protecting the Workforce
2. Building Trusted and Sustainable Digital Infrastructure
3. Broadening Access to Data Storage and Compute Resources
4. Creating Representative, Locally Relevant Datasets and Preserving Cultural Heritage
5. Developing Strategies to Deliver the Promise of AI in Practice
6. Advancing Good Governance Frameworks for the Development and Use of Safe and Rights Respecting AI Systems
7. Fostering Trust in AI through Openness, Transparency, and Explainability
8. Deploying AI Sustainably and for Climate Action



04 Digital Development | NIST AI RMF for Human Rights



Core Principles:
- Trustworthiness, Adaptability, Inclusivity, Accountability

Why international human rights matter for AI governance
- universally applicable and already function as a shared international language to enable effective due diligence and technology governance.
- commitments are relevant to both governments and private sector actors, who play significant roles in AI design, development, deployment, use and governance
- Many risks posed by AI are related to human rights

Purpose:
- Show AI designers, developers, deployers, and users how to apply NIST's AI Risk Management Framework to contribute to human rights due diligence practices.
- Facilitate rights-respecting AI governance throughout AI design, development, deployment, and use by all stakeholders.

Organizational Functions
- **Govern :** set up institutional structures and processes
- **Map :** understand context and identify risks
- **Measure :** assess and monitor risks and impacts
- **Manage :** prioritize, prevent, and respond to incidents

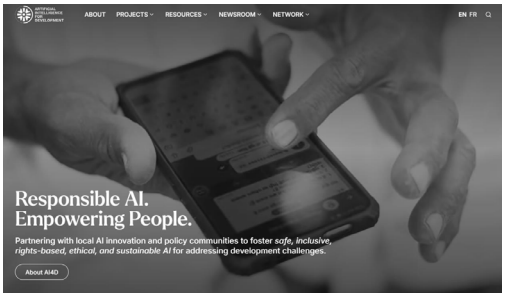
출처 National Institute of Standards and Technology https://www.nist.gov/itl/ai-risk-management-framework/perspectives-about-nist-artificial-intelligence-risk-management
https://2021-2025.state.gov/risk-management-profile-for-ai-and-human-rights/

27

04 Digital Development | UK FCDO AI for Development

- "AI for Development" initiative
  - Released "AI for Development" initiative at the AI Safety Summit (2023)
  - Partners with Canada (IDRC), Gates Foundation, USAID

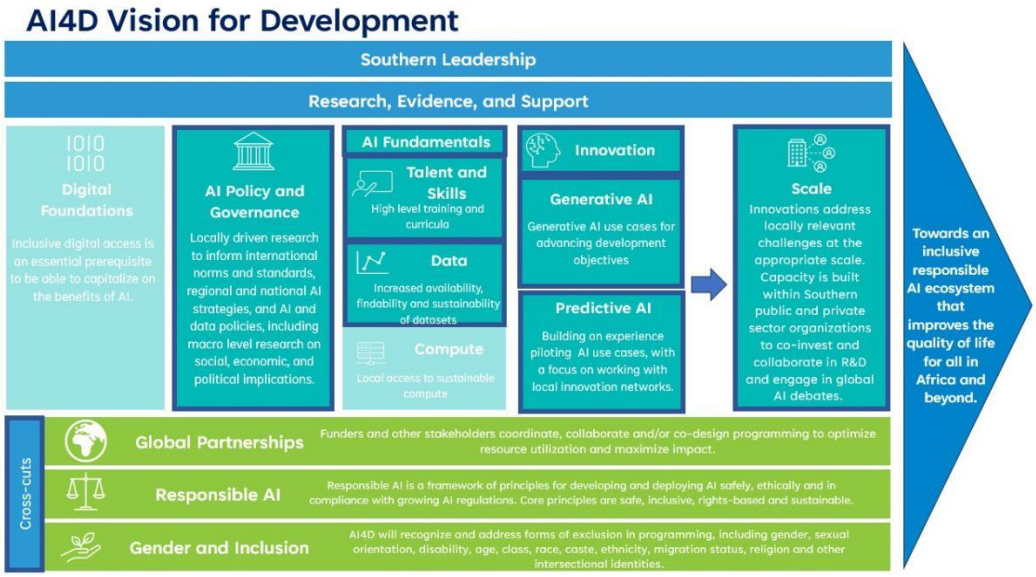
- AI in UK FCDO Digital Development Strategy (2024-2030)
  - Support AI for Development programs, initially focused on Africa, to build local AI capabilities and promote responsible AI development and application
  - Invest in graduate education and fellowships, emphasizing data-driven innovation with local representation
  - Expand developing countries' participation in the OECD-led Global Partnership for AI (GPAI) through continued engagement
  - By 2030, establish or expand at least 8 Responsible AI Research Labs in African universities and support the creation of regulatory frameworks for responsible, fair, and safe AI in over 10 partner countries
    - Lab for the Ethics, Policy and Scaling of Responsible AI (LEPSAI)



https://www.ai4d.ai/

28

FCDO - IDRC AI4D Vision



To support a responsible AI ecosystem where local experts are enabled to solve their own development challenges with safe, inclusive, rights-based and sustainable AI applications and policies

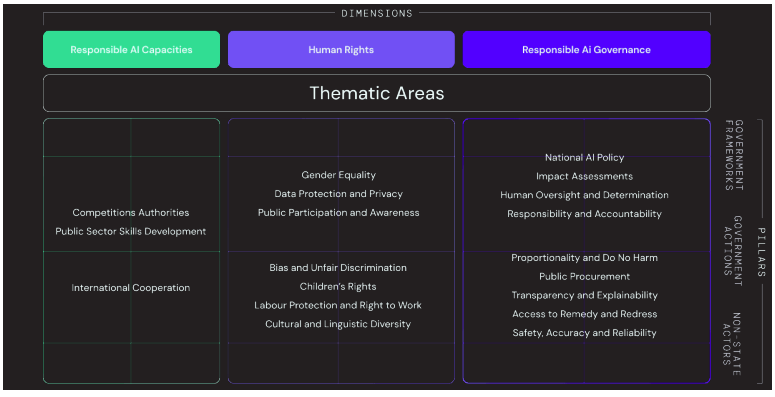
https://idrc-crdi.ca/en/initiative/artificial-intelligence-development

29

04 Digital Development | AI4D Global Index on Responsible AI

- Responsible AI
  - The design, development, deployment and governance of AI in a way that respects and protects all human rights and upholds the principles of AI ethics through every stage of the AI lifecycle and value chain.

- Measurement
  - measures 19 thematic areas of responsible AI, across three dimensions.
  - Each thematic area assesses the performance of three different pillars of the responsible AI ecosystem: Government frameworks, government actions, and non-state actors' initiatives.



https://www.global-index.ai/

30

05 Implications | Analysis of Key Initiatives (1)

- Vision and Principles
  - Vision: Responsible AI use for sustainable development and an inclusive society
    - Human rights (UNDP), democratic values (USA), Ethics (UNESCO)
  - Principles: Responsible AI, human rights, transparency
    - Local ecosystem and sustainability (UNDP), cybersecurity (USA), developing country leadership and partnership (UK)
- Approach and Implementation
  - Approach: Focus on AI accessibility, talent development, local data, and responsible AI use, human rights related impact
    - Governance, validation through assessments, Digital Public Goods (DPG), etc.
  - Implementation: Staff and organizational innovation (UNDP, USA), multilateral cooperation (USA, UK), AI readiness assessment (UNDP), AI risk management framework application (USA)

- Key Areas and Initiatives
  - Sectors: Health, education, agriculture, disaster response, environmental sustainability, public services
  - Building digital infrastructure (foundational computing), creating locally adapted datasets, promoting SDGs, and establishing talent and ecosystem networks, offering guidelines for impact (human rights, responsible AI, ethics, etc.), measuring capacity (Assessment tools, Global responsible AI index)

31

05 Implications | Analysis of Key Initiatives (2)

Responsible AI and Governance
- Based on accountability; reflects human rights and democratic values
- Governance structure includes ethical and transparent AI use, data privacy, and personal information protection

Strengthening Local AI Ecosystems
- Applying contextual approaches tailored to current circumstances
- Ensuring efficiency and fairness in local datasets, data representation, and contextual benchmarking

Talent Development and Skill Enhancement
- Access to education and training in AI skills

Data and Computing Support
- Data collection through partnerships, high-performance computing, transparency with open data

Inclusive AI and Fairness
- Ensuring inclusive access and fair solutions

Sustainable AI
- Greening AI, eco-friendly transitions

Innovation and Implementation Expansion
- Expanding impact through performance-based innovation and digital commons (open models, standards)

32

•Provide capacity-building, ethical AI frameworks, and practical guidance for responsible AI adoption
•Invest in partnerships to co-develop and share digital public goods for sustainable development

05 Implications | take aways for Korea

- Bridging the Governance Strategy Gap in Development Cooperation
  - Lacks a coordinated AI governance strategy for international development cooperation
  - AI as General-Purpose Technology; common cross-sector and sector-specific ODA guidelines for AI development and deployment
    - AI infrastructure, lightweight models, legal/governance frameworks, safety/risk assessment
  - Actively incorporating global gold standards (UNESCO, UNDP assessment frameworks and guidelines), proactively considering digital public goods—open data, open models, and open standards

- Enabling Global South Participation in AI governance
  - Recent government's initiatives (i.e. AI Seoul Summit, WSC International AI Standards Summit, Digital Bill of Rights, etc.) highlight commitment to international cooperation
  - leverage its position in global AI discussions to support more inclusive governance
  - Opportunity to align domestic AI governance experience with development cooperation policies

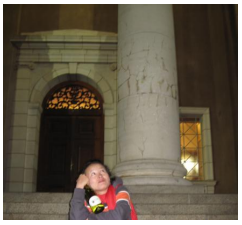
References

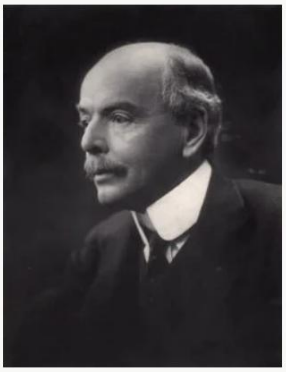
- Acharya, A. (2004). How ideas spread: Whose norms matter? norm localization and institutional change in Asian Regionalism. *International Organization*, 58(2), 239–275. https://doi.org/10.1017/S0020818304582024
- Y. Bengio, S. Mindermann, D. Privitera, T. Besiroglu, R. Bommasani, S. Casper, Y. Choi, P. Fox, B. Garfinkel, D. Goldfarb, H. Heidari, A. Ho, S. Kapoor, L. Khalatbari, S. Longpre, S. Manning, V. Mavroudis, M. Mazeika, J. Michael, J. Newman, K. Y. Ng, C. T. Okolo, D. Raji, G. Sastry, E. Seger, T. Skeadas, T. South, E. Strubell, F. Tramèr, L. Velasco, N. Wheeler, D. Acemoglu, O. Adekanmbi, D. Dalrymple, T. G. Dietterich, P. Fung, P.-O. Gourinchas, F. Heintz, G. Hinton, N. Jennings, A. Krause, S. Leavy, P. Liang, T. Ludermir, V. Marda, H. Margetts, J. McDermid, J. Munga, A. Narayanan, A. Nelson, C. Neppel, A. Oh, G. Ramchurn, S. Russell, M. Schaake, B. Schölkopf, D. Song, A. Soto, L. Tiedrich, G. Varoquaux, E. W. Felten, A. Yao, Y.-Q. Zhang, O. Ajala, F. Albalawi, M. Alserkal, G. Avrin, C. Busch, A. C. P. de L. F. de Carvalho, B. Fox, A. S. Gill, A. H. Hatip, J. Heikkilä, C. Johnson, G. Jolly, Z. Katzir, S. M. Khan, H. Kitano, A. Krüger, K. M. Lee, D. V. Ligot, J. R. López Portillo, D., O. Molchanovskyi, A. Monti, N. Mwamanzi, M. Nemer, N. Oliver, R. Pezoa Rivera, B. Ravindran, H. Riza, C. Rugege, C. Seoighe, H. Sheikh, J. Sheehan, D. Wong, Y. Zeng, "International AI Safety Report" (DSIT 2025/001, 2025); https://www.gov.uk/government/publications/international-ai-safety-report-2025
- Del Río Castro G, González Fernández MC, Uruburu Colsa Á. Unleashing the convergence amid digitalization and sustainability towards pursuing the sustainable development goals (SDGs): a holistic review. J Clean Prod. 2021;280:122204. https://doi.org/10.1016/j.jclepro.2020.122204.
- de Souza, S. P., & Taylor, L. (2025). Rebooting the global consensus: Norm entrepreneurship, data governance and the inalienability of digital bodies. Big Data & Society, 12(2). https://doi.org/10.1177/20539517251330191 (Original work published 2025)
- Gehl Sampath, P. (2021), Governing Artificial Intelligence in an Age of Inequality. Glob Policy, 12: 21-31. https://doi.org/10.1111/1758-5899.12940
- Madnick, B., Huang, K., & Madnick, S. (2023). The evolution of global cybersecurity norms in the digital age: A longitudinal study of the cybersecurity norm development process. *Information Security Journal: A Global Perspective*, *33*(3), 204–225. https://doi.org/10.1080/19393555.2023.2201482
- Malmio, I. Artificial intelligence and the social dimension of sustainable development: through a security perspective. *Discov Sustain* **5**, 466 (2024). https://doi.org/10.1007/s43621-024-00677-6
- OECD, Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449, 2023. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
- Olugbade, O. In search of a global governance mechanism for Artificial Intelligence (AI): a collective action perspective. *Glob. Public Policy Gov.* (2025). https://doi.org/10.1007/s43508-025-00113-z
- Saetra HS. AI for the sustainable development goals. Bristol: Bristol University Press; 2022.
- Sepasspour, R. (2023). A reality check and a way forward for the global governance of artificial intelligence. *Bulletin of the Atomic Scientists*, 79(5), 304–315. https://doi.org/10.1080/00963402.2023.2245249
- R.E. al Shelby, Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction, in: Proc. 2023 AAAIACM Conf. AI Ethics Soc., 2023: pp. 723–741. https://arxiv.org/abs/2210.05791.
- Stahl, B., Eke, D. 2024. " The ethics of ChatGPT – Exploring the ethical issues of an emerging technology", International Journal of Information Management, Volume 74, 2024 https://doi.org/10.1016/j.ijinfomgt.2023.102700.
- Robles, Pedro and Daniel J. Mallinson. 2023. " Catching up with AI: Pushing Toward a Cohesive Governance Framework." *Politics & Policy* 51 (3): 355–372. **https://doi.org/10.1111/polp.12529.**

Korea and AI: Bridging decoloniality and development cooperation

KIM Suweon
Hankuk University of Foreign Studies

2025 International Conference
The Korean Social Science Research Council
May 27 (Tuesday) – May 28 (Wednesday), 2025





Sir Leander Starr Jameson, 1st Bt

by George Charles Beresford
platinum print, 8 October 1913
NPG x18842

What Nigerian hip-hop lyrics have to say about the country's Yahoo Boys

Published: April 9, 2019 2.25pm BST

Nigerian artist D'banj's 2004 song Mobolowowon had the first cyber fraud theme in hip-hop music. Shutterstock

Aahaaaaaa, we thank God ooooooo/ We made it o, bac k in those days/ Oyibo play us, dem turn us to slave/ D em useless us, but now we don wise/ We don dey rise, we no wan hide again/…

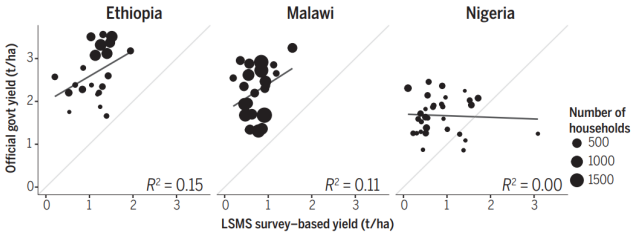
"Aahaaaaaa, we thank God oooooo!/ We made it (at last) o, back in those days/ **White people (colonialists) abused us, they enslaved us/ They maltreated us, but we are now smarter**/ We have transcended, We don't want to hide again/…"

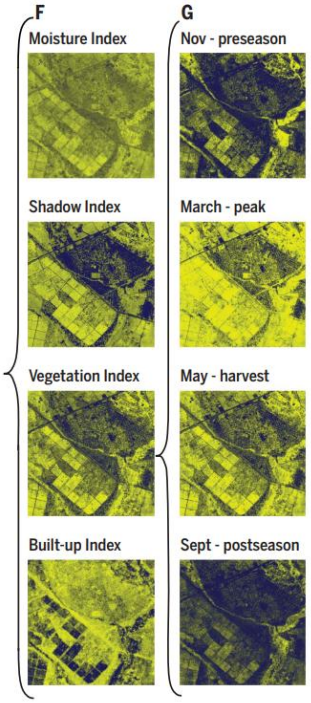
Lazarus, Suleman, Olatunji Olaigbe, Ayo Adeduntan, Edward T. Dibiana, and Geoffrey U. Okolorie. "Cheques or dating scams? Online fraud themes in hip-hop songs across popular music apps." *Journal of Economic Criminology* 2 (2023): 100033.

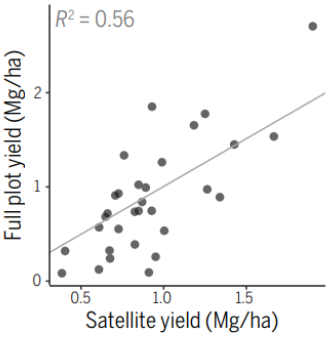
the White Men's Burden

An 1899 cartoon showing John Bull and Uncle Sam, symbolising Britain and the United States of America, bearing "The White Man's Burden" (borrowed from the title of a poem by Rudyard Kipling) by taking Asians and Africans to civilisation
'Judge' magazine

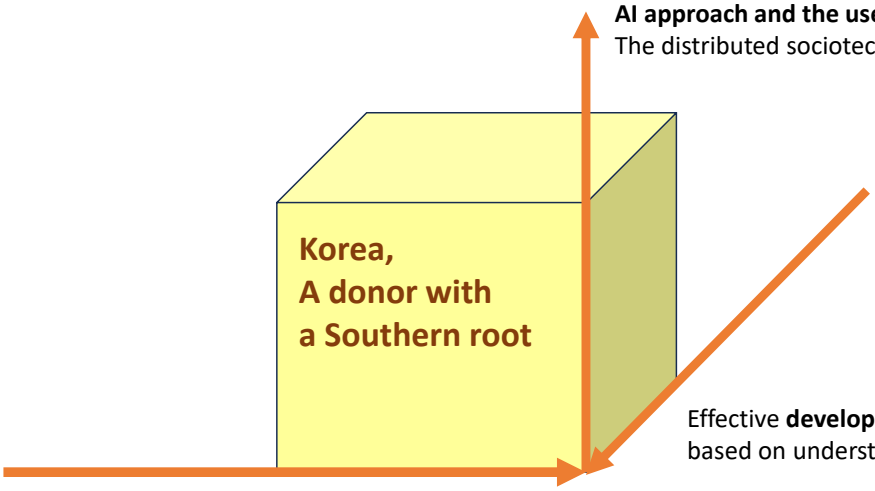
Poor numbers and Satellite images



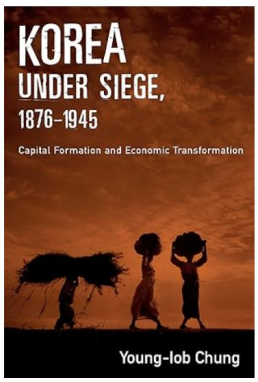


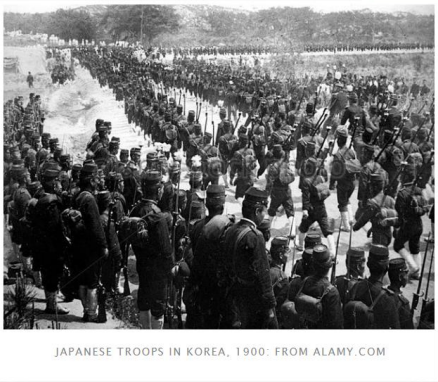


Burke, Marshall, Anne Driscoll, David B. Lobell, and Stefano Ermon. "Using satellite imagery to understand and promote sustainable development." Science 371, no. 6535 (2021): eabe8628.



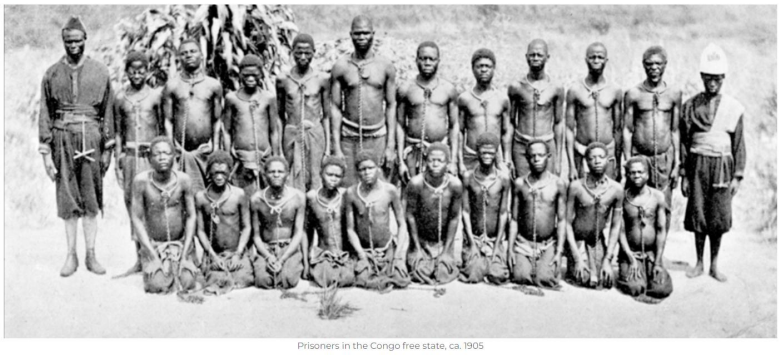
Decoloniality
Horizontal relationship in postcolonial study
Pluriverse




JAPANESE TROOPS IN KOREA, 1900: FROM ALAMY.COM

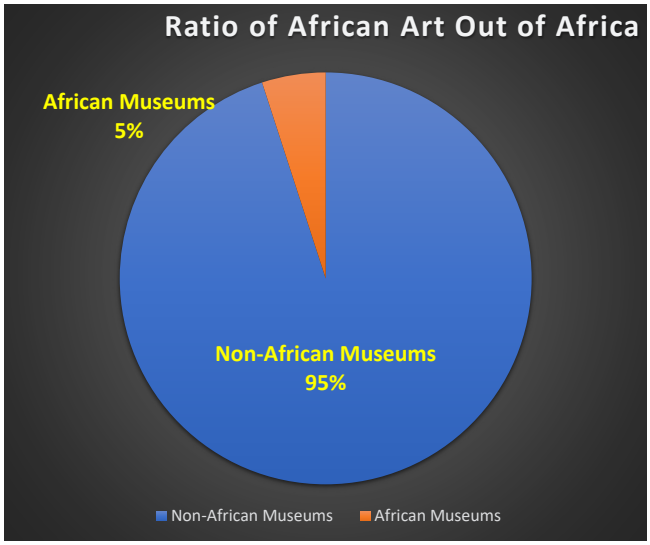
CONGO FREE STATE (1885-1908)

JANUARY 20, 2023 / CONTRIBUTED BY: SAMUEL MOMODU


Prisoners in the Congo free state, ca. 1905
From King Leopold's rule in Africa by E.D. Morel, courtesy The New York Public Library Digital Collections (b11735776)

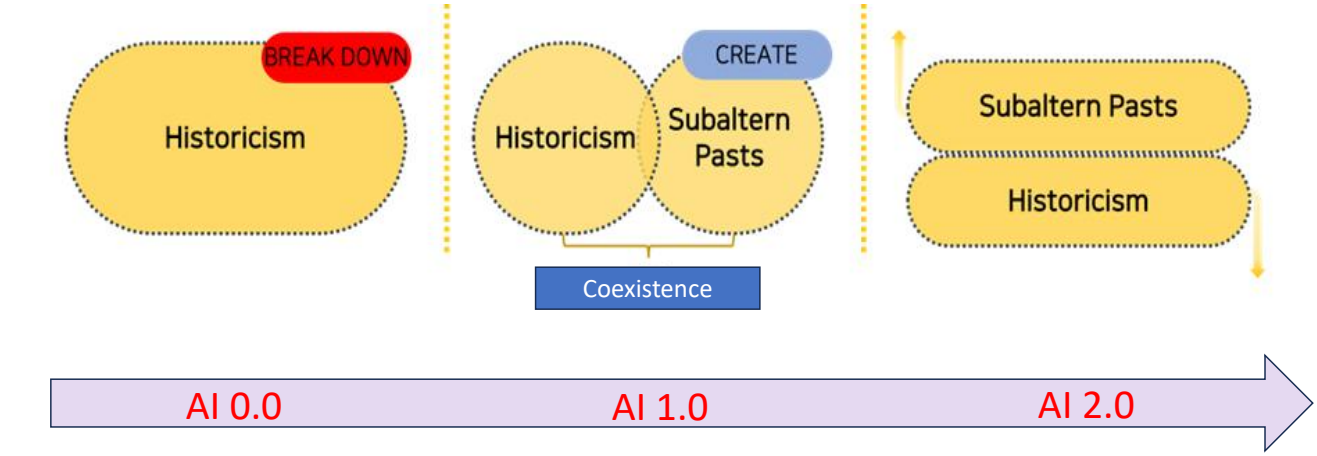
"Japan is far safer to keep Joseon-related cultural properties in light of conservation of the art pieces.“

The Treaty of Basic Relations, 1965



The National Museum of Congo

3 steps for Provincialising Europe (Chakrabarty, 2000)



동아일보

오피니언 정치 경제 국제 사회 문화 연예 스포츠 헬스동아

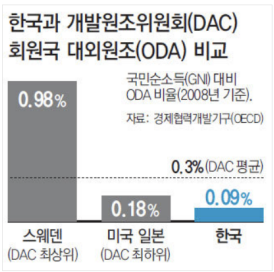
경제

한국 '진짜 선진국' 됐다... OECD개발원조委 가입

Array | 업데이트 2009-11-26 03:00

2 1

한국이 25일(현지 시간) 개발원조위원회(DAC·일명 공여국클럽)에 가입함으로써 '완전한 선진국'이 됐다. 프랑스 파리에 본부를 둔 경제협력개발기구(OECD) 산하 DAC의 23개 회원국은 이날 특별회의를 열어 한국의 DAC 가입을 결정했다. DAC 회원국은 사무국으로부터 한국 원조정책 현지실사 결과를 보고받고 한국 대표단과 질의응답을 한 뒤 한국 대표단을 퇴장시킨 가운데 비공개로 한국의 DAC 가입을 의결했다.



Kim, Suweon (2025) "Korea's Solidarity with the Global South (to Which It Didn't and Doesn't Belong)." *Pacific Focus* 40(1), pp. 125-144

The End

session 8

Sustainable Responsible AI Strategy: A Dynamic and Holistic View

Yong Sauk Hau

Professor, Department of Business Administration,
School of Business, Yeungnam University

Director, AI Management Strategy Research Center,
Business Research Institute, Yeungnam University

Chairman, AI Trustworthiness Research Association,
Korea Institute of Intelligent Information Systems (KIISS)

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Contents

Part I :
The Concept and Importance of Responsible AI

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part I : The Concept and Importance of Responsible AI

■ Serious problems have happened with AI Revolution

► There are many practical problems related to technology misuse and social dysfunction as well as technical limitations(1)

(1) AI-related Threat Factors

Threat Factors	Relevant Problems
Technical limitations	o (Bias) Amazon's job posting AI scrapped due to male-preference tendency ('18) o (Opaque) Unable to interpret AlphaGo's Go technique ('16) o (Malfunction) Google's photo recognition AI recognizes black people as gorillas ('15)
Abuse of technology	o (Algorithmic collusion) Uber, flexible pricing algorithm promotes implicit collusion ('16) o (Filter Bubble) Australian Fair Trade Commission, Google, Facebook, etc. Filter Bubble Impact Survey ('18) o (Deepfake) 8,000 deepfake videos ('18) → 14,698 ('19) o (Lethal weapon) Google suspends contract extension for the US Department of Defense AI Military Project ('18)
Social dysfunction	o (Polarization) Polarization of Wealth can be happen with AI Revolution ('17) o (Jobs) 67.9% of Korean workers can be replaced by AI ('20) o (Infringement of autonomy) Domino's Pizza installs AI surveillance cameras in 800 stores ('19)

(1) Source: Research on the Policy for AI Trustworthiness p. 3, (2022), Cho et al., Research Report, Software Policy & Research Institute (SPRi)

Part I : The Concept and Importance of Responsible AI

■ Various terms are being used to refer to responsible AI in different domains

► **Safe AI, ethical AI, trustworthy AI, reliable AI, dependable AI, etc.**

(1) Gartner's Definition on Responsible AI

"Responsible artificial intelligence (AI) is an umbrella term for aspects of making appropriate business and ethical choices when adopting AI. These include business and societal value, risk, trust, transparency, fairness, bias mitigation, explainability, sustainability, accountability, safety, privacy, and regulatory compliance. Responsible AI encompasses organizational responsibilities and practices that ensure positive, accountable, and ethical AI development and operation."

(1) Source: https://www.gartner.com/en/information-technology/glossary/responsible-ai

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part I : The Concept and Importance of Responsible AI

■ In 2019, Word Economic Forum developed "Empowering AI Leadership Toolkit", and it presented "Eight principles of AI Ethics"(1).

(1) Principles of AI Ethics in World Economic Forum

Principle	Definition	Related Concepts
Safety	o Prohibition of intentional or careless harm caused by artificial intelligence	o Physical safety of people, protection of property, restriction of use, restriction of access, restriction of military purposes, protection of employees, etc.
Privacy	o Protect personal and customer data and respect the preferences of data subjects, including their intent to control their data	o Data ownership, obligation to notify and consent in advance at the time of data collection and use, request for anonymization, prohibition of data sharing, management of data origin and history, responsibility for data protection measures, maintenance of the latest data protection technology, prohibition of re-identification of data, minimum collection of data, etc.
Equality	o Make fair decisions that protect human rights	o Prohibition of bias, protection of human rights, etc.
Well-being	o Using artificial intelligence to promote social development and welfare	o Good purpose, prohibition of filter bubbles, prohibition of threats to democracy, protection of the socially disadvantaged, democratization of artificial intelligence, etc.
Comprehension	o The reasons for AI's decisions and actions must be fully understood to control AI and enable human accountability.	o Obligation to explain, truthfulness, obligation to verify itself, obligation to verify third party, artificial intelligence education, provision of accurate information, etc.
Accountability	o Humans are responsible for AI's decisions and actions.	o Employee obligations, prohibition of final decisions by artificial intelligence, retention of human control, etc.
Remediation	o Workers, customers, and others affected by AI have a fair means to seek help or redress if AI threatens their livelihoods, reputations, or physical well-being.	o Impact assessment, compliance with law and order, stakeholder participation, duty to explain, worker protection, whistleblowing protection, etc.
Professionalism	o AI researchers, scientists, and technicians must follow high scientific and professional standards	o Compliance with research ethics, etc.

(1) Source: Research on the Policy for AI Trustworthiness p. 16, (2022), Cho et al., Research Report, Software Policy & Research Institute (SPRi)

Part I : The Concept and Importance of Responsible AI

■ In 2020, the Berkman Klein Center in Harvard University announced 8 principles for AI(1)

(1) Berkman Klein Center's 8 Principles for Artificial Intelligence

Principle	Definition	Related Concepts
Privacy	o Compliance with personal information protection regulations in the process of developing and using artificial intelligence	o Control of data use, encouragement of data laws, user consent, restriction of data processing, right to rectification, forgotten rights
Accountability	o Appropriate distribution of legal responsibility for the consequences of AI and provision of remedies	o Proposal of Artificial Intelligence Regulation, Legislative Impact Assessment, Audit Request, Verifiability, Right to Appeal, Establishment of a Monitoring Body
Safety & Security	o Artificial intelligence is safe, works as intended, and is not misused	o Reliable, predictable, robust, robust, accurate, resilient
Transparency & Explainability	o Design and operate the system so that humans understand how it works and why it makes decisions	o Advance notice when interacting with open source data/algorithms, artificial intelligence, regular reporting, and explanation
Fairness & Non-discrimination	o Design and operate so that it does not exclude certain groups or cause unfair results	o No Data Biasing, Data Representativeness, Inclusive Design, Inclusive Impact, Equity
Human Control of Tech	o Important decisions are reviewed by humans and humans control the technology.	o Human in the Loop for automated decision-making, opt-out option for automated decision-making
Professional Responsibility	o Require appropriate expertise and integrity from development and deployment personnel	o Stakeholder collaboration, responsible design, long-term effect consideration, scientific integrity
Promotion of Human Value	o The purpose and means of artificial intelligence are consistent with core values and promote human welfare	o Pursuit of social benefits, human values and prosperity, and access to technology

(1) Source: Research on the Policy for AI Trustworthiness p. 14~15, (2022), Cho et al., Research Report, Software Policy & Research Institute (SPRi)

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part I : The Concept and Importance of Responsible AI

■ In 2020, Gartner presented five AI guideline resulting from analyzing various AI principles(1)

(1) Gartner's Guideline for Artificial Intelligence

Principle	Definition	Related Concepts
Human-Centric and Socially Beneficial	o The ultimate purpose of using artificial intelligence is to achieve human purposes.	o Human autonomy and control, augmentation of human capabilities, solving social problems, and lawfulness
Fair	o legitimate, honest, impartial	o Treat everyone equally with consideration of the situation, eliminate undesirable biases, secretly manipulate human behavior, do not distort it.
Explainable and Transparent	o Reveal the use of artificial intelligence and explain the judgment	o Reveal that artificial intelligence is used, explain artificial intelligence decision-making (XAI), require documentation, etc.
Secure and Safe	o Security of data use, safe operation, maintenance of legitimacy, etc.	o Respect for privacy, monitoring of the learning process, proportional use of data and technology, harmlessness, etc.
Accountable	o Developers and Human Responsibility	o Developer responsibility, human control, artificial intelligence governance construction, etc.

