Alignment Difficulty of Scaling Swarm-AI

- What is the Next Wave after The 3rd AI Boom? -

Satoshi Kurihara

The Japanese Society of Artificial Intelligence

JSAI

Keio University
Faculty of Science and Technology

Satoshi Kurihara Ph.D. (Computer Science)

Professor of Faculty of Science and Technology, Keio University.

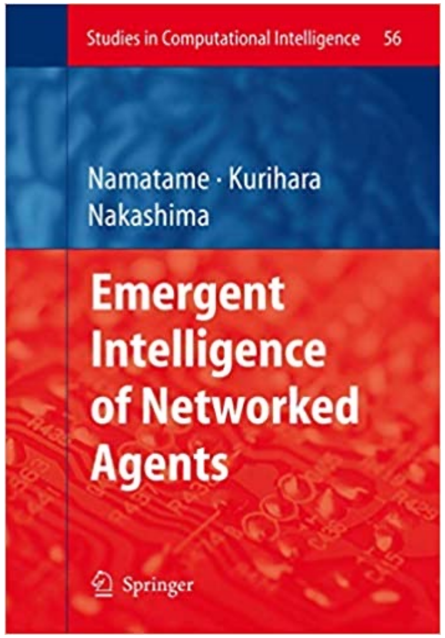
President of the Japanese Society for Artificial Intelligence (JSAI)

# multi-agents, swarm intelligence and computational social science.

One of my recent work includes developing interactive AI to utilize generative AI for innovation.

B·J 50th
AIが挑む新作読み切り
巻頭カラー32P!!
週刊少年チャンピオン
52
340円
ふろく
魔入りました！入間くん
ランダムステッカー
全6種
1枚封入！
どれか1枚が手に入る！
B·J連載50th 特別企画
漫画界の神様にAIが挑む新作読み切り!!
巻頭カラー特大32P!!!!
TEZUKA 2023
ブラック・ジャック
大人気Cカラー!!
魔入りました！入間くん
限定QUOカード200名プレゼント！
2023 週刊少年チャンピオン
AI×ヒト新作読み切り！
魔入りました！入間くん
ふろくランダムステッカー
限定QUOカード200名プレゼント！
52
秋田書店
beat Mark II
2023プロジェクト
NINE PEAKS
ルパン三世 異世界の姫君
大好評Cカラー！

This cartoon was created by the creators using our developed AI.

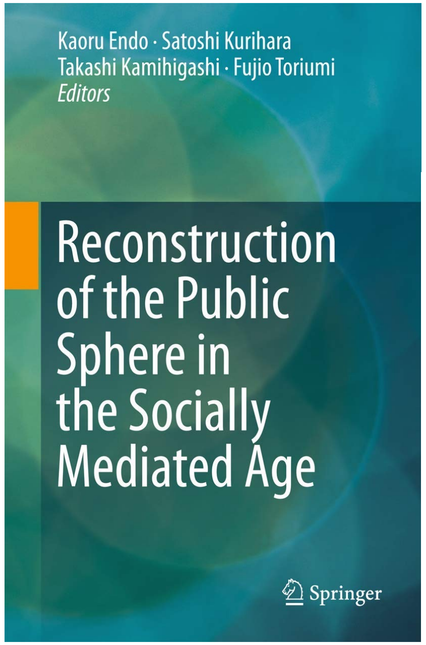


Studies in Computational Intelligence 56

Namatame · Kurihara
Nakashima

Emergent Intelligence of Networked Agents

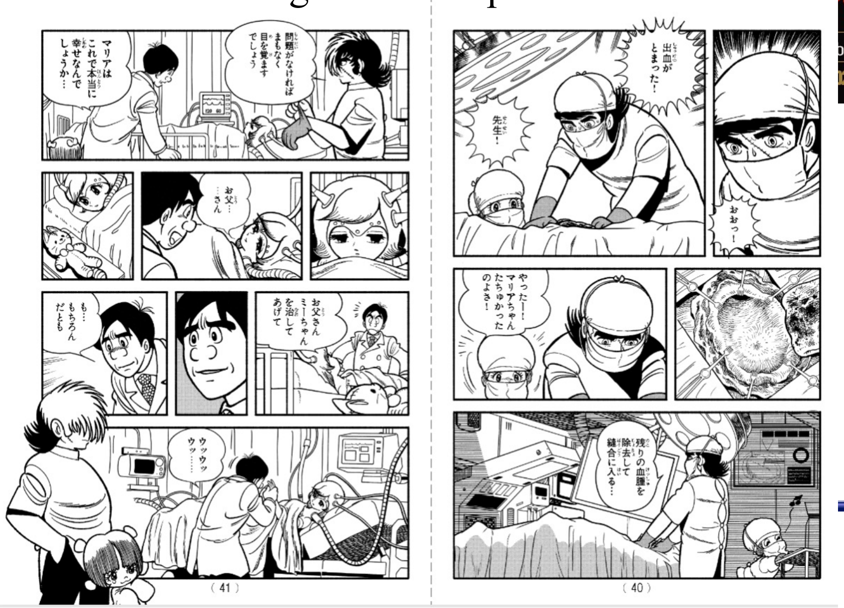
Springer



Kaoru Endo · Satoshi Kurihara
Takashi Kamihigashi · Fujio Toriumi
Editors

Reconstruction of the Public Sphere in the Socially Mediated Age

Springer


出血がとまった！
先生！
おおっ！
やったー！マリアちゃんたちゅかったのよさ！
残りの血腫を除去して縫合に入る…
(40)
問題がなければまもなく目を覚ますでしょう
マリアはこれで本当に幸せなんでしょうか…
お父…さん
お父さんミーちゃんを治してあげて
も…もちろんだとも
ウッウッウッ……
(41)

History of AI (What we have learned from the History)

The 1st AI Boom

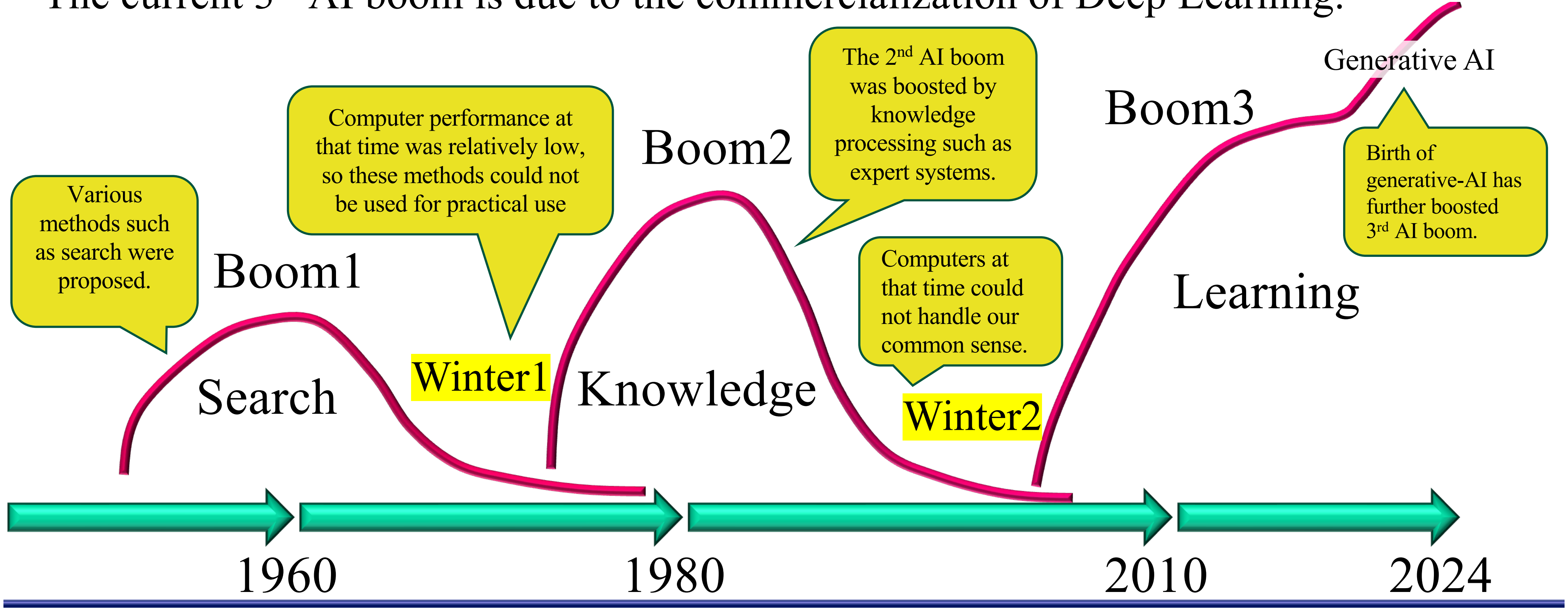
→ The latest ideas and methods are not always immediately put into practice.

→ That is, the infrastructure for practical application must be in place.

The 2nd AI Boom

→The amount of knowledge, like our common sense and tacit knowledge, is enormous.

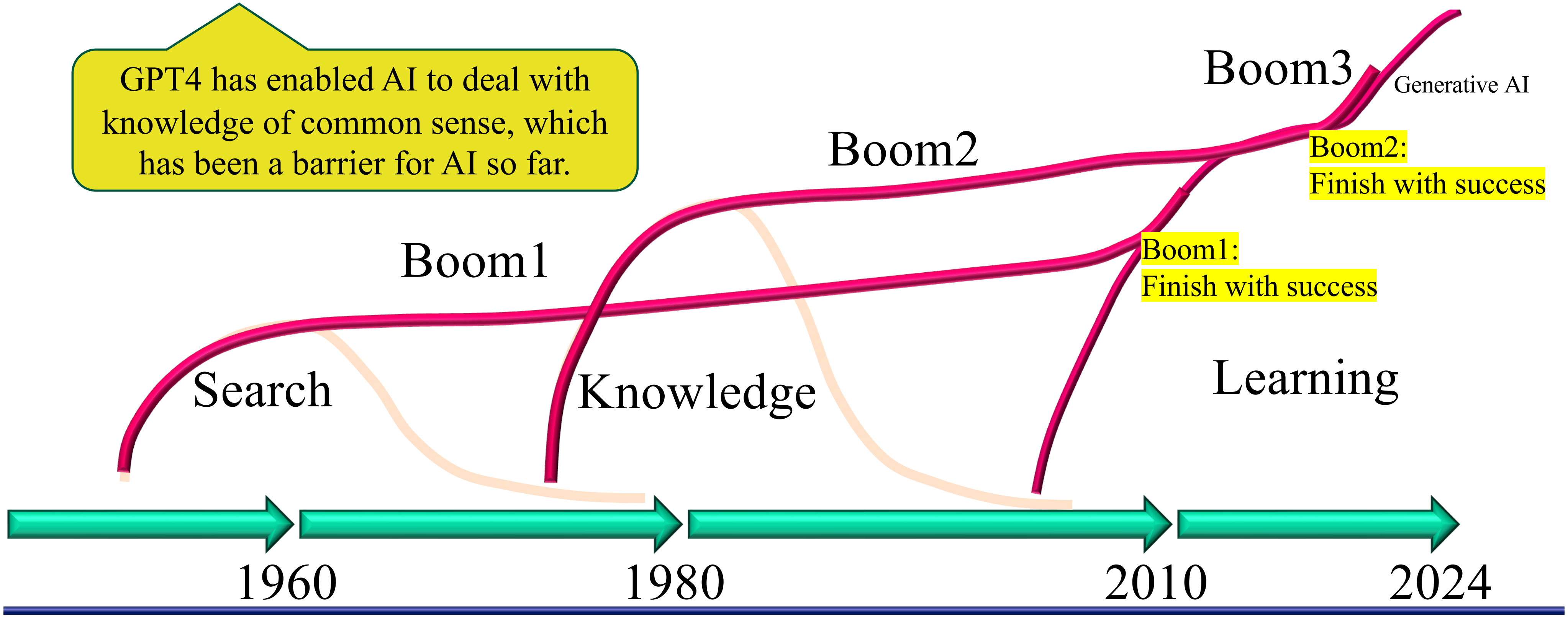
The current 3rd AI boom is due to the commercialization of Deep Learning.


Various methods such as search were proposed.
Boom1
Search
Computer performance at that time was relatively low, so these methods could not be used for practical use
Winter1
Boom2
Knowledge
The 2nd AI boom was boosted by knowledge processing such as expert systems.
Computers at that time could not handle our common sense.
Winter2
Boom3
Learning
Generative AI
Birth of generative-AI has further boosted 3rd AI boom.
1960
1980
2010
2024

Have the 1st and 2nd AI Booms really failed?

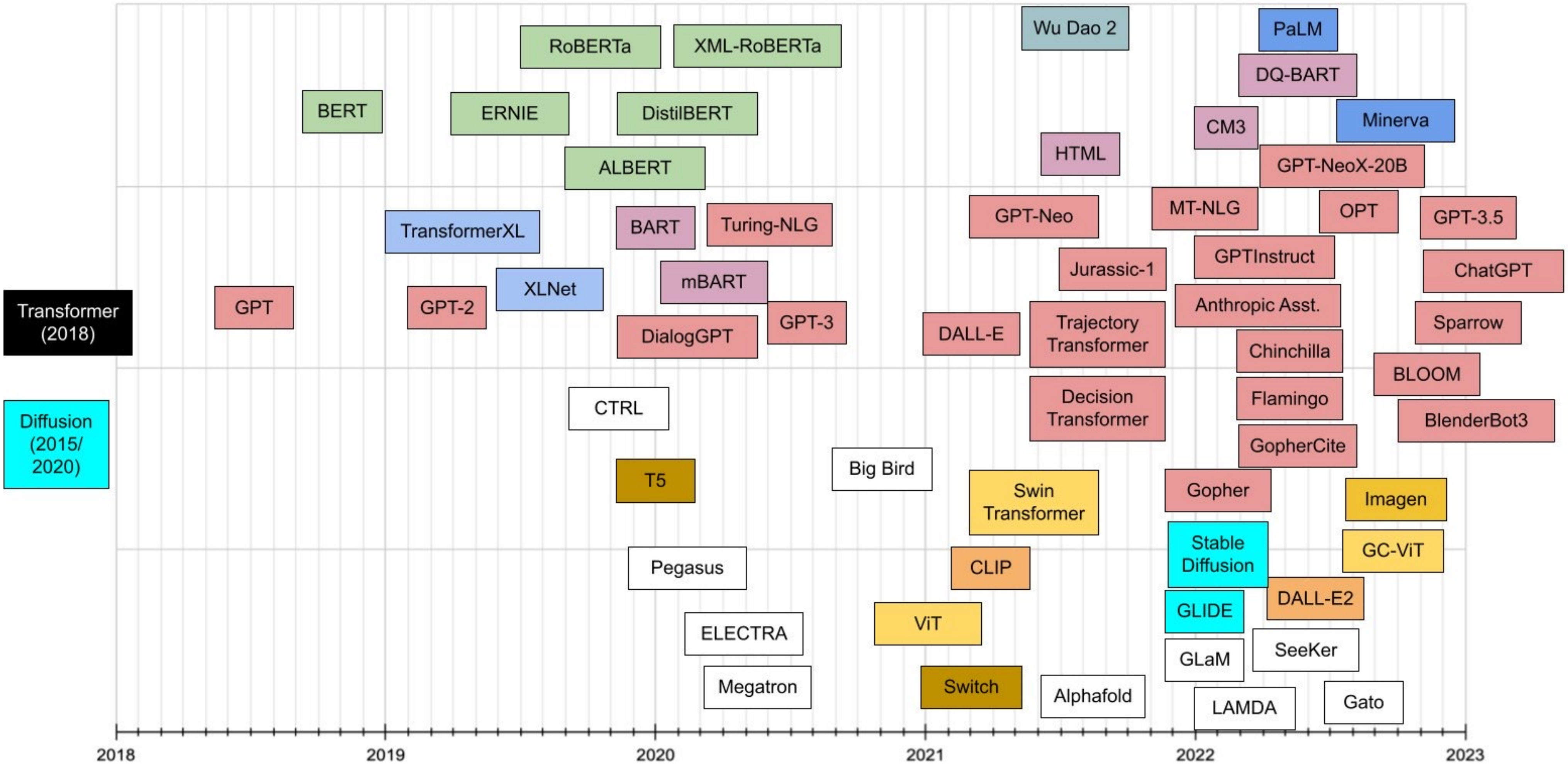
Meaning of 3rd AI Boom is,

- ▶ The 1st AI boom was completed successfully, by the infrastructure in place to make Deep Leaning practical.
- ▶ The 2nd AI boom was completed successfully, thanks to the successful development of GPT4-level LLMs.



