POTENTIAL RISKS AND NEW DIRECTIONS OF GENERATIVE ARTIFICIAL INTELLIGENCE

Prof. BYUN, SUNYONG

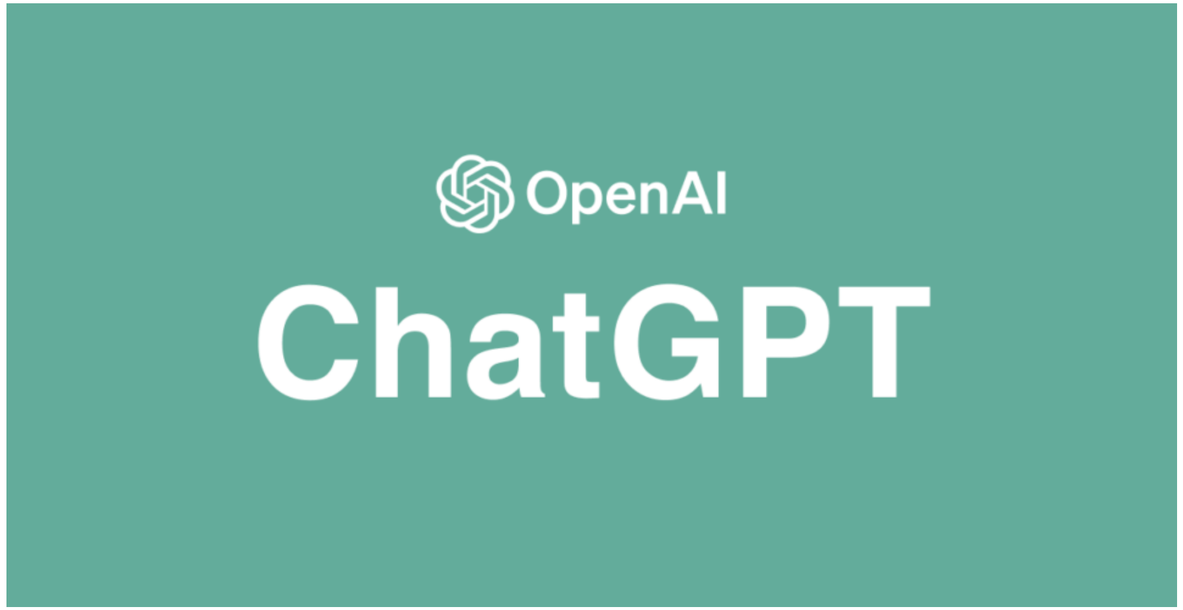
Seoul National University of Education

Table of Contents

Directional conflict brought about by generative artificial intelligence

the First Shock of the 4th Industrial Revolution

the Second Shock of the 4th Industrial Revolution



- While the AlphaGo shock sparked innovation in artificial intelligence technology, the ChatGPT shock triggered the fear of artificial intelligence.
- As ChatGPT became popular, people were surprised by the technological changes and felt new fears.

- Several recent incidents related to OpenAI clearly illustrate this fear : Sam Altman, CEO of OpenAI, was fired by the board of directors and was recruited by Microsoft. A Few days later he was returned to OpenAI due to collective action by employees.

- The Italian government had requested a ban on access to ChatGPT due to concerns about personal information rights infringement. About a month later, the ban was lifted. This comes as ChatGPT's developer, OpenAI, has introduced improvements to how it handles personal data. What is interesting is the case of Sweden, which shows a different pattern. The Swedish government stresses the need to reintroduce the 'traditional method of education' using paper books, as digital devices have a negative impact on children's literacy.

ethical vacuum

Therefore, this study discusses the right direction of AI growth and management in the generative AI era, taking into account the increasing social impact of AI. We propose three types of AI that are needed to address the ethical issues of Super-Massive AI: **Restricted AI**, **Ethical AI**, and **Sustainable AI**, which we call **RES**ponsible AI.

Restricted AI

All technologies always have both positive and negative aspect.