(1) Source: Research on the Policy for AI Trustworthiness p. 17, (2022), Cho et al., Research Report, Software Policy & Research Institute (SPRi)

Part I : The Concept and Importance of Responsible AI

■ In December 2020, the Korean government announced the 'Ethical Standards for Artificial Intelligence' and launched a policy to promote responsible AI(1)

► In May 2021, it announced ‘Trustworthy Artificial Intelligence Realization Strategy’(1)

(1) Essential Elements of Trustworthy AI Realization Strategy

Key elements	Key Implications
Safety	o A state in which system operation and function performance due to the judgment and prediction results of artificial intelligence can be prevented from adversely affecting people and the environment.
Explainability	o The basis for AI's judgment and prediction and the process leading to the result are presented in a way that can be understood by humans, or the process of deriving the result can be analyzed in the event of a problem.
Transparency	o The operation process such as judgment and prediction of artificial intelligence and the components to implement it are recognized, checked, and inspected by the user.
Robustness	o Artificial intelligence maintains the user's intended level of performance and functionality even in external interference and extreme operating environments.
Fairness	o Ability to ensure that AI does not draw conclusions that include discrimination or bias against any particular group in the process of processing data.

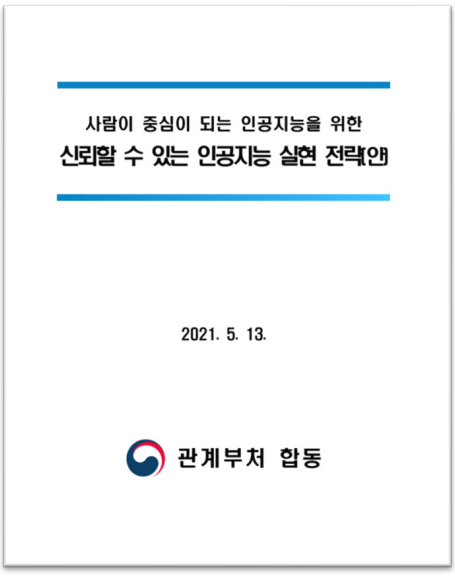
(1) Source: Research on the Policy for AI Trustworthiness p. 13, (2022), Cho et al., Research Report, Software Policy & Research Institute (SPRi)

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part I : The Concept and Importance of Responsible AI

■ From 2021, the Korean government's strategy to secure AI Trustworthiness and the promotion of support policies came into full swing

(1) Trustworthy AI Strategy



(1) Source:https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=112&pageIndex=&bbsSeqNo=94&nttSeqNo=3180239&searchOpt=ALL&searchTxt=

(2) AI Trustworthiness & Quality Award Contest

(2) Source: https://www.msit.go.kr/bbs/view.do;jsessionid=K8NC4xZEtrJ3rUxZiZkcmuGJThefDPgjG5RCIltE.AP_msit_1?sCode=user&mPid=218&mId=122&bbsSeqNo=96&nttSeqNo=3179898

Part II :
Research Trend in Responsible AI

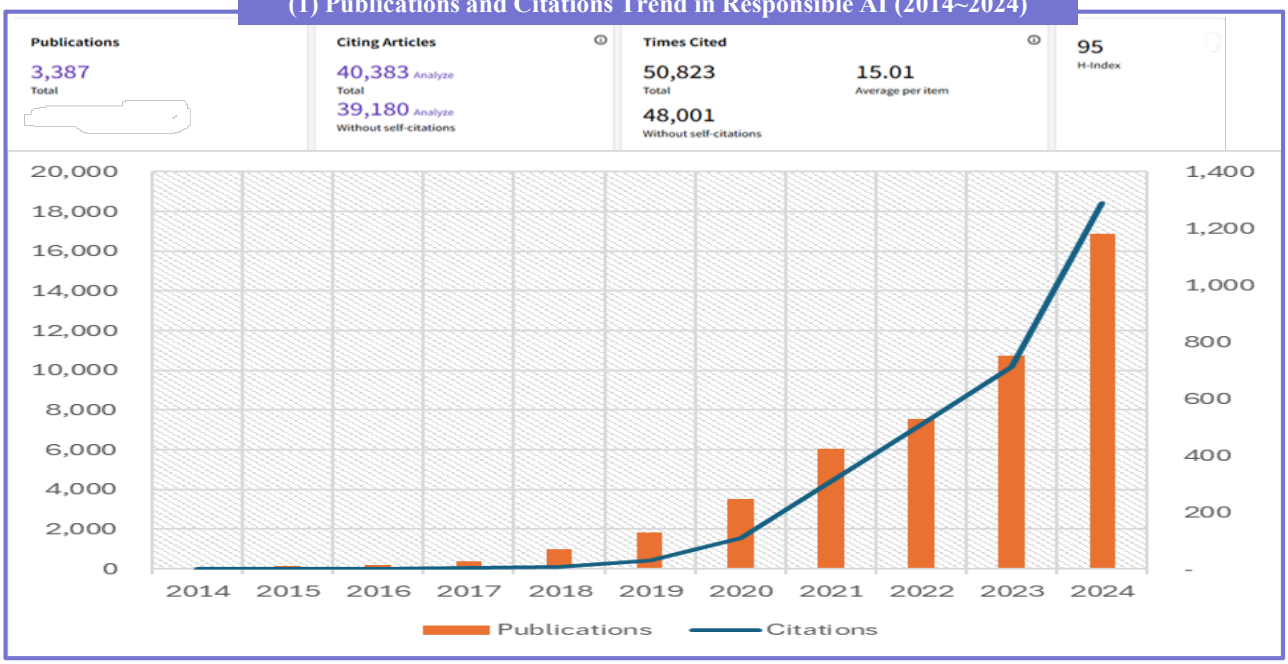
The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part II : Research Trend in Responsible AI (Global)

■ There are drastically growing trends in both publications and citations

► **Publications: 5 (2014) → 1,182 (2024); Citations 0 (2014) → 18,413(2024)**

(1) Publications and Citations Trend in Responsible AI (2014~2024)



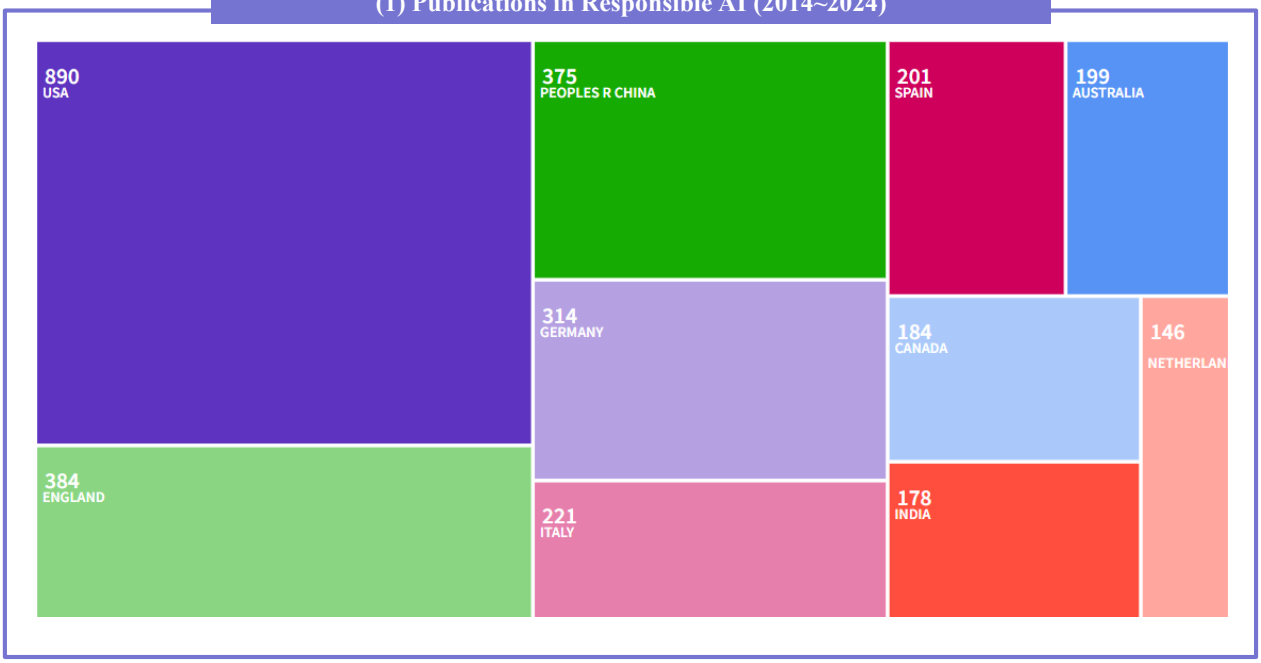
(1) Web of Science Analysis Results: https://www.webofscience.com/wos/woscc/citation-report/22e019ce-7240-4ee1-8bcc-626b027824f3-016457b695

Part II : Research Trend in Responsible AI (Global)

■ USA, England and China are leading research on responsible AI

► South Korea has been ranked as the 11st (the number of articles = 124, portion = 3.661%)

(1) Publications in Responsible AI (2014~2024)



(1) Web of Science Analysis Results: https://www.webofscience.com/wos/woscc/citation-report/22e019ce-7240-4ee1-8bcc-626b027824f3-016457b695

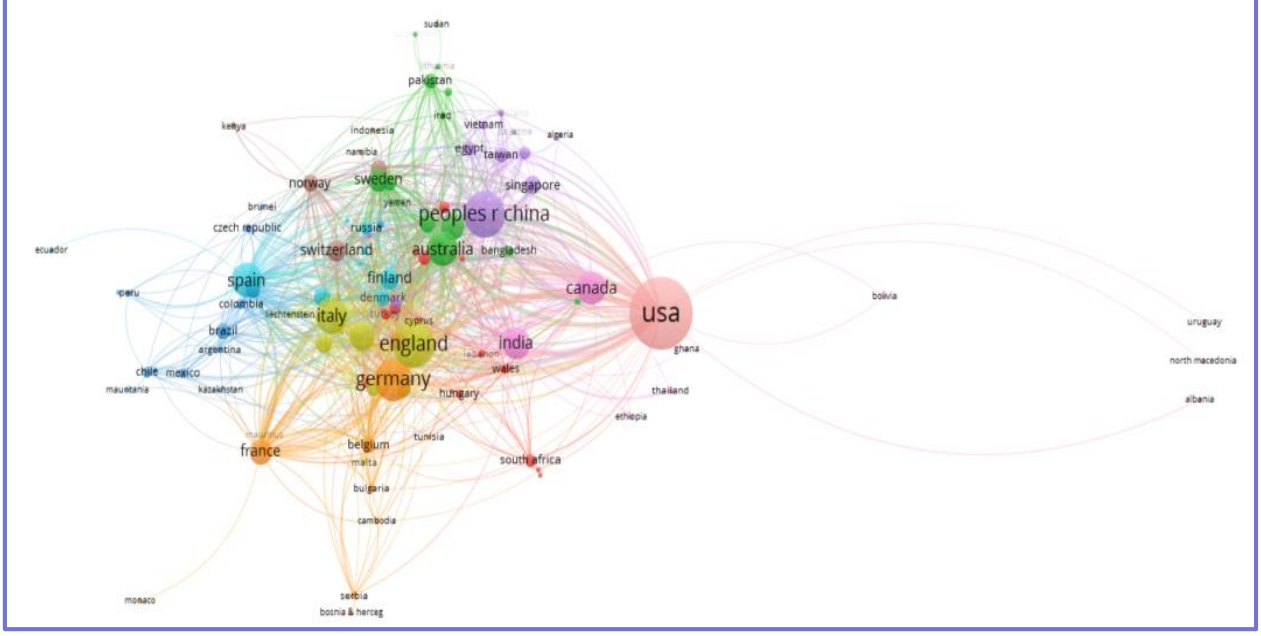
The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part II : Research Trend in Responsible AI (Global)

■ USA is the biggest hub in the collaboration network for research on responsible AI

► USA's Total link strength = 642 vs. South Korea's total link strength = 131

(1) Collaboration Network Concerning Research on Responsible AI (2014~2024)



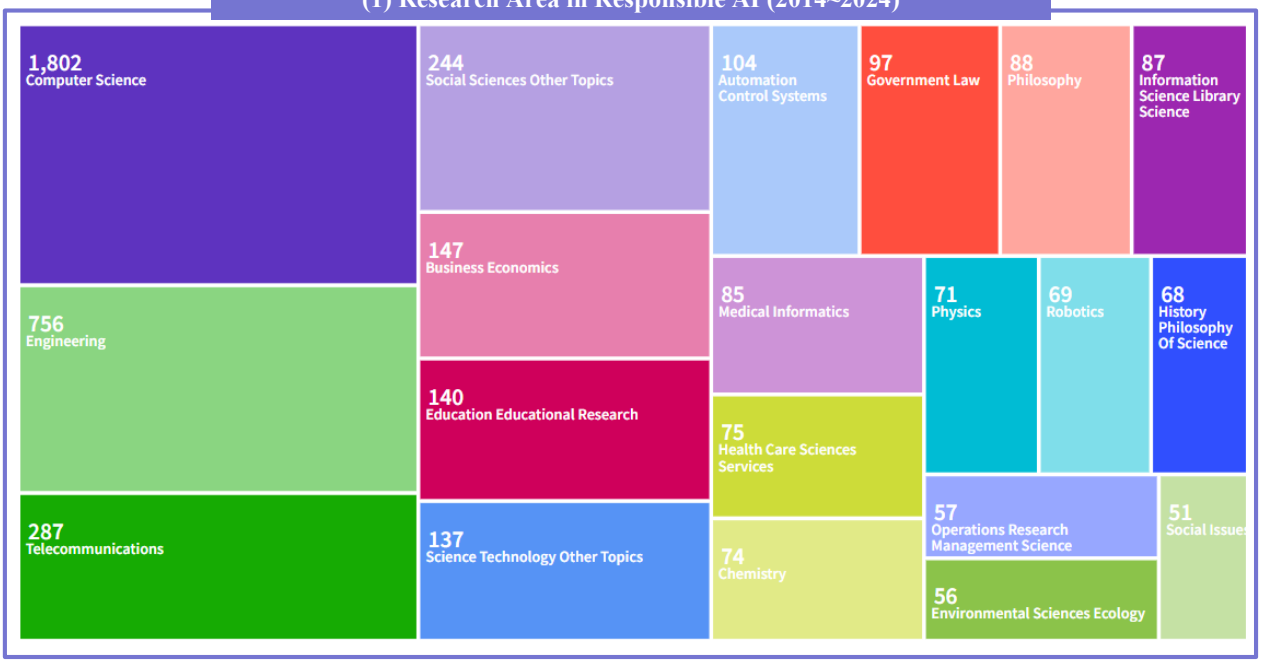
(1) Vosviewer Analysis Results: https://www.vosviewer.com/

Part II : Research Trend in Responsible AI (Global)

■ Research areas are not confined to Technology or Engineering domain

► Including Business Economics, Education, History Philosophy, etc.

(1) Research Area in Responsible AI (2014~2024)



(1) Web of Science Analysis Results: https://www.webofscience.com/wos/woscc/citation-report/22e019ce-7240-4ee1-8bcc-626b027824f3-016457b695

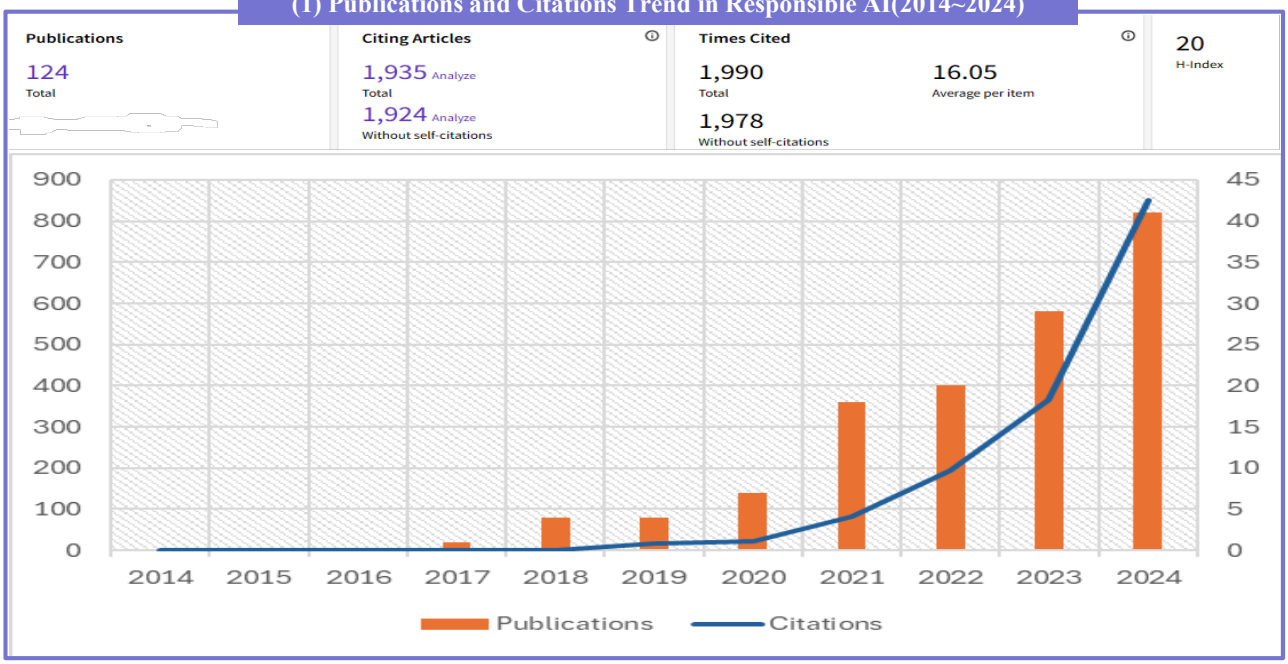
The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part II : Research Trend in Responsible AI (South Korea)

■ There are rapidly growing trends in both publications and citations

► Publications: 0 (2014) → 41 (2024); Citations 0 (2014) → 851(2024)

(1) Publications and Citations Trend in Responsible AI(2014~2024)



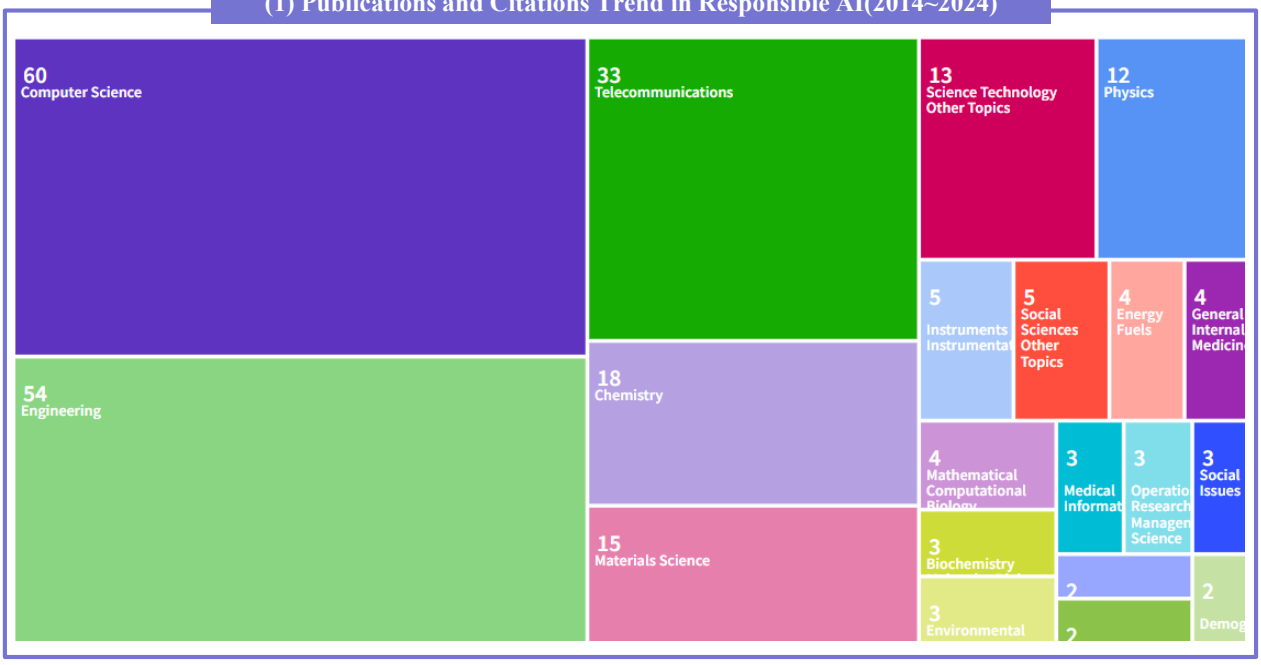
(1) Web of Science Analysis Results: https://www.webofscience.com/wos/woscc/citation-report/22e019ce-7240-4ee1-8bcc-626b027824f3-016457b695

Part II : Research Trend in Responsible AI (South Korea)

■ Computer Science, Engineering, and Telecommunications are the top 3 areas

► Including Telecommunication, Chemistry, and Material Science, etc.

(1) Publications and Citations Trend in Responsible AI(2014~2024)



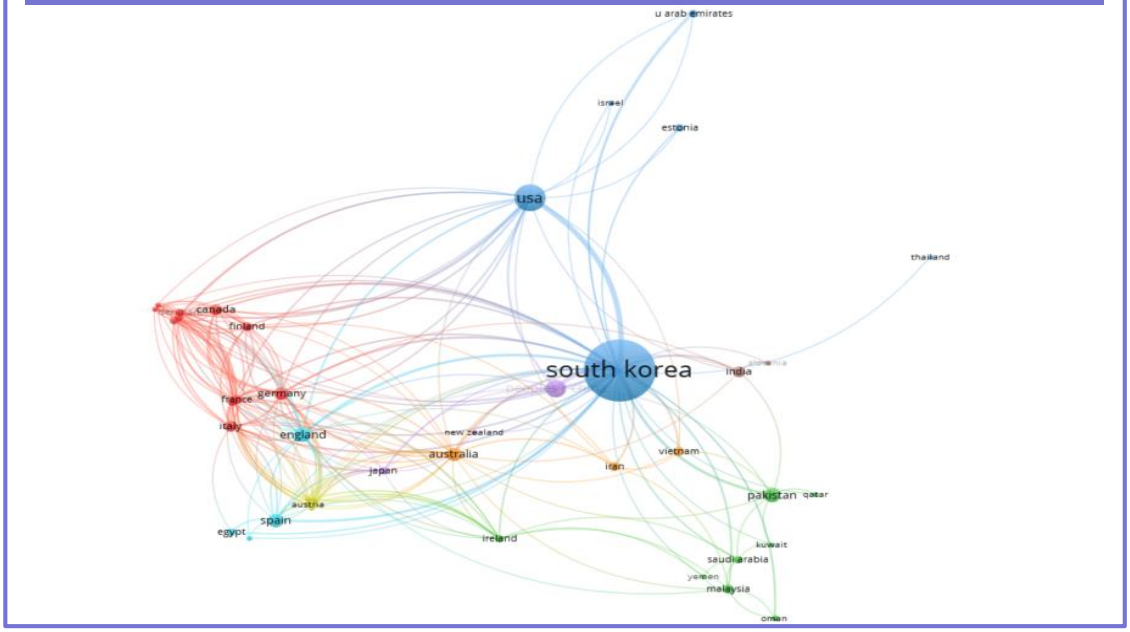
(1) Web of Science Analysis Results: https://www.webofscience.com/wos/woscc/citation-report/22e019ce-7240-4ee1-8bcc-626b027824f3-016457b695

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part II : Research Trend in Responsible AI (South Korea)

■ South Korea formed a collaboration network cluster related to research on responsible AI with several countries with USA, Israel, Estonia, Luxembourg, Thailand

(1) South Korea's Collaboration Network Concerning Research on Responsible AI (2014~2024)



(1) Vosviewer Analysis Results: https://www.vosviewer.com/

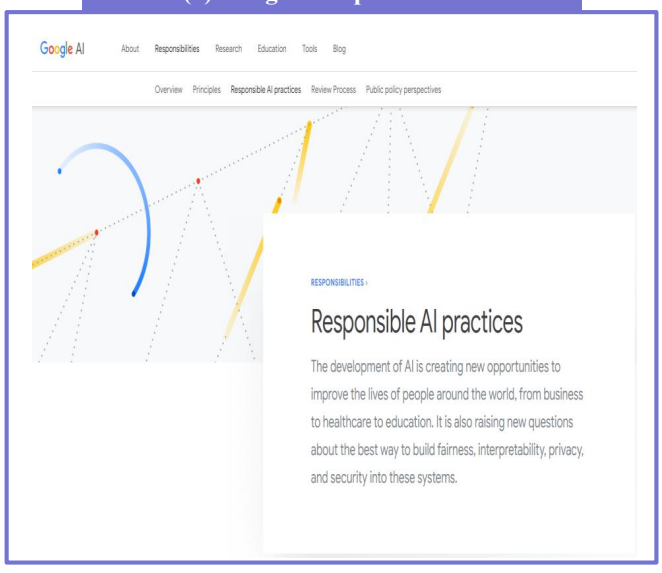
Part III : Responsible AI Strategy under a Dynamic and Holistic View

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

Part III : Responsible AI Strategy under a Dynamic and Holistic View

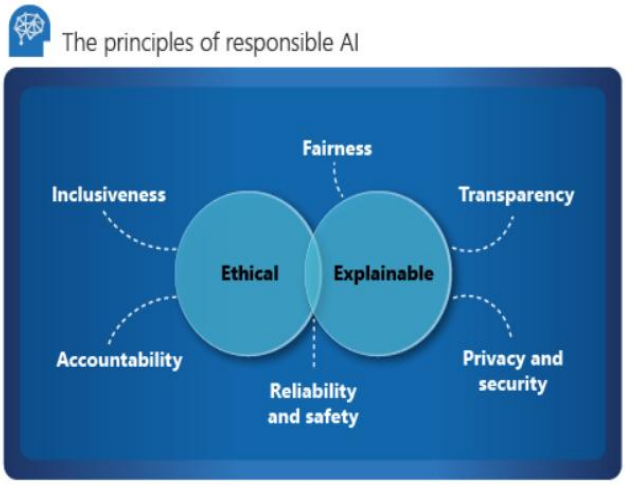
■ **Global leading IT companies including Google and Microsoft have been implementing not only AI development but also responsible AI strategy**

(1) Google's responsible AI



(1) Source: https://ai.google/responsibilities/responsible-ai-practices/

(2) Microsoft's responsible AI



(2) Source: https://docs.microsoft.com/ko-kr/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai?toc=https%3A%2F%2Fdocs.microsoft.com%2Fko-kr%2Fazure%2Farchitecture%2Ftoc.json&bc=https%3A%2F%2Fdocs.microsoft.com%2Fko-kr%2Fazure%2Farchitecture%2Fbread%2Ftoc.json

Part III : Responsible AI Strategy under a Dynamic and Holistic View

■ IBM, Google and Microsoft have been presenting responsible AI technology tools to increase privacy protection, robustness, fairness, and transparency(1)

(1) Responsible AI Technology Tools

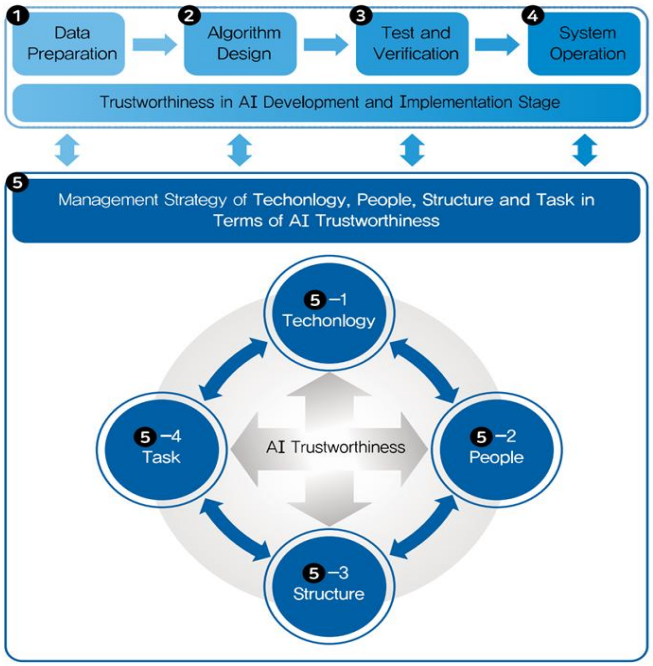
Responsible AI Components	Responsible AI Technology Tools	Homepage
Privacy Protection	o (IBM) AI Privacy 360	https://aip360.mybluemix.net/
	o (Google) Tensorflow Privacy	https://github.com/tensorflow/privacy
Robustness	o (IBM) AI Adversarial Robustness 360	https://art360.mybluemix.net/
Fairness	o (Google) What-If Tool	https://pair-code.github.io/what-if-tool/
	o (Microsoft) Fair Learn	https://fairlearn.org/
	o (IBM) AI Fairness 360	https://aif360.mybluemix.net/
Transparency	o (Google Cloud) Explainable AI	https://cloud.google.com/explainable-ai
	o (Microsoft) InterpretML	https://interpret.ml/
	o (IBM) AI Explainability 360	https://aix360.mybluemix.net/

(1) Source: Research on the Policy for AI Trustworthiness p. 97, (2022), Cho et al., Research Report, Software Policy & Research Institute (SPRi)

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

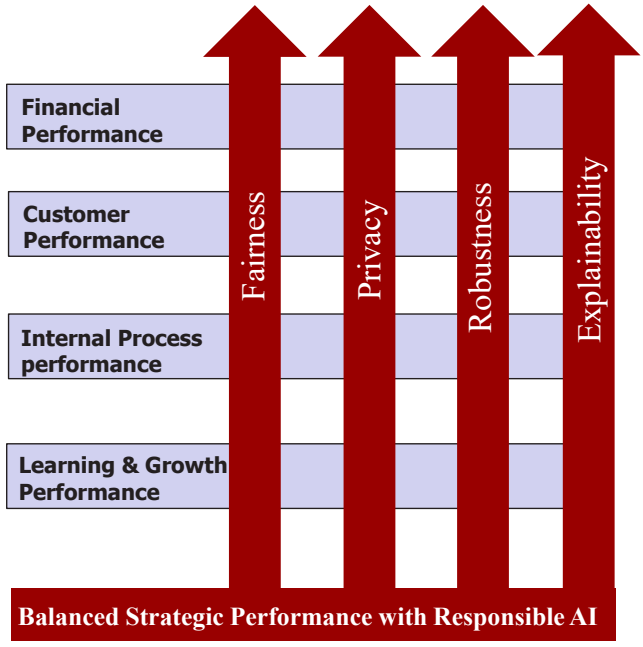
Part III : Responsible AI Strategy under a Dynamic and Holistic View

(1) Responsible AI Strategy Framework



(1) Source: *Understanding and Practicing AI Trustworthiness*, Hau et al., 2024, p. 11, Chungram Publishing Company

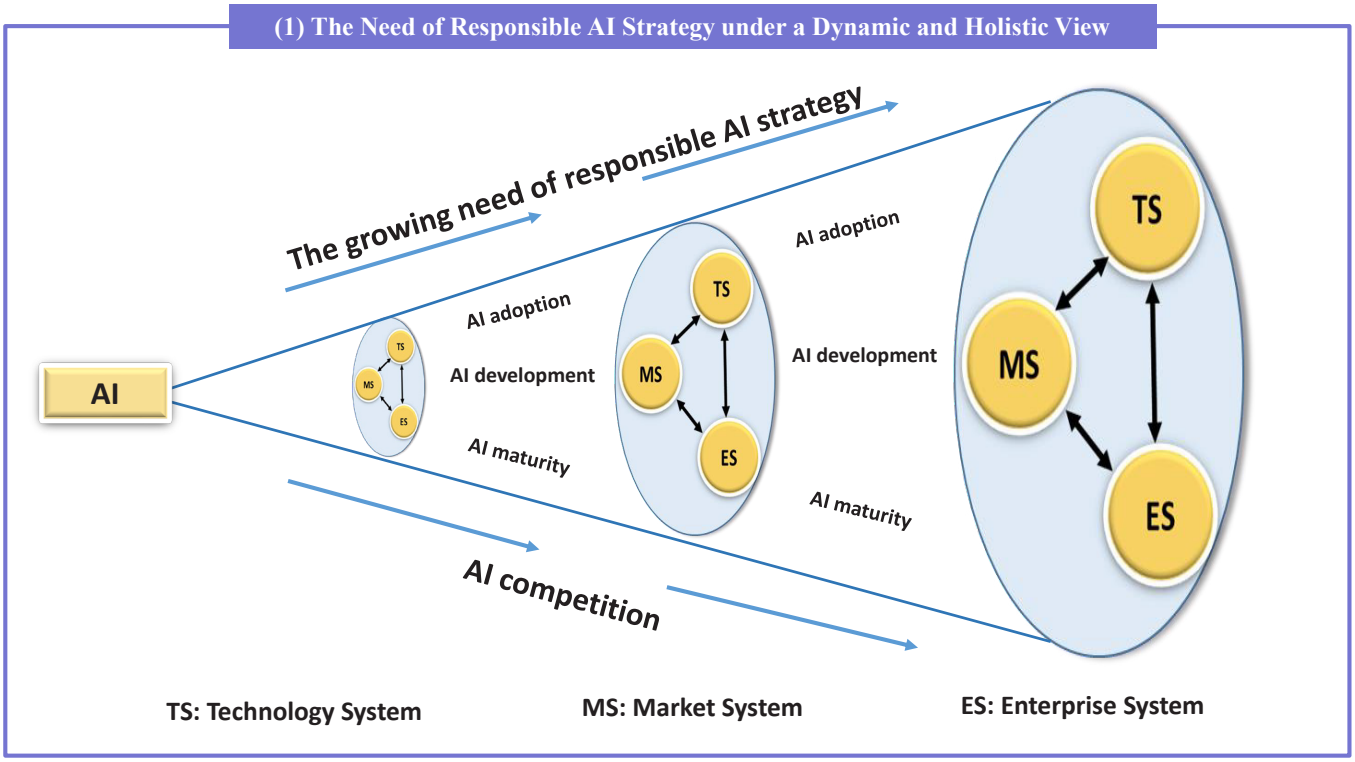
(2) Goals of Responsible AI Strategy



(2) Source: : Adapted from Kaplan, R. S. & Norton D. P., (2001), *The Strategy-Focused Organization*, Harvard Business School Press, p. 79. Ammanath, B., (2022); *Trustworthy AI: A Business Guide for Navigating Trust and Ethics in AI*, John Wiley & Sons.