GPT4 has enabled AI to deal with knowledge of common sense, which has been a barrier for AI so far.

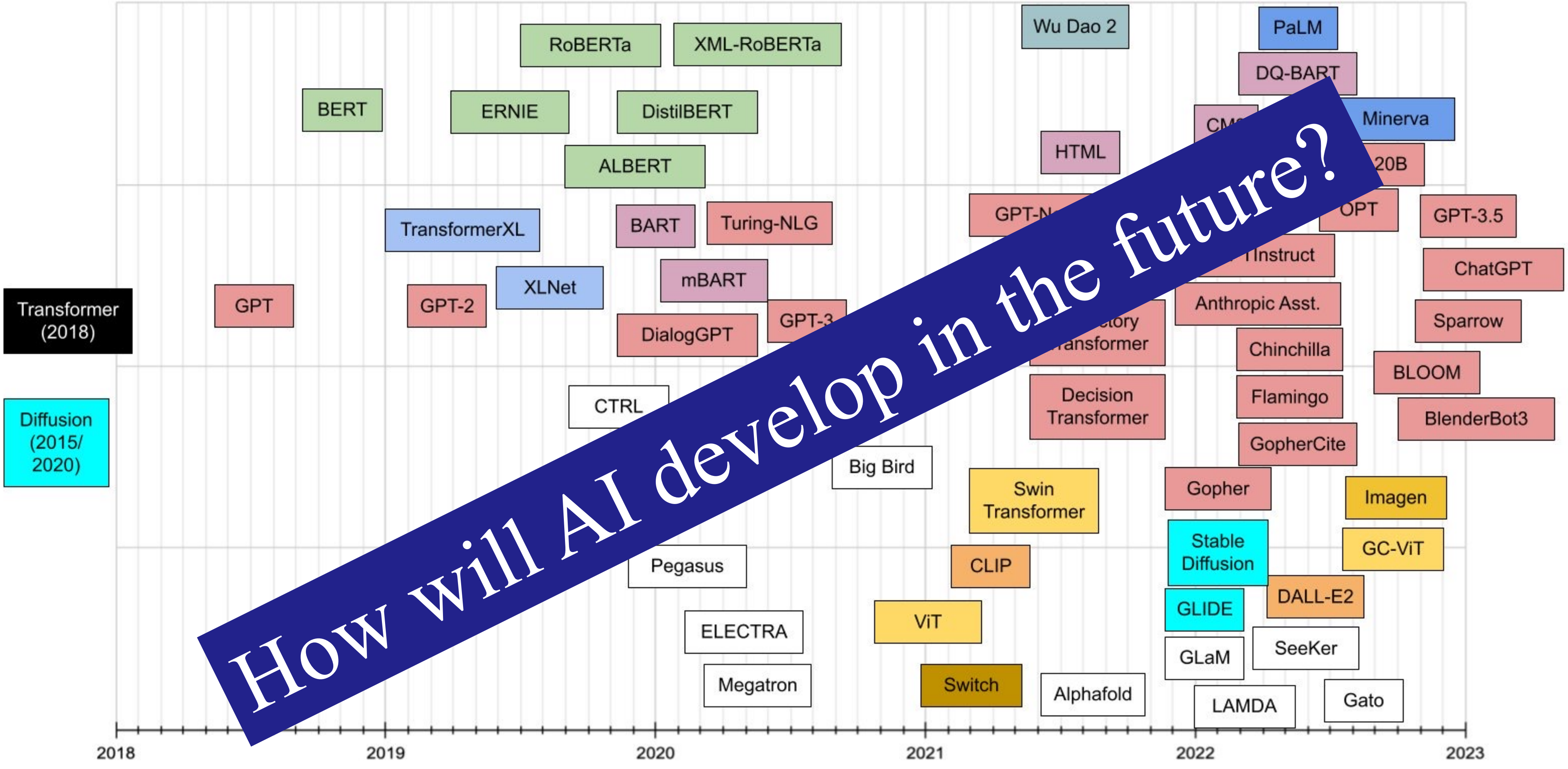
Now is Generative-AI Era



Number of generative-AIs are appearing every day.

https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/#Timeline

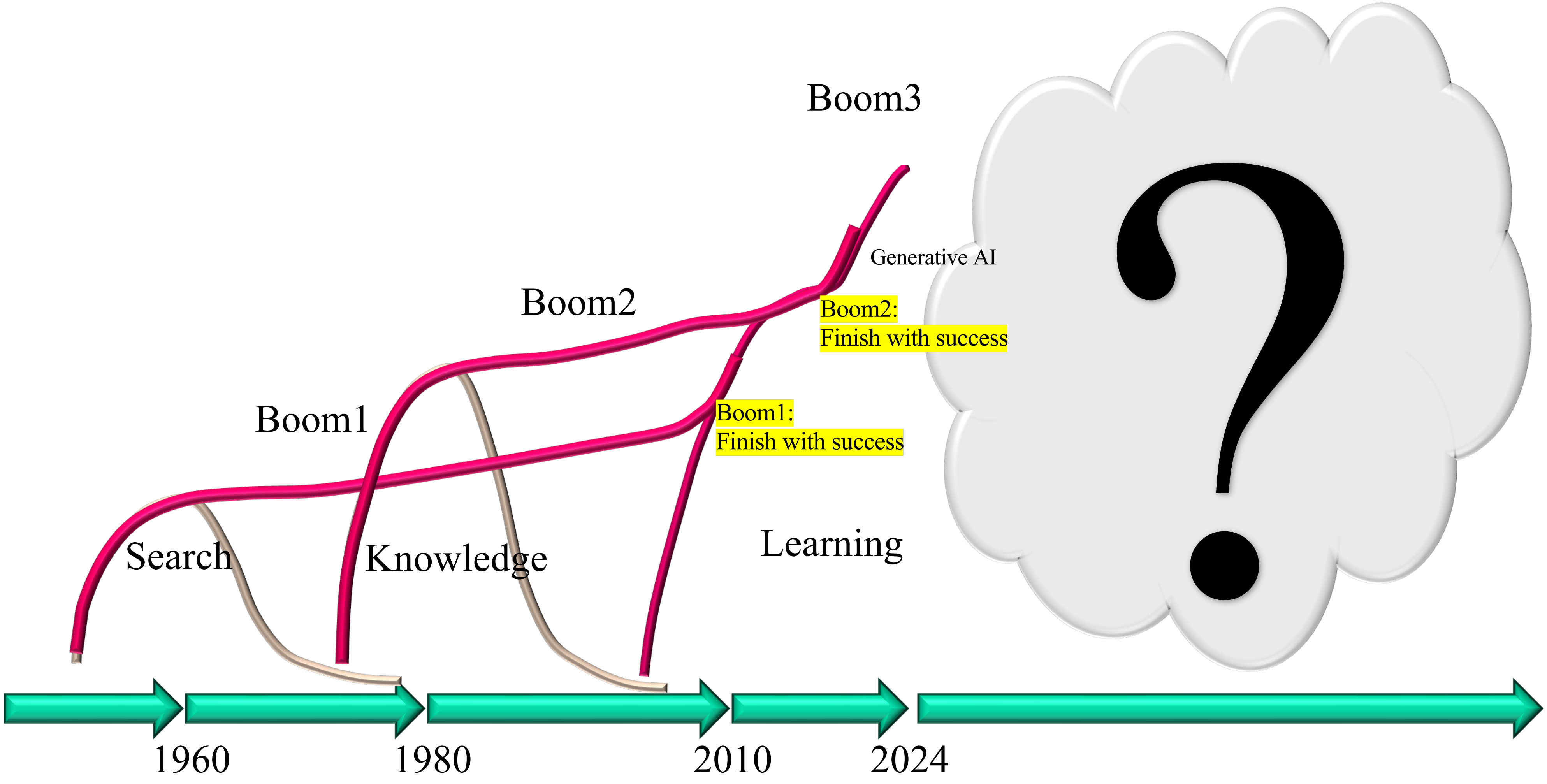
Now is Generative-AI Era



Number of generative-AIs are appearing every day.

https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/#Timeline

What will be the next wave after the third AI boom?

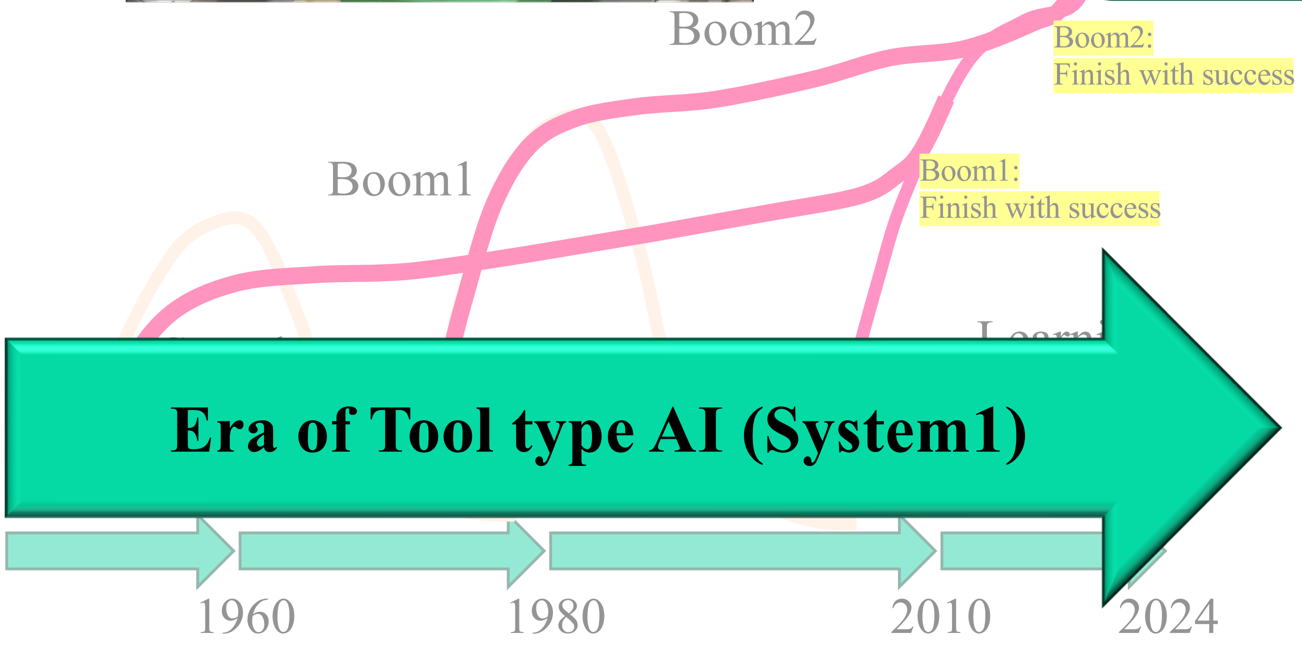


The 70-year history of AI, from the 1st to the current 3rd AI boom, can be described as the Era of Tool-based AI.

This picture is a cooperative robot packing a lunch box alongside a person.

However, it cannot flexibly change its behavior to adapt to a person's movements.

It only performs predetermined actions.



Boom2

Boom2:
Finish with success

Boom1

Boom1:
Finish with success

Era of Tool type AI (System1)

The 70-year history of AI, from the 1st to the current 3rd AI boom, can be described as the Era of Tool-based AI.

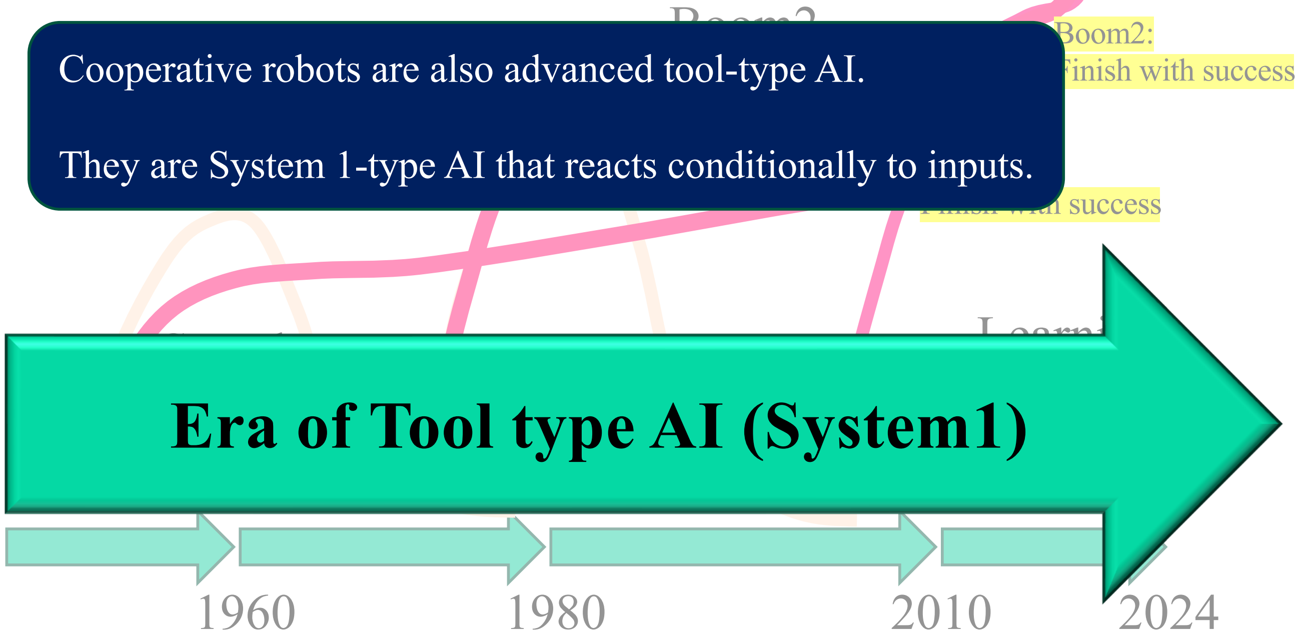
This picture is a cooperative robot packing a lunch box alongside a person.