Restricted AI

2. 1 Technological ethics approach

2. 2 Systemic approach

2. 1 Technological ethics approach

- the implementation of ethically appropriate methods for technical processes to ensure the safety and ethics of AI
- focusing on the data for AI training and artificial intelligence algorithms and models.

■ **The first consideration is what to do with data for AI training.**

training data → outcome

biased or personalized training data

A Chatbot uses discriminatory language
An AI produces output that reveals personal information

"the need for ethical purification of data and the need to establish ethical procedures in data collection (Byun, 2023a: 192)" has been emphasized.

2. 1 Technological ethics approach

- **The first consideration is what to do with data for AI training.**

When defining data ethics, ‘data rights’ and ‘**data responsibility**’ can be presented as key concepts (see Byun, 2023a: 200).

the importance of identifying and protecting data subjects

"Data responsibility is related to how to respond to ethical issues arising from data. Ultimately, data responsibility can be seen as another essential requirement because it serves as the basis for achieving the ultimate goal by allowing data subjects to fully consider other core values of data ethics (Byun, 2023a: 205). "

Responsibility must be applied to all procedures, including the creation, collection, processing, use, evaluation and disposal of data.

How?

2. 1 Technological ethics approach

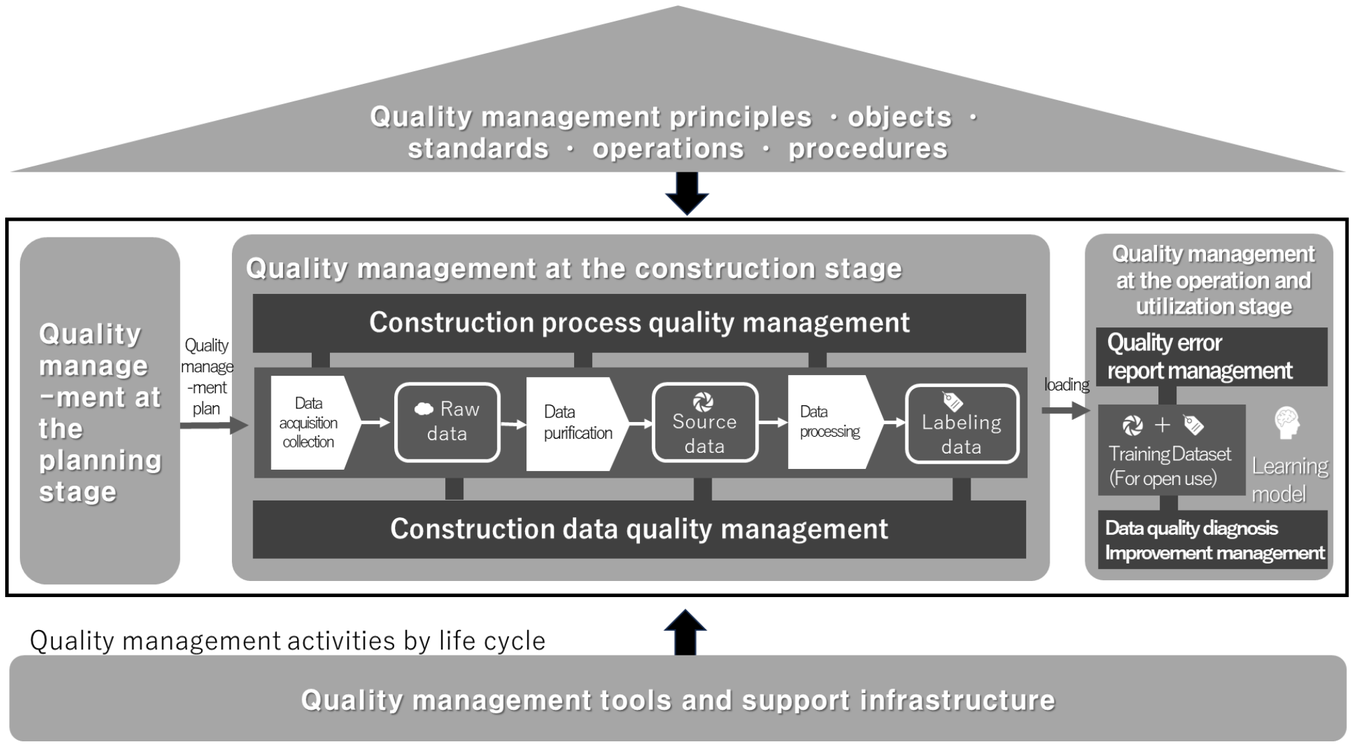
■ **The first consideration is what to do with data for AI training.**