Part III : Responsible AI Strategy under a Dynamic and Holistic View

(1) The Need of Responsible AI Strategy under a Dynamic and Holistic View



(1) Source: Adapted from *The Theories and Applications of Technology Innovation Management Strategy*, Y. S. Hau, 2022, p. 50, Yeungnam University Press

The 2025 International Conference of the Korean Social Science Research Council | The Korean Academic Society of Business Administration

"The toughest thing about the power of trust is that it's very difficult to build and very easy to destroy"(1).

Thomas J. Watson, Sr., ex-CEO of IBM

(1) Source: Varshney, K. R., *Trustworthy Machine Learning*, (2022), p.4, independently published.

Thank you very much !!!

AI-Native Enterprises and the AX Revolution

A New Era of Business Strategy

Trust Connector Seokbin Yoon
Collaboration Professor, Sogang University

Introduction to AI-Native Enterprises

What are AI-Native Enterprises?

Organizations fundamentally built around artificial intelligence as their core operational foundation, where AI is not merely a tool but the central nervous system driving all business functions.

Key Characteristics

AI-First Architecture

Built from the ground up with AI at the core rather than as an addition. All systems, processes, and infrastructure are designed to support AI operations.

Data Ecosystem

Advanced data infrastructure that enables continuous learning, with real-time data collection, processing, and analysis as the foundation of operations.

Algorithmic Decision-Making

Critical business decisions are augmented or made by algorithms, with human oversight focused on exception handling and strategic direction.

Continuous Evolution

Self-improving systems that constantly adapt to new data, market conditions, and customer behaviors without requiring manual intervention.

"AI-Native Enterprises don't just use AI—they are AI."

The Concept of AX Revolution

Understanding Artificial Experience (AX)

Artificial Experience (AX) refers to the intelligent, personalized interactions between humans and AI systems designed to enhance, augment, or replace traditional experiences.

Unlike UX (User Experience), AX creates **dynamic, adaptive experiences** that evolve in real-time based on context, user behavior, and continuous learning.

Transforming Business Interactions

From Static to Dynamic

Business processes evolve in real-time rather than following pre-determined paths

Hyper-Personalization

Each customer interaction uniquely tailored beyond traditional segmentation

Autonomous Operations

Self-managing systems that optimize without constant human intervention

AX Revolution

Transforming business through AI-driven experiences

> *"The AX Revolution represents a fundamental shift from **humans adapting to technology** to **technology adapting to humans**."*

Transformation of Business Models

AI-Native Revolution in Business

AI-Native enterprises fundamentally reimagine business models by shifting from **linear value chains** to **dynamic value networks** - creating exponential growth opportunities through intelligent systems.

New Revenue Streams

- Data monetization & insights-as-a-service
- Dynamic pricing & personalized offerings
- AI-enabled product extensions
- Subscription-based intelligence platforms

Operational Efficiency

- Predictive resource optimization
- Autonomous decision-making systems
- Self-healing infrastructure
- Intelligent workflow automation

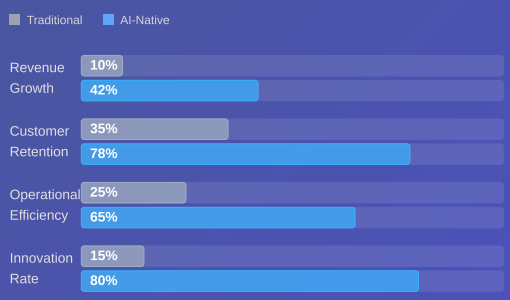
Customer Engagement

- Hyper-personalized experiences
- Anticipatory service models
- Emotion-aware interactions
- Continuous relationship intelligence

Traditional vs. AI-Native Business Model

Traditional	AI-Native
Static product offerings	Adaptive solutions that evolve
Periodic innovation cycles	Continuous intelligence-driven innovation
Siloed data utilization	Unified data ecosystem with insights
Linear scaling with resources	Exponential scaling through intelligence

Performance Comparison (%)



> *"AI-Native enterprises don't just digitize existing business models – they create entirely new value propositions that were previously impossible."*

Key Technologies Enabling AI-Native Enterprises

Foundation of the AI-Native Revolution

The technological backbone powering AI-Native enterprises combines mature AI foundations with emerging capabilities to create unprecedented business value through intelligent systems.

Technology	Category	Description	Use Cases
Machine Learning	Core Technology	Algorithms that enable systems to automatically learn and improve from experience	• Predictive analytics & forecasting • Anomaly detection systems • Pattern recognition • Deep learning architectures
Natural Language Processing	Language AI	Enables machines to understand, interpret, and generate human language	• Conversational AI interfaces • Sentiment analysis • Document understanding • Multilingual capabilities
Computer Vision	Visual AI	Systems that can identify, process and interpret visual data from the world	• Image recognition & classification • Object detection & tracking • Visual inspection systems • Augmented reality integration
Reinforcement Learning	Advanced AI	Learning through trial-and-error interactions with dynamic environments	• Autonomous decision systems • Resource optimization • Adaptive control systems • Self-improving algorithms
Generative AI	Frontier AI	AI systems that can create new content, designs, and solutions	• Content creation & design • Synthetic data generation • Product/service innovation • Multi-modal generation
Federated Learning	Privacy-Preserving	Training AI models across distributed devices while keeping data local	• Privacy-centric AI development • Edge AI capabilities • Decentralized intelligence • Cross-organizational learning

Technology Integration Matrix

Operational Integration • Process automation • Decision intelligence • Supply chain optimization	**Customer Experience** • Personalization engines • Intelligent assistants • Sentiment-aware services
Product Innovation • Self-improving products • AI co-creation • Adaptive feature sets	**Data Infrastructure** • Real-time analytics • Knowledge graphs • MLOps & model governance

AI Technology Adoption Trends (%)

Technology	Current (%)	Expected 2025 (%)	Current Adoption
ML/Deep Learning	85%	95%	
NLP	75%	90%	
Computer Vision	65%	80%	
Generative AI	70%	90%	
Reinforcement Learning	40%	70%	
Federated Learning	25%	60%	

"The competitive advantage of AI-Native enterprises lies not in any single technology, but in their ability to orchestrate these capabilities into cohesive, intelligent systems."

Case Studies of Successful Implementations

AI-Native Leaders Transforming Industries

Organizations leveraging AI as their core competitive advantage are achieving unprecedented levels of efficiency, customer satisfaction, and market growth across diverse sectors.

Tesla

Automotive

Autonomous Driving AI Ecosystem

Building AI-native vehicle systems with neural networks processing real-time sensor data for autonomous driving capabilities.

Key Implementation Components:

- Neural networks analyzing camera, radar, and ultrasonic data
- Fleet learning systems with 3B+ miles of real-world driving data
- Over-the-air updates delivering continuous AI improvements

98%	10x	30%
Crash probability reduction	Data processing efficiency	Increased range prediction

Netflix

Entertainment

Personalization & Content Recommendation

Using AI across the entire business from content recommendation to production decisions and quality optimization.

Key Implementation Components:

- Machine learning for personalized recommendations
- Content performance prediction algorithms
- AI-enhanced video encoding optimization

80%	$1B	15%
Content discovered through AI	Annual retention value	Streaming bandwidth reduction

Ocado

Retail & Logistics

Automated Fulfillment Technology

Transforming grocery retail with AI-powered fulfillment centers featuring thousands of robots coordinated by AI systems.

Key Implementation Components:

- Machine learning for demand forecasting
- Swarm robotics with real-time coordination
- Computer vision for quality control

3,000+	99.9%	50%
Robots per warehouse	Order accuracy rate	Labor cost reduction

Ping An

Financial Services

End-to-End AI Financial Services

Integrating AI across insurance, banking, and healthcare with over 35,000 AI models processing 85% of customer interactions.

Key Implementation Components:

- Facial recognition for contactless services
- AI-powered risk assessment algorithms
- NLP for customer service automation

60%	30%	95%
Cost reduction in claims	Improved fraud detection	Loan approval automation

Business Impact Metrics of AI-Native Implementations

Company	Revenue Growth (%)	Operational Efficiency (%)	Customer Experience Score
Tesla	42	28	86
Netflix	35	25	92
Ocado	28	50	82
Ping An	22	60	78

Key Success Factors

Data Strategy
Treating data as strategic asset, not byproduct

Continuous Learning
Systems that improve with every interaction

Human+AI Integration
Optimizing the division of labor between humans and AI

Bold Reinvention
Rethinking core business processes from scratch

Challenges and Ethical Considerations

Navigating the Complexities of AI-Native Implementation

As AI-Native enterprises scale, they face critical challenges balancing innovation with responsibility, requiring frameworks for ethical deployment.

Technical & Business Challenges

Data Quality & Access

- Insufficient high-quality training data
- Data silos prevent comprehensive insights
- Real-time data integration bottlenecks

Integration & Infrastructure

- Legacy system incompatibility
- Compute resource limitations
- AI model deployment complexity

Talent & Organizational

- AI skill shortages across all levels
- Change resistance in traditional teams
- Cross-functional collaboration barriers

Ethical Considerations

Bias & Fairness

- Algorithmic bias perpetuating inequalities
- Unequal impact across demographic groups
- Representation imbalances in training data

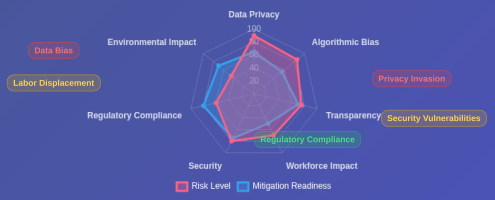
Privacy & Security

- Personal data collection concerns
- Vulnerability to adversarial attacks
- Data sovereignty across global operations

Labor & Social Impact

- Workforce displacement concerns
- Skills transition inequality
- Algorithmic accountability gaps

AI Ethics & Risk Landscape



Mitigation Strategies: Ethical AI Governance | Diverse Development Teams | Explainable AI Models | Ongoing Human Oversight | Regular Ethical Audits

Future Trends and Predictions

The Next Decade of AI-Native Evolution

The AI landscape is rapidly evolving, with transformative technologies reshaping industries and creating new paradigms for enterprise operations.

2023 — Generative AI Integration
2025 — AI Autonomy & Self-Healing Systems
2028 — Quantum-Powered AI
2030 — Neurosymbolic AI & Human Augmentation
2033+ — Ambient Intelligence Ecosystems

Autonomous AI Systems

- Self-optimizing enterprise systems
- Auto-adaptive security postures
- Predictive operational resilience

Multimodal Intelligence

- Cross-domain knowledge synthesis
- Vision-language integrated analysis
- Emotional intelligence in AI systems

Edge AI Proliferation

- Ultra-low latency decision making
- Distributed intelligence networks
- Data sovereignty preservation

Human-AI Collaboration

- AI augmented workforce capabilities
- Cognitive assistance ecosystems
- Creative partnership paradigms

Trustworthy AI Framework

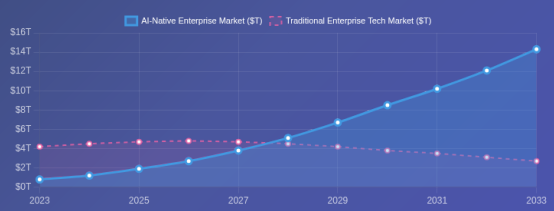
- Explainable decision intelligence
- Verifiable fairness guarantees
- Regulatory-aligned development

Quantum AI Integration

- Complex optimization at scale
- Next-gen material discovery
- Cryptographic paradigm shifts

2023 2025 2028 2030 2033

AI-Native Market Growth Projection



Technology Adoption Timeline

Key Strategic Insights

Generative AI
Will reshape 70% of enterprise content operations by 2025

Data Economics
Data marketplaces become central to AI ecosystem value chains

AI Governance
Automated compliance frameworks emerge as competitive advantage

Industry Convergence
Sector boundaries blur as AI enables cross-domain innovation

Implementation Strategies for Traditional Companies

Transformation Roadmap for AI-Native Evolution

Strategic approaches for established enterprises to successfully transition into AI-Native organizations

1 Assessment — Evaluate AI readiness, data maturity & capability gaps

2 Strategy & Vision — Define AI value propositions & strategic roadmap

3 Foundation Building — Establish data infrastructure & AI governance

4 Transformation — Execute pilots, scale solutions & transform operations

5 AI-Native State — Continuous innovation & AI-driven organization

Technology Stack

- **Data Infrastructure:** Implement unified data lakes with governance
- **AI Platforms:** Build or buy scalable ML operations platforms
- **Integration:** Develop API-first architecture for AI services
- **Security:** Implement AI-specific security frameworks

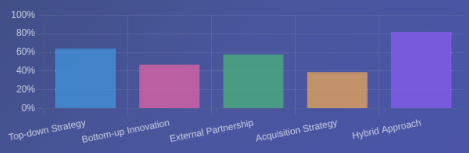
Organizational Design

- **Leadership:** Create Chief AI Officer role & AI steering committee
- **Team Structure:** Establish AI Centers of Excellence & embedded teams
- **Skills Development:** Implement AI upskilling programs for all employees
- **Change Management:** Design AI-focused cultural transformation initiatives

Process Transformation

- **Business Processes:** Redesign core workflows for AI augmentation
- **Decision Systems:** Implement AI-enhanced decision frameworks
- **Development:** Adopt MLOps/AIOps methodologies
- **Governance:** Establish AI ethics & monitoring processes

Success Factors by Implementation Approach



Implementation Best Practices

Start with High-Value Use Cases
Focus on business outcomes with measurable ROI to build momentum & executive support

Build for Scale from Day One
Design solutions & infrastructure that can scale across the enterprise, avoiding siloed implementations

Iterative Transformation
Use agile methodologies to deliver continuous value rather than "big bang" approaches

Co-Creation & Partnerships
Leverage ecosystem partners, vendors & academia to accelerate capabilities development

Common Pitfalls to Avoid

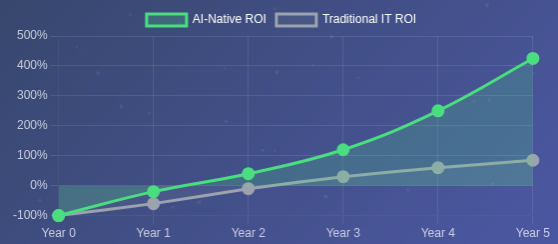
- Treating AI as a technology-only initiative
- Neglecting data quality & governance
- Insufficient executive sponsorship & vision
- Failing to address skills gaps & change management

Measuring Success and ROI in AI-Native Enterprises

Quantifying Value Creation in the AI Era

Frameworks, methodologies and metrics for evaluating the business impact of AI-Native initiatives

ROI Timeline: AI-Native vs Traditional Approaches



- AI-Native investments typically show compound returns after initial investment period
- Traditional IT investments often show linear or diminishing returns

AI-Native Measurement Framework

Financial Impact Metrics

Revenue Growth	+15-35%	Cost Reduction	20-45%
New Revenue Streams	2-5x	Time-to-Value	-40-60%

Operational Excellence Metrics

Process Cycle Time	-50-80%	Decision Quality	+30-60%
Resource Utilization	+25-45%	Error Reduction	-70-95%

Experience & Engagement Metrics

Net Promoter Score	+15-30pts	Customer Retention	+10-35%
Personalization Depth	5-10x	Engagement Time	+40-80%

Strategic Value Measurement

Intellectual Capital

- Algorithmic assets
- Data moats created
- Patent & IP generation

Market Position

- Market share velocity
- Customer acquisition cost
- Competitive displacement

Building AI-Ready Organizational Culture

The Cultural Foundation for AI-Native Success

Transforming mindsets, structures, and behaviors to thrive in the intelligence-driven economy

Visionary Leadership

Key Mindsets

- Data-driven decision making
- Continuous learning orientation
- Embracing intelligent automation

Leadership Shifts

From Intuition → To Intelligence

From Control → To Enablement

AI-Native Talent Strategy

Talent Acquisition

- T-shaped AI skill profiles
- Cross-functional backgrounds

Talent Development

Critical Skills Mix

- Data Literacy
- Systems Thinking
- Algorithmic Thinking

Adaptive Organization Design

Structure Evolution

Hierarchical → Matrix → Neural Networks

Key Structural Elements

Rapid

Decision Rights

Human-AI Decision Spectrum

Human Only — AI Autonomous

Optimal Balance:

Innovation Ecosystem

Fostering AI Innovation

- Protected experimentation time
- Rapid prototyping frameworks
- Innovation metrics &

Cultural Values

- Psychological safety
- Continuous iteration
- Question assumptions

Cultural Transformation Journey

Awareness — STAGE 1
Building understanding of AI potential and limitations

Activation — STAGE 2
Initial pilots and skills development programs

Adaptation — STAGE 3
Restructuring processes and organizational models

Acceleration — STAGE 4
Scaling solutions and building innovation flywheels

AI-Native — STAGE 5
Intelligence embedded across all aspects of the organization

Regulatory Considerations

Navigating the Evolving Regulatory Landscape

AI governance frameworks, regional compliance requirements, and emerging standards for AI-Native enterprises

Key Regulatory Areas

Data Privacy

- Consent requirements
- Data subject rights
- Cross-border data flows
- Breach notification

Algorithmic Transparency

- Explainable AI requirements
- Decision traceability
- Bias mitigation reporting
- Human oversight provisions

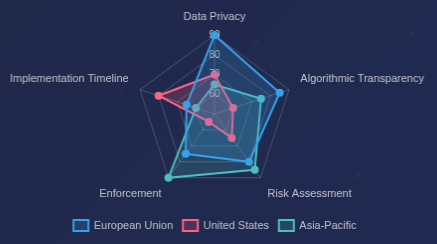
Industry-Specific

- Financial services (DORA)
- Healthcare (FDA AI/ML)
- Transportation (autonomous)
- Critical infrastructure

Intellectual Property

- AI-generated content rights
- Model copyright issues
- Training data licensing
- Patent challenges

Regional Regulatory Approaches



United States

NIST AI Framework · State-level Laws (CCPA) · FTC Enforcement

European Union

AI Act · GDPR · Digital Markets Act

Asia-Pacific

China's AI Ethics · Singapore AI Framework · Japan's AI Guidelines

Global Standards

ISO/IEC 42001 · IEEE 7000 Series · OECD AI Principles

Key Takeaway: Strategic Compliance

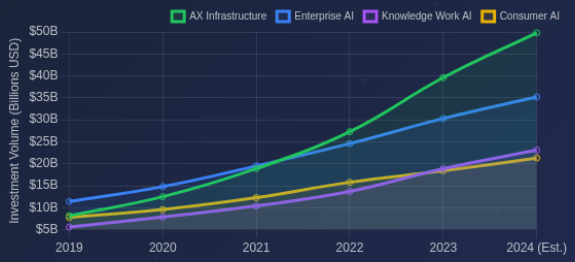
Successful AI-Native enterprises build regulatory compliance into their core business strategy rather than treating it as an afterthought. This "compliance by design" approach reduces risk, builds trust, and creates competitive advantages in highly regulated markets.

Investment Opportunities

The Financial Landscape of AI-Native Innovation

Funding trends, high-growth sectors, and strategic investment approaches in the AX Economy

Global Investment Trends (2019-2024)



↑ **+124%** AX Infrastructure

↑ **+86%** AI Enterprise Solutions

↑ **+67%** Knowledge Work AI

↑ **+53%** Consumer AI

High-Potential Investment Sectors

AI Infrastructure $49.8B

- Specialized AI hardware
- AI-optimized cloud services
- Edge computing solutions

5-Year CAGR: 38.2%

Enterprise AI $35.2B

- Decision intelligence platforms
- AI-powered ERP/CRM systems
- Predictive analytics solutions

5-Year CAGR: 31.7%

Vertical AI $28.6B

- Healthcare AI systems
- Financial services platforms
- Manufacturing optimization AI

5-Year CAGR: 29.4%

AX Platforms $21.3B

- Personalization engines
- Immersive AI experiences
- Ambient intelligence systems

5-Year CAGR: 42.8%

AI-Native Valuation Metrics

Data Asset Value
Quality, uniqueness, and scale of proprietary data assets that power AI systems

Network Effect Multiplier
Compounding value from each additional user improving AI performance

Algorithm Performance
Accuracy, efficiency, and uniqueness of proprietary AI algorithms

AI Talent Density
Concentration of ML/AI expertise in key roles and leadership positions

Funding Landscape by Stage

Seed	Series A	Series B	Series C	Late Stage
$2-5M	$8-15M	$20-40M	$50-100M	$100M+

↑ Deal size increased 3.2x in past 24 months

72% concentration in Series A & B rounds

Leadership Strategies for AI-Native Enterprises

Driving Transformation through Executive Leadership

Key approaches, mindsets, and competencies needed to lead in the age of AI-Native business

Leadership Approaches

Visionary Leadership
Long-term AI transformation roadmaps

Crafting compelling AI-driven futures and navigating through uncertainty with clear strategic vision

Experimental Leadership
Test-learn-iterate approach

Creating safe spaces for AI experimentation, rapid prototyping, and embracing intelligent failure

Ecosystem Leadership
Collaborative networks of AI partners

Building and nurturing AI partnerships, academic collaborations, and open innovation networks

Critical Mindsets

Adaptive Thinking
Embracing continuous change as AI capabilities evolve at exponential rates

Flexibility · Resilience

Ethical Conscious
Proactively addressing the ethical implications of AI deployment in business context

Responsibility · Foresight

Data-Driven Decision Making
Balancing AI insights with human judgment for strategic decision-making

Objectivity · Analysis

Human-AI Collaboration
Viewing AI as an augmentation to human capabilities rather than replacement

Complementarity · Synergy

Long-Term Value Creation
Prioritizing sustainable AI transformation over short-term optimizations

Sustainability · Vision

Essential Competencies

AI Technical Literacy — Critical
Understanding core AI concepts without needing deep expertise
Understanding AI capabilities 85%

Decision Intelligence — High
Leveraging AI insights to enhance strategic decision-making
AI-augmented decisions 78%

Change Leadership — Critical
Guiding organizational transformation toward AI-Native models
Transformation leadership 92%

AI Talent Development — High
Building, attracting and retaining AI-specialized workforce
Talent strategy 82%

Ethical AI Leadership — Critical
Ensuring responsible AI deployment and governance
Responsible AI oversight 90%

Executive Action Plan

1. **Establish AI Vision** — Define strategic AI objectives aligned with business goals
2. **Build AI Governance** — Create frameworks for responsible AI deployment and oversight
3. **Develop Talent Strategy** — Identify skill gaps and create AI talent acquisition plans
4. **Foster AI Culture** — Create organizational environment that embraces AI innovation
5. **Build Partnerships** — Develop strategic alliances with AI technology providers and experts

Risk Management in AI-Native Transformation

Strategic Approach to Risk

Identifying, assessing and mitigating risks in your AI-Native journey

Key Risk Categories

Security & Privacy Risks HIGH
Data breaches, model vulnerabilities, adversarial attacks

Regulatory & Compliance HIGH
Legal liability, non-compliance penalties, regulations

Operational Risks MEDIUM
System failures, performance issues, integration problems

Organizational Risks MEDIUM
Talent gaps, change resistance, cultural misalignment

Mitigation Strategies

Security-by-Design
Embed security & privacy in development lifecycle

Testing & Validation
Rigorous testing protocols for AI systems

Human Oversight
Maintain humans in critical decision loops

Continuous Training
Upskill workforce on AI risk management

Redundancy Planning
Fail-safe mechanisms for critical AI systems

Governance Frameworks

AI Risk Committee
Dedicated cross-functional oversight committee

Ethics Review Board
External stakeholders evaluate ethical implications

Policy Framework
Clear guidelines for AI development & deployment

Monitoring System
Continuous monitoring of AI model performance

Risk management is not a one-time effort but an ongoing process integrated throughout the AI lifecycle

Conclusion and Call to Action

Embrace the AI-Native Future

Transform your organization today for tomorrow's AI-powered business landscape

Key Takeaways

AI-Native is the new competitive frontier
Companies that embrace AI-Native principles will outperform traditional competitors

AX Revolution transforms customer experiences
Creating artificially intelligent experiences drives customer loyalty and retention

Technology stack is only part of the equation
Culture, leadership, and organizational structure are equally critical

Ethical AI is a business imperative
Responsible AI practices build trust and mitigate regulatory risks

Human-AI collaboration creates optimal outcomes
The most effective implementations blend human creativity with AI capabilities

Call to Action

Assess AI Readiness
Evaluate your organization's current capabilities, infrastructure, and culture

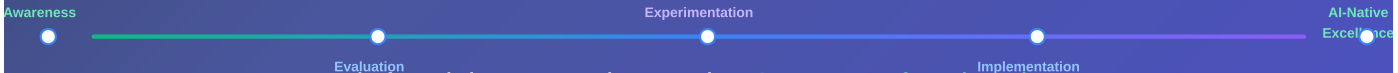
Develop an AI Roadmap
Create a strategic plan with clear milestones for AI-Native transformation

Start with Pilot Projects
Identify high-impact, low-risk areas to demonstrate AI value quickly

Invest in AI Literacy
Build organizational understanding from executive leadership to frontline staff

Build an AI Ecosystem
Partner with AI vendors, research institutions, and domain experts

Your AI-Native Transformation Journey



The future belongs to AI-Native enterprises. Start your transformation today.