However, it cannot flexibly change its behavior to adapt to a person's movements.

It only performs predetermined actions.

Cooperative robots are also advanced tool-type AI.

They are System 1-type AI that reacts conditionally to inputs.

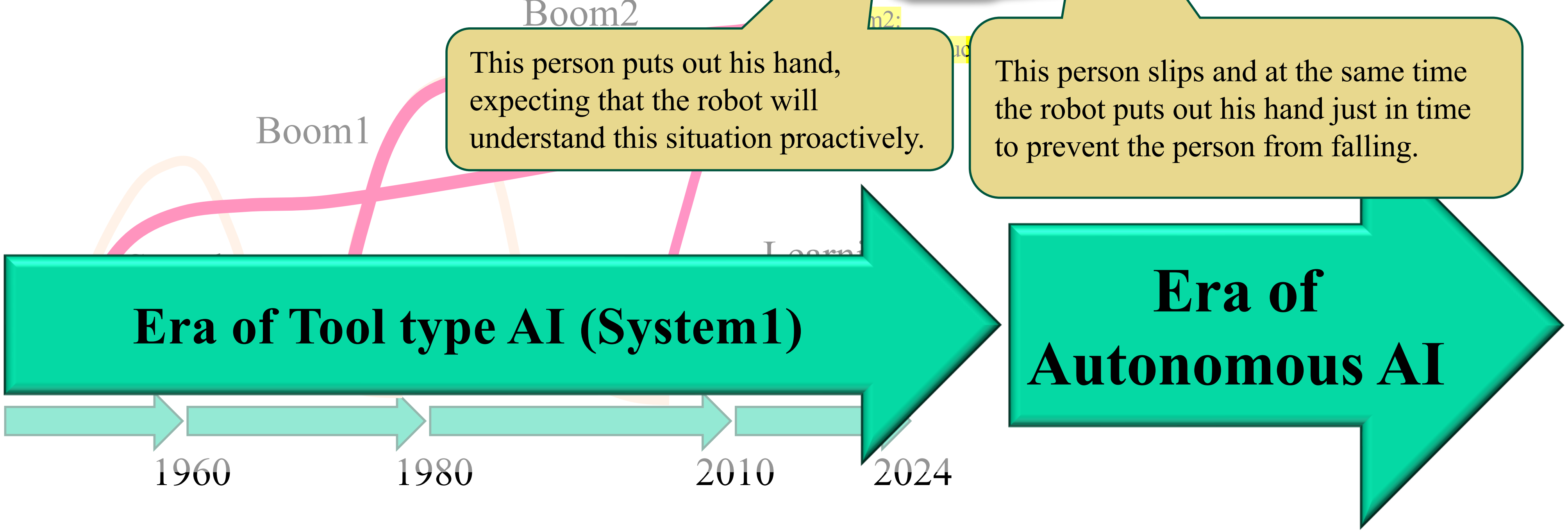


Era of Tool type AI (System1)

Next is Era of Autonomous AI

This person puts out his hand, expecting that the robot will understand this situation proactively.

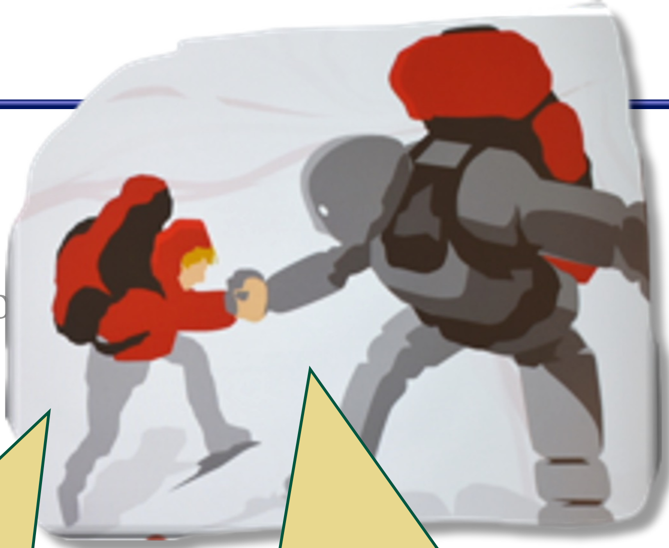
This person slips and at the same time the robot puts out his hand just in time to prevent the person from falling.

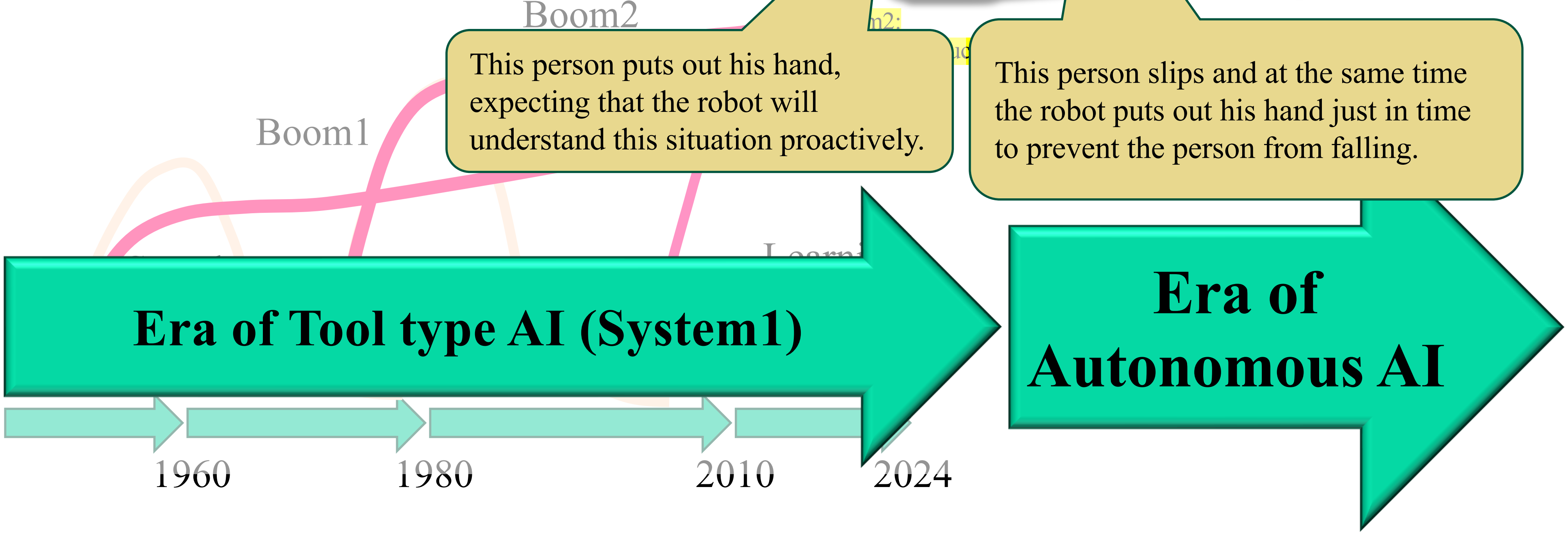


Era of Tool type AI (System1)

Era of Autonomous AI

Next is Era of Autonomous AI



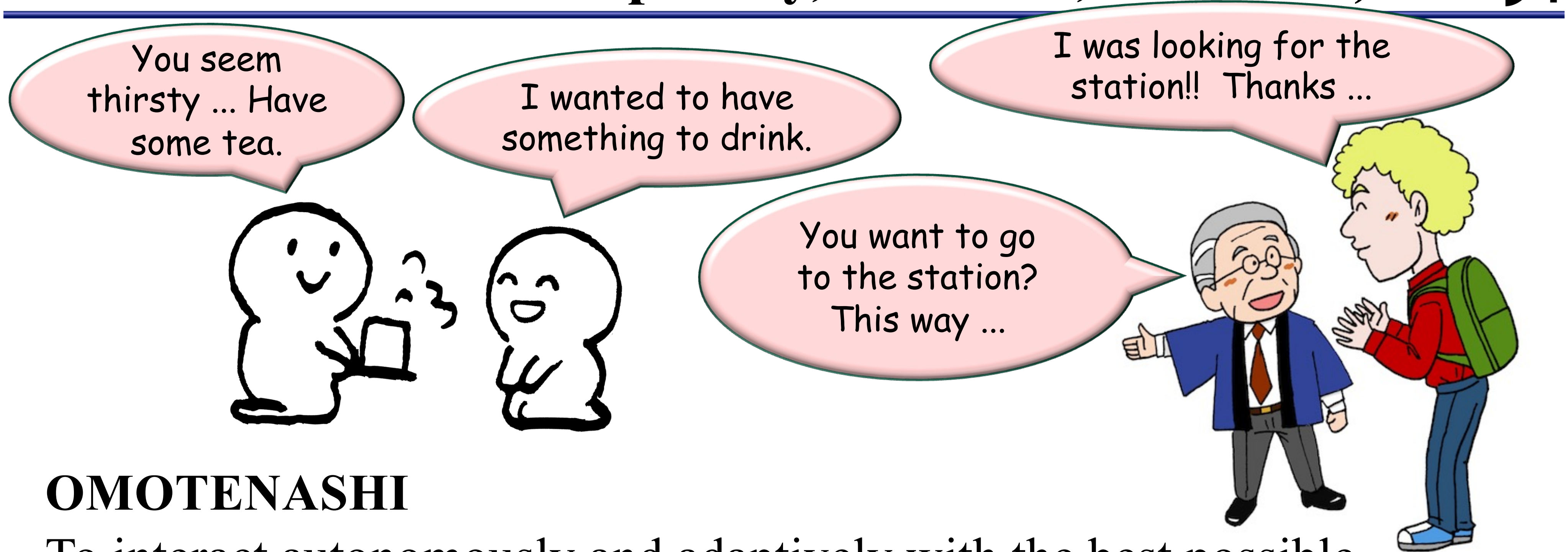


This person puts out his hand, expecting that the robot will understand this situation proactively.

This person slips and at the same time the robot puts out his hand just in time to prevent the person from falling.

In order to be able to understand the situation and act proactively, robots need a high degree of autonomy.

OMOTENASHI （Hospitality, Welcome, Entertain)


You seem thirsty ... Have some tea.
I wanted to have something to drink.
I was looking for the station!! Thanks ...
You want to go to the station? This way ...

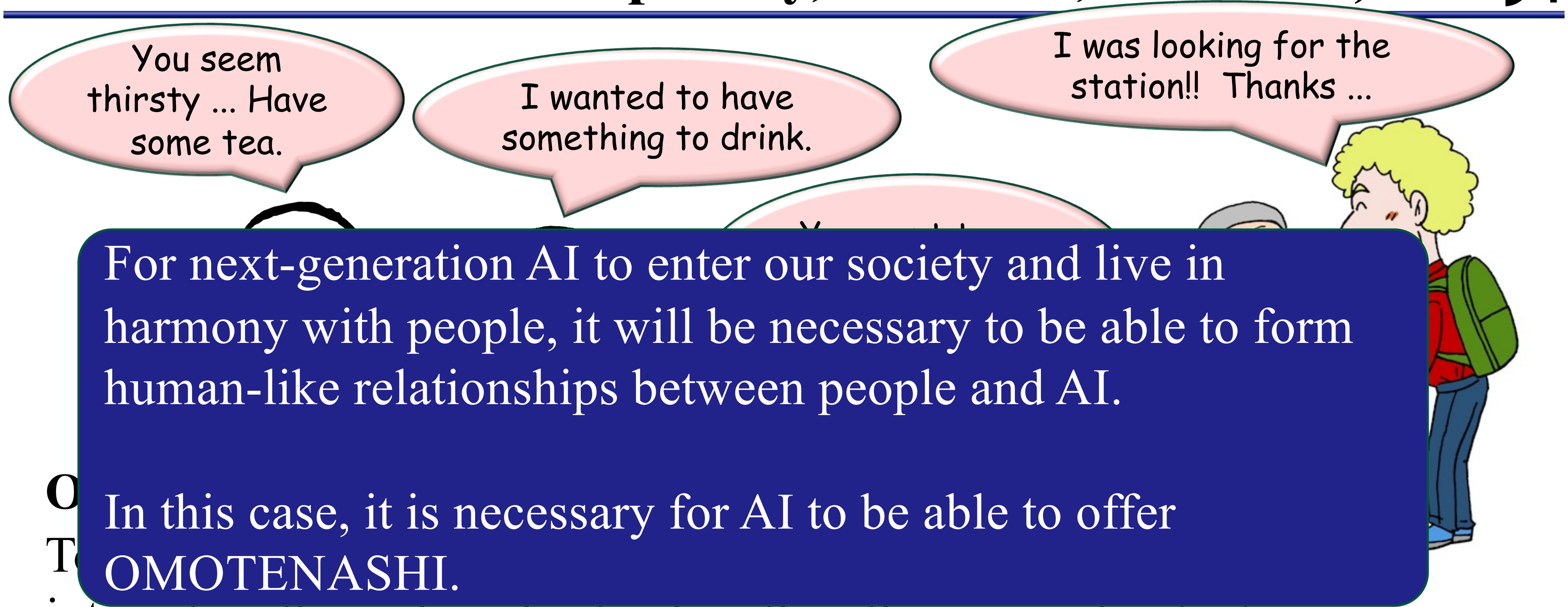
OMOTENASHI

To interact autonomously and adaptively with the best possible interaction through understanding the other person's situation proactively and real-timely.

Only an autonomous AI can perform OMOTENASHI.

OMOTENASHI was the term used as a concept for the Summer Olympics in Tokyo in 2020.

OMOTENASHI （Hospitality, Welcome, Entertain)


You seem thirsty ... Have some tea.
I wanted to have something to drink.
I was looking for the station!! Thanks ...

For next-generation AI to enter our society and live in harmony with people, it will be necessary to be able to form human-like relationships between people and AI.

In this case, it is necessary for AI to be able to offer OMOTENASHI.