Byun et al. (2023)

- It needs the comprehensiveness of data regulation by including both consumers and producers as data subject.
- it ensures that all data subjects involved are continuously accountable throughout the data processing process.

Data Quality Management Framework for Artificial Intelligence Learning

Ministry of Science and ICT and the National Information Society Agency (NIA) in Korea.



2. 1 Technological ethics approach

- **The second consideration is the technical restrictions used by Large Language Models (LLMs) such as ChatGPT.**

Recently, LLMs such as OpenAI's GPT, Anthropic's Claude, and Meta's LLaMA have applied the 'Reinforcement Learning from Human Feedback (RLHF)' methodology.

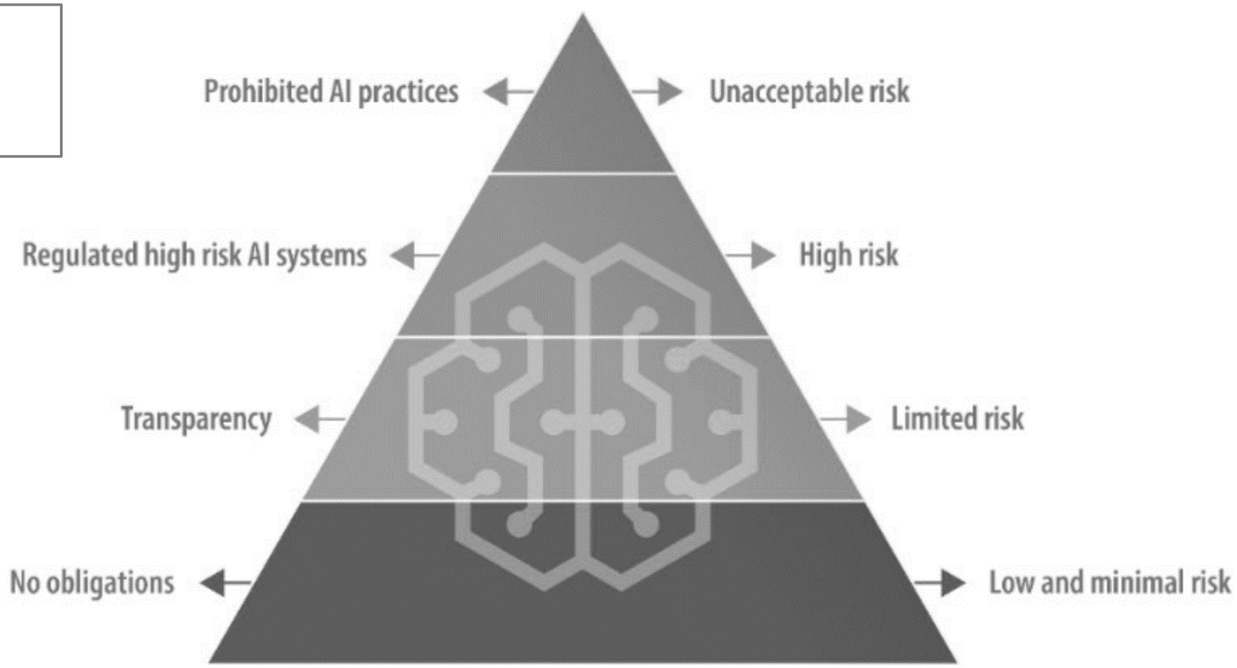
The main characteristic of RLHF is that humans can develop artificial intelligence by providing direct feedback on each value and actively reflecting their preferences.