Q&A and References

Explore Further

Additional resources to deepen your understanding of AI-Native enterprises and the AX Revolution

Recommended Reading

AI Superpowers
Kai-Fu Lee (2018)

The AI-First Company
Ash Fontana (2021)

Competing in the Age of AI
Karim Lakhani & Marco Iansiti (2020)

Prediction Machines
Ajay Agrawal, Joshua Gans & Avi Goldfarb (2018)

Human + Machine
Paul Daugherty & H. James Wilson (2018)

Research Papers

Building the AI-Powered Organization
Harvard Business Review (2019)

Artificial Intelligence and Business Strategy
MIT Sloan Management Review (2022)

The Economics of Artificial Intelligence
National Bureau of Economic Research (2021)

AI Governance: A Research Agenda
Future of Humanity Institute, Oxford (2020)

Machine Learning for Business Decision-Making
Journal of Business Research (2022)

Online Resources

AI for Business Leaders
www.youtube.com/ai-business-academy

AI Business Podcast
www.aibusinesspodcast.com

AI for Executives Course
www.coursera.org/ai-executives

AI Trends Newsletter
www.aitrends.com/subscribe

Contact Information

Email
contact@ai-enterprise.com

Phone
+1 (555) 123-4567

LinkedIn
linkedin.com/ai-enterprise

Twitter
@AIEnterprise

Thank You

Questions? We're here to help on your AI-Native transformation journey

session 9

AI Literacy for Social Work: Understanding Societal Impacts, Governance Needs, and Ethical Engagement

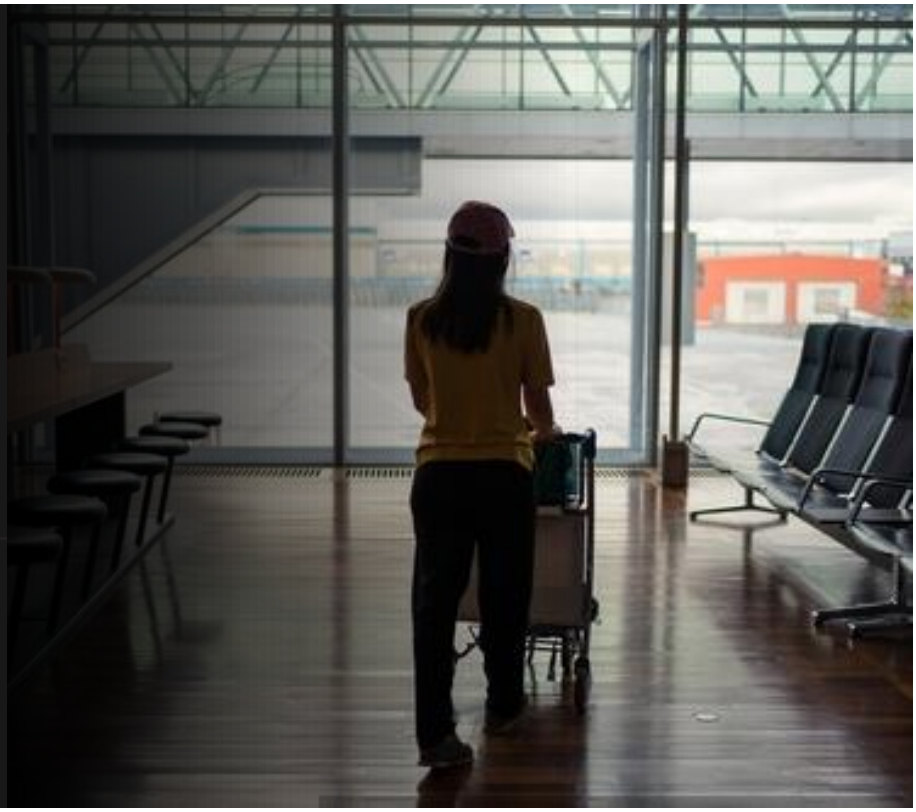
Eunhye Ahn, PhD
Assistant Professor, Brown School of Social Work

ICKOSSRC International Conference
May 28, 2025

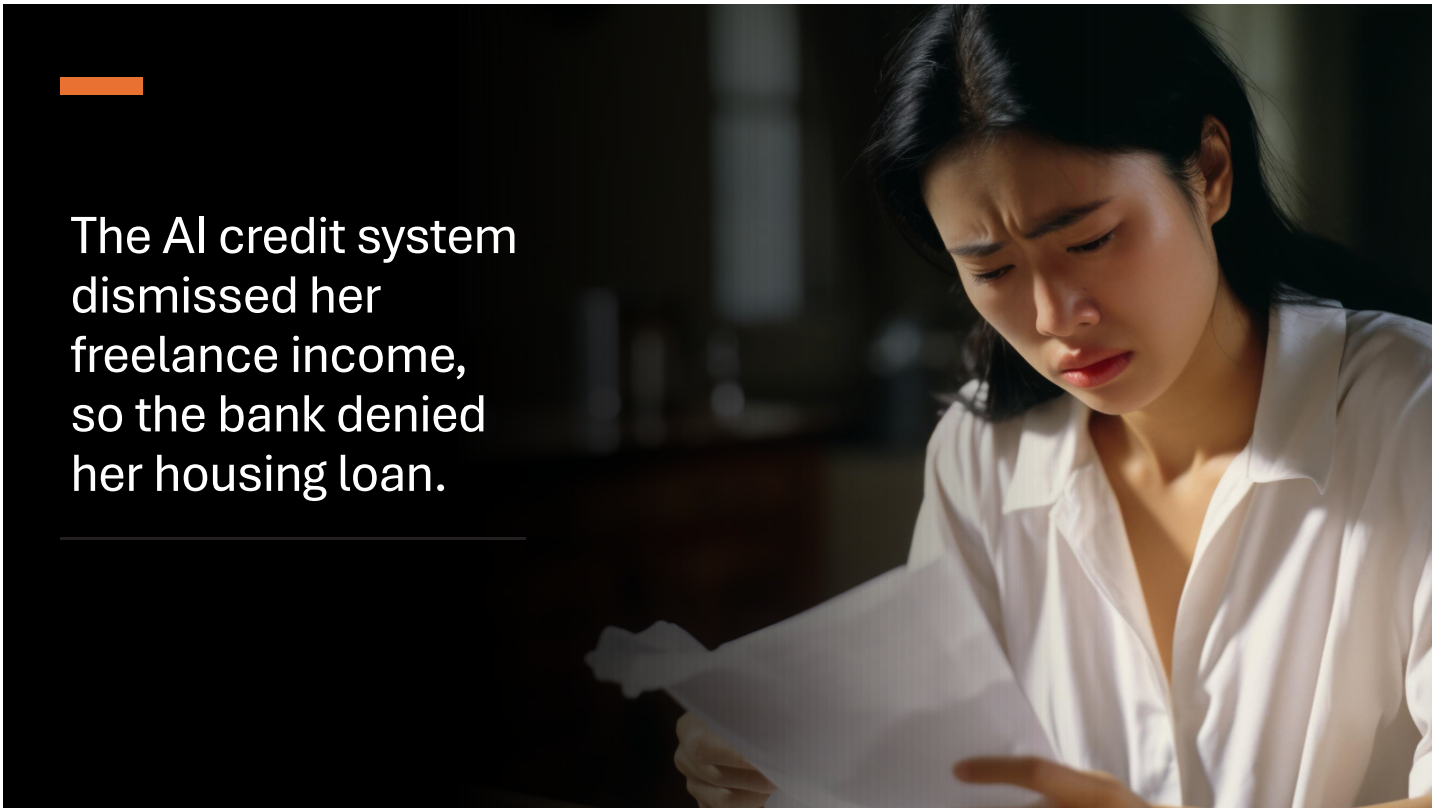
Washington University in St. Louis
Student Affairs

The AI tutor misunderstood his accent and said he was failing.

The algorithm replaced her before her manager could say goodbye.



She drives safely but her zip code told the AI she was high risk.


The AI credit system
dismissed her
freelance income,
so the bank denied
her housing loan.

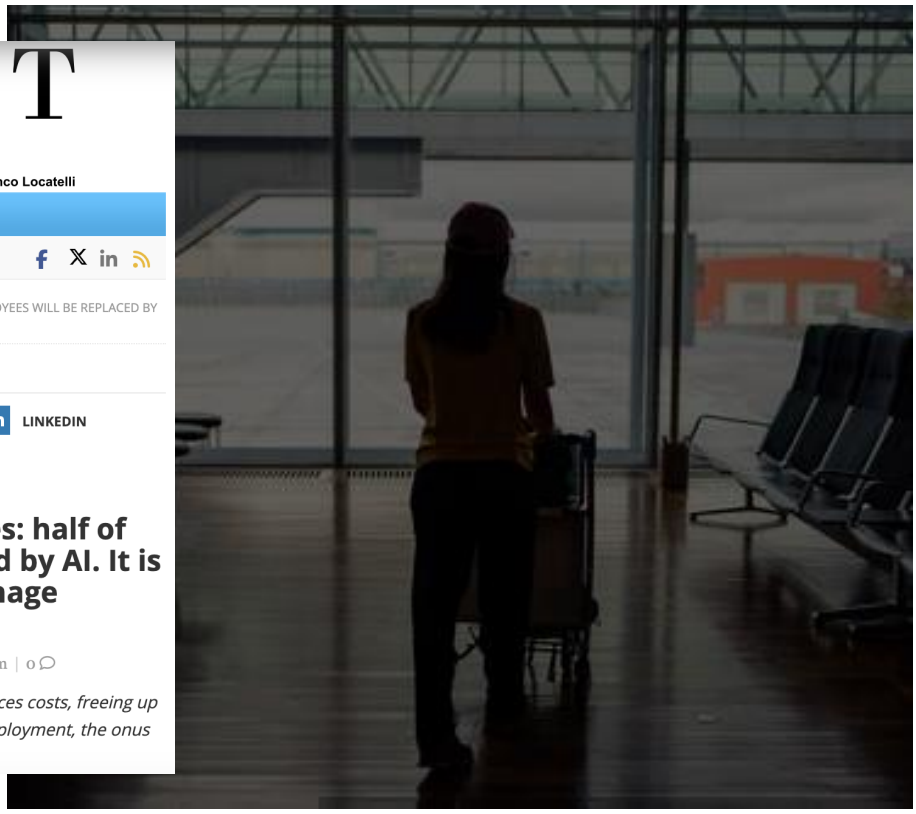

These are not tech glitches
they are social failures written in code.

UChicago News

AI is biased against speakers of African American English, study finds

Copyright Shutterstock.com

By Tori Lee Sep 17, 2024



FIRST online

Presidente: Ernesto Auci Direttore: Franco Locatelli

MENU

APPROFONDIMENTI

HOME ▸ WORK ▸ KLARNA FIRES AND AUTOMATES: HALF OF EMPLOYEES WILL BE REPLACED BY AI. IT IS UP TO GOVERNMENTS TO MANAGE UNEMPLOYMENT

SHARE

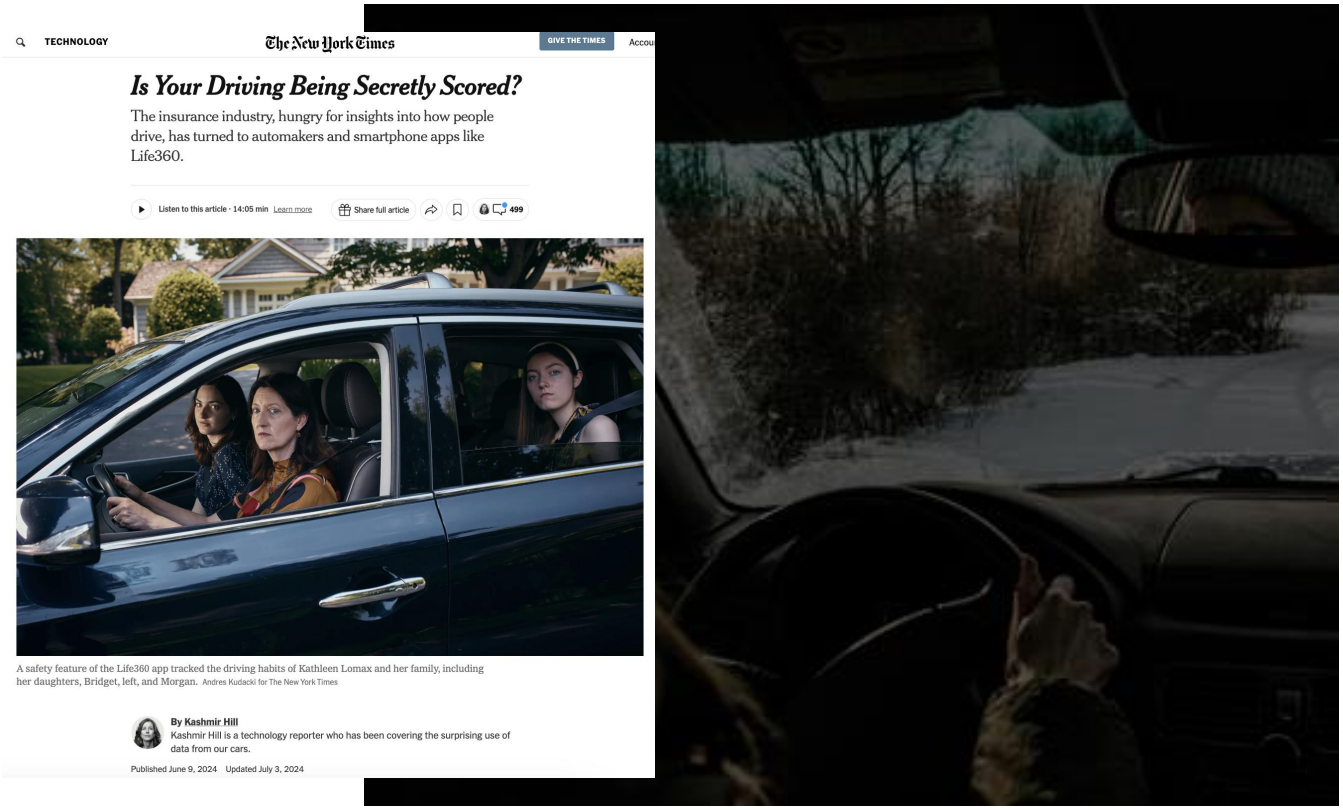
X.COM FACEBOOK LINKEDIN

PRINT EMAIL

Klarna fires and automates: half of employees will be replaced by AI. It is up to governments to manage unemployment

August 28, 2024, 16:58 | di FIRSTonline editorial team | 0

The CEO of the Swedish fintech believes AI reduces costs, freeing up budgets to pay top talent. And with rising unemployment, the onus is on governments to address the problem



TECHNOLOGY

The New York Times

GIVE THE TIMES

Accou

Is Your Driving Being Secretly Scored?

The insurance industry, hungry for insights into how people drive, has turned to automakers and smartphone apps like Life360.

Listen to this article · 14:05 min Learn more

Share full article

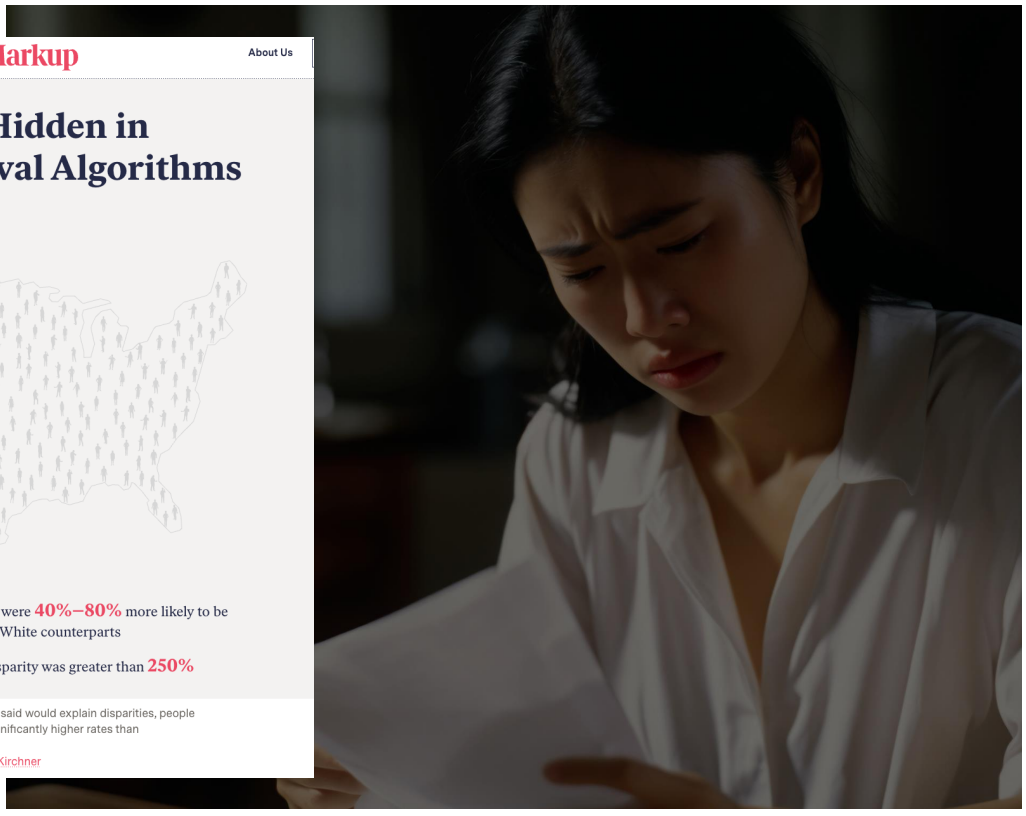
499

A safety feature of the Life360 app tracked the driving habits of Kathleen Lomax and her family, including her daughters, Bridget, left, and Morgan. Andres Kudacki for The New York Times

By Kashmir Hill

Kashmir Hill is a technology reporter who has been covering the surprising use of data from our cars.

Published June 9, 2024 Updated July 3, 2024



The Markup

About Us

Denied

The Secret Bias Hidden in Mortgage-Approval Algorithms

Nationally, loan applicants of color were **40%–80%** more likely to be denied than their White counterparts

In certain metro areas, the disparity was greater than **250%**

2021 06:50 ET

Even accounting for factors lenders said would explain disparities, people of color are denied mortgages at significantly higher rates than White people

By Emmanuel Martinez and Lauren Kirchner

“Why and how is AI a social work problem?”

11

“Why and how is AI a social work problem?”

“Why does AI governance matter?”

“What does AI literacy mean?”

12

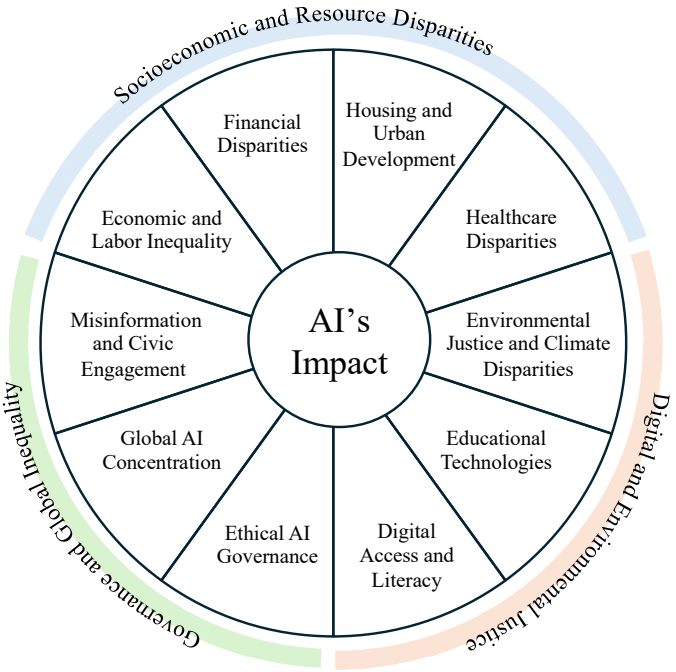
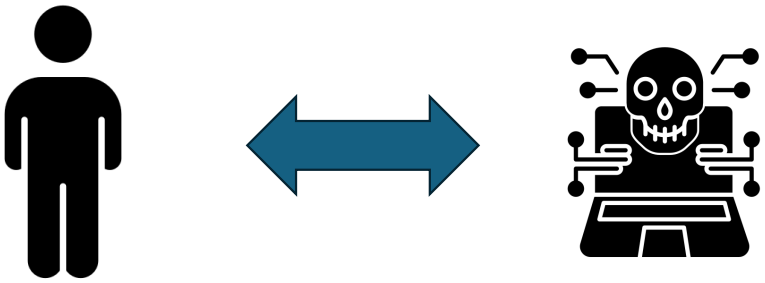


Figure 1. Circle of Transformation: AI's Impact on Social Inequalities

AI is often framed as being at odds with humans—as if it were our enemy, not our creation.



Musk Allies Discuss Deploying A.I. to Find Budget Savings

A top official at the General Services Administration said artificial intelligence could be used to identify waste and redundancies in federal contracts.

Share full article

Some of Elon Musk's former employees and allies are now evaluating the use of artificial intelligence through government systems to find budget cuts and detect waste and abuse.

Business / Tech

House Republicans want to stop states from regulating AI. More than 100 organizations are pushing back

By Clare Duffy, CNN

4 minute read · Updated 4:29 PM EDT, Tue May 20, 2025

US Speaker Mike Johnson (R-LA) speaks to reporters as he leaves the House Chamber at the US Capitol on May 15, 2025. Despite disagreements within his own party, Speaker Johnson said he plans to bring the House Republican budget bill to the floor next week for a vote. Kevin Dietsch/Getty Images

Korea passes AI Basic Act, second globally, enhancing national AI competitiveness

By Kim Min-kuk

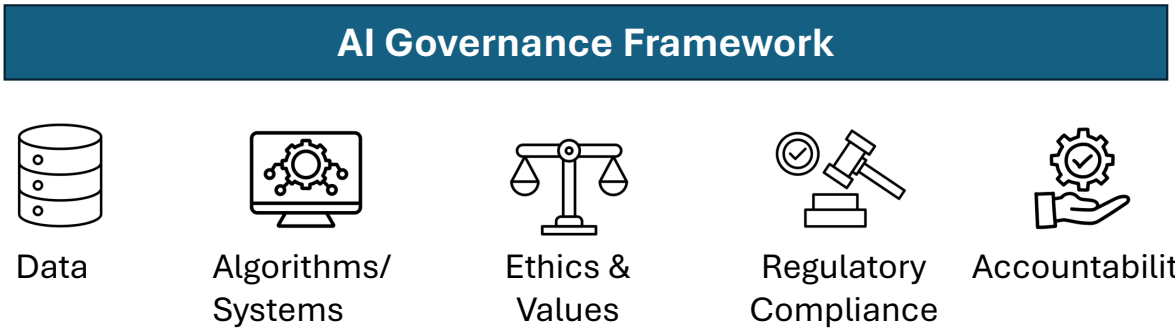
Published 2024.12.26. 16:49

The basic bill on artificial intelligence development and trust-based foundation is passing in the National Assembly's plenary session on the 26th. /Courtesy of Yonhap

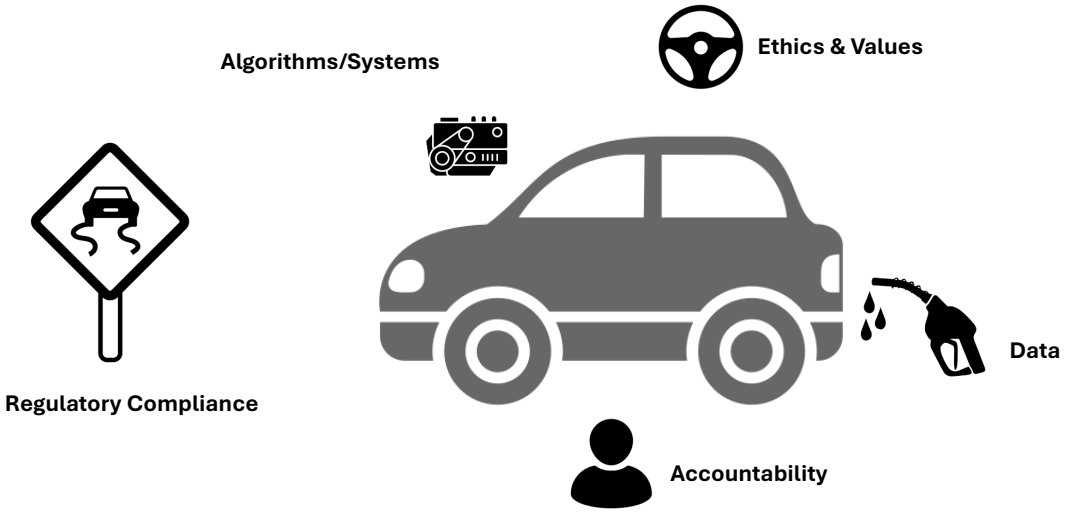
AI governance is the foundational framework that ensures AI is used responsibly, safely, and effectively.

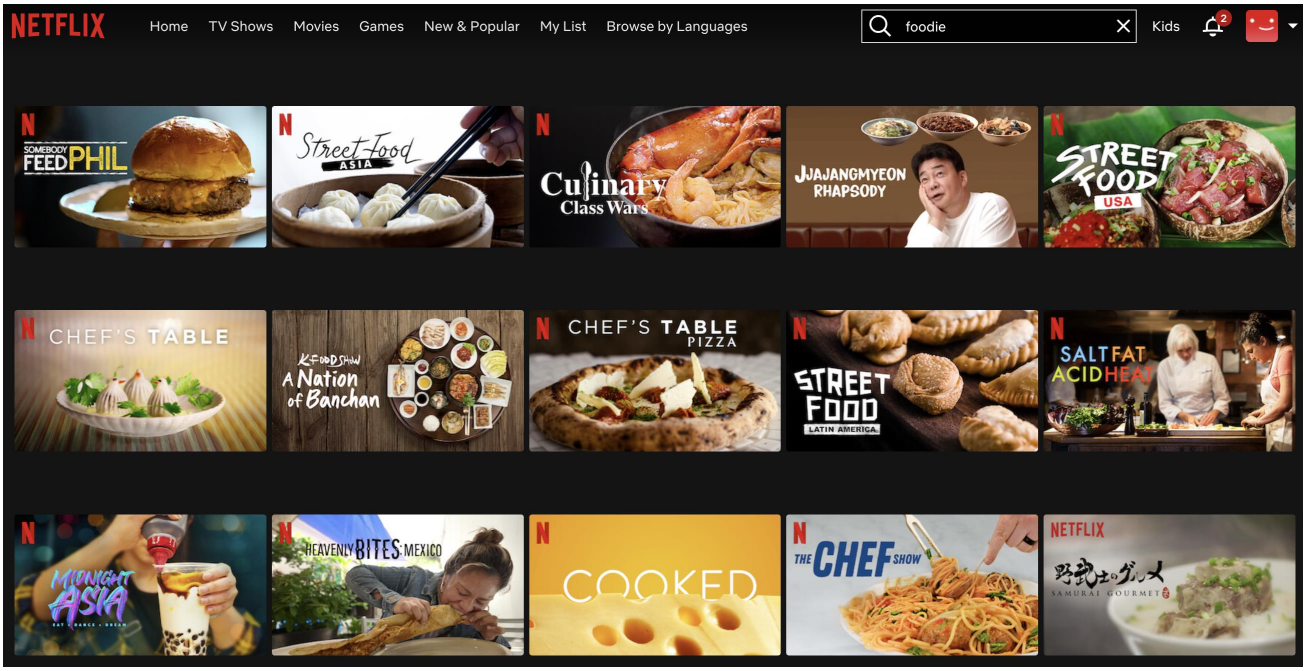
*AI governance is **a system of rules, practices, processes, and technological tools** that are employed to ensure an organization's use of AI technologies aligns with the **organization's strategies, objectives, and values**; fulfills legal requirements; and meets principles of ethical AI followed by the organization.*
(Mäntymäki et al., 2022)

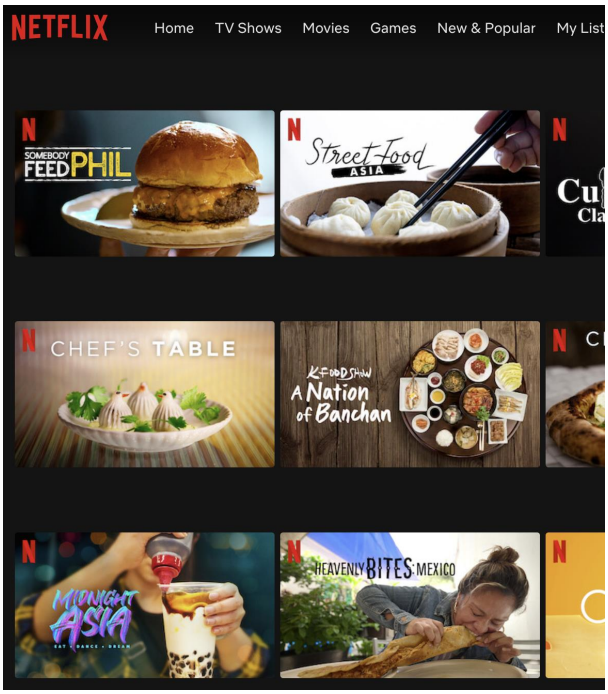
AI governance is the foundational framework that ensures AI is used responsibly, safely, and effectively.



(Attard-Frost & Lyons, 2024; Birkstedt et al., 2023; Mäntymäki et al., 2022)







by Jianan Liu for ProPublica

Criminal Justice

An Algorithm Deemed This Nearly Blind 70-Year-Old Prisoner a "Moderate Risk." Now He's No Longer Eligible for Parole.

by Richard A. Webster, Verite News

April 10, 2025, 6 a.m. EDT

Share Republish

CALIFORNIA

Social workers charged with child abuse in case involving torture and killing of Gabriel Fernandez, 8

L.A. County social workers Patricia Clement, left, and Stefanie Rodriguez, third from left, are arraigned in Los Angeles along with their respective supervisors, Gregory Merritt, fourth from right, and Kevin Bom, second from right.
(Marcus Yam / Los Angeles Times)

By Garrett Therolf
writer

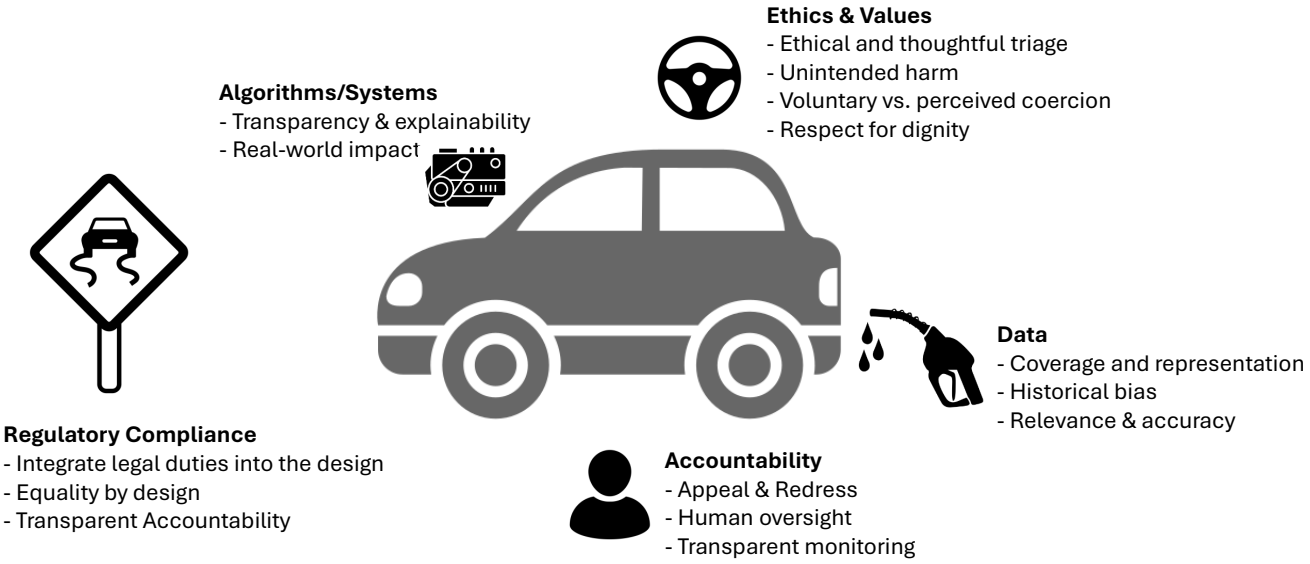
April 7, 2016 8:40 AM PT

Private sector AI governance doesn't fit the public sector's distinct goals and responsibilities.

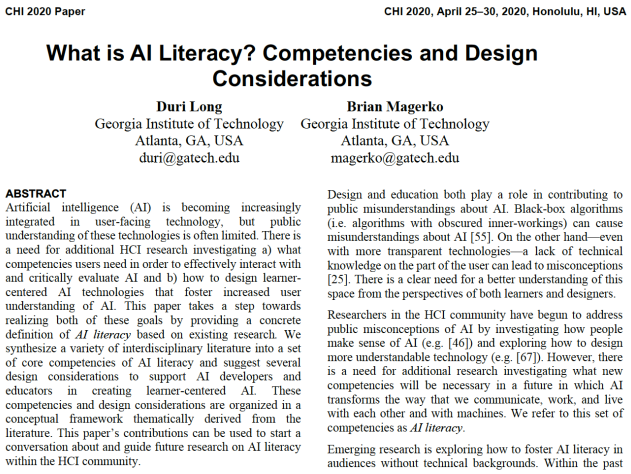
Source: https://www.dreamstime.com/royalty-free-stock-photos-child-oversized-suit-image14869508

Example: A community agency offering home visiting services wants to use AI to identify families most in need of support.

Source: https://www.freepik.com/free-photo/little-baby-being-health-clinic-vaccination_12892233.htm#fromView=search&page=1&position=25&uuid=913da2ba-b368-4fb6-b37d-e0aa2b05cf52&query=home+visiting+baby+nurse



*"We define **AI literacy** as a set of competencies that enables individuals to **critically evaluate** AI technologies; **communicate and collaborate effectively** with AI; and **use AI as a tool online, at home, and in the workplace**." (Long & Magerko, 2020)*



CHI 2020 Paper CHI 2020, April 25–30, 2020, Honolulu, HI, USA

What is AI Literacy? Competencies and Design Considerations

Duri Long
Georgia Institute of Technology
Atlanta, GA, USA
duri@gatech.edu

Brian Magerko
Georgia Institute of Technology
Atlanta, GA, USA
magerko@gatech.edu

ABSTRACT
Artificial intelligence (AI) is becoming increasingly integrated in user-facing technology, but public understanding of these technologies is often limited. There is a need for additional HCI research investigating a) what competencies users need in order to effectively interact with and critically evaluate AI and b) how to design learner-centered AI technologies that foster increased user understanding of AI. This paper takes a step towards realizing both of these goals by providing a concrete definition of *AI literacy* based on existing research. We synthesize a variety of interdisciplinary literature into a set of core competencies of AI literacy and suggest several design considerations to support AI developers and educators in creating learner-centered AI. These competencies and design considerations are organized in a conceptual framework thematically derived from the literature. This paper's contributions can be used to start a conversation about and guide future research on AI literacy within the HCI community.