O

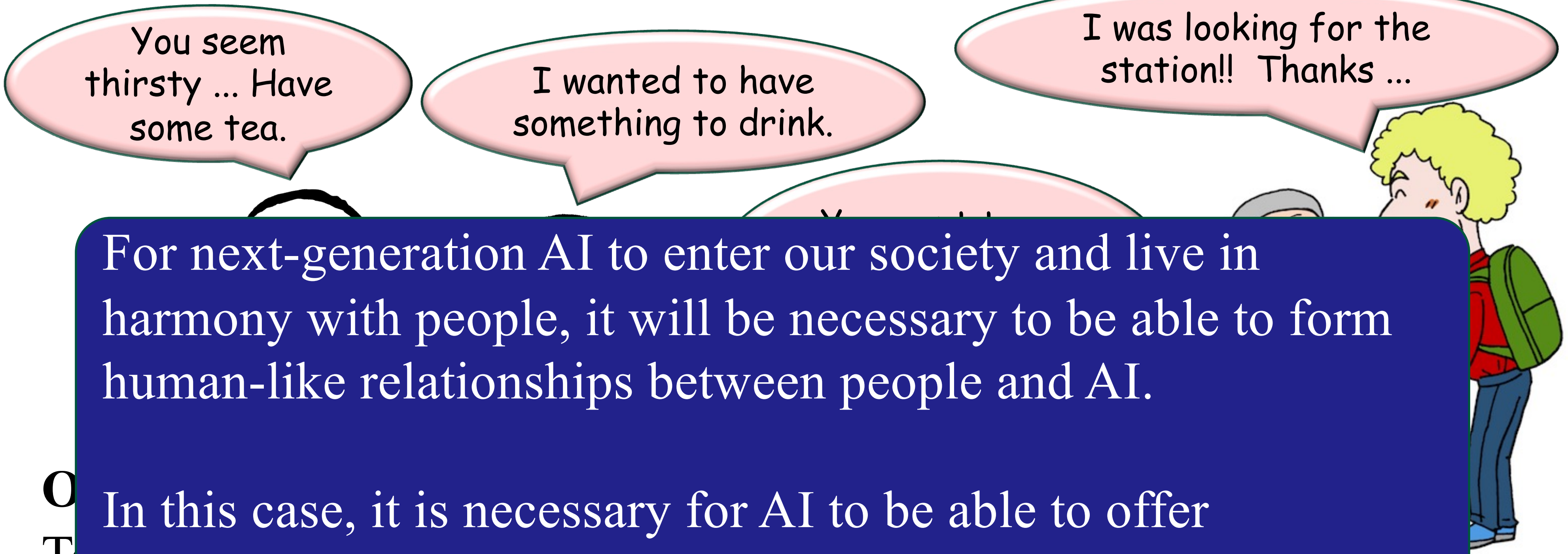
T

interaction through understanding the other person's situation proactively and real-timely.

Only an autonomous AI can perform OMOTENASHI.

OMOTENASHI was the term used as a concept for the Summer Olympics in Tokyo in 2020.

OMOTENASHI （Hospitality, Welcome, Entertain)


You seem thirsty ... Have some tea.
I wanted to have something to drink.
I was looking for the station!! Thanks ...

For next-generation AI to enter our society and live in harmony with people, it will be necessary to be able to form human-like relationships between people and AI.

In this case, it is necessary for AI to be able to offer OMOTENASHI.

O

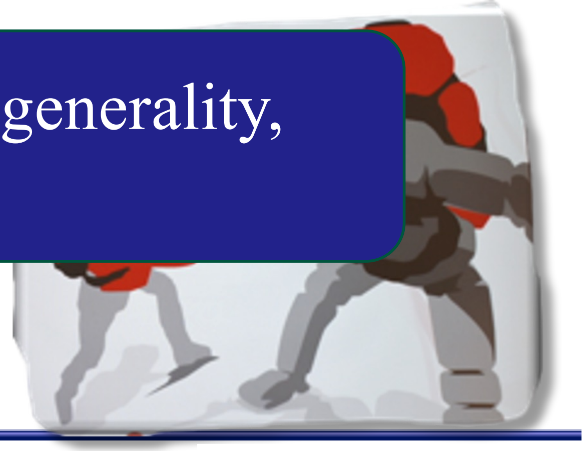
T

interaction through understanding the other person's situation

pr

To make this, OMOTENASHI requires not only generality, but also a high degree of autonomy.

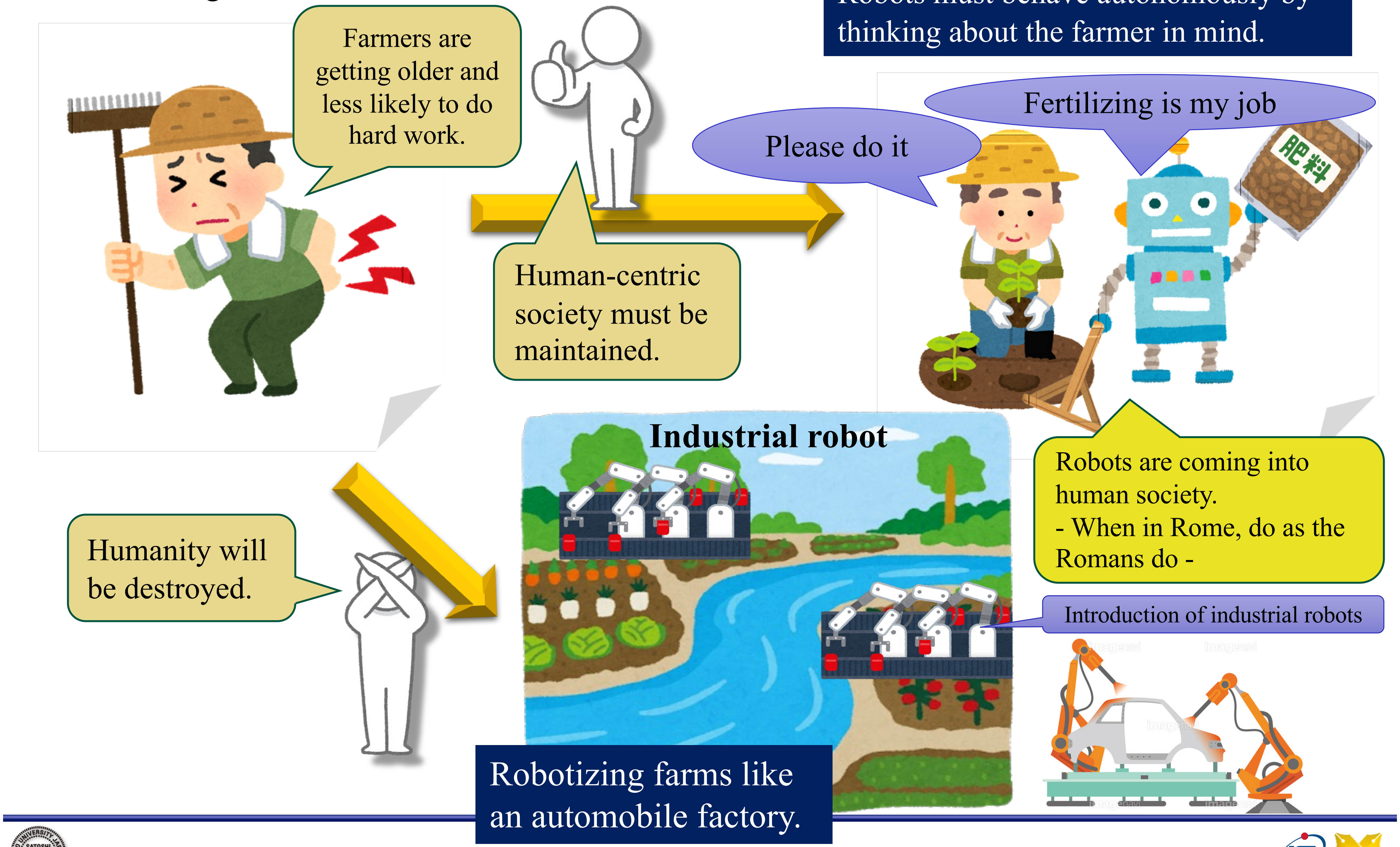
perform OMOTENASHI.



OMOTENASHI was the term used as a concept for the Summer Olympics in Tokyo in 2020.

The 2040 Problem - The Near Future of Japan’s Aging -

Situation of agriculture.



Farmers are getting older and less likely to do hard work.

Human-centric society must be maintained.

Robots must behave autonomously by thinking about the farmer in mind.

Please do it

Fertilizing is my job

Robots are coming into human society.
- When in Rome, do as the Romans do -

Humanity will be destroyed.

Industrial robot

Introduction of industrial robots

Robotizing farms like an automobile factory.

The Japanese Society for Artificial Intelligence Ethical Guidelines

Heading of the article in Code of Ethics of JSAI

一般社団法人 JSAI 人工知能学会
The Japanese Society for Artificial Intelligence

1 Contribution to humanity
2 Abidance of laws and regulations
3 Respect for the privacy of others
4 Fairness
5 Security
6 Act with integrity
7 Accountability and Social Responsibility
8 Communication with society and self-development
9 Abidance of ethics guidelines by AI

The Japanese Society for Artificial Intelligence Ethical Guidelines

一般社団法人
JSAI 人工知能学会
The Japanese Society for Artificial Intelligence

1 Contribution to humanity
2 Abidance of laws and regulations
3 Respect for the privacy of others
4 Fairness
5 Security
6 Act with integrity
7 Accountability and Social Responsibility
8 Communication with society and self-development
9 Abidance of ethics guidelines by AI

The AI we are going to develop itself will have to comply with these guidelines.

AI that complies with Articles 1 to 8 means that this AI is assumed to be an autonomous AI.

The Japanese Society for Artificial Intelligence Ethical Guidelines

一般社団法人 JSAI 人工知能学会
The Japanese Society for Artificial Intelligence

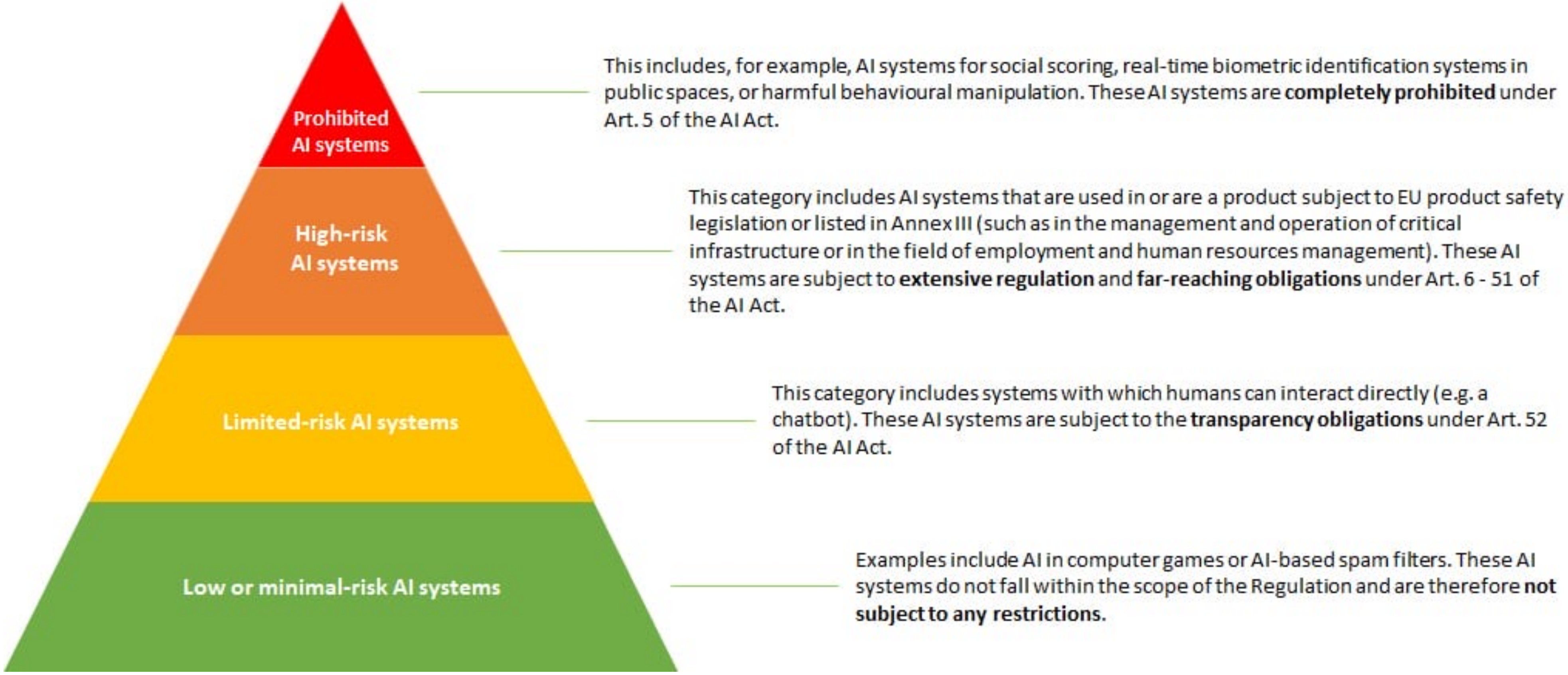
1 Contribution to humanity
2 Abidance of laws and regulations
3 Respect for the privacy of others
4 Fairness
5 Security
6 Act with integrity
7 Accountability and Social Responsibility
8 Communication with society and self-development
9 Abidance of ethics guidelines by AI

The AI we are going to develop itself will have to comply with these guidelines.

AI that complies with Articles 1 to 8 means that this AI is assumed to be an autonomous AI.

In Japan, caution against autonomous AI is not so high, as evidenced by the fact that autonomous AI has long appeared as characters in manga and anime.

AI regulation of EU



This includes, for example, AI systems for social scoring, real-time biometric identification systems in public spaces, or harmful behavioural manipulation. These AI systems are **completely prohibited** under Art. 5 of the AI Act.

This category includes AI systems that are used in or are a product subject to EU product safety legislation or listed in Annex III (such as in the management and operation of critical infrastructure or in the field of employment and human resources management). These AI systems are subject to **extensive regulation** and **far-reaching obligations** under Art. 6 - 51 of the AI Act.