2. 2 Systemic approach

- institutionalize responsibility for AI developers and consumers to ensure that technical regulation is fully functional
- The legal system and evaluation system

Artificial Intelligence Act (AI Act)
→ risk-based approach

The AI Act codifies the law by focusing on the risks that can arise from the use of AI. This shows that AI regulation is moving towards eliminating the negative consequences of AI.



2. 2 Systemic approach

- institutionalize responsibility for AI developers and consumers to ensure that technical regulation is fully functional
- The legal system and evaluation system

AI Impact Assessment (AIA)

AIA is an evaluation tool and guideline for recognizing, preparing for, and eliminating various risks of artificial intelligence systems.

↓

criteria evaluation and autonomy evaluation

a method of evaluating the risk level of an artificial intelligence system based on major standards of artificial intelligence ethics such as responsibility, transparency, and bias

typifying and evaluating the moral and ethical autonomous judgment ability of artificial intelligence due to the technical characteristics of artificial intelligence systems that are produced to solve various problem situations through autonomous decisions.

Ethical AI

- the background and foundation for restricted AI
- a technological and social buffer

Ethical AI

3. 1 Normative approach

3. 2 Organizational approach

3. 1 Normative approach

The normative spread social awareness and prepare countermeasures against potential risks.

Ethics guidelines for trustworthy AI(2019, EU)

These guidelines present the concept of trustworthy AI, ethical principles, key elements, and a list of evaluations.

‘Lawful AI’, ‘Ethical AI’, and ‘Robust AI’

respect for human autonomy, prevention of harm, fairness, and explainability

3. 1 Normative approach

Recommendation of the Council on Artificial Intelligence(OECD, 2019)

- AI as ‘trustworthy AI’ and confirms that the initial perspective of artificial intelligence ethics focuses on forming and maintaining social trust in artificial intelligence technology.
- The recommendations largely specify the principles of responsibility, national policy and international cooperation.
- Five principles of responsibility are listed: ‘inclusive growth, sustainable development, well-being’, ‘human-centered values and fairness’, ‘transparency and explainability’, ‘robustness, security, safety’, and ‘accountability’.

3. 1 Normative approach

Recommendation on the Ethics of Artificial Intelligence(UNESCO, 2021)

- Chapter 3 of this document provides guidance on four values and ten principles related to artificial intelligence ethic.
- The four values are ‘respect, protection and promotion of human rights, fundamental freedoms and human dignity’, ‘living in a peaceful, just and interconnected society’, ‘prospering environment and ecosystem’ and ‘realizing diversity and inclusion’.
- There are 10 principles that include ‘safety and security’ and ‘transparency and explainability’.

3. 1 Normative approach

Artificial Intelligence (AI) Ethical Standards(Ministry of Science and ICT in Korea, 2020)

- It sets ‘humanity’ as the highest value and establish the ontological position of the relationship between humans and artificial intelligence.
- In addition, 10 core requirements are provided, including the human dignity, the public good of society, and the principle of the purposefulness of technology, guaranteeing human rights, protecting privacy, and respecting diversity.

3. 1 Normative approach

Ethics guidelines for trustworthy AI(2019, EU)

Recommendation of the Council on Artificial Intelligence(OECD, 2019)

Recommendation on the Ethics of Artificial Intelligence(UNESCO, 2021)

Artificial Intelligence (AI) Ethical Standards(Ministry of Science and ICT in Korea, 2020)

the Four Essences of Ethical Artificial Intelligence

1. Artificial intelligence must respect human dignity and rights.
2. Ethical AI must protect the public good of society.
3. Ethical AI should enhance human capabilities.
4. Ethical AI needs to go beyond technical and instrumental rationality in order to pursue a techno-ethical goodness.