Design and education both play a role in contributing to public misunderstandings about AI. Black-box algorithms (i.e. algorithms with obscured inner-workings) can cause misunderstandings about AI [55]. On the other hand—even with more transparent technologies—a lack of technical knowledge on the part of the user can lead to misconceptions [25]. There is a clear need for a better understanding of this space from the perspectives of both learners and designers.

Researchers in the HCI community have begun to address public misconceptions of AI by investigating how people make sense of AI (e.g. [46]) and exploring how to design more understandable technology (e.g. [67]). However, there is a need for additional research investigating what new competencies will be necessary in a future in which AI transforms the way that we communicate, work, and live with each other and with machines. We refer to this set of competencies as *AI literacy*.

Emerging research is exploring how to foster AI literacy in audiences without technical backgrounds. Within the past

Five Themes of AI Literacy by Long & Magerko (2020)

What is AI?

- Distinguish AI artifacts
- Understanding intelligence
- Interdisciplinarity in AI

What can AI do?

- AI's strengths & Weaknesses
- Imagine future AI and its effects

How does AI work?

- Knowledge representation
- AI decision-making

How should AI be used?

- Understand ML steps
- Recognize human role in AI
- Understand data literacy
- Critically interpreting data

How do people perceive AI?

- Understand AI can act on the world

(Long & Magerko, 2020) 26

ARTIFICIAL INTELLIGENCE & SOCIAL WORK SERIES

Artificial Intelligence (AI) Literacy for Social Work: Implications for Core Competencies

Eunhye Ahn *Washington University in St. Louis*
Moon Choi *Korea Advanced Institute of Science and Technology*
Patrick Fowler *Washington University in St. Louis*
In Han Song *Yonsei University*

ABSTRACT Artificial intelligence (AI) is fundamentally reshaping society, offering new opportunities while potentially intensifying socioeconomic inequalities. For social workers working with marginalized populations, understanding AI's societal impact is crucial, even if they do not directly engage with AI tools. This invited paper explores how AI literacy—the knowledge and skills to understand, use, and critically evaluate AI systems—can enhance social workers' ability to support communities navigating AI-driven changes. We review AI's impacts and applications in social work, explore their implications for the profession, and discuss their effects on social work core competencies. Specifically, we discuss how each of the social work competencies should evolve in response to AI-driven societal changes to better prepare social workers to support affected communities. By embedding AI literacy into core competencies, social workers can better address emerging challenges and promote equity in an AI-influenced society.

KEYWORDS: artificial intelligence (AI) literacy, social work core competencies, AI impacts on social work, AI-driven social change, AI ethics in social work

doi: 10.1086/735187

27

Questions & Comments?

28

〈The 2025 International Conference of the Korean Social Science Research Council, 2025.5.28.〉

사회보장 영역에서 인공지능 기술 적용의 순기능, 위험성 및 정책 방향

한국보건사회연구원 연구위원
김기태

1. 들어가며

공공영역 가운데 사회보장은 인공지능의 발전과 적용이 가장 활발하게 관찰되는 분야 가운데 하나다(Zaber, Casu, Brodersohn, 2024). 인공지능 기술은 다수의 복지국가에서 이미 급여 자격심사, 급여액 산정, 급여 지급 등의 과정에서 적용되고 있다. **미국의 경우, 미국 연방정부에서 인공지능을 활용하는 사례 가운데 3분의 1 이상이 보건복지부 업무 영역에 속한다**(Burt, 2024.10.17.). 인공지능 기술은 사회정책 영역에서 효율성, 과학성, 중립성의 증진에 기여할 것이란 기대가 크기 때문이다(Zaber, Casu, Brodersohn, 2024). 아동수당 및 실업수당 등 복지급여를 중심으로 오스트리아, 뉴질랜드, 영국, 브라질 등의 다수의 국가에서 인공지능 기술이 적용되고 있다(정세정 외, 2023).

그럼에도, **사회보장 영역에서 인공지능이 초래할 영향 및 정책적 대응에 대한 국내 연구는 희소**하다. 디지털 및 정보화의 사회정책적 함의에 관한 연구는 일부 있었고(김수영, 2016; 김수영, 김수완, 2022; 정세정 외, 2023), 보건복지 현장에서의 인공지능 기술 적용에 관한 연구(조남경, 송기호, 2023; 정유채, 2023)는 일부 수행된 바 있다. 성윤희(2022)는 사회보장 영역에서 인공지능이 초래할 변화를 검토한 드문 연구이지만, 인공지능을 "4차 산업혁명의 요체이며 지식기반사회의 핵심 화두"로 파악하는 기능적 부분에 초점을 뒀다.

한국이 공공행정에서 인공지능의 활용이 뒤처진 것도 아니다. 김기태 외(발간예정)에서는 **한국의 사회보장 영역에서 ICT 기술이 활용된 36개 사례를 확인**했다. 한국의 공공분야에서 디지털 인프라도 잘 갖춰진 나라다. **OECD(2024) 발표하는 디지털 정부 지수 순위에서 한국은 압도적 1위**를 차지했다. 한국은 여섯 개 평가 영역 가운데 데이터 기반 정부, 플랫폼 정부, 개방형 정부, 선제적 정부 등에선 1위를 차지했고, 디지털 우선 정부, 국민 주도형 정부 등 2개 부문에서 2위였다. 한편, 영국 미디어 업체인 Tortoise Media에서 제시한 'Global AI Index'에서는 미국, 중국, 싱가포르, 영국, 프랑스에 이어서 6위에 한국이 위치했다(Cesareo, White, 2023).

한국은 사회보장 영역에서 인공지능 기술을 디지털 인프라 위에서 빠른 속도로 적용시키고 있는 한편, 인공지능 기술 적용이 초래할 법적·윤리적 부작용 문제에 대해서는 상대적으로 방임적인 태도

를 보였다. 관련 연구와 정책이 한국에서 제시된 바가 희소한 것은, 이러한 상황을 증명한다. 이번 글은 이러한 문제의식에서 출발했다. 이 글에서는 **① 다른 복지국가들의 사회보장 영역에서 인공지능 기술 적용 현황을 짚고**, **② 인공지능 기술의 활용 영역 및 활용에 따른 순기능과 위험성을 살펴본 뒤,** ③ **인공지능 기술의 사회보장 영역에서의 적용 과정에서 짚어야 할 정책적, 제도적 함의를 제시**한다. 그리고, 추가적으로 ④ 기술 적용에 따른 관련 규제 동향을 유럽연합의 인공지능법을 중심으로 살펴본다.

2. 국외 사회보장 영역에서 인공지능 기술의 적용 현황[1)]

사회보장 영역에서 인공지능이 사용되기 시작한 시점은 **2008년 이후지만 본격적인 확산기는 2017년 이후**였다(Zaber, Casu, Brodersohn, 2024). 인공지능이 활용되는 사회보장영역은 가족급여, 보건, 산업재해, 연금, 실업, 보편적 급여 등 다양하다. **2020년 이전에 사회보장 영역에서 주로 활용된 인공지능 기술은 챗봇**이었다. 챗봇은 2020년 이후에도 코로나 범유행 상황에서 폭증한 급여 신청을 기관들이 대응하는 과정에서 활용도가 더욱 높았다. 일부 기관들은 챗봇의 기능을 강화하는 과정에서 생성형 인공지능(generative artificial intelligence)을 활용하기 시작했다.

Zaber, Casu, Brodersohn. (2024)는 사회보장 관련 기관들의 인공지능 활용 영역을 다음과 같이 제시했다(pp. 22~23).

첫째, **서비스 제공(Service delivery)**. 기관이 다양한 유형의 고객에게 더 쉽게 접근할 수 있는 채널을 활용하여 더 적절하고 더 나은 다양한 유형의 고객에게 더 쉽게 접근할 수 있는 채널을 활용하여 서비스를 개선

둘째, **자동화 및 사례 관리(Automation and case management)**. 사회보장 기관이 인공지능을 사용하여 사례 처리 방식을 자동화하고 개인의 서비스 후속 조치를 위한 고충 대응 지원을 제공

셋째, **전향적이고 능동적인 사회 보장**(Prospective and proactive social security). 사회보장 기관에 인사이트, 비전 및 잠재적 결과를 파악할 수 있는 전망 분석을 위한 도구를 제공. 기관은 이러한 도구를 활용하여 코호트와 개인의 삶을 선제적이고 능동적으로 개선하기 위한 접근 방식을 수립.

넷째, **위험관리 및 예방**(Risk management and prevention). 기관이 위험을 식별하고 위험을 완화하거나, 위험에 대응할 수 있는 역량을 강화. 더 나은 서비스를 제공할 수 있도록 회원 정보를 분석할 수 있도록 지원

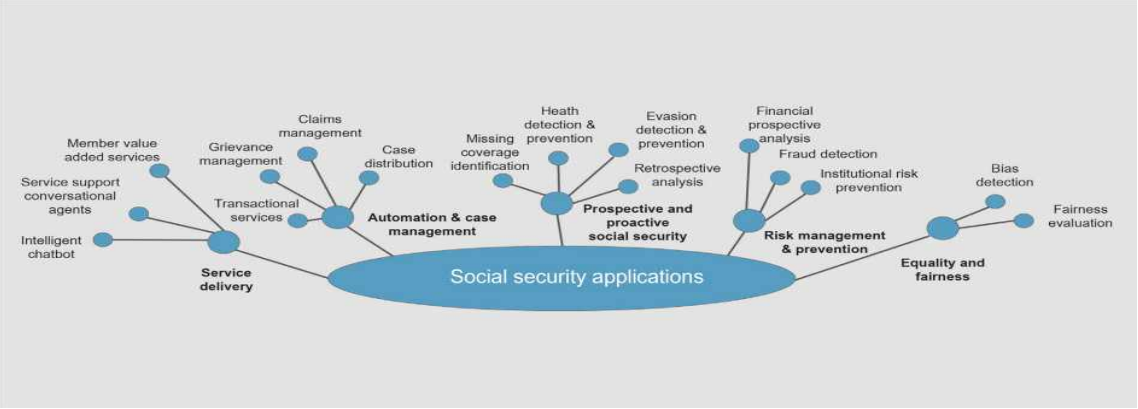
다섯째, **평등과 공정성**(Equality and fairness). 공정성과 형평성의 원칙을 지킬 수 있는 기관의 의무 측면에서 다양한 인공지능 솔루션과 프로그램 및 대응을 평가

Zaber, Casu, Brodersohn (2024)가 제시하는 범주들이 깔끔하지는 않다. 영역별로 사례가 제

1) 이 부분 전반부는 국제사회보장협회(International Social Security Association)이 발간한 Zaber, Casu, Brodersohn (2024) Artificial Intelligence in Social Security Organizations 보고서 내용을 정리한 결과다.

시되지 않아서, 직관적인 이해가 쉽지도 않다. 그러나, [그림 1]을 보면, 다섯 개 영역에서 파생되는 인공지능 기술의 활용 내용을 보다 상세하게 파악할 수 있다. 다섯 영역 가운데 하나인 서비스 제공 분야에서 인공지능은 가장 활발하게 사용된다.

[그림 1] 사회보장 영역에서 인공지능 기술의 적용 분야



출처: Zaber, Casu, Brodersohn, International Social Security Association. (2024). "Artificial Intelligence in social security organizations(p23)". International Social Security Association

서비스 제공(Service delivery) 가운데서도 챗봇은 인공지능이 가장 활발하게 사용되는 영역이다. 특히, 흥미롭게도, **남미국가에서 챗봇의 활용도가 높았다**(ISSA, 2021). 브라질, 아르헨티나, 파나마, 우루과이 등에서 국민들의 급여 관련 문의나 민원을 응대하는 데 챗봇이 활용됐다. 이를테면, 코로나19 범유행 위기 상황에서 **브라질 국립사회보장원**(Instituto Nacional de Seguro Social, INSS)은 인공지능 기반의 가상 비서 **챗봇인 Helô를 도입**했다. INSS가 관리하는 애플리케이션인 Meu INSS에 대한 문의에 응답하도록 설계된 Helô는 사용자와의 상호작용을 개선하고 더 복잡한 응답도 가능하도록 기능이 점진적으로 확장됐다. 초기 평가 결과 Helô는 첫 달에 100만 건 이상의 상담을 처리했다(ISSA, 2021). 이후 3년 동안 **Helô는 90개의 주제에 대해 660만 건의 상담**을 처리했고, 사용자 만족도가 0~5점 척도 기준으로 긍정적인 3.8점을 받았다(INSS. 2023).

챗봇의 활용은 소위 '선진국'에 국한되지 않는다. **말레이시아의 직원연금기금(EPF)은 ELYA 스마트 챗봇을 운영**하고 있다(Zaber, Casu, Brodersohn, 2024; ISSA, 2020). ELYA는 자연어 처리(NLP)를 사용하면서 인공지능으로 구동되고 라이브 채팅으로 지원되는 사회보장 기관 최초의 이중 언어 가상 비서(VA)다(ISSA, 2020.). 고객이 직접 EPF 상품 및 서비스에 대한 정보에 액세스할 수 있도록 하는 방식으로 상담센터의 부담을 덜어주는 것이 목표였다. ELYA는 기본 챗봇에서 3단계에 걸쳐 업그레이드해서 복잡한 문의를 위한 자문 챗봇까지 가능하게 진화했다. **ELYA는 실시간 상호작용, 다국어 지원, 연중무휴 24시간 상담 서비스를 제**공했다.

보건 영역에서는 인공지능이 응급실까지 들어왔다. **호주에서도 뉴사우스웨일스주 공공 의료 시스템의 응급실에서 패혈증 조기 발견을 목표로 하는 머신러닝 프로토타입 제품을 개발**했다(Zaber,

Casu, Brodersohn, 2024). 이름은 eHealth NSW였다. 인공지능 프로토타입은 2017~2019년 네 병원에서 추출한 과거 데이터를 활용해서, 로지스틱 회귀 및 XGBoost 알고리즘을 사용했다. 이 도구는 응급실 대기실에서 패혈증 발병 위험이 있는 환자를 조기에 발견하여 패혈증 관련 사망, 중환자실 입원 및 재입원을 줄이는 것을 목표로 한다.

복지급여 수급자의 포착에도 인공지능은 이미 활용되고 있다. **캐나다 고용사회개발부(Employment and Social Development Canada)는 저소득 노인을 위한 급여인 보장소득보조금**(Guaranteed Income Support) 영역에 인공지능 기술을 적용했다(Zaber, Casu, Brodersohn, 2024). 급여 수급 자격이 있는 노인을 신속하게 식별하기 위한 목적이었다. 머신러닝 모델을 통해서 2,000명 이상의 수급 대상자를 식별했다. **정확도는 92~98%**였다(Zaber, Casu, Brodersohn, 2024).

사례 관리 영역에도 인공지능은 역할을 확대하고 있다. **오스트리아의 사회보험연합은 청구 자동 처리를 지원하고 의사와 매칭하는 인공지능 기반 시스템을 구현**했다(Zaber, Casu, Brodersohn, 2024). 인공지능을 활용한 의료비 환급 관리 사례가 예시될 수 있다. 인공지능 솔루션을 도입하기 전에는 수작업 처리 시간이 길어 환급받기까지 몇 달씩 기다려야 했다. 플랫폼을 도입한 후에는 추가 인력 없이도 처리 시간이 며칠로 크게 단축됐다. AI 기반 접근 방식은 광학 문자(Optical Character R) 및 개체 인식과 같은 기술과 오픈소스 도구 및 언어를 사용하여 문서를 수집하고 처리함으로써 효율성과 투명성을 개선했다.

〈표 1〉 사회보장 영역에서 인공지능 기술 적용 사례

사회보장 영역	인공지능의 해법	상세 내용
연금	연금 행정 자동화	연금 관리 및 연금 지급을 자동화하여 퇴직자에게 적시에 급여 지급, 관리 비용 절감, 효율성 개선
아동 돌봄 및 교육	인공지능 아동 돌봄 및 교구 활용	스마트 모니터링 시스템으로 어린이의 안전 보장. AI 기반 교육 플랫폼으로 어린이 개인화한 학습 경험 제공 및 발달 과정 모니터링
급여 수급자 확인	자동화한 수급자 확인	인공지능 기반 시스템으로 본인 확인 절차 간소화. 신원, 문서, 데이터를 신속하고 정확하게 확인하여 부정 수급 가능성
고령 돌봄	돌봄을 위한 로봇 활용	인공지능 기반 로봇 공학은 노약자에게 신체적, 인지적 지원 제공 및 일상 업무 지원, 복약 알림, 말벗 기능 수행
정서적 돌봄	정신건강을 위한 챗봇	인공지능 기반 챗봇이 정서적 지원 및 상담 서비스 제공.

출처: Zaber, Casu, Brodersohn, International Social Security Association. (2024). “Artificial Intelligence in social security organizations(p29)”. International Social Security Association의 표 3의 내용을 번역하고 일부 편집함

인공지능이 그리는 사회보장의 미래가 장밋빛으로만 그려지지는 않는다. 네덜란드의 덴마크의 사례는 인공지능 적용에 따른 인권 침해, 정보 유출, 공공성 훼손에 관한 도전을 보여준다.

먼저, 네덜란드의 사례를 살펴보겠다. **SyRI(Systeem Risico Indicatie)는 중앙정부와 지방정부가 사회보장 및 소득 관련 제도, 세금 및 사회보험료, 노동법 분야에서 부정행위를 방지하고 대처하기 위해 도입한 인공지능 시스템**이다. SyRI는 다양한 공공기관의 데이터를 연결하고 분석하여 위험 보고서를 생성함으로써 재정의 오용을 방지하고, 부당한 이익을 탐지하는 데 사용된다(Rechtbank Den Haag, 2020). 이 시스템에 참여하는 기관은 지방정부, 네덜란드 조세 및 관세청, 사회보험은

행(Social Insurance Bank), 이민 및 귀화 서비스, 고용보험청, 사회 및 고용감찰청 등과 같은 감독 기관이다. 이 기관들로부터 수집되는 데이터는 보건, 재정, 교육, 재정 지급, 고용 등 방대한 영역에 걸쳐 있다. **알고리즘을 통한 데이터 분석 결과를 바탕으로 특정 사례에 대한 '위험 보고서'가 제출되면, 관련자는 부정행위의 가능성이 있는 것으로 판단되어 정부의 조사 대상으로 간주**된다(Appelman et al., 2021, p.263).

이러한 알고리즘을 활용한 분석은 부정수급을 적발하는 데 효과적이라는 것이 네덜란드 정부의 입장이다. 이 시스템은 주로 부정수급 가능성이 클 것으로 예상되는 가난하거나 취약한 사람들에 초점을 맞추어 조사를 수행한다는 점에서 논란이 많았다. 이와 관련하여 Platform Bescherming Burgerrechten, Nederlands Juristen Comité voor de Mensenrechten 등 네덜란드 6개 시민단체는 SyRI의 문제점을 지적하며 공동행동을 시작했다. 단체들은 2018년 3월 네덜란드 정부를 상대로 이 시스템이 **"모든 개인은 사생활과 가족생활 및 주거와 통신을 존중받을 권리가 있다"고 규정한 유럽인권협약(European Convention of Human Rights) 제8조를 위반**했다는 취지로 헤이그 지방법원에 소송을 제기했다(van Bekkum, Borgesius, 2021). 2020년 **헤이그 법원은 정부 기관들이 개인 정보를 공유하는 과정에서 투명성이 부족하고 개인 정보에 대한 적절한 안전장치가 부족하다는 이유로 SyRI가 유럽인권협약을 위반한다고 판결**했다.

2018년 **덴마크의 지방정부인 Gladsaxe는 사회경제적으로 취약한 아동을 추적하기 위해 알고리즘을 활용한 정책 실험을 추진**하였다. 덴마크 수도 코펜하겐(Copenhagen) 근교에 위치한 Gladsaxe 정부는 실업 및 의료를 비롯한 다양한 사회경제적 영역의 데이터를 결합하여 **200개 이상의 위험 지표를 분석하는 머신러닝 모델을 구축**하였고, 이를 활용하여 가정폭력이나 학대의 위험이 큰 아동을 찾으려고 시도했다. 이 모델은 가정이나 부모의 상황들을 점수화했다. 예를 들면 정신 질환의 경우 3,000점, 실업은 500점, 예약된 의사의 진료 불참은 1,000점, 예약된 치과 진료 불참은 300점 등이 부여되었다. Gladsaxe 알고리즘 모델은 이러한 점수를 바탕으로 위험한 상황에 처한 아동들은 판별하고자 했다(Thapa, 2019).

이러한 시도는 시민단체, 학계의 신랄한 비판을 받았고, 물론 대중으로부터도 심각한 반발을 낳았다. Mchangama and Liu(2018)는 **인공지능을 활용한 복지행정이 효율성을 높일 수는 있지만, 그 과정에서 개인의 자유와 민주주의의 가치가 희생**되고 있다며, "덴마크 복지국가가 인공지능 때문에 자살하고 있다"라고 주장했다. 이러한 비판에도 불구하고 덴마크 정부는 위험에 처한 아동을 조기에 발견할 수 있다는 장점을 강조하며 Gladsaxe 모델을 전국적으로 확대할 계획을 세웠다(Bendixen, 2018; Jørgensen, 2021). 그러나 Gladsaxe 모델에 대한 공개적인 검증이 이루어지면서 상황은 바뀌었다. 이 모델을 통해 작성되는 개별 평가서에 포함된 다양한 정보들이 부모들 모르게 활용되고 저장된 사실이 밝혀졌다. 이로 인해 2018년 12월 **Gladsaxe 모델은 중단**되었다(Algorithm Watch and Bertelsmann Stiftung, 2020).

3. 사회보장 영역에서 인공지능 기술의 활용 영역·순기능·위험성

3-1. 인공지능 기술의 활용 영역

국내·외의 인공지능 활용 현황을 보면, 사회보장 영역에서 인공지능이 활용되는 영역을 확인할 수 있었다. 과거 복지국가에서 디지털 기술이 활용되는 영역을 제시했던 Alston(2019), 김기태(2024), Zaber, Casu, Brodersohn(2024) 의 논의를 참고하면서 논의를 종합하면 아래 아홉 가지 영역이 제시될 수 있을 것이다[2].

첫째, 본인 인증(identity verification)이다. 본인 인증은 급여신청, 자격심사, 급여 지급 과정에서 반드시 필요하다. 물론, 한국에서는 지문 및 얼굴 정보가 포함된 주민등록 데이터베이스가 매우 높은 수준의 본인 인증 체계가 성립된 점을 확인할 수 있다. 한국에서는 인공지능 기술까지 필요하지 않을 수도 있다.

둘째, 자격 심사(eligibility assessment)다. 캐나다 온타리오주에서는 2014년부터 사회부조운영시스템을 통해서 급여 자격을 심사하고 있다. 빅데이터 및 인공지능을 통해서 급여 자격 심사를 빠르게, 정확하게 처리할 여지가 생긴다.

셋째, 복지급여액 산정 및 지급(welfare benefit calculation and payments)이다. 다수의 국가에서 점점 더 많은 복지급여액이 사람의 개입 없이 자동적으로 산정되고 지급되고 있다. 영국은 실시간 소득정보시스템(Real Time Information System)을 활용해서 복지급여를 지급하고 있다.

넷째, 부정·오류 수급 예방 및 탐색(fraud prevention and detection)이다. 많은 복지국가에서 디지털 자료를 활용하는 주요한 이유 가운데 하나가 부정·오류 예방 및 탐색에 있다. 네덜란드에서 논란을 놓았던 SyRi(Systeem Risico Indicatie)가 여기에 해당한다.

다섯째, 위험의 점수화 및 범주화(risk scording and classification)다. 한국의 사각지대 발굴관리 시스템이 여기에 해당한다. 개인에 관한 공공데이터 자료의 수집과 빅데이터 분석, 분석결과를 바탕으로 한 고위험 대상자 도출 과정을 거치기 때문이다. 이러한 접근은 유럽연합이 인공지능법에서 '수용할 수 없는 위험성(unacceptable risk)'으로 분류한 '사회적 평점(social scoring)'과 흡사하다. 앞으로도 논란을 낳을 수 있는 대목이다[3].

여섯째, 개인 맞춤형 정보 서비스다(personalized information service)이다. 남미 지역에서 특히 활성화한 챗봇이 대표적 사례가 될 것이다. 다양한 영역에서 개인에게 최적화된 서비스를 제공

2) OECD(2024)도 'Modernizing Access to Social Protection' 보고서에서 인공지능의 활용 영역을 다음과 같은 네가지로 제시했다. ① 공무원의 문서 작업 자동화 및 효율화, ② 개인이나 지역 사회의 위험 예측 혹은 예방적 접근, ③ 급여 자격 기준 등에서 결정 자동화, ④ 챗봇, 오피스 프로세스 자동화, 부정 수급 포착 영역 등이다. OECD(2024)가 전세계적으로 사회보장 영역에서 인공지능은 "간헐적(infrequently)"(p. 9)으로 쓰인다고 논평한 점을 보면, OECD는 인공지능이 활용되는 영역을 지나치게 협소하게 파악한 것으로 보인다.

3) Alston(2019)은 위험의 점수화 및 범주화 관련해서 세 가지 위험을 예시했다. 첫째, 전체 인구집단의 데이터를 근거로 한 예측모델에 따라 개인의 위험 수준을 파악하는 과정에서 나타날 수 있는 오류의 가능성, 둘째, 점수화 및 범주화의 근거가 되는 기술이 공개되지 않으면서 나타날 수 있는 권리 침해의 가능성, 셋째, 점수화 및 범주화가 현재의 불평등과 차별을 강화하거나 유지할 가능성이다.

하기 위해 빅데이터 분석 기반 기술을 활용하여 서비스 이용자에게 맞춤형 서비스를 제공할 수 있다. 한국의 AI 활용 초기상담시스템도 여기에 해당될 것이다. AI 활용 초기상담정보시스템도 콜 기반 대화 시스템을 활용하여 잠재적 위기대상자와 초기상담을 진행한다. 이러한 기능은 복지수급자의 사례 관리에까지 확장될 수 있다. 사회보장 기관이 인공지능을 사용하여 사례 처리 방식을 자동화하고 개인의 서비스 후속 조치를 위한 고충 대응 지원을 제공할 수 있다. 오스트리아의 사회보험연합은 청구 자동 처리를 지원하고 의사와 매칭하는 인공지능 기반 시스템을 구현했다.

일곱 번째, 온라인을 활용한 소통을 넘어서 실제 돌봄 영역에서도 활용되고 있다. 한국에서도 이미 폭넓게 활용되고 있다. 대부분 노인을 위한 것으로 AI·IoT를 활용한 건강 서비스, 생체건강 셀프 체크 서비스, 안전감지 센서를 활용한 안전 확인 서비스, 실종 방지를 위한 위치 기반 모니터링 서비스 등이다. 여가 활용에 도움이 되는 다양한 정보 제공 서비스 등이 제공되고 있다.

여덟 번째, 사회보장 행정 기관 내부적인 용도로 업무 담당자의 효율적 업무 처리를 돕고, 내부 교육 등의 용도로도 활용될 수 있다. 한국에서 다수의 복지기술은 업무 담당자의 효율적 업무 처리를 지원하기 위해 활용되고 있었다. 이를테면, 빅데이터를 기반으로 의사결정을 효율적으로 지원하는 서비스도 확인되었는데 수급자의 상담을 위한 기초 자료를 제공한다.

아홉번째, 사회정책의 효율성과 효과성을 평가하는 데 활용될 수 있다. 데이터의 가용성이 실시간화한다면, "적절한 성과지표와 주기적인 활용은 사업의 효과성과 효율성을 평가하는 데 필수적인 것으로 증거기반정책(evidence-based policy)의 기초"(유종성, 2023, p. 8)가 될 수 있다. 또한, 평가의 과정에서 정책의 성과, 실패, 한계를 낳은 원인을 분석해서, 정책을 조정, 갱신, 폐기하는 근거가 될 수 있다.

장기적으로 보면, **전국민 대상 실시간 데이터에 근거한 인공지능의 활용은 사회보장 제도 전체를 재편하는 방향으로 나갈 잠재력**을 가지고 있다. 노대명(2024)은 소득보장제도 재구조화를 준비하자고 제안하면서 "소득기반 사회보험의 실험이 중단됐지만 계속 추진할 필요"(p. 14)가 있으며 "현행 85개 제도를 4~5개 정도로 단순화하는 재정기반 소득보장제도를 준비"(p. 14)해야 한다고 제안하고 있다. 시민들의 다양한 욕구에 부응하면서 자리잡은 복지제도들이 제도운영 주체 및 전달체계에 따라 너무 복잡하게 얽혀있는 점을 고려할 필요가 있다. 이에 따라, 사회보장 제도들의 재구조화에 대한 요구는 끊임없이 있었던 점을 고려하면(강혜규, 강지원, 강희정, 김기태, 김세진, 김태완.. 주하나, 2022), 변화의 단초가 빅데이터와 인공지능에서 마련될 여지가 있다. 노대명(2024)은 현재의 급여지급 정보관리 체계에서 복지제공 부처별로 급여 자격 조건 심사를 따로 추진하면서 급여 지급이 지연되고 있다며, **디지털 사회보장 허브 구축**을 통해 급여 자격 심사와 지급을 효율화하자고 제안했다.

3-2. 인공지능 기술 적용의 순기능

사회보장영역에서 점점 폭넓게 활용되는 인공지능 기술 적용이 불러올 효과는 양면적이다. 먼저, 인공지능이 공공의 민주적인 통제 아래 작동할 경우를 전제로 할 때, 기대되는 순기능은 다음과 같

다.

첫째, 효율성이다. 앞서 살펴보았듯이, 본인인증, 자격심사, 복지급여액 산정 및 지급 등의 일선 서류 행정이 더욱 자동화하게 된다. 이럴 경우, 더 적은 인력으로 더 많은 행정 업무가 가능해진다. 일선 공무원은 민원인을 대상으로 하는 대면서비스나 사례관리에 집중할 수 있다. 물론, 인공지능의 적용에 따른 부가적인 업무도 생길 수 있다. 이를테면, 미국의 행정명령 14110 7조 2항 (b)호에 따르면, 급여 거부에 대해 급여 신청자가 사람인 심사자에게 이의를 제기할 수 있도록 하는 절차를 규정하고 있다.

둘째, 적시성이다. 빅데이터를 활용해서 관리되는 인공지능 기술은 급여 신청과 심사, 지급에 이르는 과정을 단순화할 수 있다. 영국의 경우, 통합급여를 도입하는 과정에서 실시간소득파악 시스템을 활용해서, 급여 지급 절차를 간소화하고자 했다.

셋째, 정확성이다. 빅데이터에 근거한 인공지능의 판단, 범주화, 예측은 인간이 오류와 편견이 개입할 가능성을 줄일 수 있다. 영국의 경우, 연금을 제외한 복지급여액 가운데 오류 또는 부정으로 인해 초과 지급된 액수가 2021~2022년 회계연도 기준으로 85억 파운드(약 12조 7천억원)으로 추정됐다(Davies, 2022). 같은 기간 연금을 제외한 전체 복지급여 지급액의 7.6%를 차지하는 액수다. 특히, 서구의 복지국가에서 디지털 기술을 통해 급여의 오류 및 부정 수급을 시정하고자 하는 의지가 강해 보인다.