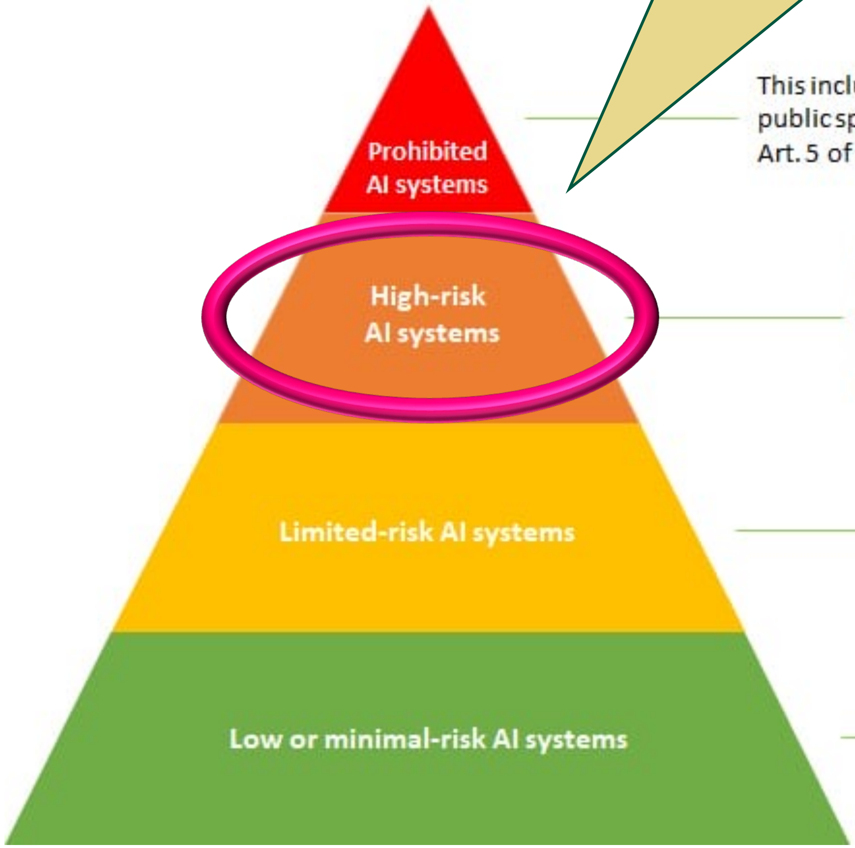
This category includes systems with which humans can interact directly (e.g. a chatbot). These AI systems are subject to the **transparency obligations** under Art. 52 of the AI Act.

Examples include AI in computer games or AI-based spam filters. These AI systems do not fall within the scope of the Regulation and are therefore **not subject to any restrictions.**

The four categories of AI hazards in the EU AI Regulation Act.

AI regulation of EU

Generative AI is included in high-risk AI systems.



This includes, for example, AI systems for social scoring, real-time biometric identification systems in public spaces, or harmful behavioural manipulation. These AI systems are **completely prohibited** under Art. 5 of the AI Act.

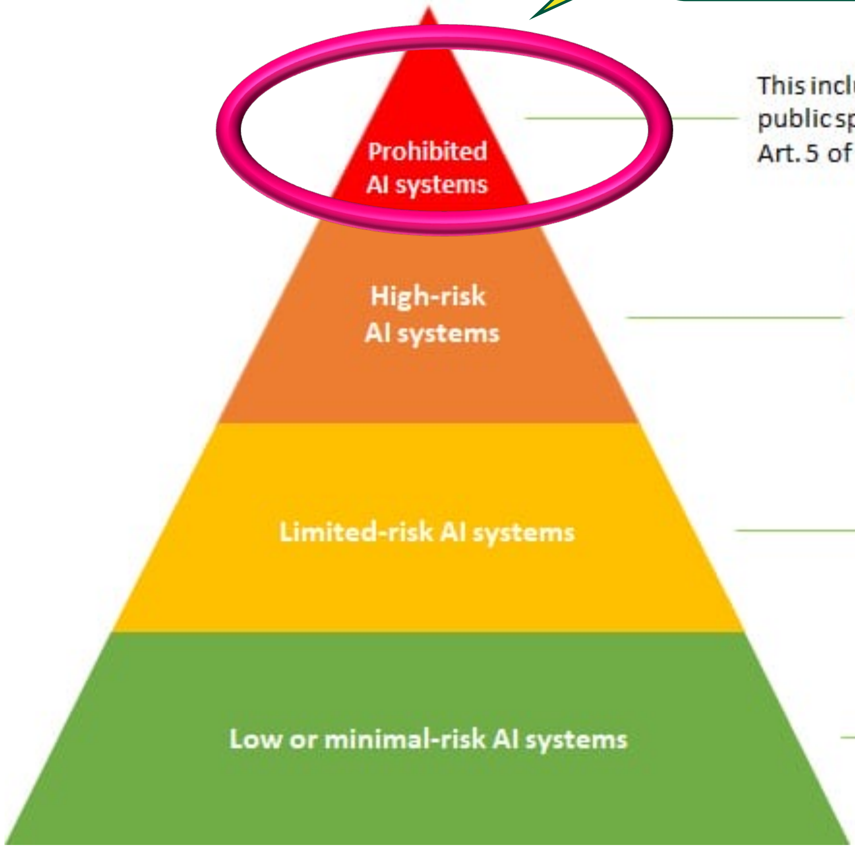
This category includes AI systems that are used in or are a product subject to EU product safety legislation or listed in Annex III (such as in the management and operation of critical infrastructure or in the field of employment and human resources management). These AI systems are subject to **extensive regulation** and **far-reaching obligations** under Art. 6 - 51 of the AI Act.

This category includes systems with which humans can interact directly (e.g. a chatbot). These AI systems are subject to the **transparency obligations** under Art. 52 of the AI Act.

Examples include AI in computer games or AI-based spam filters. These AI systems do not fall within the scope of the Regulation and are therefore **not subject to any restrictions.**

AI regulation of EU

Autonomous AI is assumed to be classified as the highest-risk AI, as it will interact actively with humans and be strongly involved in human thinking.



This includes, for example, AI systems for social scoring, real-time biometric identification systems in public spaces, or harmful behavioural manipulation. These AI systems are **completely prohibited** under Art. 5 of the AI Act.

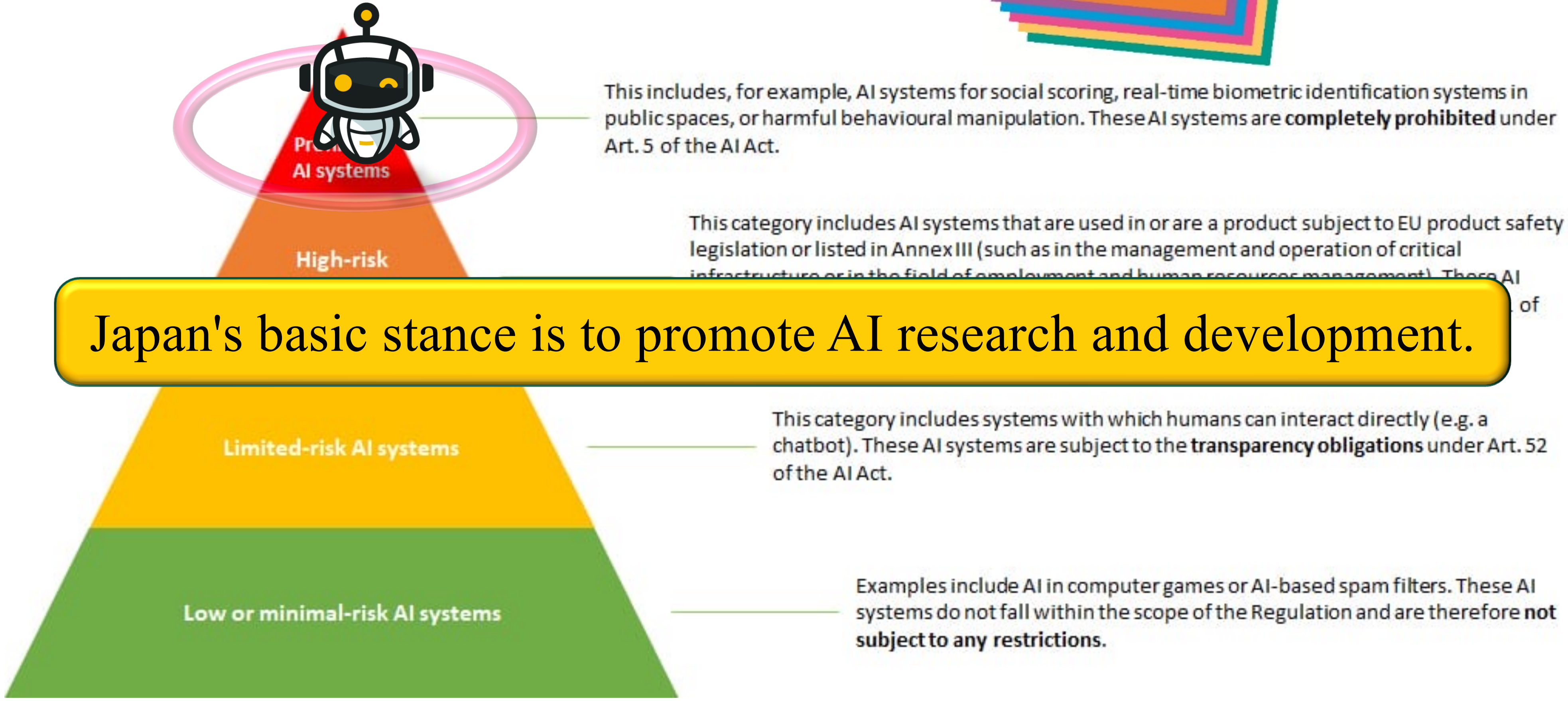
This category includes AI systems that are used in or are a product subject to EU product safety legislation or listed in Annex III (such as in the management and operation of critical infrastructure or in the field of employment and human resources management). These AI systems are subject to **extensive regulation** and **far-reaching obligations** under Art. 6 - 51 of the AI Act.

This category includes systems with which humans can interact directly (e.g. a chatbot). These AI systems are subject to the **transparency obligations** under Art. 52 of the AI Act.

Examples include AI in computer games or AI-based spam filters. These AI systems do not fall within the scope of the Regulation and are therefore **not subject to any restrictions.**

AI regulation of EU

G7 2023 HIROSHIMA SUMMIT



This includes, for example, AI systems for social scoring, real-time biometric identification systems in public spaces, or harmful behavioural manipulation. These AI systems are **completely prohibited** under Art. 5 of the AI Act.

This category includes AI systems that are used in or are a product subject to EU product safety legislation or listed in Annex III (such as in the management and operation of critical infrastructure or in the field of employment and human resources management). These AI

of

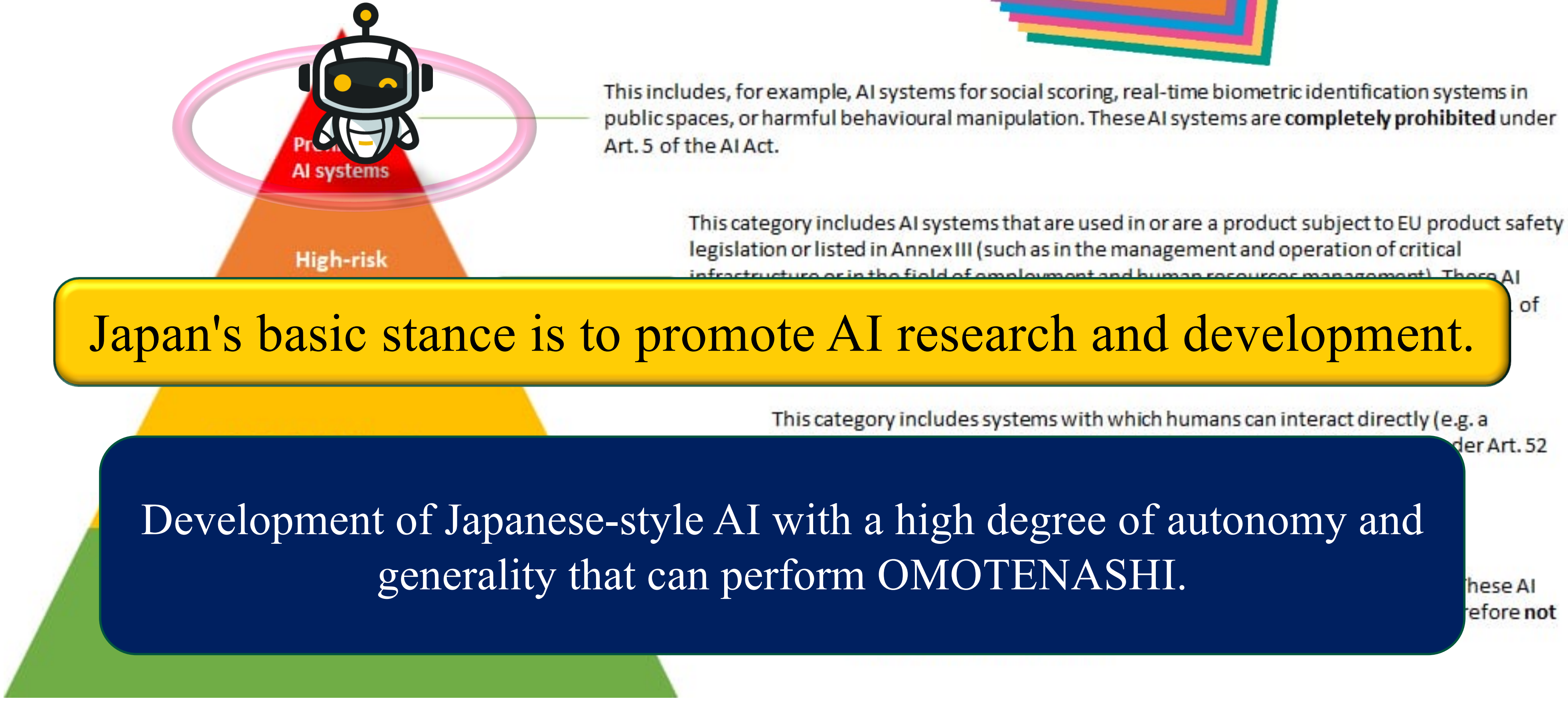
Japan's basic stance is to promote AI research and development.

This category includes systems with which humans can interact directly (e.g. a chatbot). These AI systems are subject to the **transparency obligations** under Art. 52 of the AI Act.

Examples include AI in computer games or AI-based spam filters. These AI systems do not fall within the scope of the Regulation and are therefore **not subject to any restrictions.**

AI regulation of EU

G7 2023 HIROSHIMA SUMMIT



This includes, for example, AI systems for social scoring, real-time biometric identification systems in public spaces, or harmful behavioural manipulation. These AI systems are **completely prohibited** under Art. 5 of the AI Act.

This category includes AI systems that are used in or are a product subject to EU product safety legislation or listed in Annex III (such as in the management and operation of critical infrastructure or in the field of employment and human resources management). These AI

of

Japan's basic stance is to promote AI research and development.