3. 2 Organizational approach

a methodology that identifies how organizations design, develop, distribute, manage, operate, and utilize artificial intelligence ethically.

Example 1 : Google

- Google presents ‘responsibility’ as its core value.
- The goals of artificial intelligence are public interest, prohibition of unfair bias, safety, responsibility, protection of personal information, scientific excellence, and suitability.

Example 2 : Microsoft

- responsible AI
- MS is setting six goals: ‘Accountability’, ‘Transparency’, ‘Fairness’, ‘Reliability and Safety’, ‘Privacy and Security’, ‘Inclusiveness’.

3. 2 Organizational approach

Example 3 : IBM's AIM

- IBM is designing and applying the AIM(AI Maturity) framework, and IBM explains AIM as "a measure of how mature AI is within an industrial application".
- IBM's AIM combines business capabilities and technical capabilities to present seven measurement criteria and three levels of scale.
- The seven measurement criteria are 'business impact', 'customer value', 'technological advancement', 'reliability', 'ease of use', 'AI operating model', and 'data'
- The three-level scale is 'Silver (1 points)', 'Gold (2 points)', and 'Platinum (3 points)'

Sustainable AI

Big tech companies have not considered energy consumption and environmental pollution that occur during the learning process of Super-Massive Generative artificial intelligence models.

Sustainable AI

4. 1 Ecological approach

4. 2 Mutual cooperative approach

4. 1 Ecological approach

- The ecological approach is a core approach to achieving sustainable AI.
- For example, in the learning process of artificial intelligence, it is necessary to build a data center, and environmental issues with such data centers have recently been raised.

International Energy Agency (IEA)

the electricity used by data centers

2022	2026
460 TWh	1000 TWh

≒ Japan’s electricity consumption

4. 1 Ecological approach

- The discussion of "Green AI" is introduced to address this ecological crisis.
- Schwartz et al. (2020) refer to conventional artificial intelligence as "Red AI" due to its large carbon footprint, and propose “Green AI” as an alternative to Red AI.
- Green AI (Verdecchia et all, 2023)

  = “It is using AI to mitigate human impact on the natural environment in terms of natural resources”

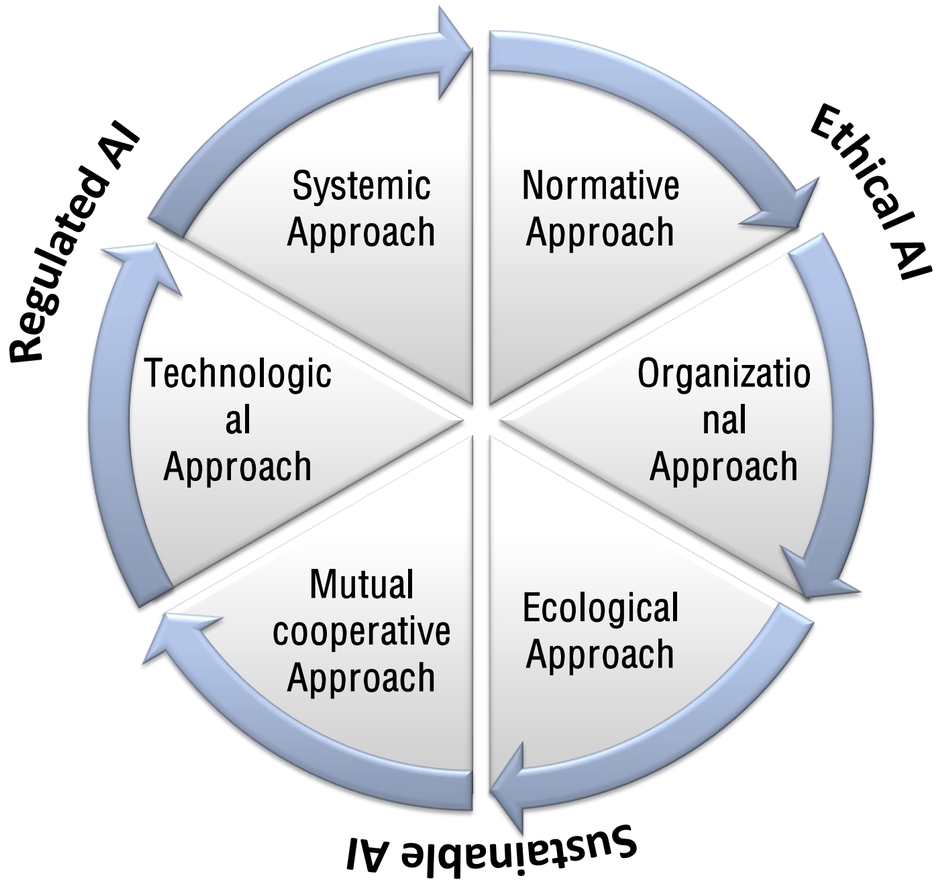
  =“It mitigates the impact that AI itself may have on the natural environment”