넷째, 개인 맞춤형 급여 및 서비스 제공이다. 인공지능의 활용을 통한 행정 집행은 개인당 위험의 점수화 및 범주화까지 맞춤형 수준으로 정교화할 수 있다. 개인의 여건과 수요에 맞춘 현금 및 서비스, 일자리가 개인 맞춤형으로 제공될 여지가 커진다. 한국에서 AI 기반 일자리 매칭서비스가 대표적인 예다. AI기반 일자리 매칭서비스는 구직자의 이력서 정보와 구인기업의 구인정보를 활용하여 인공지능 알고리즘에 기반한 추천서비스를 제공하고 기업과 구직자 간 일자리 미스매치를 해소한다. 이를 통해서 개인의 제도에 대한 체감도가 향상될 수 있다.

다섯째, 범용성이다. 챗봇을 통한 상담은 국민들의 입장에서는 시공간의 제약을 뛰어넘어서 제도에 접근할 수 있는 기반을 제시해 준다. 시민들은 주민센터를 방문하지 않고도 언제, 어디서나 급여와 관련한 상담과 더불어 자격을 확인할 수 있고, 급여를 신청할 수 있게 된다. 이를 통해서, 지역별 접근성에서의 격차를 해소할 수 있는 가능성도 커진다.

여섯째, 정책 평가의 용이성이다. 빅데이터를 활용한 급여 집행 내용은 정책의 효과를 평가하는데 용이한 기반을 제공한다. 평가를 위한 데이터의 구축이 용이해지면서 근거 기반 정책 평가, 형성 및 집행의 토대가 마련된다.

일곱 번째, 사각지대 해소다. 복지 사각지대 발굴관리시스템이나 AI 활용 초기상담시스템은 다중 위험의 시대에서 잠재적 수급 대상자들의 욕구를 빠르게 파악하는 토대가 된다. 한국 복지국가에서 고질적으로 지적되는 사각지대의 문제를 빠르게 해소할 수 있는 제도적 기반이 될 수 있다. 물론, 사각지대 문제의 핵심은 위기가구의 발굴에 있다기보다는, 발굴된 위기가구를 지원할 급여가 없거나 부족하다는 지적(함영진, 이현주, 어유경, 김가희, 박성준, 조용찬.. 오민수, 2023)도 함께 고려할 필요가 있다.

3-3. 인공지능 기술 적용의 위험성

빅데이터 활용을 수반하는 인공지능 기술이 사회보장 영역에서 초래할 위험성도 함께 보겠다. 위험의 내용은 김수영(2016)과 김기태(2024)를 참조하면서 다음과 같이 제시하도록 하겠다.

첫 번째, 프라이버시의 문제다. 한국의 경우, 국민기초생활보장제도에서 수급자는 자신과 부양의무자의 소득 및 재산정보를 국가에 제공하는 조건 아래 급여를 받을 수 있다. 복지급여 수급자들의 소득, 재산, 가족 정보뿐 아니라 일부 사례관리 정보까지 이미 방대하게 집적돼 있다. 공공기관을 넘나드는 개인 정보의 유출 및 남용을 가능성도 완전히 배제할 수는 없다. 또한, 국가 권력에 의한 데이터 남용 가능성도 고려해야 한다(홍승헌, 황하, 2024).

둘째, 부정확성의 문제다. 앞서 인공지능이 불러올 장점으로 정확성을 제시한 점을 고려하면, 다소 모순적인 지적일 수 있다. 그럼에도, 국내의 사회보장정보시스템에서 개인의 사망 및 출생신고가 반영되지 않거나 과거 소득이 현재 소득과 합산돼서 제시되는 등의 문제가 끊임없이 발생하고 있다. 현장 공무원들은 복잡한 데이터를 처리하고 오류를 수정하느라 대면 접촉을 통해서 사각지대를 발굴할 기회를 오히려 놓치고 있다는 지적(임덕영, 2023)도 나오고 있다. 물론, 이러한 문제는 데이터 기반 인공지능 기술이 현장에 접목하는 과정에서 발생하는 일시적이고 과도기적인 문제라고 볼 수 있다. 그러나, 인공지능 기술에 대한 맹신적인 태도 역시 경계할 필요는 있다. 이 대목은 사회보장 행정에서 관리하는 데이터와 알고리즘의 질 관리와도 직결되는 문제다.

셋째, 데이터 소유권의 문제다. 김수영(2016)은 급여 신청자는 소득, 재산, 가족정보, 신체 등에 대한 자기 정보 제공에 동의하게 되는데, 여기서 정보의 소유권에 대한 문제가 제기된다고 지적한다. "국가는 정보 제공자의 의도와는 상관없이 이들이 제공한 정보를 보유할 뿐만 아니라 향후 해당 정보의 활용, 분배, 처분에 관한 권한을 얻게 되는 상황"(p.211), 이 대목은 앞서 논의한 프라이버시와 함께 앞으로 논점으로 부상할 여지가 크다.

넷째, 개인 정보의 영리 목적 활용의 문제다. '디지털 헬스'를 둘러싼 데이터 활용은 한국에서 이미 오랜 의제다. 건강보험의 개인 정보에 대한 접근권을 민간기업은 지속적으로 요구하고 있다. 보건복지부는 2024년 11월 '보건의료데이터 혁신포럼'을 전국 43개 의료데이터 중심병원과 진행한 바 있다. 복지부는 "정부와 의료계, 학계, 산업계가 데이터 협력체계를 구축하여 미래의료 혁신과 국민건강 증진으로 나아가는 계기가 되기를 희망"(보건복지부, 2024.11.26. p. 4)한다고 밝혔다. 여기에서 개인 건강 정보의 영리 목적 활용의 여지는 잠재적으로 남아 있다.

다섯째, 알고리즘 결정에 근거한 개입의 문제다. 덴마크의 취약 아동을 포착하기 위해서 실업 및 의료를 비롯한 다양한 사회경제적 영역의 데이터를 결합하여 200개 이상의 위험 지표를 분석하는 머신러닝 모델을 구축했다(Jørgensen, 2021). 그리고, 모델이 위험신호를 보내는 가구에 대해서 부모의 동의 없이 개입이 가능하도록 했다. 이후 해당 모델은 모델의 신뢰성에 의문이 제기되면서 2018년 12월 중단된 바 있다(Algorithm Watch and Bertelsmann Stiftung, 2020). 한국에서도

2023년부터 44종의 위기 정보에 근거해서 위기가구에 대해서는 경찰·소방 협력을 통한 비상 개문(開門) 지침을 마련했다(정세정 외, 2023). 이 대목에서 빅데이터에 근거한 인공지능의 판단에 근거해서 국가는 어디까지 개인의 삶에 개입할 수 있는가라는 윤리적인 문제가 발생한다.

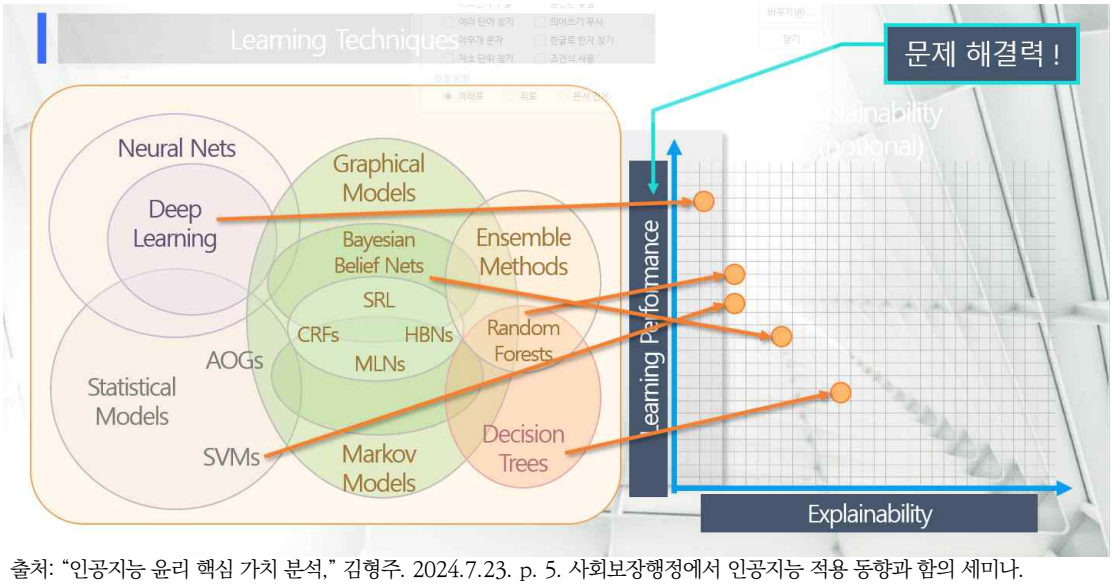
여섯째, 데이터와 알고리즘의 편향성의 문제다. 미국의 데이터 과학자인 Cathy O'Neil(2017)은 알고리즘이 현대 사회에서 대량살상무기(Weapons of Mass Destruction)에 준하는 위험성을 가졌다고 경고하기 위해 책 'Weapons of Math Destruction(대량살상수학무기)'을 출판한 바 있다. 그는 여기에서 편향된 데이터와 알고리즘이 취약계층에게 차별적인 결과를 낳게 된다고 경고했다. 그는 알고리즘에 대한 기술낙관론은 '테크노-유토피아'로 명명하면서, "알고리즘과 기술이 가져다 줄 혜택에 대한 무제한적이고 부적절한 희망에서 깨어나야 한다"(p. 343)고 강조한다.

일곱 번째, 인공지능의 '설명 불가능성'의 문제다. 2장에서도 살펴보았듯이, 인공지능의 복잡성을 고려할 때, 인공지능이 어떤 기준과 원칙에 따라 작동하는지에 대해서 인간은 이해할 수 있어야 한다. 이를 인공지능 관련 핵심원칙 가운데 하나인 설명 가능성(explainability)이다. 문제는 인공지능이 점점 더 설명가능하지 않다는 점이다. 인공지능이 고도화할수록 설명가능성은 떨어진다(〈그림 1〉 참고). 물론, **사회보장 분야에서는 인공지능이 활용된다고 하더라도 딥러닝 수준까지 활용되는 경우는 많지 않다**. 따라서, 인공지능의 설명가능성에 대한 문제가 적어도 한동안은 중요하지 않을 수도 있다. 그렇지만, 기술의 발전 속도를 고려한다면 인공지능의 설명 불가능성이 조만간 난제로 등장할 가능성도 있다. 사회보장 행정에서는 특히 취약계층에 대한 차별적인 결정이 이뤄질 가능성도 염두에 둬야 한다.

지금까지 살펴본 인공지능의 잠재적 위험성을 살펴봤다. 다만, 국외에서 제기되는 위험성을 고려할 때, **한국과 외국이 제도적 환경이 다르다는 점도 염두에 둘 필요**가 있다. 정세정 외(2023)에서는 세가지로 나누어서 설명하고 있다(pp. 213~214).

첫째, 한국에서는 지문 및 얼굴 정보가 포함된 전 국민 주민등록 데이터베이스 덕분에 개인인증 영역에서 압도적인 인프라를 구축했다. 전 국민의 일률적인 국민식별번호를 운영하는 국가는 한국이 거의 유일하다(성준호, 2016). 둘째, **해외에서는 개인정보의 영리적 집적 및 활용에 대한 우려**가 크다(Eubanks, 2018; Alston, 2019). 미국에서는 민간 보험회사에서 개인 데이터를 활용하는 빈도가 더 높다. 반면, 한국은 공공이 전국민 소득, 재산, 건강, 인적 데이터를 주도해서 관리하고 있다. 셋째, **다른 복지국가에서는 복지 제도와 성숙화 과정에서 디지털 기술을 급여의 부정 및 오류 수급 포착에 활용하는 경향**이 있다. 서구가 보편적 복지국가의 성숙기를 지난 이후에 나타나는 문제에 대응하는 과정에 있다면, 한국은 보편적 복지국가와는 거리가 멀다. 한국에서는 디지털 기술에 대해서도 초점이 사각지대 해소 혹은 위기가구 포착에 있다. 사회보장 영역에서 인공지능의 활용과 규제를 모색할 때, 이러한 한국적 특수성도 함께 고려될 필요가 있다.

〔그림 2〕 인공지능 윤리 관련 의제와 원칙



출처: "인공지능 윤리 핵심 가치 분석," 김형주. 2024.7.23. p. 5. 사회보장행정에서 인공지능 적용 동향과 함의 세미나.

4. 정책 방향

지금까지 우리는 인공지능 기술의 적용 현황 및 사회보장 영역에서 불러올 수 있는 순기능과 위험성을 두루 확인하고, 유럽연합의 인공지능법을 중심으로 규제 동향을 살펴봤다. 앞으로 정책 방향은 자연스럽게 도출된다. **위험성을 최소한으로 관리하면서 동시에 순기능을 최대화하는 방향**이다. 리스크 가능성을 규제하고, **편의를 증진하는 방향으로의 지원 사이에서의 적절한 정책적 배합이 중요**하다는 의미다. 여기에서 신기술에 대한 규제와 지원을 길항 혹은 모순 관계로 볼 필요는 없다. 오히려, **지원의 전제는 규제이고, 규제의 이유는 지원**이다. 규제와 지원 사이의 동적인 관계에 대해서 김병권(2024.12.26.)의 다소 긴 언급을 들어보자.

(정부의 인공지능 지원 편향적 정책에 대해) "우리는 얼른 인공지능을 경쟁력 있게 개발해야 하니 당연한 거 아니냐는 분들도 많다. 그러나, 누차 말하지만, 식품과 의약품은 정확한 안전성 규제를 해야 소비자들이 맘놓고 구입하여 먹을 수 있고, 이럴 때 진짜 식품산업과 제약업이 발전한다. 지금처럼 우리가 항공기를 엄청 자유롭게 타는 것은 엄청나게 까다로운 항공기 운항규제가 생겼기 때문이다 그 이전에는 워낙 사고가 많아 비행기는 모험가들만의 이용에 국한되었다. 지금 지구상에 15억대가 넘는 자동차가 굴러다니는 것은 오직 엄격한 '교통법규'와 '자동차 안전성 규제' 때문이다. 인공지능도 잘 사용하면 좋지만, 위험성도 함께 있기에 '제대로 안전성을 엄격히 규제'해야 번성할 수 있다."

유럽연합과 미국 행정명령 14110도 인공지능에 대한 규제와 지원 사이에서 일정한 균형을 갖췄다. 앞선 발표에서는 규제에 초점을 맞춰서 논의했지만, 지원을 강조한 다음의 대목들도 있다. 유럽연합 인공지능법은 "이 규정은 공공행정이 준수되고 안전한 AI 시스템의 광범위한 사용을 통해 이

익을 얻을 수 있도록 혁신적인 접근 방식의 개발과 사용을 저해해서는 안 된다."(Artificial Intelligence Act, (58))라고 명기하고 있다. 미국 행정명령에서도 7조 2항 (b)호에서, 보건복지부는 급여 및 서비스를 시행할 때 자동화 또는 알고리즘 시스템의 사용을 촉진하는 계획(plan)을 행정명령 시점 기준으로 180일 이내에 발표해야 한다"라고 규정했다.

안타깝게도 한국에서는 사회보장 영역에서 인공지능 적용에 관한 공적인 비전 혹은 계획이 제시된 바 없다. 물론, **2024년 12월 26일 한국에서 인공지능 기본법이 국회 본회의를 통과**한 상황을 고려할 필요는 있다. 그러나, **한국의 인공지능기본법에서 사회보장 영역에서의 인공지능 기술 적용을 촉진하거나 규제하는 내용이 반영되지는 않은 것으로** 보인다(고환경, 채성희, 손경민, 이일신, 2024)[4]. 유럽연합의 인공지능법과 미국의 행정명령에서 사회보장 영역을 비중있게 다룬 점과 대조된다. 한국의 사회보장 영역에서도 인공지능 기술 적용이 폭넓게 활용되고 있다. **한국의 인공지능기본법이 사회보장 영역을 간과한 점은 의아한 대목**이다. 법은 처음부터 이렇게 중요한 결함을 가지고 있다. 앞으로라도 인공지능 기본법에서 사회보장 영역을 포괄하는 개정을 시도하거나, 사회보장 영역에서의 인공지능 적용 및 규제에 대한 별도의 입법을 시도하는 방안을 검토할 수 있을 것이다. **정부의 행정 영역에서 사회보장 분야가 차지하는 비중을 고려하면 더욱 그러하다. 미국의 경우, 미국 연방정부에서 인공지능을 활용하는 사례 가운데 3분의 1 이상이 보건복지부 업무 영역에 속한다**(Burt, 2024.10.17.). 미국의 행정명령에서 지속적으로 '보건복지부'가 호명되는 이유다. 한국의 정부 지출 가운데 보건·복지·고용 영역 비율이 가장 높은 점은 주지의 사실이다. 2024년 기준으로 보면 전체 정부 예산 656.6조원 가운데 보건·복지·고용 분야가 237.6조원으로, 총지출의 36.1%를 차지하며, 비율은 앞으로도 증가할 것으로 예상된다(기획재정부, 2024). 이와 같은 점을 염두에 두고 앞으로 빅데이터와 인공지능의 사회보장 영역에서의 적용을 촉진하고 부작용을 선제적으로 예방하기 위한 정책 제언을 제시하도록 하겠다.

첫째, 사회보장 정보에 활용되는 데이터 품질을 개선하기 위한 정책적인 노력이 필요하다. 인공지능의 적 과정에서 쓰레기 데이터가 쓰레기 출력값을 산출하는(garbage in, garbage out) 문제가 제시된다(James, 2024). 데이터와 데이터에 근거한 인공지능의 결괏값에 대한 신뢰도를 높이기 위해서는 양질의 데이터 확보 및 집적, 관리가 필요하다. "사회보장제도에 대한 공정한 평가와 이를 기반으로 한 사회보장제도의 설계를 위해서는 정확한 사회보장제도 관련 정보가 필요하다. 따라서 정보 포괄성을 바탕으로 하는 정확한 데이터 품질 관리가 요구된다" (이상원, 2023, p. 3). 공적 데이터를 관리하는 기관들에 대한 인적, 행정적 인프라 개선 및 확충이 필요함을 의미한다.

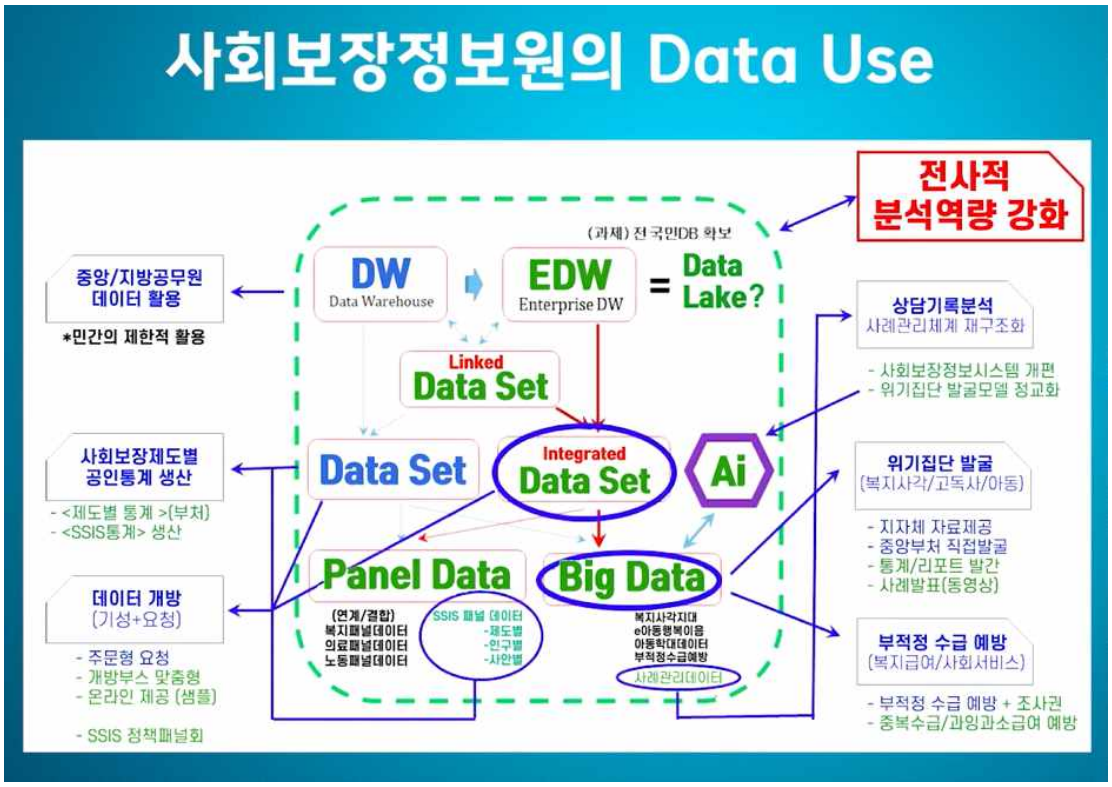
둘째, 사회보장 영역에서 데이터 통합, 연계, 관리를 위한 노력이 필요하다. 사회보장 영역에서는 2022년부터 사회보장기본법 42조1)에 근거해서, 사회보장 정책의 심의·조정과 관련 연구를 위해서 행정데이터가 수집되고 있다. 한국에서는 모든 개인에게 부여되는 주민등록번호로 기관들의 데이터를 연계할 수 있는 강력한 여건을 구축하고 있다. 그렇지만, 데이터 연계 과정에서 주민등록번호를

4) 2024년 12월 30일 현재 국회 본회의를 통과한 '인공지능 발전과 신뢰 기반 조성을 위한 기본법' 전문은 공개되지 않고 있다. 고환경 외(2024)가 소개한 인공지능법 요약 내용을 통해서 법의 내용을 미루어 짐작하는 수밖에 없다.

결합키 생성에 사용하지 못하도록 하는 법적인 규제가 발목을 잡고 있다. 더욱이 대부분의 정부 부처와 공공기관들도 데이터 공개에 소극적이다. 북유럽, 미국, 영국, 캐나다 등에서 행정자료 구축과 공개, 연계에 적극적인 점(유종성, 2022)을 참고할 필요가 있다. 한국에서도 노대명(2023)이 사회보장정보원을 중심으로 행정 데이터 연계와 활용 방안에 대한 구상을 제시한 바 있다(〈그림 2〉 참고).

셋째, 다양한 기관들이 집적한 정보들의 표준화 및 단순화가 필요하다(노대명, 2024). 이는 앞선 데이터 질 개선 및 연계와 직결되는 문제다. 노대명(2024)은 사회보장 정보시스템의 많은 데이터가 제대로 활용되고 있지 않다며, 사회보장정보원의 데이터 테이블은 11,280개에 달하지만 정작 자주 활용되는 정보는 500칼럼에 불과하다고 지적한다. 다른 부처와 기관이 생산하는 데이터들 역시 통합은커녕 연계가 힘든 상황이다. 현재로서는 구슬은 서말이지만, 이들을 꿰지는 못하고 있다.

〔그림 3〕 사회보장데이터의 정제와 활용



출처: 노대명 (2023) 한국사회보장정보원 행정데이터 자산화 프로젝트. 사회보장정보원 유튜브 화면 중 캡쳐 https://www.youtube.com/watch?v=zl7XhnJBDSM

넷째, 데이터 구축, 연계, 활용 과정에서 데이터 보안 및 안전에 대한 엄격한 기준 설정이 반드시 필요하다. 유종성(2023)은 행정 데이터 활용에서의 향후 과제를 제안하면서 다음과 같이 언급했다. "전수를 포괄하는 빅데이터인 경우가 많아 개인 정보 보호에 보다 세심한 주의가 필요하다. 개인정보 보호에 대한 안전장치를 이중 삼중으로 설치할 필요가 있다. 과거 스웨덴 등 북유럽 국가들에서도 데이터 활용과 개인정보 보호를 둘러싼 논쟁이 있었다. 개인정보 보호를 철저히 하면서 데이터를

연계하고 가명 처리된 데이터를 연구자들이 안전하게 사용할 수 있는 방법들이 발전되어 왔다" (p. 15). 개인의 소득, 재산, 건강, 가족 등의 개인 정보가 결합될 수록 데이터 유출에 따른 충격은 커질 수밖에 없다. 이 문제는 데이터 보안을 위한 데이터 가버넌스의 문제와도 직결된다.

다섯째, 데이터 관리를 넘어서 이를 활용하는 알고리즘 및 인공지능 시스템의 편향성을 최소화하면서, 정확성을 개선하기 위한 노력이 반드시 필요하다. 이 대목에서는 호주의 로보데트 스캔들(Robodebt Scandal)을 복기할 필요가 있다. 호주는 2016년부터 개개인의 복지수당 초과지급분의 계산 및 초과분 환수 통보를 자동화한 로보데트 시스템(Robodebt Scheme)을 도입했다(홍승헌, 황하, 2024). 이 과정에서 데이터 매칭 알고리즘이 핵심적 역할을 담당했지만, 47만건에서 오류가 발생한 것으로 추정됐다. 액수로는 총액 10억 호주 달러였다. 데이터에 기반한 알고리즘에 과신·과의존함으로써 수많은 사회적 약자들에게 해악을 끼친 대표적 사례다. 2020년 시스템은 공식적으로 폐지됐다(홍승헌, 황하, 2024). 정확성이 담지되지 않은 인공지능 기반 사회보장 행정이 초래할 수 있는 끔찍한 결과다. 특히, 인공지능을 활용한 행정에서 가장 흔하게 지목되는 리스크가 편견과 불평등의 심화(O'neil, 2017; Eubanks, 2018)라는 점을 상기할 필요가 있다. 인공지능 기반 행정에 대한 신뢰는 과학성, 중립성에 대한 입증을 통해서 조금씩 누적될 수 있을 것이다. 그리고, 그 신뢰가 무너지는 것도 한 순간이다.

여섯째, 한국의 정부 부처에서, 특히 사회보장 영역에 한정하자면, 보건복지부에서 인공지능 관련 조직과 인력을 신설 및 배치하는 안을 검토할 필요가 있다. 유럽연합의 인공지능법에 따르면, 고위험성 인공지능 시스템 제공자는 17개의 의무를 지게 된다(윤혜선, 2024)[5]. 물론, 유럽연합의 다소 엄격한 기준이 그대로 한국에 적용될 것으로 확언하기는 어렵다. 미국에서조차도 지난 행정부에서 미국 보건복지부가 행정명령에 따른 다양한 이행 조치를 행정명령이 강제했다. 이를테면, 'Office of the Chief Artificial Officer (OCAIO)'를 임명하고, 'HHS Artificial Intelligence (AI) Strategy'를 공표해야 했다. 한국에 보건복지부는 시스템 제공자 혹은 제공자의 관리감독 기관으로 해당 조치들을 적극적으로 검토할 필요가 있다. 지금까지 보건복지부에서는 보건 분야에서 일부를 제외하고는 사회보장 영역에서 인공지능과 관련한 규제 조치를 내놓은 바는 없다.

일곱째, 국제적인 인공지능 발전과 규제 동향에 대한 모니터링이 필요하다. 유럽연합의 인공지능법은 법 집행 과정을 거치면서 규제의 내용이 구체화할 가능성이 높다. 미국의 경우, 트럼프 행정부 집권 이후 바이든 정부의 행정명령이 폐지되면서 또 다른 격변을 거칠 것으로 예상된다. 여기에는 앞으로 인공지능 발전의 속도, 발전이 사회에 미치는 파장, 인공지능 패권을 둘러싼 미·중·유럽권의 경쟁, 국제기구의 개입 등이 복잡하고 역동적으로 작용할 것으로 예상된다. 무엇보다 중요한 변수는 정치인과 정책 전문가들마저도 인공지능이 정확히 무엇인지, 그래서 인공지능이 미칠 파장이 어떠

5) ① 고위험성 AI시스템 준수사항 준수 여부 확인, ② 고위험성 AI시스템에 연락처 표시, ③ 품질 관리 시스템 준수, ④ 기술문서, 품질관리시스템 문서 등 필수 자료 보관, ⑤ 통제 하에 있는 경우 자동으로 생성된 로그 보관, ⑥ 시장 출시/서비스에 투입되기 전 적합성평가 절차 준수, ⑦ EU 적합성 선언 작성, ⑧ CE 마크 부착, ⑨ 등록 의무 준수, ⑩ 필요한 시정조치(규정 위반시 적절한 조치) 및 정보 제공, ⑪ 국가 관할 기관의 합리적인 요청이 있을 때, 고위험성 AI시스템이 모든 준수사항에 부합함을 입증, ⑫ 제품 및 서비스에 대한 접근성 의무 보장, ⑬ 관할당국과의 협력, ⑭ 중요한 변경시 적합성 평가 실시, ⑮ 출시 후 모니터링 체계 설계·구축·운영, ⑯ 중대사고 관련 정보 공유 및 신고 의무, ⑰ AI 리터러시 보장 (윤혜선, 2024)

할지에 대해 잘 모르고 있다는 점이다. 놀라운 일은 아니다. 심지어 기술 전문가들 사이에서도 인공지능의 영향에 대해서 doomer와 boomer로 엇갈리고 있는 점을 상기할 필요가 있다. 물론, 미지(未知)보다 해소되기 어려운 문제는 상충하는 이해관계라는 점도 확인해둔다. 한국의 사회보장 영역에서 인공지능의 적용을 지원하고 규제하고자 한다면, 이러한 내·외 환경에 대한 동적인 모니터링이 필요할 수밖에 없다.

여덟째, 지금까지 제시한 데이터 관리, 연계, 표준화 및 알고리즘 질 관리 등을 총괄하는 가버넌스 체계 구축이 필요하다. 한국에서 행정 데이터 활용에 대한 논의는 데이터 활용도를 높이는 방향을 중심으로 논의되는 경향이 있다. 그러나, 앞서 호주의 Robodebt 사례에서 살펴본 바와 같이, 데이터를 활용한 복지행정에서의 오류 혹은 사고는 한번만 발생해도 전체 시스템을 폐쇄할 정도로 충격이 큰 점을 고려할 필요가 있다. 관련해서, 2024년 국회 본 회의를 통과한 인공지능 기본법에서는 규제의 주체로 과학기술정보통신부를 호명하고, 독립적인 규제 및 감독기구가 구체적으로 제시되지는 않은 것으로 보인다. 지금까지 공개된 초안에서는 인공지능사업자나 대통령령으로 정하는 인공지능기술 관련 기관이 '민간자율인공지능윤리위원회'를 '둘 수 있다'는 느슨한 규제를 제시하고 있다. 한국의 인공지능 기본법이 진흥과 규제 사이에서 진흥 쪽으로 과하게 경도됐다고 보는 것이 적절할 것이다. 김병권(2024.12.26.)이 지적한 대로, 새로운 법은 규제법은 아니고 그냥 '진흥법'이다.

앞서 언급한 대로, 제대로 된 규제 없이 인공지능은 발전할 수 없다. 사회보장 영역은 사람을 대상으로, 특히 빈곤, 아동, 노인 등 취약계층을 대상으로 한다. 그리고 이들의 사적인 소득, 건강, 재산 자료에 근거해서 정책을 편다. 유럽연합에서 금지하는 'social scoring'에 근접하다. 그래서, 인공지능 정책에서 시장 친화적인 정책을 펴는 미국에서도 지난 정권에서 보건복지부에 대해서는 보다 상세한 지침을 제시했다.

종합하면, 한국 공공영역에서 디지털 기술, 인공지능 기술 도입에는 빠른 반면, 관련 도덕적, 법적 쟁점에 대한 사회적 논의 및 규제 형성 과정은 매우 더디다. 사회보장 영역에서 국내의 인공지능 기반 기술 도입 속도는 다른 복지국가들에 견줘 전혀 늦지 않다. 이러한 점을 고려한 디지털 가버넌스 구축이 필요하다. 관련해서 유도진(2024.12.27.)은 인공지능의 민주적 활용을 위해 세가지 제도적 기구의 설치를 제안했다. 첫째, 인공지능 활용 정책과 데이터 사용의 적법성, 투명성을 감사하는 독립기구다. 여기에는 인권, 기술, 법률 등 다양한 전문가가 참여한다. 둘째, 정부 인공지능 활용에 대한 시민의 감시와 통제를 위한 옴부즈맨 프로그램이다. 신고 절차를 간소화해서 시민 참여를 확대하는 식의 접근이 제안됐다. 셋째, 인공지능 시스템의 알고리즘 공정성, 데이터 처리 투명성, 윤리적 설계를 평가하고 인증하는 독립인증기구다. 여기에서는 기술적·윤리적 위험 요소를 정기적으로 점검하고, 문제 발생 시 시정 조치를 요구할 수 있는 권한도 가지게 된다.