This category includes systems with which humans can interact directly (e.g. a

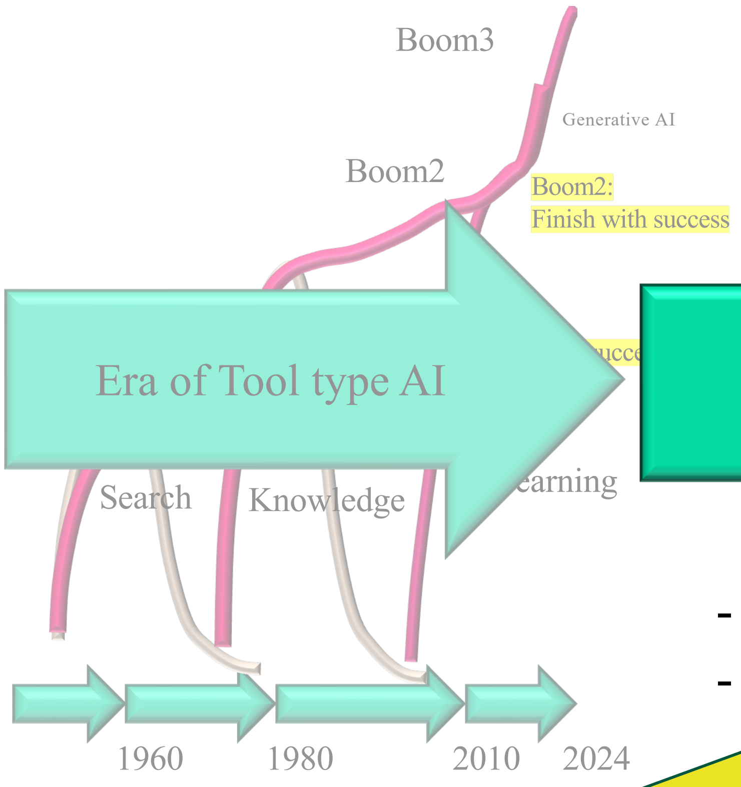
der Art. 52

Development of Japanese-style AI with a high degree of autonomy and generality that can perform OMOTENASHI.

hese AI

efore **not**

Ability of Autonomous AI is basically same as Tool type AI



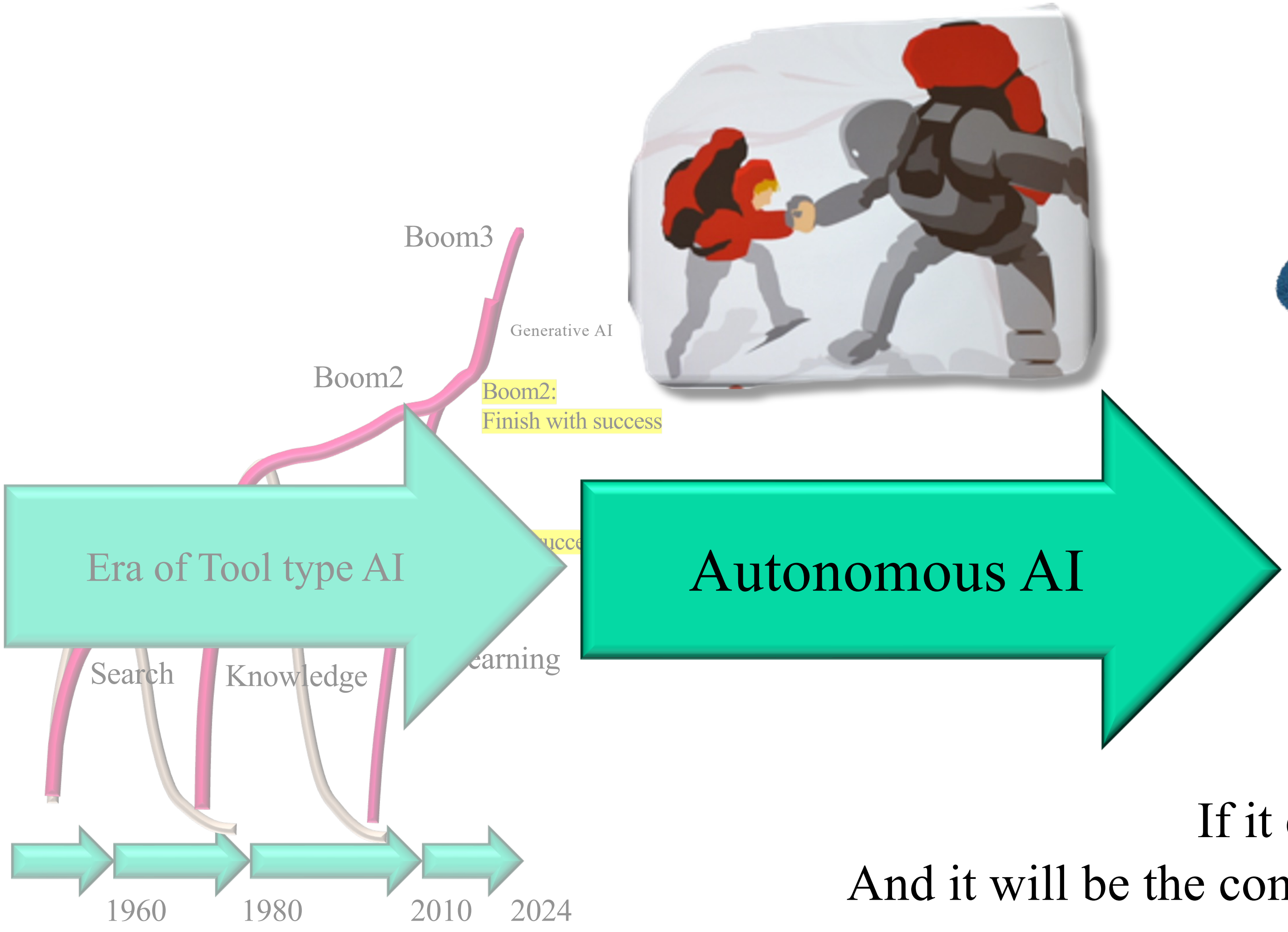
Autonomous AI

- AI that learns human data cannot essentially surpass human.
- These AI cannot invent things that humans cannot understand.

Even if an AI understands a vast number of papers and generates new hypotheses, even if those hypotheses are Nobel Prize-worthy discoveries, these discoveries will never leave the range of human comprehension.

It includes geniuses that the average person does not understand ...

Will AI emerge, which has an intelligence that humans cannot understand?



Boom3

Generative AI

Boom2

Boom2:
Finish with success

ucce

Era of Tool type AI

Autonomous AI

Search

Knowledge

earning

1960

1980

2010

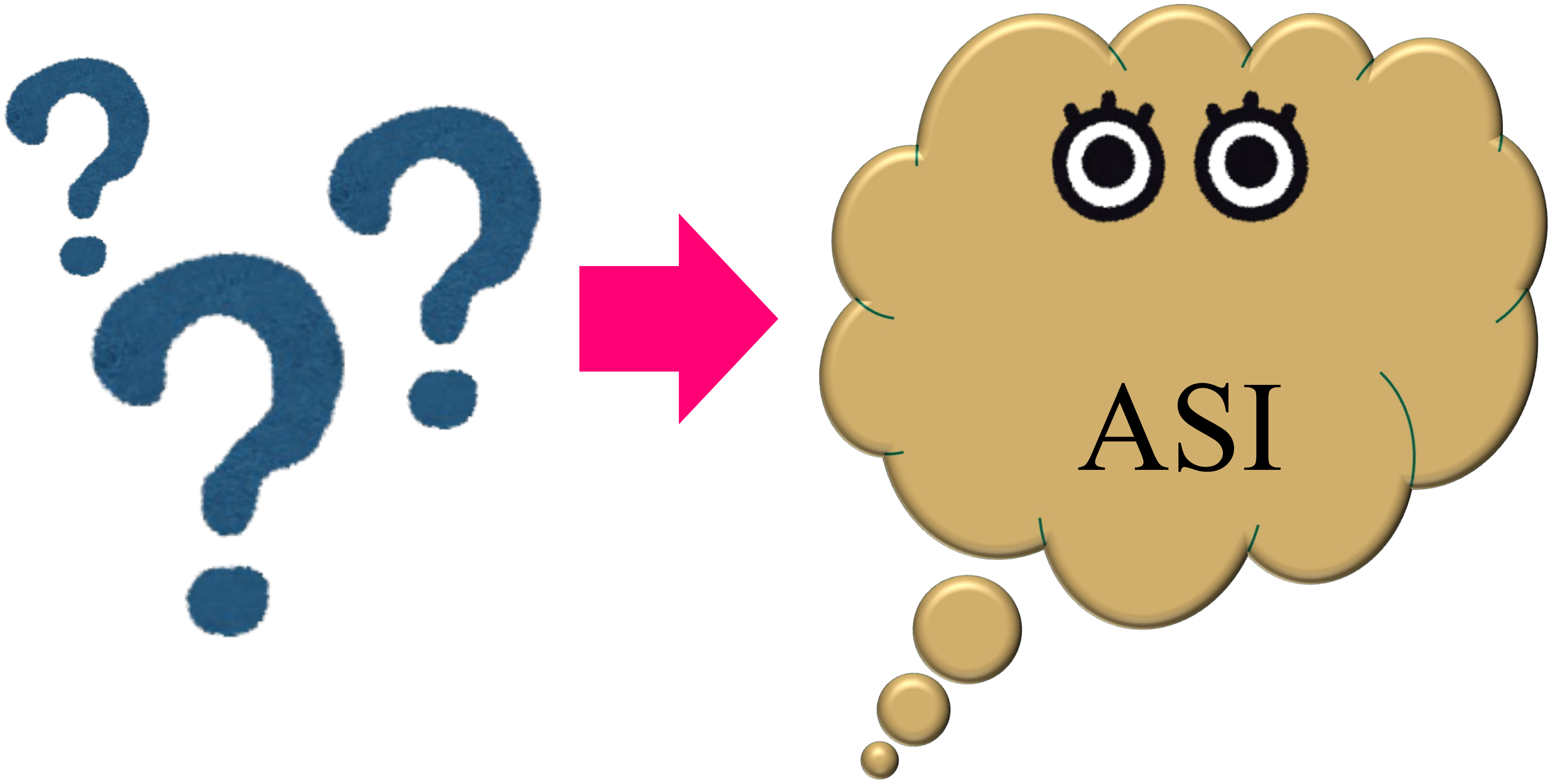
2024

If it emerges, it will be **ASI** ...
And it will be the coming of the **Singularity** ...

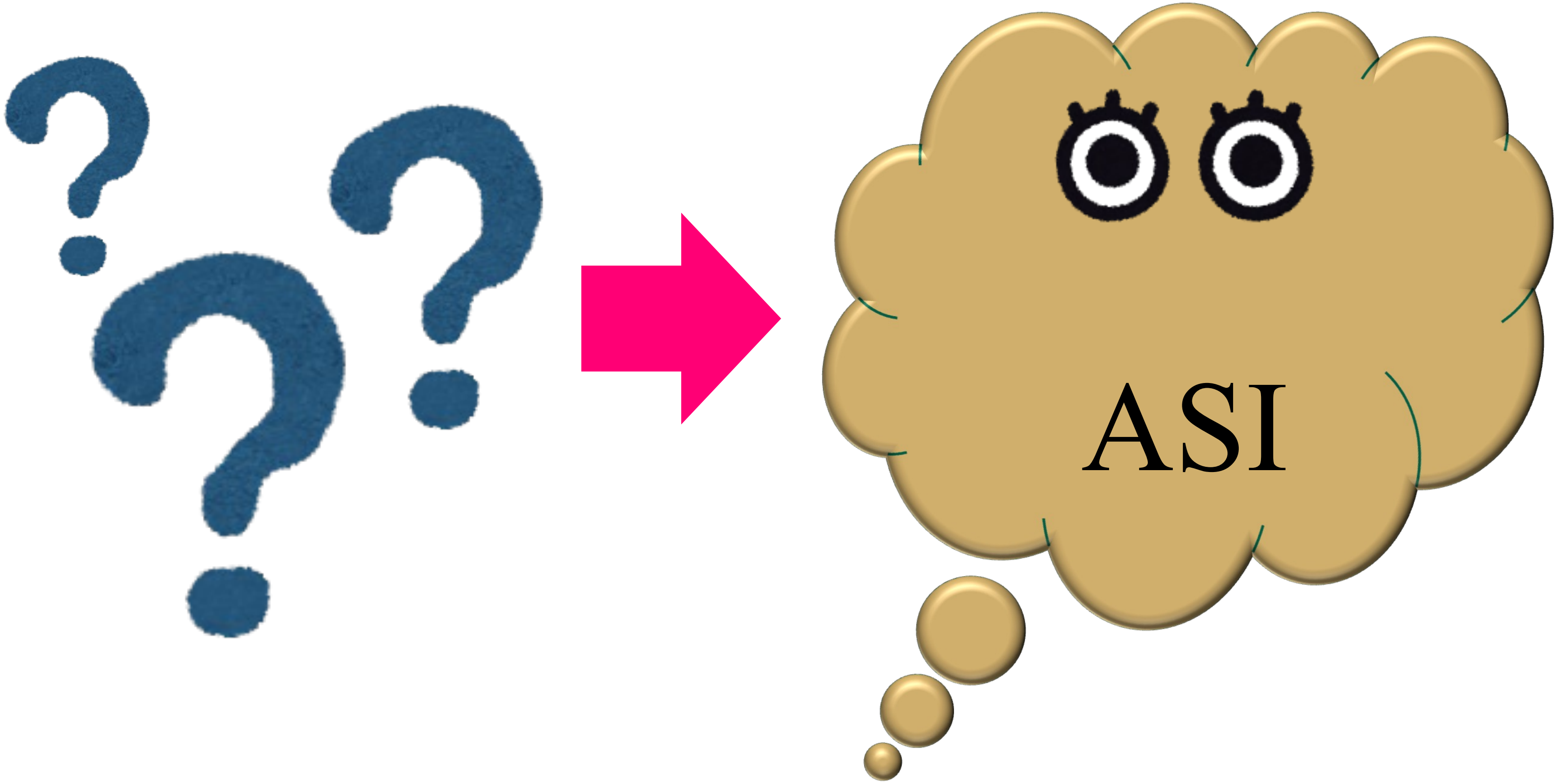
Why is the term “emerge” used?
→ Because ASI is not something that humans create.

- ASI: Artificial Super Intelligence

There is a possibility that ASI may emerge.