4. 2 Mutual cooperative approach

- Social problems that may arise due to artificial intelligence require international cooperation.
- The G7 established the ‘Hiroshima AI Process’ in May 2023, and in October of that year, ‘International Guidelines and Code of Conduct for Advanced Artificial Intelligence Development Organizations’ was finally agreed upon.
- The agreement proposes ‘international guiding principles’ that require risk assessment and management of artificial intelligence development companies and a ‘code of conduct’ that urges users to cultivate artificial intelligence literacy.

- In November 2023, at the AI Safety Summit, 28 countries, including the United States, China, South Korea, and the EU, agreed to the Breslow Declaration.
- It states that highly capable “frontier AI” poses potentially enormous risks that the world must manage in a coordinated effort.

New direction in the era of generative artificial intelligence: RESponsible AI & AI CItizenship Education



RESponsibility about AI
RESponsibility of AI itself

The first is 'RESponsibility about AI', which means that all members including developers, users related to artificial intelligence fulfill their responsibilities to manage the potential risks of artificial intelligence.

The second is 'RESponsibility of AI Itself', which means that AI itself should be viewed as a responsible entity in order to enhance human dignity, the common good of society, and the sustainability of the natural environment.

The method must be specified and implemented on the basis of responsibility.

Digital Citizenship -> AI Citizenship:

The need for AI citizenship education

- The world is now changing very quickly due to the emergence of a new technology called AI.

- The growing role of AI in civil society
  - AI society as a complement to representative democracy (recall of direct democracy and participatory democracy)
  - Increasing the reality of AI democracy: debates using metaverse, the emergence of digital politicians
  - Possibility of transition of AI from a democracy assistant to a decision-maker

- AI is playing the role of transforming the traditional state-centered citizenship framework into the global citizenship framework.

- Beyond the level of replacing analog with digital, AI is expanding from the role of an assistant in knowledge production and consumption to the role of the Subject.

- As society changes from the industrial age to the digital age, digital literacy and digital ethics are necessary for the digital age, so in the future society, AI literacy and AI ethics will be essential to use AI correctly.

A new paradigm in global citizenship : The need for AI citizenship education

- a qualitative difference between the existing Internet-based society and AI-based society.

- Because AI systems are changing the platform itself beyond the level of hardware and software. We need to find out what competencies or values are required of citizens by the emergence of such a system. AI citizenship education that teaches the ability to find solutions and decide alternatives to various social problems that may arise in a future society where AI systems begin to operate in earnest is necessary.

Competencies emphasized in AI citizenship education

- Competencies to manage AI systems and operations so that human dignity and autonomy are respected in human-AI interactions

- Competencies to prevent the harm that AI systems can cause to humans as much as possible in advance, and to hold and realize responsibility for the development, production, management, and use of such AI systems in the aftermath

- Competencies to establish procedures and standards that can institutionally guarantee the explainability of AI systems

- Competencies of fairness to be conscious of data bias and the resulting bias and to solve the problem of inequality that may arise from it.

THANK YOU