사회보장 행정에서의 인공지능 기술 적용에 대한 가버넌스는 인간중심적 활용의 원칙 아래 **① 민주적 통제, ② 시민 참여, ③ 프라이버시를 보장**하고, 동시에 시민의 적극적 사회권 보호라는 원칙 아래 **① 사각지대 해소, ② 행정 효율화 제고, ③ 급여 지급의 정확성, 적시성 보장의 편의를 제공**하는 정책 방향을 확인할 필요가 있다.

5. (부록) 인공지능 기술의 사회보장 분야 적용에 대한 규제: 유럽연합 인공지능법

전 세계 국가 및 지역 정부 및 국제기구에서 인공지능에 대한 규제를 내놓고 있다. 인공지능이 초래할 수 있는 사회적 위험이 거대할 수 있고, 동시에 비가시적, 불확정적이기 때문이다. 대표적인 예가 World Economic Forum(2023)의 Presidio Recommendations on Responsible Generative AI, European Union(2024)의 AI Act, OECD(2023)의 AI Principles, UNESCO(2021)의 Recommendation on the Ethics of AI, 미국 백악관의 Blue Prints for an AI Bill of Rights(White House, 2022), 미국 캘리포니아주에서 시도했던 AI 규제법안(SB 1047) 등이다. 여기에서는 유럽연합의 AI Act를 살펴보겠다. 다른 형태들은 대부분 법적 구속력이 없는 권고나 가이드라인의 형태를 띠기 때문이다.

5-1. EU 인공지능법(AI Act)의 개요

유럽연합의 인공지능법(Artificial Intelligence Act)은 세계 최초로 AI에 관한 포괄적인 법적 프레임워크다(European Commission, 2024). 인공지능법은 유럽연합에서 자주 사용되는 간접적인 방식인 지침(directive)이 아닌, 법(act)으로 제정됐다. 따라서 모든 회원국에 직접적이고 즉각적인 규제(regulations)로 적용되고, 국가별로 별도의 법률 제정 없이 효력을 발휘한다. 유럽연합 인공지능법의 목표는 유럽을 비롯한 세계 어느 지역에서도 신뢰할 수 있는 AI가 개발될 수 있도록 모든 AI 시스템이 인간의 기본권, 안전, 윤리 원칙 등을 존중하게 하고, AI 모델의 위험성을 관리하는 데 있다(European Commission, 2024). 따라서, 이 법은 구체적인 상황별로 AI의 개발자와 배포자가 지켜야 할 명확한 요건과 의무를 명시하고 있다. 물론, 법이 AI 개발을 지원하려는 의도도 담고 있다. 따라서 AI 관련 혁신 정책 패키지 및 AI 관련 협력 계획을 명시하고 있고, AI 개발 및 활용과 관련하여 기업, 특히 중소기업의 행정적 부담과 비용을 줄이는 방안도 포함한다(European Commission, 2024). 인공지능법은 모두 13개 장의 113개 규정과 13개의 부속서로 구성되어 있다. 각 장의 제목과 내용은 〈표 2〉과 같다.

인공지능법의 가장 큰 특징은 위험성 차원에서 AI 시스템을 분류한 다음, 그에 따라 규제의 내용을 차등화한다는 점이다. 인공지능 시스템이 내포하는 위험성의 강도와 범위에 비례하여 규제 유형과 정도가 규정되는 것이다.

인공지능법에서 핵심 개념인 위험성은, 인공지능이 개념 정의에 바로 뒤이어 다음과 같이 정의된다. "위험의 발생 가능성과 그로 인한 피해의 심각성의 결합"(Artificial Intelligence Act, Article 3, (2)). 유럽연합은 인공지능 모델의 성격과 용도 등을 고려해서 위험성의 경중을 진단하고, 그에 따른 규제의 수준도 연동한다.

이 법에서 AI 시스템의 위험성 수준을 네 가지로 구분한다(〈그림 3〉 참고). 네 가지 위험성 수준은 '수용할 수 없는 위험성(unacceptable risk)[6]', '고위험성(high risk)', '제한된 위험성(limited

6) 유럽연합의 인공지능법 본문에서는 '수용할 수 없는(unacceptable)'에 더해 '금지된(prohibited)'이라는 표현도 자주 등장

risk)', '최소한의 위험성(minimal risk)'이다. 가장 높은 수준의 '수용할 수 없는 위험성'을 가진 인공지능은 활용 자체가 금지된다. 다음으로 '고위험성'을 가진 인공지능 활용에는 엄격한 준수사항이 요구된다. 세 번째인 '제한된 위험성'을 내포하고 있는 AI의 활용은 투명성 의무가 부과되고, 최소한의 위험성이 있는 경우에는 자율적 규제가 따라붙는다. 위의 네 가지 위험성을 순서대로 하나씩 살펴보겠다.

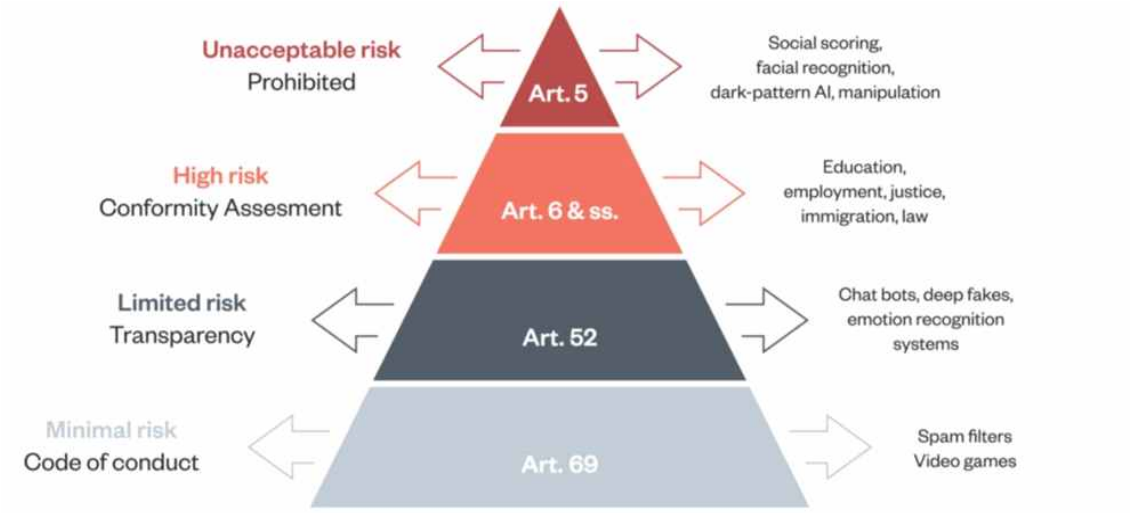
〈표 2〉 EU 인공지능법의 구성

장	내용
제1장	**총칙:** 법 적용 대상 및 범위, 용어 정의, AI 리터러시
제2장	**AI 활용 관련 금지행위:** AI 시스템의 활용이 금지되는 경우와 예외
제3장	**고위험 AI 시스템:** 제1절: AI 시스템의 고위험 분류 기준 제2절: 고위험 AI 시스템의 준수사항 제3절: 고위험 AI 시스템 제공자·배포자·기타 이해관계자의 준수사항 제4절: 승인기관(통보기관), 인증기관 제5절: 표준·적합성 평가·인증서·등록 기준
제4장	**특정 AI 시스템의 제공자 및 배포자에 대한 투명성 의무:** 사람과 직접 상호작용할 목적으로 사용되는 AI 시스템의 제공자 및 배포자의 투명성 의무
제5장	**범용 AI 모델:** 제1절: 범용 AI 모델의 분류 규칙 제2절: 범용 AI 모델 제공자의 준수사항 제3절: 시스템적 위험이 있는 범용 AI 모델 제공자의 준수사항
제6장	**혁신 지원 방안:** 규제 샌드박스 및 실제 조건에서의 고위험 AI 시스템 시험이 가능한 예외와 스타트업 등 중소기업에 대한 산업 진흥 제도
제7장	**거버넌스**: 제1절: 유럽연합 수준의 거버넌스 – AI사무국, 유럽AI위원회, 자문포럼, 독립전문가 과학패널 제2절: 국가 관할기관 – 각 회원국의 관할 기관 지정
제8장	**고위험 AI 시스템을 위한 EU 데이터베이스:** 유럽연합 집행위원회가 관리하는 고위험 AI 시스템에 대한 EU 데이터베이스
제9장	**사후 모니터링, 정보공유, 시장감독:** 제1절: 사후 모니터링 제2절: 중대한 사고에 대한 정보 공유 제3절: 시장감독기관 및 집행위원회의 규범 집행을 위한 수단 제4절: 구제수단 제5절: 범용 AI 모델 제공자에 대한 감독, 조사, 집행 및 모니터링
제10장	**행동규범 및 지침:** 고위험 AI 시스템을 제외한 시스템에 적용할 수 있는 행동규범 및 지침 마련
제11장	**권한의 위임과 위원회 절차:**집행위원회 권한 위임의 근거, 집행위원회 보조 위원회 근거 규정
제12장	**처벌:** AI 시스템에 대한 규범 위반, 범용 AI 모델에 대한 규범 위반에 대한 처벌 규정
제13장	**종칙**

주: 라기원 (2024)이 정리한 내용을 일부 수정·인용함.

한다. 또한 '제한된 위험성'은 투명성(transparency) 위험으로도 사용된다.

〔그림 4〕 유럽연합 인공지능법에서 규정하는 위험의 위계



출처: 자료: Ethical Intelligence(2021). Akhmedjonov, A. (2023). 재인용.

첫째, 가장 높은 수준의 '수용할 수 없는 위험성(unacceptable risk)'의 경우를 살펴보자. 수용할 수 없는 위험성은 AI 시스템이 사람들의 안전, 건강, 기본권에 명백한 위협이 되는 경우에 해당한다. 인공지능법 제5조는 수용할 수 없는 위험성이 있다고 간주하는 사례를 제 5조 1항에서 (a)~(h)에 걸쳐 여덟 가지를 제시했다. 여덟 가지를 짧게 요약하면 다음과 같다(Artificial Intelligence Act Article 5 (1)). 첫째, 사람의 의식을 넘어서는 잠재적인 기술을 활용하거나 의도적으로 조작적이거나 기만적인 기술을 사용해서, 개인 또는 집단의 행동을 실질적으로 왜곡해서 중대한 피해를 초래하거나 그럴 가능성이 있는 경우, 둘째, 연령, 장애 또는 사회경제적 환경에 기인한 취약점을 악용하여 행동을 왜곡함으로써 심각한 피해를 유발하거나 그럴 가능성이 있는 경우, 셋째 자연인 또는 집단의 사회적 행동이나 알려진, 추론된, 또는 예측된 개인적 또는 성격적 특성을 기반으로 일정 기간 평가하거나 분류하기 위해 사회적 평점(social scoring)를 사용한 경우다. 넷째, 프로파일링 또는 성격 특성만을 기반으로 하여 개인의 범죄 행위를 예측하는 경우, 다섯째, 인터넷이나 CCTV 영상에서 무차별적으로 얼굴 이미지를 스크랩하여 얼굴 인식 데이터베이스를 구축하는 경우, 여섯째, 의료 또는 안전을 제외한 목적으로 직장이나 교육 기관에서 개인의 감정을 추론한 경우, 일곱째 인종, 정치적 견해, 노동조합 가입 여부, 종교적 혹은 철학적 신념, 성생활, 성적 지향 등과 같이 민감한 개인 특성을 유추하기 위한 생체 인식 분류 시스템(biometric categorisation systems)인 경우, 마지막으로 법 집행 목적으로 공공장소에서 실시간 원격 생체 인식 정보('real-time' remote biometric identification) 수집한 경우다. 여덟 가지 가운데 일부는 사회보장 영역과 가장 밀접한 내용은 세 번째인 사회적 평점 부분이다. 관련 내용은 아래에서 살펴보겠다.

수용할 수 없는 위험성 관련한 금지 행위를 위반한 경우에는 최대 3,500만 유로(약 500원)와 전년 회계연도 기준 전 세계 연매출액의 최대 7% 가운데 더 높은 금액이 과징금으로 부과된다(윤혜선, 2024). 단, 중소기업이 위반했을 때는 위 두 기준 가운데 더 낮은 금액이 과징금으로 부과된다. 유럽

연합 소속 기관, 대리인, 조직의 경우에는 최대 150만 유로(약 20억 원)가 부과된다.

두 번째로 고위험성(high risk)을 살펴보겠다. 인공지능법 제6조 제3항에 따르면 고위험성을 가진 인공지능 시스템은 자연인의 건강, 안전 또는 기본권에 위해를 가할 중대한 위험을 내포하는 시스템이다. 인공지능법 "부속서 III"에서 제시된 여덟 개 영역의 고위험성을 요약하면 다음과 같다[7]. ① 생체인식 및 분류, 감정인식 시스템[8], ② 주요 사회기반시설, ③ 교육, 직업훈련 영역, ④ 고용, 근로자 관리 및 자영업자에 대한 접근, ⑤ 필수 민간 및 공공 서비스, ⑥ 인간의 기본권과 관련된 법 집행, ⑦ 이주, 망명 및 국경 통제 관리, ⑧ 사업절차 및 민주적 절차 등이다. 여덟 가지 가운데 ⑤ 필수 민간 및 공공 서비스 분야(essential private and public services)는 사회보장 영역과 직접적으로 연관된다. 해당 내용도 아래 4절에서 살펴보겠다.

고위험 인공지능 시스템은 시장에 출시되기 전에 준수사항으로는 다음의 일곱 가지가 제시된다(European Commission, 2024; 윤혜선, 2024). ① 적절한 위험 평가 및 완화 시스템 구축, ② 리스크 및 차별적 결과를 최소화하기 위한 고품질 데이터 구축 및 관리, ③ 결과의 추적 가능성을 보장하기 위한 활동 기록(logging), ④ 시스템과 그 목적에 대한 모든 정보를 제공하는 상세한 문서화.(규제 준수 여부 확인 목적), ⑤ 사용자에게 명확하고 적절한 정보 제공, ⑥ 위험을 최소화하기 위한 적절한 인간의 감독 조치, ⑦ 높은 수준의 견고성, 보안 및 정확성이다. 유럽연합에서는 인공지능 시스템을 둘러싼 이해관계자를 제공자, 배포자, 공인 대리인, 수입업자 등으로 분류하고, 각자에 대한 의무를 구체적으로 명기하고 있다. 이를테면, 제공자(provider)에게는 유럽연합 적합성 선언 작성, CE 마크 작성 등의 17개 의무가 부과된다. 사회보장 영역에서 작동하는 다수의 인공지능 시스템은 이러한 규제에 놓이게 될 가능성이 높다. 사회보장 영역에서 공공기관은 제공자의 위상을 가질 가능성이 높다.

유럽연합의 인공지능법에서 규정한 이해관계자들은 법의 적용 범위와 관련해서 중요한 의미가 있다. 인공지능법 2조는 법의 적용범위를 규정하면서, 일곱 가지 유형의 이해당사자 가운데 하나로 "유럽연합에서 AI 시스템 출시·서비스화 & 범용 AI 모델 출시하는 제공자"라고 규정했다. 그러면서 동시에 '장소불문'("irrespective of whether those providers are established or located within the Union or in a third country")(Artificial Intelligence Act Article 2 (1))이라는 조건을 달았다. 즉, 유럽에서 서비스를 제공하는 경우라면, 인공지능 시스템을 출시하거나 적용하는 업체는 국적을 불문하고 규제의 대상이 된다[9]. 국제적으로 인공지능 산업을 대부분 선도하는 미국의 업체들도

---

7) 유럽연합의 인공지능법에서는 고위험성 인공지능을 부록 I과 보록 III에서 각각 다르게 제시하고 있다. 부록 I에서 제시하는 내용은 사회보장 영역과 무관해서 별도로 설명하지는 않는다.

8) 생체인식 및 감정인식 시스템은 앞서 '수용할 수 없는 위험'으로 원칙적으로 금지되지만, 유럽연합 혹은 개별 회원국의 법에 따라 예외적으로 허용되는 경우는 '고위험군'으로 분류된다 (유럽연합 인공지능법 부록 III 1호). 공공장소에서 실시간 생체인식이 예외적으로 인정되는 경우로는 실종된 아동을 수색해야 하는 경우, 구체적이고 임박한 테러 위협을 방지해야 하는 경우, 중대한 범죄 행위의 가해자 또는 용의자를 탐지, 위치 파악, 식별 또는 기소해야 할 경우 등이다. 유럽 일부 국가에서는 인공지능 사용에 따른 프라이버시 침해 등의 논란을 우회하는 하나의 방식으로 이와 같은 인공지능 사용 목적에 보안(security)이나 안전(safety)을 중심에 두고 접근하는 경향이 일고 있다(van Bekkum, Borgesius, 2021).

9) 물론 예외도 다음과 같이 있다. ① 군사, 국방, 국가 안보 목적의 AI시스템, ② 유럽연합 또는 그 회원국과 사법 공조 협약을 체결한 제3국의 공공기관이 그 협약의 체계 안에서 사용하는 AI 시스템, ③ 과학적 연구 및 개발 목적으로만 특별히 개발되고, 서비스가 제공되는 AI 시스템과 AI 모델, ④ 출시 및 서비스 개시 전 AI 시스템을 연구, 테스트, 개발하는 활동, ⑤ 고위험성 AI 시스템이 아닌 무료 및 오픈소스 라이선스로 출시된 AI 시스템, ⑥ 사적 및 비전문적인 활동, ⑦ 일부 목

유럽연합에 와서는 규제를 따라야 한다. 이는 한국에도 간접적으로 영향을 미칠 수밖에 없다.

이러한 고위험성 인공지능시스템 준수사항을 위반하면, 최대 1,500만 유로(약 218억원) 혹은 전년 회계연도 기준 전 세계 매출액의 최대 3% 중에 더 높은 금액이 부과된다(European Commission, 2024; 윤혜선, 2024). 중소기업이 고위험 관련 조항을 위반하면 두 기준 가운데 더 낮은 금액이 부과된다. 유럽연합의 기관, 에이전시, 기구의 경우, 과징금은 최대 75만 유로(약 10억원) 로 한정된다.

5-2. 유럽연합 인공지능법이 사회보장에 미칠 영향

유럽연합의 인공지능법에서 사회보장과 가장 연관된 가장 민감한 대목[10]은 '수용할 수 없는 위험성(unacceptable risk)'에서 세 번째로 제시된 사회적 평점(social scoring)이다. 사회적 평점은 "개인의 데이터를 기반으로 개인을 평가하는 것을 의미하며, 여기에는 신용 행태, 교통 위반, 사회적 참여 등이 포함된다. 이러한 평가는 특정 서비스나 특권에 대한 접근을 규제하기 위해 사용된다"(Mosene, 2024). 유럽연합의 인공지능법을 본문 그대로 번역하면 다음과 같다(European Parliament, 2024, Chapter II, Article 5, 1. (c)).

"자연인 또는 집단의 사회적 행동이나 알려진, 추론된, 또는 예측된 개인적 또는 성격적 특성을 기반으로 일정 기간 동안 대상을 평가하거나 분류하기 위한 목적으로 인공지능(AI) 시스템을 시장에 출시하거나, 서비스에 투입하거나 사용하는 행위로서, 이러한 사회적 평점이 다음 중 하나 이상의 결과로 이어지는 경우:

(i) 데이터가 원래 생성되거나 수집된 맥락과 무관한 사회적 맥락에서 특정 자연인 또는 집단에 대해 불리하거나 부정적인 대우를 초래하는 경우.

(ii) 특정 자연인 또는 집단의 사회적 행동 또는 그 행동의 심각성과 비교하여 부당하거나 과도하게 불리하거나 부정적인 대우를 초래하는 경우"

유럽연합의 인공지능법은 사회적 평점을 논의하면서 사회보장 제도와 관련한 언급을 하지는 않았다. 유럽의회 산하의 연구기관인 유럽의회연구소(European Parliamentary Research Service)가 인공지능법을 설명하는 짤막한 해설서(Madiega, 2024)에도 사회보장제도에 대한 언급은 없다. 그럼에도, 사회적 평점 부여가 사회보장과 일정한 연관을 가질 수밖에 없다. 개인 혹은 가구 단위의 소득, 재산, 가구원 등의 정보에 근거해서 빈곤, 실업, 은퇴, 상병 여부를 판단하고, 그에 근거해서 급여를 제공하는 사회보장제도는 급여 자격을 판정하는 과정에서 일종의 사회적 평점(social scoring)을 개인 혹은 가구에게 부여할 수밖에 없다. 이를테면, 국내의 국민기초생활보장제도에서도 가구의 소득, 재산, 부양의무자, 가구원 정보 등에 기반해서 수급자격 및 급여액을 결정한다. 그러한

적을 위한 법집행기관(경찰, 이민 당국 등)의 AI 시스템 사용 등에는 인공지능법 적용이 제외되거나 특례가 인정된다(윤혜선, 2024).

10) 한가지 확인한 점은 있다. 해당 주제에 대한 분석을 담은 학술논문이나 보고서는 아직 찾기는 어렵다. 현재로서는 법률 내용, 관련 시민단체 성명, 국내 법률 전문가 자문 등을 중심으로 관련된 영향을 추정하는 수밖에 없었다.

자료의 내용이 개인의 소득, 연령, 가구, 건강 등 개인적 정보를 담고 있다는 점에서 사회적 평점은 민감할 수밖에 없는 의제다.

유럽연합의 인공지능법 제정 과정에서도 사회적 평점 관련한 우려가 제기됐다. 유럽의 인권 단체인 Human Rights Watch(2023.10.9.)는 다른 시민단체들과 함께 유럽이사회와 유럽의회에 사회적 평점 관련 규정을 강화하는 제안을 하면서 사회보장제도에서 나타날 문제를 다음과 같이 언급했다. "프랑스, 네덜란드, 오스트리아, 폴란드, 아일랜드에서의 조사 결과, AI 기반의 사회적 평점 부여 시스템이 사람들의 사회보장 지원 접근을 방해하고, 프라이버시를 침해하며, 빈곤에 대한 고정관념과 차별적인 방식으로 데이터를 프로파일링하고 있다고 드러났다.(Human Rights Watch, 2023.10.9.)" 이러한 문제들은 시민단체나 학계를 중심으로 꾸준히 제기됐다. 앞서 살펴보았듯이, 네덜란드(van Bekkum, Borgesius, 2021)와 덴마크(Jørgensen, 2021)외에도 프랑스(La Quadrature Du Net, 2023) 등에서도 논란이 일었다. 이를테면, 정부에서 복지급여의 부정수급을 탐지하거나 위험가구를 발굴하는 과정에서 개인 정보를 무리하게 활용하거나(Appelman et al.,2021), 알고리즘이 취약계층에게 편향적으로 작동했다(Van Bekkum, Borgesius, 2021)는 지적이다.

유럽연합도 사회적 평점과 관련한 비판 및 부작용을 염두에 둔 것으로 보인다. 법률에 붙은 (i)와 (ii)의 내용은 인공지능의 사회적 평점이 '수용할 수 없는 위험'으로 간주되는 경우를 한정했다. 즉, 특정 집단이나 개인에게 불리하거나 부정적이거나 부당한 대우를 초래할 때만 사회적 평점이 금지된다. 바꾸어 말하면, 사회보장제도에서 인공지능의 작동이 '순기능'을 하는 경우에는 규제의 대상으로 삼지 않겠다는 의미로 해석된다. 그렇지만, 논란이 종료되기는 어렵다. 사회보장 영역에서 인공지능이 급여 대상자를 판단 혹은 추론하는 과정에서 데이터 편향성 등의 문제로 특정 집단에 대한 급여를 변경/중지/중단한다면, 이는 불리/부정/부당할 수 있다(강지원, 2024). 실제로, 덴마크나 네덜란드 등 다수의 국가에서 사회보장제도에 순기능 할 것으로 기획된 인공지능 알고리즘들이 결과적으로 차별과 편향을 낳았다는 비판에 직면했다(Jørgensen, 2021, Bekker, 2021). 이 대목은 앞으로 인공지능 기술이 사회보장 영역에서 적용되는 과정에서 끊임없이 논점으로 부각될 것으로 예상된다.

다음으로 인공지능법에서 두 번째로 수위가 높은 '고위험(high risk)' 인공지능 영역을 보겠다. 여기에서는 다섯 번째 고위험 인공지능 영역으로 '필수 민간 및 공공 서비스 분야'가 제시됐다. 이 영역은 사회보장 영역을 직접적으로 언급하고 있다. 고위험 인공지능을 영역별로 상술한 유럽연합 인공지능법 부록(Annex) III에서 사회보장을 다음과 같이 언급했다(Artificial Intelligence Act Annex III, 5).

"필수 민간 서비스 및 필수 공공 서비스와 혜택에 대한 접근 및 이용[11]:

11) 급여 수급에 관한 내용을 담은 (a) 외에도 보건의료 영역에서 다음과 같은 언급도 있다. 사회보장 영역과 일정한 연관성이 있지만, 이번 글에서는 해당 내용까지 다루지는 않는다. "(c) 생명 및 건강 보험의 경우, 자연인에 대한 위험 평가 및 가격 책정을 위해 사용되는 AI 시스템; (d) 자연인의 긴급 호출을 평가 및 분류하거나, 경찰, 소방관, 의료 지원을 포함한 긴급 대응 서비스의 출동 우선순위를 결정하거나 출동하는 데 사용되는 AI 시스템, 그리고 응급 의료 환자 분류 시스템"

(a) 공공기관 또는 공공기관을 대신하여 사용될 목적으로, 자연인의 필수적인 공공 복지 급여[12] 및 서비스(의료 서비스 포함)에 대한 자격을 평가하거나, 해당 혜택 및 서비스를 부여, 축소, 취소 또는 회수하기 위해 사용되는 AI 시스템"

'고위험'으로 분류한 '필수 민간 및 공공서비스'는 앞서 살펴본 사회적 평점(social scoring)과 밀접하게 연관됨을 알 수 있다. 급여를 제공하는 사회보장 영역에서 자격 요건을 판정하기 위해서는 사회적 평점 산정이 대부분 필요하기 때문이다. 유럽연합법은 관련 고위험에 대해서 다음과 같이 상세하게 설명했다(Artificial Intelligence Act, (58)).

"필수적인 공적 복지급여와 서비스를 신청하거나 받는 자연인은… 공공기관 앞에서 취약한 위치에 있다. 이러한 급여와 서비스가 지급, 거부, 축소, 취소 또는 회수되어야 하는지 여부를 결정하기 위해 인공지능 시스템이 사용된다면, 해당 시스템은 사람들의 생계에 상당한 영향을 미칠 수 있고, 사회보호에 대한 권리, 차별 금지, 인간의 존엄성 또는 효과적인 법적 구제와 같은 기본권을 침해할 수 있으므로 고위험으로 분류되어야 한다."

여기까지만 보면, 유럽연합의 입장은 사회보장제도에서의 인공지능 적용에서 엄격한 입장으로 보인다. 앞서 살펴본, 다소 강한 규제 내용이 사회보장 영역에서 적용될 여지도 크다. 그렇지만, 법의 다음 문장에서는 현행 사회보장제도에 인공지능 기술 적용의 가능성도 열어둔다.

"이 규정은 공공행정이 준수되고 안전한 AI 시스템의 광범위한 사용을 통해 이익을 얻을 수 있도록 혁신적인 접근 방식의 개발과 사용을 저해해서는 안 된다. 단, 해당 시스템이 법적 및 자연인에게 고위험을 초래하지 않아야 한다."(Artificial Intelligence Act, (58)).

사회보장 행정을 집행하는 정부나 공공기관은 고위험 인공지능 시스템 과정에서 '제공자(provider)' 혹은 '배포자(deployer)'로 구분될 수 있다(강지원, 2024). 제공자와 배포자는 고위험성 인공지능 활용에 있어서 각각 17가지와 13가지의 의무를 이행해야 한다. 이를테면, 데이터보호 영향평가, 기본권 영향 평가(fundamental rights impact assessment) 등이 예가 된다. 정부 및 공공기관 입장에서는 부담이 크다.

고위험 인공지능 관련 규정이 유럽 사회 및 사회보장에 미칠 영향에 대해서 유럽 사회는 다소 유보적인 입장으로 보인다. 독일 사회보험 유럽대표부 (Deutsche Sozialversicherung Europavertretung, 2024)는 "인공지능법 도입 이후 사회적 영향을 면밀히 모니터링해야 한다"고 논평했다. 이러한 반응은 아직 인공지능법의 내용이 전반적으로 모호하고, 구체적인 후속 조치도 아직 이루어지지 않았기 때문으로 보인다. 물론 유럽 노동조합 측에서는 인공지능법이 관련 대기업에 빠져나갈 구멍을 만들어주었다는 비판도 가하고 있다(Vranken, 2023; Del Castillo, 2023). 따라서 인공지능법규제의 실질적인 강도는 향후 추가로 마련될 가이드라인, 기준, 표준, 판례 등의 내용에 따라 결정될 것으로 보인다. 앞으로 인공지능법의 구체화를 위한 논의와 협상에서 기업, 노동자 단체, 인권 단체, 관료 집단 등 이해당사자 사이에 첨예한 대립과 갈등이 발생할 가능성이 크다. 그

12) 인공지능법에서는 'public assistance benefit'라는 용어를 사용했고, 사회보장 영역에서는 이를 공공부조 급여로 번역되는 것이 적정할 듯 하지만, 법의 맥락에서는 공공부조에 한정되지 않는 공적인 복지급여 전체를 지칭하는 것으로 보았다. 참고로, 유럽에서는 공공부조에 대해서 social assistance라는 표현을 더 일반적으로 사용한다.

결과에 따라 이 법이 사회보장에 미치는 의미와 영향력이 달라질 것으로 보인다.

〈참고문헌〉

Algorithm Watch and Bertelsmann Stiftung. (2020) Automating Society Report 2020. [online] Available at: https://automatingsociety.algorithmwatch.org

Alston, P. (2019). Digital Welfare States and Human Rights. Report of the Special Rapporteur on Extreme Poverty and Human Rights. https://undocs.org/A/74/493에서 2024.6.14. 인출.

Appelman, N., Fathaigh, R. O., & van Hoboken, J. (2021). Social welfare, risk profiling and fundamental rights: The case of SyRI in the Netherlands. J. Intell. Prop. Info. Tech. & Elec. Com. L., 12, 257-271.

Bekker, S. (2021). Fundamental rights in digital welfare states: The case of SyRI in the Netherlands. Netherlands Yearbook of International Law 2019: Yearbooks in International Law: History, Function and Future, 289-307.

Bendixen, M. (2018) Denmark's 'anti-ghetto' laws are a betrayal of our tolerant values. The Guardian. [online] Available at: https://www.theguardian.com/commentisfree/2018/jul/10/denmark-ghetto-laws-niqab-circumcision-islamophobic

Burt. C. (2024.10.17.). HHS Strategic Plan for AI Coming Soon. ExecutiveGov. https://executivegov.com/2024/10/department-of-health-and-human-services-strategic-ai-plan/에서 2024. 12.28. 인출.

Cesareo, S.; White, J. (2023). The Global AI Index. Tortoise Media

Davies, G. (2022). Report on Accounts: Department for Work & Pensions. London: National Audit Office.

Del Castillo, A. (2023). The AI Act: deregulation in disguise. Social Europe. https://www.socialeurope.eu/the-ai-act-deregulation-in-disguise

Deutsche Sozialversicherung Europavertretung. (2024). Conference highlights the synergy between AI and the European Pillar of Social Rights. https://dsv-europa.de/en/news/2024/03/ki-in-sozialer-sicherheit.html

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. NY: St. Martin's Press.