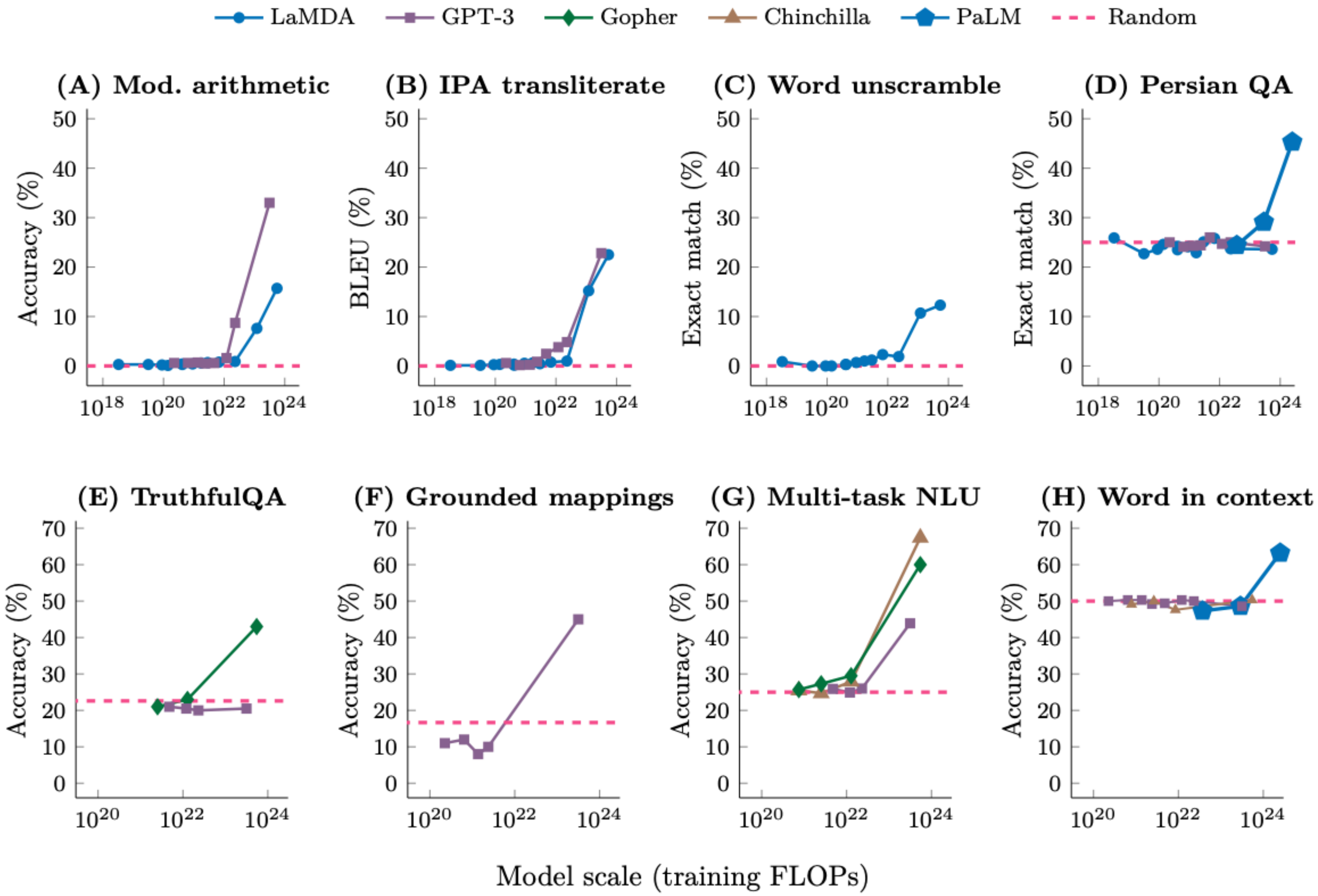


There is a possibility that ASI may emerge.



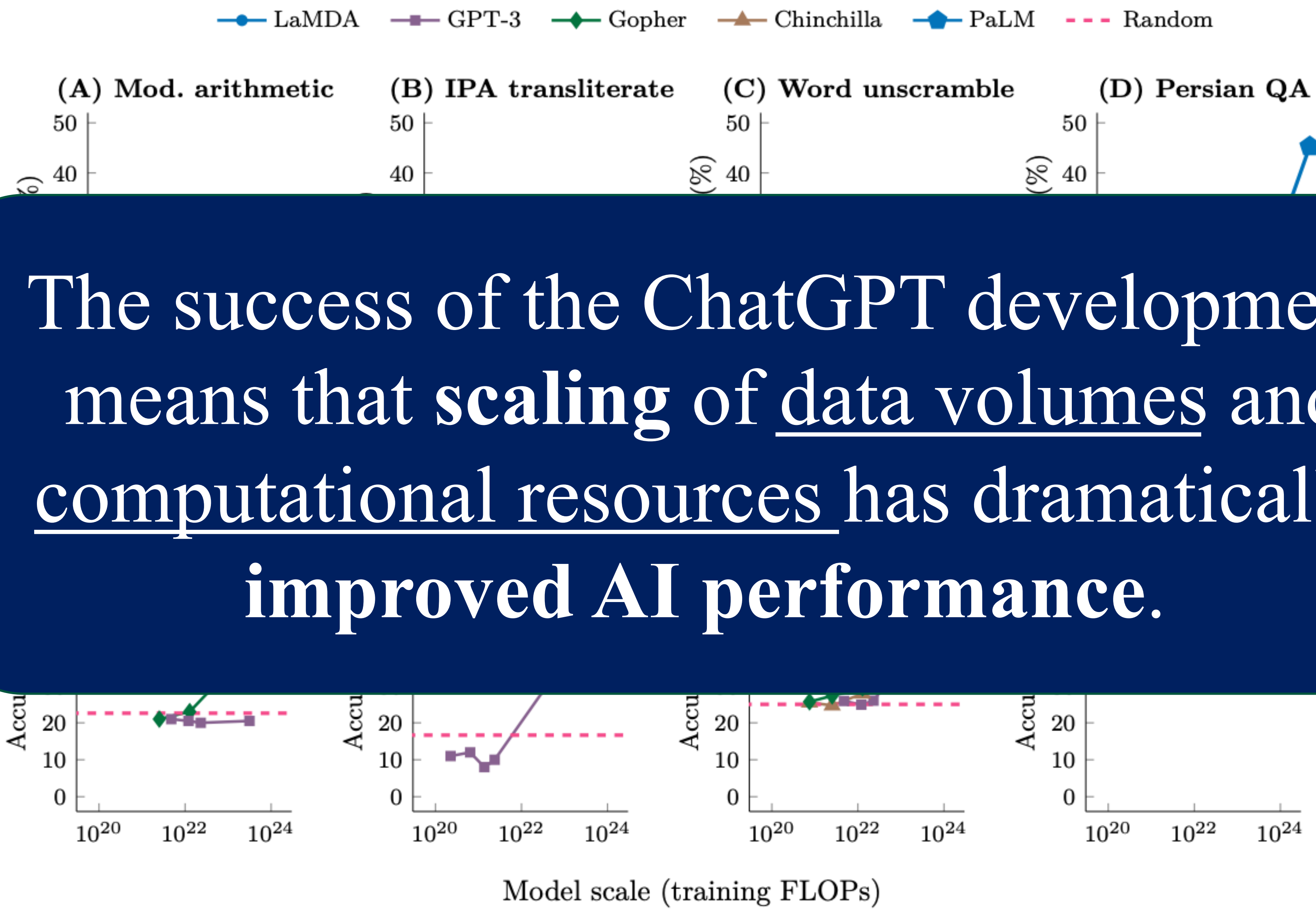
A clue to the emergence of ASI is **scaling.**

Quality changes through scaling



https://arxiv.org/pdf/2206.07682.pdf

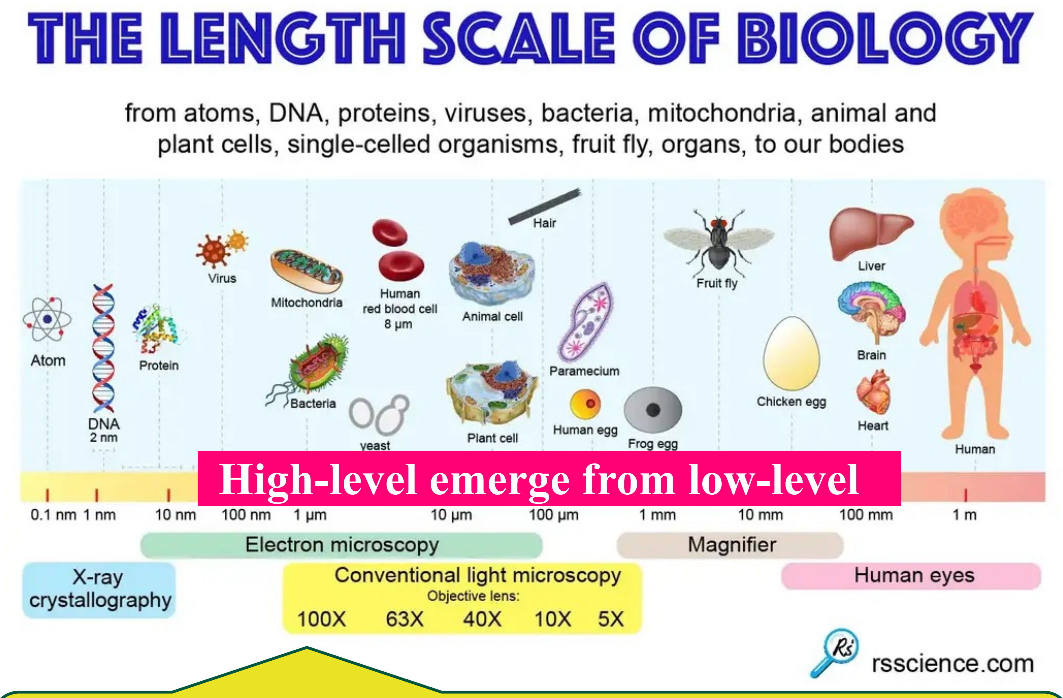
Quality changes through scaling



The success of the ChatGPT development means that **scaling** of data volumes and computational resources has dramatically **improved AI performance.**

https://arxiv.org/pdf/2206.07682.pdf

Large quality changes emerge by scaling



High-level emerge from low-level

Life has high intelligence like humans by emergence through scaling.

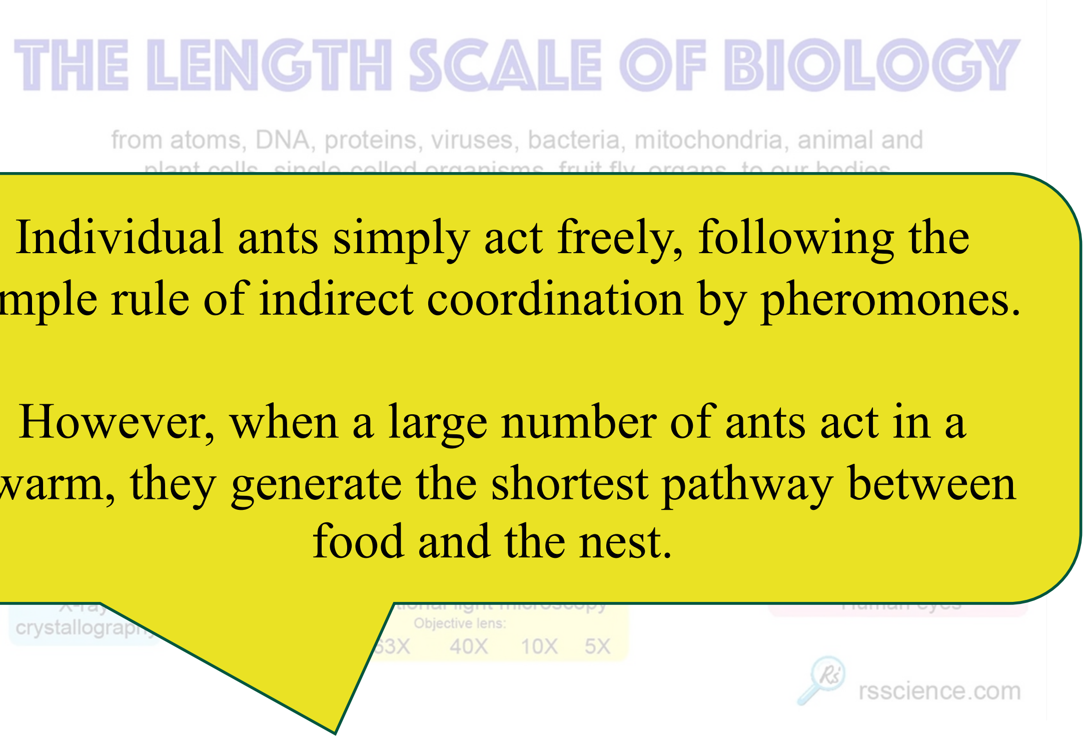
Nature is also composed of diverse scales.

American Museum of Natural History
UNIVERSE
Virgo Supercluster
GALAXIES
Our Local Group of Galaxies
Milky Way Galaxy
STARS
Our Stellar Neighborhood
Sun
PLANETS
Our Planetary System
Earth

The quality change through scaling is the essence of the world …

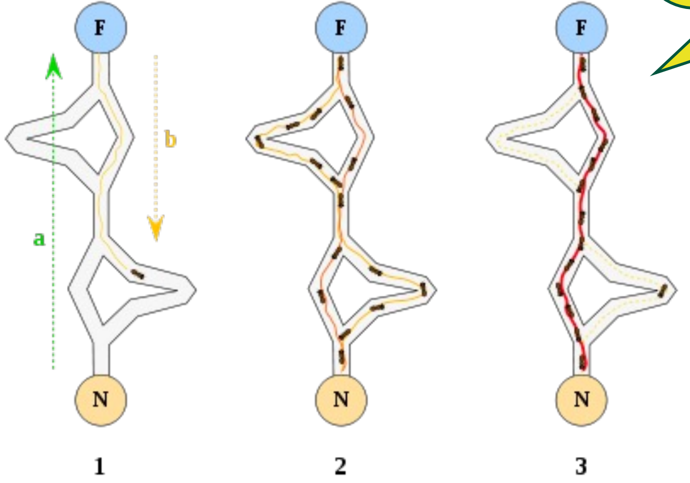
https://www.reddit.com/r/space/comments/390yia/this_diagram_shows_our_cosmic_address_at_a_glance/
https://singularity-bio.jp/amateras/

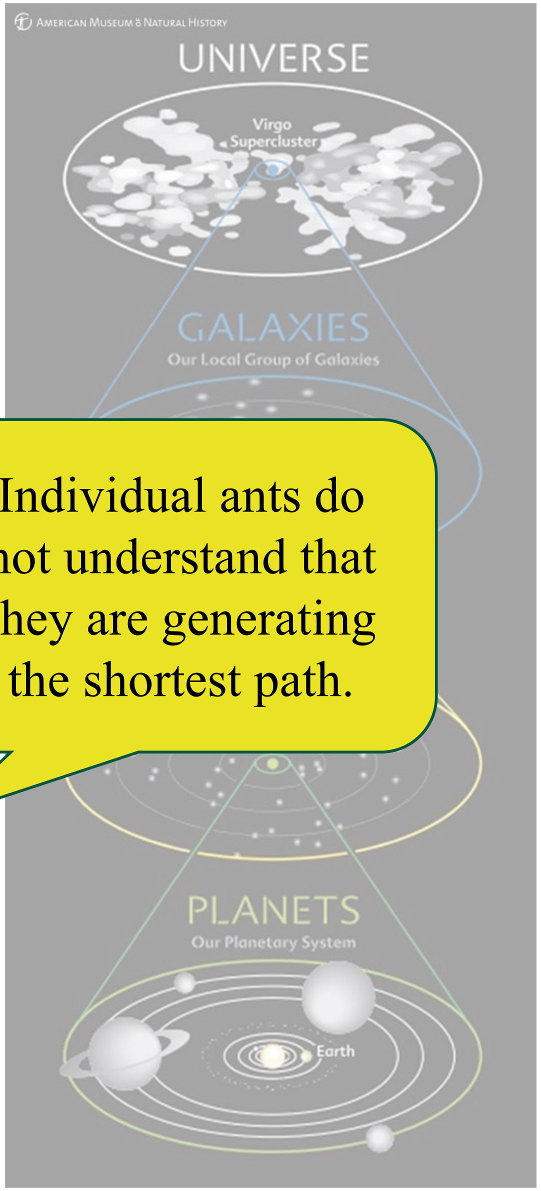
Example of shortest line emergence as the number of ants scales


THE LENGTH SCALE OF BIOLOGY
from atoms, DNA, proteins, viruses, bacteria, mitochondria, animal and plant cells, single-celled organisms, fruit fly, organs, to our bodies
X-ray crystallography
tional light microscopy
Objective lens:
63X 40X 10X 5X
Human eyes
rsscience.com

Individual ants simply act freely, following the simple rule of indirect coordination by pheromones.