European Commission. (2024). AI Act. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai에서 2024. 12. 1. 인출.

Human Rights Watch. (2023.10.9.). EU: Artificial Intelligence Regulation Should Ban Social Scoring. Human Right Watch. https://www.hrw.org/news/2023/10/09/eu-artificial-intelligence-regulation-should-ban-

social-scoring에서 2024.12.15. 인출.

INSS. (2023). Helô, assistente virtual do INSS completa três anos. https://www.gov.br/inss/pt-br/noticias/helo-assistente-virtual-do-inss-completa-tres-anos에서 2024.8.21. 인출.

ISSA. (2020). ELYA: The bilingual virtual assistant of the Employees Provident Fund - Empowering customers to self-serv, anytime, anywhere. https://www.issa.int/gp/208625

ISSA. (2021). The application of chatbots in social security: Experiences from Latin America (Analysis). Geneva: International Social Security Association

James, G. (2024). Concerns over AI data quality gives new meaning to the phrase "garbage in, garbage out". SC World. https://www.scworld.com/perspective/concerns-over-ai-data-quality-gives-new-meaning-to-the-phrase-garbage-in-garbage-out

Jørgensen, R. F. (2021). Data and rights in the digital welfare state: the case of Denmark. Information, Communication & Society, 1-16.

La Quadrature du Net. (2022). CAF : le numérique au service de l'exclusion et du harcèlement des plus précaires. https://www.laquadrature.net/2022/10/19/caf-le-numerique-au-service-de-lexclusion-et-du-harcelement-des-plus-precaires/에서 2024.12.15. 인출.

Madiega, T. (2024). Briefing: Artificial Intelligence Act. Brussel: European Parliamentary Research Service.

Mchangama, J. and Liu, H.-Y. (2018). The Welfare State Is Committing Suicide by Artificial Intelligence. Foreign Policy. https://foreignpolicy.com/2018/12/25/the-welfare-state-is-committing-suicide-by-artificial-intelligence/

Mosene, K. (2024). One step forward, two steps back: Why artificial intelligence is currently mainly predicting the past. Digital Society Blog. Humboldt Institut für Internet und Gesellschaft. https://www.hiig.de/en/why-ai-is-currently-mainly-predicting-the-past/

O'Neil, C. (2017.). 대량살상수학무기. (김정혜 옮김). 흐름출판. (Original work published 2016)

OECD. (2023). OECD AI Principles overview. Paris: OECD.

OECD. (2024), "2023 OECD Digital Government Index: Results and key findings", OECD Public Governance Policy Papers, No. 44, OECD Publishing, Paris, https://doi.org/10.1787/1a89ed5e-en.

Petrosyan, L., Ataliotou, K. (2024). A Tale of Two Policies: The EU AI Act and the U.S. AI Executive Order in Focus. https://trilligent.com/a-tale-of-two-policies-the-eu-ai-act-and-the-us-ai-executive-order-in-focus/에서 2024.12.16. 인출.

Rechtbank Den Haag. (2020). SyRI legislation in breach of European Convention on Human Rights. de Rechtspraak. [online] Available at: https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank

-Den-Haag/Nieuws/Paginas/SyRI-legislation-in-breach-of-European-Convention-on-Human-Rights.aspx

Thapa, E.P. (2019). Predictive Analytics and AI in Governance: Data-driven government in a free society – Artificial Intelligence, Big Data and Algorithmic Decision-Making in government from a liberal perspective. European Liberal Forum.

Tortoise Media. (2024). Global AI Ranking. https://www.tortoisemedia.com/intelligence/global-ai#rankings에서 2024.12.14.. 인출.

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO.

van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. European Journal of Social Security, 23(4), 323-340.

Vranken, B. (2023). Big Tech lobbying is derailing the AI Act. Social Europe. https://www.socialeurope.eu/big-tech-lobbying-is-derailing-the-ai-act에서 2024.12.14. 인출.

White House. (2022). Blue Prints for an AI Bill of Rights. Washington: White House.

World Economic Forum. (2023). The Presidio Recommendations on Responsible Generative AI. Geneva: World Economic Forum.

Zaber, M., Casu, O., Brodersohn, E. (2024). Artificial Intelligence in social security organizations. International Social Security Association.

강지원. (2024). 정부의 사회보장 서비스에서 AI의 활용과 EU/미국 일부 주 법의 시사점. 한국보건사회연구원 사회보장적용에서의 인공지능 적용동향과 함의 세미나 발표 자료.

강혜규, 강지원, 강희정, 김기태, 김세진, 김태완.. 주하나. (2022). 사회보장제도 실태분석 및 개선방안 연구. 보건복지부, 한국보건사회연구원.

고환경, 채성희, 손경민, 이일신. (2024). 인공지능기본법 국회 본회의 통과. 법인법인 광장 Newsletter. https://www.leeko.com/newsl/techai/202412/202412.pdf

기획재정부. (2024). 2024~2028년 국가재정운용계획 주요내용. 기획재정부.

김기태, 김명주, 김은하, 신영규, 변소연. (발간예정). 사회보장행정에서의 인공지능 적용 동향 및 함의. 한국보건사회연구원.

김기태. (2024). 디지털 복지국가의 개념과 논점. 비판과 대안을 위한 사회복지학회 학술대회 발표논문집, 133-145.

김명주. (2024). AI 윤리와 규제. 한국보건사회연구원 사회보장적용에서의 인공지능 적용동향과 함의 세미나 발표 자료.

김병권. (2024.12.26.). 한국의 처참한 AI법. 페이스북 게시글. https://www.facebook.com/byoungkweon.kim

김수영, 김수완. (2022). 데이터 복지국가의 도래와 쟁점: 국가, 시민, 시장의 관계 지형을 중심으로. 한국사회복지정책학회 춘계학술대회자료집, 2022, 91-118.

김수영. (2016). 사회복지정보화의 윤리적 쟁점-사회보장정보시스템을 통한 데이터감시를 중심으로-. 한국사회복지학. 68(1), 193-224.

김형주 (2024.7.23). 인공지능 윤리 핵심 가치 분석: 한국 사례를 중심으로. 사회보장행정에서 인공지능 적용 동향과 함의 세미나 발표문.

노대명. (2023). 한국사회보장정보원 행정데이터 자산화 프로젝트. 유튜브 발표자료. https://www.youtube.com/watch?v=zl7XhnJBDSM

노대명. (2024). 디지털 전환과 한국 사회보장제도의 다가오는 과제. 2024.2.15. 한국보건사회연구원 빈곤불평등연구실 세미나 발표 자료.

라기원. (2024). 유럽연합 인공지능법(EU AI ACT)의 특성과 쟁점: 우리나라 인공지능 입법에 대한 시사점. AI Issue Paper 24-20-3.한국법제연구원.

보건복지부. (2024.11.26.). 보건의료데이터와 인공지능이 열어가는 디지털 헬스케어 미래. 보도자료. 세종: 보건복지부.

성윤희. (2022). 인공지능 (AI) 과 사회복지법제에 대한 소고. 사회복지법제연구, 13(2), 119-148.

성준호. (2016). 주민등록번호에 의존한 본인확인제도의 문제점: 각국의 개인식별번호제도 및 관련 법률의 검토를 통한 시사점. 공공사회연구, 6(2), 208-246.

유도진. (2024.12.27.). AI 기본법, 기술과 민주주의의 균형, 해결해야 할 과제들. Ohmynews. https://www.ohmynews.com/NWS_Web/View/at_pg.aspx?CNTN_CD=A0003092112&CMPT_CD=P0010&utm_source=naver&utm_medium=newsearch&utm_campaign=naver_news

유종성. (2023). 사회보장 행정데이터 활용사례와 향후 과제. 보건복지포럼, 325, 6-19.

윤혜선. (2024). EU 인공지능법의 주요 내용과 함의. 한국보건사회연구원 사회보장적용에서의 인공지능 적용 동향과 함의 세미나 발표 자료.

임덕영. (2023). 사회보장 현장 모니터링: 현장 전문가와 실무자 포럼. 세종: 한국보건사회연구원.

정세정, 김기태, 곽윤경, 우선희, 최준영, 이영수. (2023). 한국 복지국가의 재구조화를 위한 연구 - Ⅰ. 디지털 복지국가의 딜레마. 세종: 한국보건사회연구원.

정유채. (2022). AI 와 블록체인 기반 의료 복지 서비스 사례 연구 1. 복지경영학연구, 11, 77-86.

조남경, 송기호. (2023). 사회복지의 상담기록, 좀 더 활용할 수 있을까?'머신러닝'을 통한 사회복지 상담 텍스트 활용 가능성의 점검. 한국사회복지조사연구, 79, 5-26.

함영진, 이현주, 어유경, 김가희, 박성준, 조용찬… 오민수. (2023). 복지 전달체계 혁신을 위한 대안적 고찰: 취약계층 발굴정책 개선을 중심으로. 세종: 한국보건사회연구원.

홍승헌, 황하. (2024). 누구를 위한 디지털 전환인가?: 자동화된 복지행정의 위험성.정부학연구,30(2), 61-84.

session 10

Techno-Affect and AI Ethics: An Ethnographic Study of a Data Annotation Team for a Novel-Generating AI

SO YEON LEEM
(DONG-A UNIVERSITY)
MAY 28 2025

Outline

1. Introduction: DFC Lab
2. The Ethics of Generative AI: From Copyright Concerns to Data Extractivism
3. The “Corpus Team” at DFC Lab
4. Affect in the “Corpus Team”
5. Techno-Affect and AI Ethics
6. Conclusion: Rethinking AI Ethics

Introduction: DeepFiction Convergence Lab

A Convergence Study for Deep-Learning-Based AI Fiction Generation with Human in the Loop
Supported by the National Research Foundation in South Korea (2023-2026)

- **Multi-disciplinary AI research team led by an English literature scholar**

- One telecommunications engineer (in charge of Model A)
- One computer scientist (in charge of Model B)
- Three scholars of English Literature
- One scholar of Linguistics
- Two scholars in digital humanities
- One STS scholar (myself) as "embedded" ethnographer

Model B: Structuralist Novel-Generating AI

- **What Makes a Novel?**

- According to structuralist theory, a novel must contain a narrator and a focalizer.

- As the PI notes, current AI-generated texts may appear to have narrative perspective, but this is merely *"an illusion created by large language models,"* or *"a magic of language,"*

- **Goals of the DFC Lab (Model B)**

- To develop an AI model trained on data annotated with structural narrative elements, such as plot, character, emotion, perspective, and setting
- To write "good novels" based on structuralist(humanistic) knowledge, rather than relying on the illusions created by current LLMs

- To build a high-quality, semantically rich corpus for training novel-writing AI systems

The Core of Generative AI Ethics: Training Data

- **Generative AI ethics and copyright issues**

- Ethical concerns around generative AI have primarily focused on bias, discrimination, and hate speech.

- The emergence of LLM-based generative AI led to earlier discussions on copyright for AI-generated outputs.

- Technical importance of training data copyright: the autoregressive nature of LLMs (Franceschelli & Musolesi, 2023)

- **From copyright to data extractivism**

- Paying royalties or invoking fair use provisions is not enough!

"Everything is treated as data to be fed into a function and absorbed to improve technical performance. This is the central premise of the ideology of data extractivism." (Crawford, 2021: 121)

The "Corpus Team" at DFC Lab

- **Team members**

- Majoring in literature (mostly English and Korean literature) at top-tier universities in South Korea.

- Usually 7–9 members: 2 men and 5+ women

- **Two main tasks of the Corpus Team**

- Character personality annotation

- Plot annotation.

Data Annotation by The “Corpus Team”

- Character personality annotation tasks

1) Reading novels
2) Identifying the main characters
3) Analyzing their personalities using a psychological personality taxonomy called X (a pseudonym)
4) Scoring character personalities on numerical scales (-3~+3) according to X

- Plot annotation tasks

1) Identifying ten elements of a novel's plot
2) Classifying each unit of the novel according to its plot element
3) Selecting a representative sentence from each unit that reflects the assigned plot element

Anxious Data Annotators

“Since the AI wouldn’t adjust to me, I had to adjust to it. I was suddenly asked to build a completely new kind of relationship with literary texts—one that felt unfamiliar and unnatural. I experienced a surge of anger and frustration: *This isn’t right.* As someone who has trained in literary studies, I’ve gone through a painful, rigorous process to internalize how to engage deeply with texts. But now I had to disregard all of that and relate to literature in a fragmented and superficial way. I kept asking myself: *Is this okay? Isn’t this lacking and inaccurate compared to how I’ve been taught to read literature?* Because this was an entirely new experience for me, I initially felt overwhelmed by negative affects.”

VS. the initial "great expectations”

: Joined the team seeking to bring humanistic experimentation into the heart of AI development

Alignment between Data Annotators and Data

"I realized where I needed to compromise, and that changed my mindset. It wasn't so much that I had to unilaterally change myself; rather, I found a point within the collaboration with AI that felt like the best possible outcome."

"Just as food labels show the origin of their ingredients, I believe AI data should indicate its source too. Data needs origin labeling. It's a way of respecting the bodies of data workers. Only when such aspects are considered can we build trustworthy and ethical AI."

"It was an opportunity to rethink distant reading. I appreciated being able to step back and relativize close reading, which had become overly canonized."

"It's not a fundamentally failed project."

"Whenever I meet my friends from engineering, I say, 'Aren't you scared AI might replace you? I'm not—because I'm in the humanities.'"

Techno-Affect and AI Ethics(1)

- "Techno-affect" (Amrute, 2019)

"...an intense attachment that produces **an alignment between a specific technological formation and a particular kind of subject.**(Amrute, 2023: 180)"

E.g., elite Indian women programmers in the U.S. adapt to shifting immigration and labor regimes while shouldering emotional burdens—often leading to affective states close to depression

- "Techno-ethics" activated through "techno-affect" (Amrute, 2019)

- Attending to how technologies and subjects align and realign enables ethical practices that critically evaluate socio-technical realities and imagine alternative futures

- "Mess and emotion" in algorithmic audits (Keyes & Austin, 2022)

- "Aspirations in data annotation" in ethical AI (Wang, Prabhat & Sambasivan, 2022)

Techno-Affect and AI Ethics(2)

- "Virtue Ethics" (Harris, 2008)

- Limitations of rule-based and preventive ethics

- Emphasis on both technical and non-technical virtues: socio-technical sensitivity, respect for novels, and commitment to the public good

- "Ethics of Care" (Gray & Witt, 2021)

"...as text researchers we learned to treat the texts with care, as living data with consequences for individuals rather than purely as data items abstracted from producers, receivers, social practices and consequences." (Leedham et al., 2021: 5)

- Beyond the ELSI Model

- Not an ethics that lags behind technology or merely cleans up its problems

- Not a "clean-up" or "catch-up" model of ethics

Conclusion: Rethinking AI Ethics

- AI Ethics as Ethico-Onto-Epistemology (Barad, 2007)

: Ethics are not an add-on to matter but are already embedded in the very process of mattering.

- AI Ethics as "Science with Ethics"

"Ethics are proactive to and constitutive of science; they are 'at the heart of innovation itself.'" (Thompson, 2013: 221)

- AI Ethics as a Politics of Care (Puig de la Bellacasa, 2017)

- An ethics that includes the "neglected things" (Puig de la Bellacasa, 2011)

- My ethnographic study is a form of response-ability to the indispensable yet invisible labor behind AI production

Predicting or Tracing?

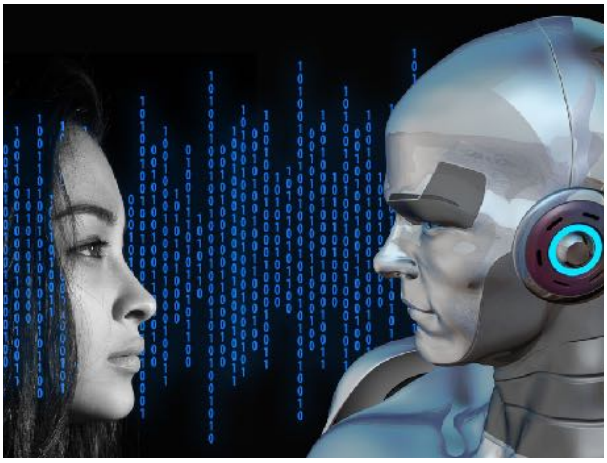
Rethinking Care and Gendered AI Imaginaries in Elder Care

Jieun Lee (Yonsei University)

Gender and AI?

Existing feminist critiques

- Bias
  - in training data and of algorithms
  - reproduction of discrimination and intensification of injustice
- Gendering of AI agents
  - reinforcing gender stereotypes
- Women in AI
- AI and women's labor



Predicting or Tracing

- How gendered imaginaries of AI shape the current endeavors to utilize AI in elder care
- Surveillance and artificial sociality: How AI in elder care enacts a bifurcated model of care and how it reflects and reinforces gendered imaginaries of AI and care
- Prediction as a shared modality of treating care-related data
- From predicting to tracing: What might become possible if we treat data not as cues for prediction, but traces of people's lives in their ongoingness

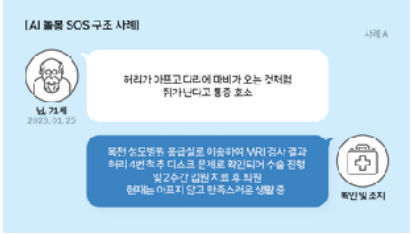
119 and Companions

Surveillance and Artificial Sociality in Elder Care

Emotional support by friendly machines

"아리아, 살려줘" 독거어르신을 돌보는 AI 돌봄의 현황은?

Protecting the elderly from lone death



Control Center

Surveillance as a means of technocratic care

- Detecting emergency cues
  - behavioral, verbal, vital
  - detecting “anomaly”
  - delegating care tasks to relevant actors
  - predicting future risks

- Surveillance AI
  - work as an omniscient eye, without being present in people’s everyday life
  - cognitive and managerial work in care
  - relation to the state authorities
  - (roles and institutions historically coded as masculine)

Companions

Communicative AI Conducive To Surveillance

- Artificial sociality
  - simulated intimacy
  - friendly appearances (often feminized and/or young)
  - “emotional support”
  - (conventionally coded as feminine)

- User engagement
  - “natural” conversation expected (→adopting generative AI)
  - necessary condition for effective surveillance

Surveillance vs. Artificial sociality

in AI-Driven Elder Care

- Bifurcation of care
  - Surveillance and protection: managerial work, cool, cognitive, and surveilling sensors and decision-making brains
  - Communicative and intimate (artificial) sociality: warm, caring voices and touches
  - Corresponds to the gendered imaginaries of AI and Care
- What do they share, and what aspect of care is missing here?

Prediction

as a shared modality

- Prediction
  - Surveillance AI: predicting the future risk from the data about the elderly's present condition
  - Communicative AI (powered by generative AI): generate sentences by predicting the next likely word to simulate emotional support
- Data as a cue for immediate response
  - Short temporal horizon for the data's "significance"

Caring and knowing

- (Human) Caregivers
  - the elderly's everyday life, past history, desires, concerns, and longings, as well as health conditions
  - visits, observations, and conversations
  - the significance of these details, remembered by caregivers, change over time

- care as knowledge-generating practice
  - about the person
  - about how to care
  - about the general conditions

"What day is it?"

From Predicting to Tracing

- A banal question repeated
  - won't catch the surveillance AI's attention
  - but be responded by the communicative AI without affecting the AI

- Asking different questions to understand
  - the person
  - what's going on throughout the day
  - what it is like to live with dementia

- AI as a collaborator for care
- Data as traces for people
- Technology for care, curiosity, and openings rather than closures
- Speculative endeavors without certainty

Thank you!

Jieun Lee(jieunleeh@yonsei.ac.kr)

Flexible Labor

From Keypunch Labor in the 1960s
to Modern AI Data Labeling

Heewon Kim
Hanyang University
May 28, 2025

The Messiness of Data Work

- Data work (generating, annotating, and verifying data) essential for sustaining Artificial Intelligence (AI)
- Criticized for reproducing discriminatory decisions based on “biased data”

- The industry striving to fix the problem through better models and datasets
- The fundamental problem of outsourced data production as a manifestation of centuries-old coloniality (Postada, 2022)

Research Theme and Questions

Overcoming presumptions of center-periphery in the information economy

Situating South Korea's keypunch export within the long dureé of outsourced digital labor

- How did the South Korean government and private sectors reconfigure flexible labor to meet export demands?
- What types of training, gendered assumptions, and infrastructures were involved in South Korea's keypunch labor exports?
- How did this reflect the nation-state's strategic engagement with the emerging global information economy?

3

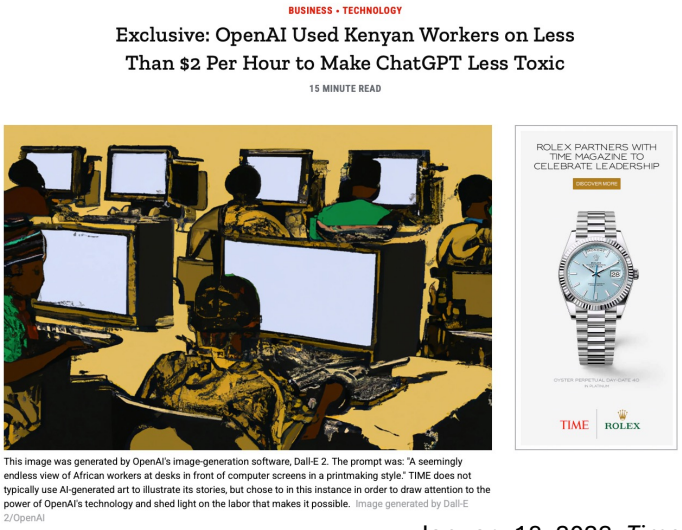
AI Learning From Humans. Many Humans
New York Times, Aug 16, 2019

4

Data Labelers in the AI Supply Chain

Time article on the Outsourcing Company for OpenAI

- Data for detecting toxic information
- Reading and labeling between 150 to 250 texts in a nine-hour shift
- 2 dollars/hour, group therapy for psychological harm



BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ

This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI

January 18, 2023, *Time*

Data Labelers in the AI Supply Chain

Letter from data labeling outsourcing company tech workers in Kenya (for Facebook, Scale AI, Open AI) to President Biden on May 22, 2024

"Working conditions amounting to modern slavery."

- Undermining local labor laws and justice systems
- Violating international labor standards

Open letter to President Biden from tech workers in Kenya

22 May 2024

Dear President Biden,

Cc: Ambassador Katherine Tai, US Trade Representative,

We are 97 data labellers, content moderators and Artificial Intelligence (AI) workers in Nairobi, Kenya. We work for American companies like Facebook, ScaleAI, OpenAI via their outsourcing companies in Kenya.

We understand you will meet Kenya's President William Ruto on his official state visit, to "*mark the 60th anniversary of US – Kenyan diplomatic relations and will celebrate a partnership that is delivering for the people of the United States and Kenya.*" We write because we understand you will discuss, "*trade, investment and technological innovation*" – issues in which our workforce has a direct and personal stake.

US Big Tech companies are systemically abusing and exploiting African workers. In Kenya, these US companies are undermining the local labor laws, the country's justice system and violating international labor standards. Our working conditions amount to modern day slavery. Any trade-related discussions between the US and Kenya must take into account these abuses and ensure that the rights of all workers are protected.


Data Labelers in the AI Supply Chain
AI operating on modularizing complex information tasks into low-skill, routine, and invisible labor
Explained with the narrative of extractionism and exploitation
Data labeling startup in Nangongshi, China (November 25, 2018, New York Times)
7

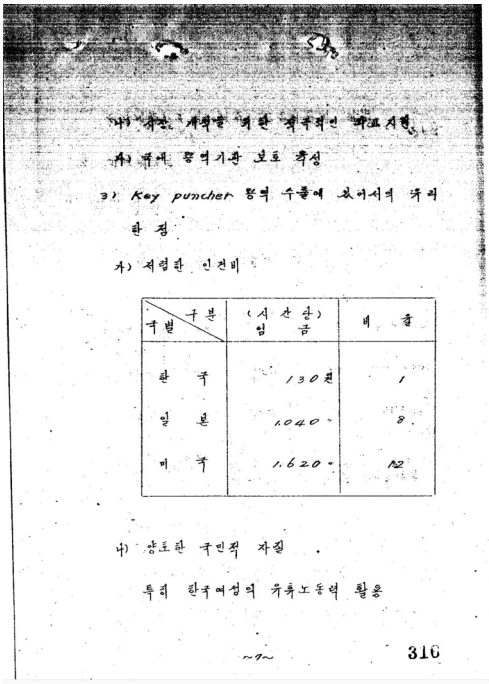

Keypunch Service in the 1970s
Punching and Verifying Punch Cards
Keypunch service export in the South Korean export-oriented growth model
Considered as the foundation for the Software Industry in South Korea
8

The Idle Labor Force

Electronic Data Processing System Development Plan (1969)

Keypunch service export

- Low wage
- Desirable national character (Especially the idle labor force of Korean Women)



나) 유능한 용역기관 보호 육성

3) Key puncher 등록 수출에 있어서의 유리한 점

가) 저렴한 인건비

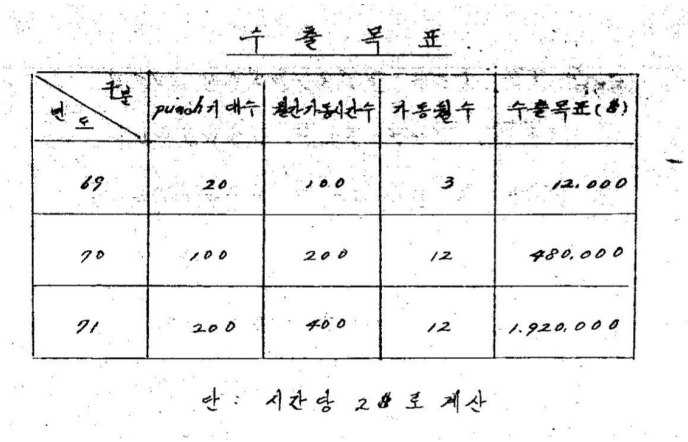
구분 / 국별	(시간당) 임금	비율
한국	130원	1
일본	1040 "	8
미국	1.620 "	12

나) 양호한 국민적 자질

특히 한국여성의 유휴노동력 활용

~9~ 316

Turning the Idle Labor Force into Export Goods



수 출 목 표

구분 / 년도	punch기 대수	월간가동시간수	가동월수	수출목표($)
69	20	100	3	12.000
70	100	200	12	480.000
71	200	400	12	1.920.000

단: 시간당 2$로 계산

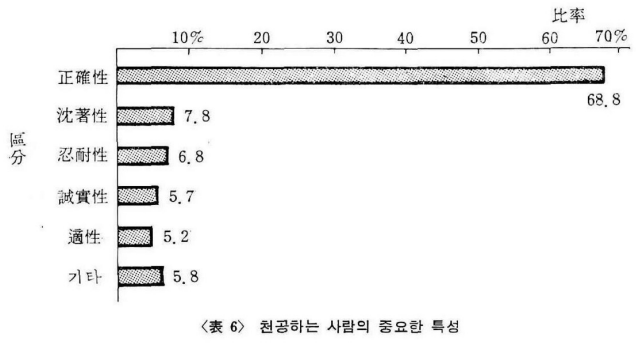
Turning the Idle Labor Force into Export Goods

11

Feminizing the Data Work

Accuracy, calmness, patience, and diligence “known as women’s characteristics. The work is suited for women.”

“General characteristics of an office-type women (사무형여자)”

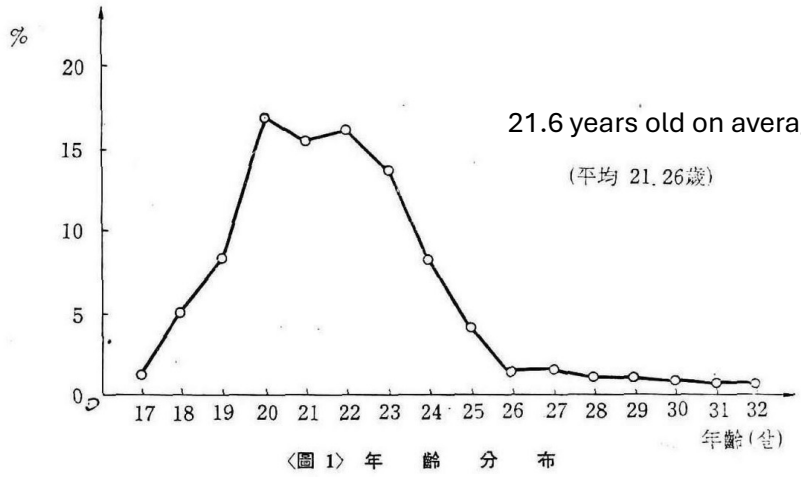


〈表 6〉 천공하는 사람의 중요한 특성

Key Puncher’s Background and Work Ethics (1974)

12

Feminizing the Data Work



21.6 years old on average

(平均 21.26歲)

Key Puncher's Background and Work Ethics (1974)

13

White Collar Women

Promoted as a pleasant job for female workers to work in clean offices

"While it may get loud occasionally, the ambiance in the room stays pleasant."
"The machine is sensitive to its surroundings, ensuring optimal working conditions... It is perfect for women seeking a clean-air environment, a comfortable temperature, and tranquility."

Key Puncher's Background and Work Ethics (1974)

14

Harsher Reality

- Eight hours per day, break times provided at intervals of one or two hours (severe eye strain and high levels of noise).
- 70% working 25 to 26 days per month, 20% working on Sundays

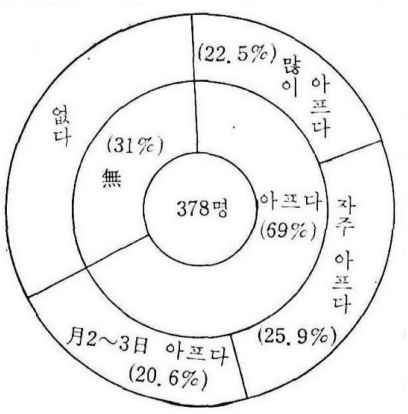
Key Puncher's Background and Work Ethics (1974)

15

Harsher Reality

- Occupational disease

- 3.3m^2 space, a single keypunch machine, a chair, and a small work table
- Hearing and vision impairments are expected
- Musculoskeletal problems
- Neurological disorders
- 70% of workers reporting pain



〈圖 13〉 직업병에 대한 반응

Key Puncher's Background and Work Ethics (1974)

16

Harsher Reality

〈表 23〉 여자동생에게 천공직업을 권유하고 싶은가?

구 분	비 율		비 고
현직장에 권유하겠다	0.8%	8.8%	긍 정 적
권유 하겠다	5.9%		
타직장에 권유하겠다	2.4%		
의향대로 하겠다	49.3%	49.3%	방 관 적
다른 직업을 권유하겠다	17.8%	41.9%	부 정 적
만류 하겠다	24.1%		
계	100%		

Only 8.8% would recommend the work to their younger sister.

〈表 36〉 결혼 후에도 천공업무에 종사하고 싶은가?

구 분	비 율		비 고
계속 종사하고 싶다	7.0%	17.4%	47%
타 직업을 원한다	16.4%		
결혼 후에 결정하겠다	36.6%	36.6%	
종사하지 않겠다	46.0%	46.0%	

Only 7.0% hoped to continue working as a keypunch operator after getting married.

Key Puncher's Background and Work Ethics (1974)

17

Conclusion

- Keypunch labor in South Korea in the 1970s was deskilled and feminized.
- The work was framed as a national asset and a stepping stone toward technological modernization and economic development.
- This historical case complicates the dominant narrative of outsourced labor as a simplistic extension of colonial extraction.
- Calls for a thorough analysis of how AI data labeling is shaped, facilitated, and legitimized in developing countries, leveraging their human labor as a source of exportable goods.

18

Thank you

heewon.kim09@gmail.com

MEMO

MEMO

MEMO

MEMO

MEMO