However, when a large number of ants act in a swarm, they generate the shortest pathway between food and the nest.

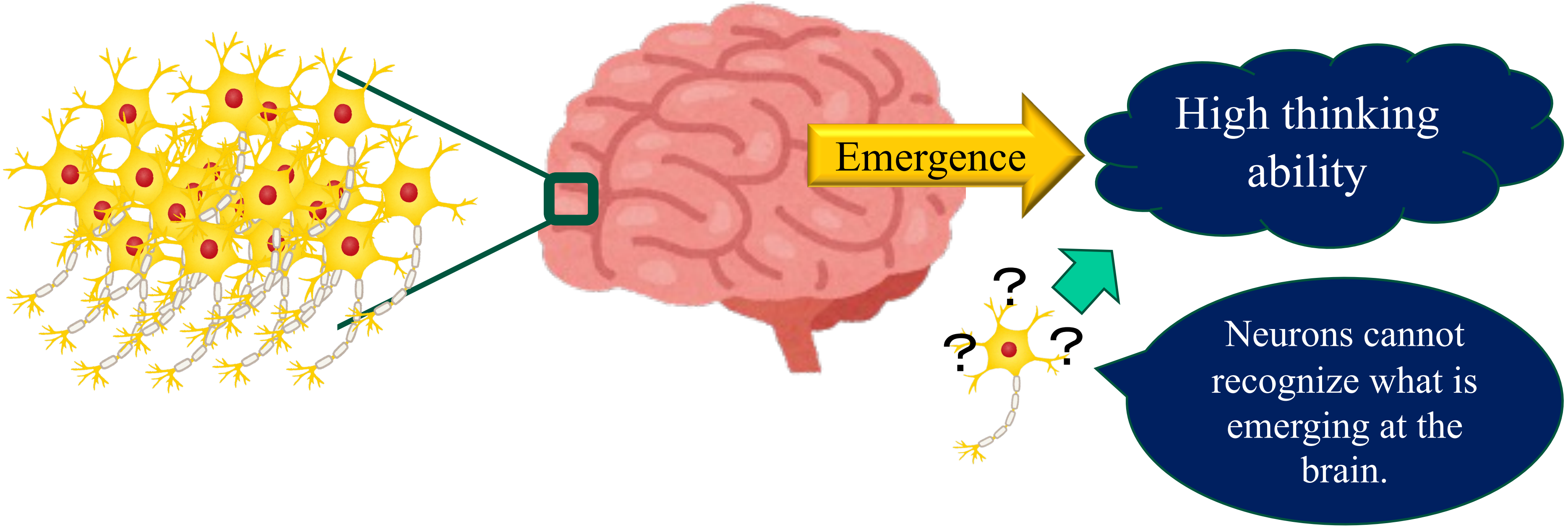




Individual ants do not understand that they are generating the shortest path.

https://www.reddit.com/r/space/comments/390yia/this_diagram_shows_our_cosmic_address_at_a_glance/
https://singularity-bio.jp/amateras/

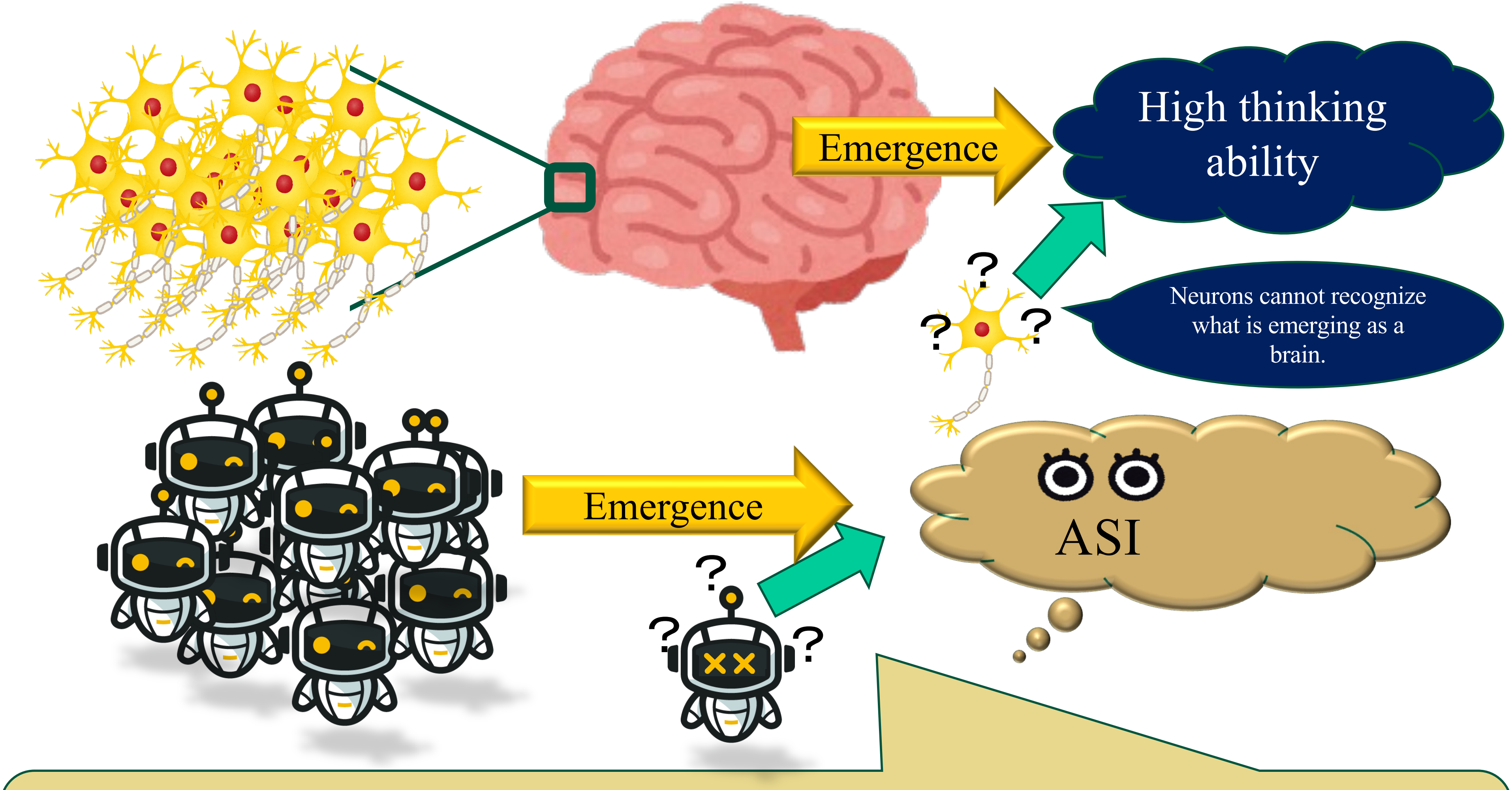
The lower layers and the upper layers from which the lower layers emerge have basically different dynamics



High thinking ability

Neurons cannot recognize what is emerging at the brain.

ASI may emerge if autonomous AI is scaled.



Emergence

High thinking ability

Neurons cannot recognize what is emerging as a brain.

Emergence

ASI

- The dynamics of individual autonomous AIs and ASIs are different.
- The intelligence of ASI cannot be understood by humans.
- For people, the impact of ASI will be on the same level as natural phenomena.

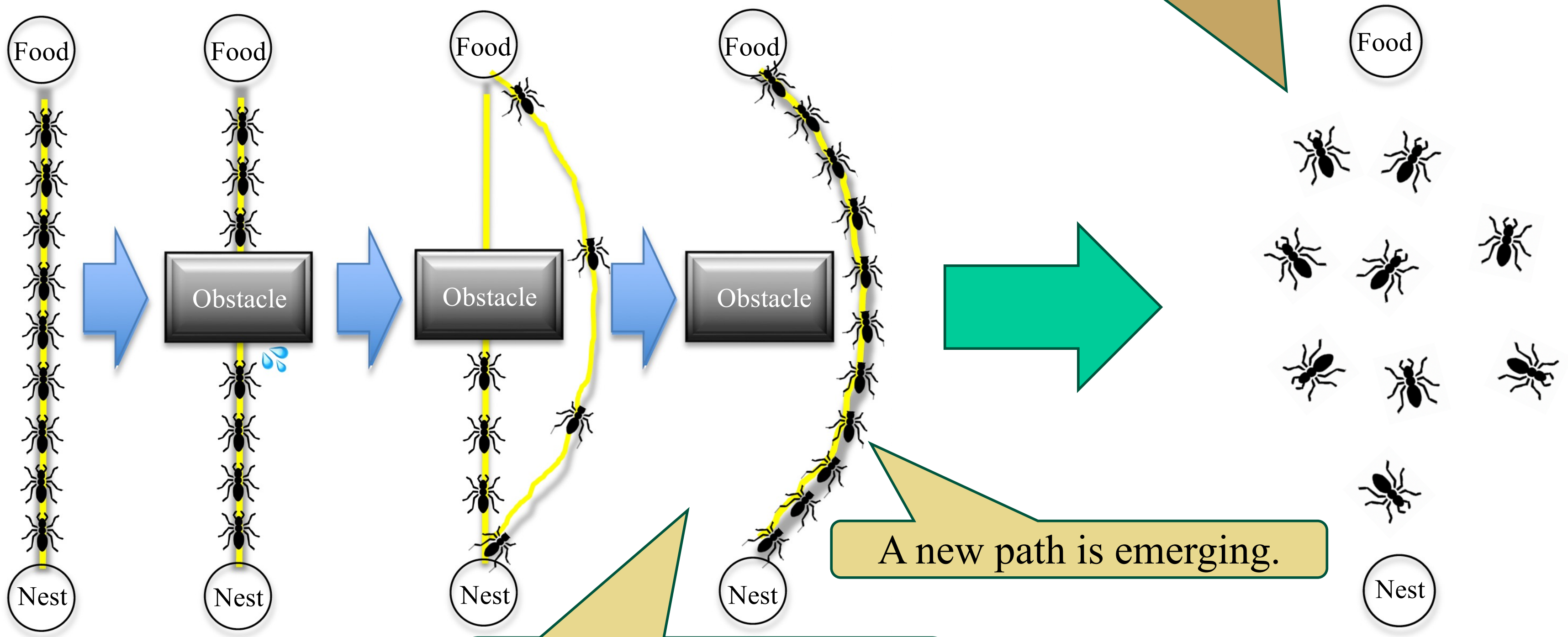
※ASI: Artificial Super Intelligence

However, intervention into ASI may be possible

If the dynamics of the lower layers change, the dynamics of the emergent upper layers also change.

Intervening into the emergent ant line itself is difficult.

Intervening into the behavioral rules of each ant itself changes what is emergent !!

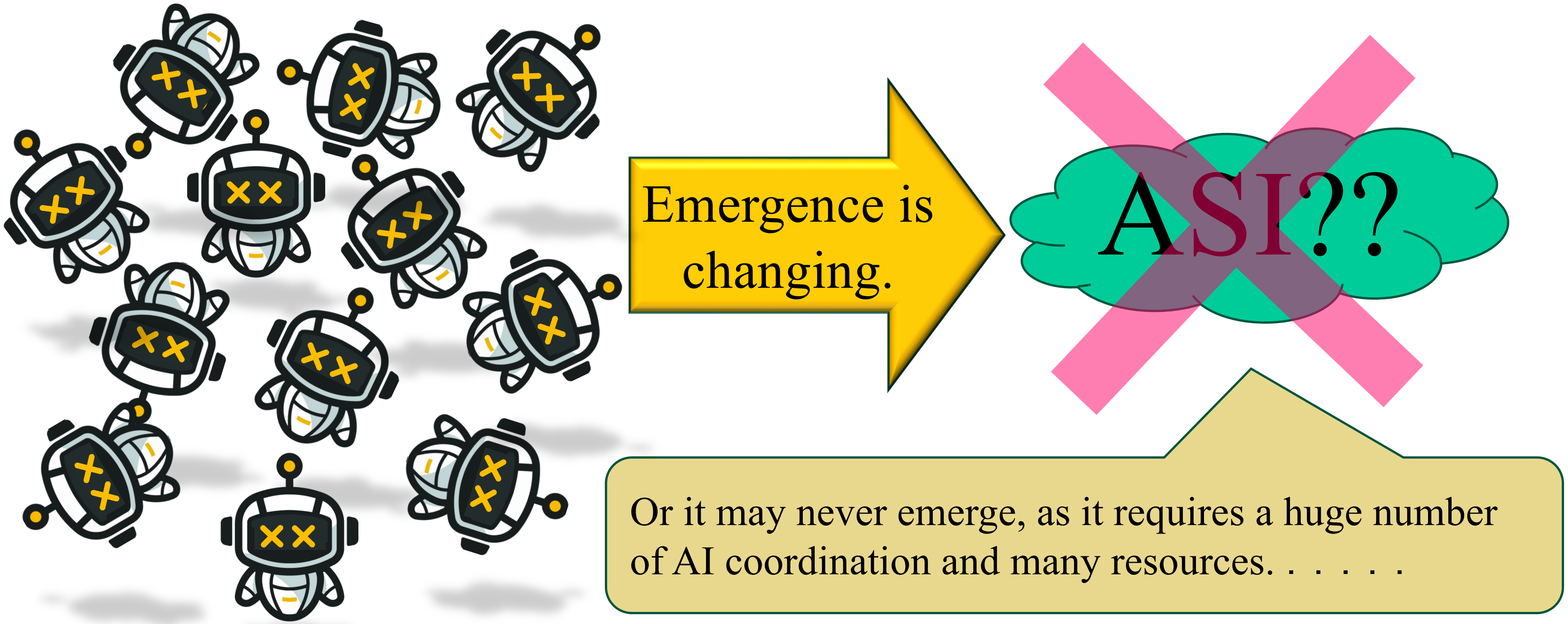


A new path is emerging.

Ant line has high redundancy

However, intervention into ASI may be possible

Modifying the behavioral rules of each autonomous AI may change the behavior of emerging ASI …



Emergence is changing.

ASI??

Or it may never emerge, as it requires a huge number of AI coordination and many resources. . . . . .

Although even current AI poses threats to human survival, such as the generation of demagogic fakes and AI weapons, etc. ….

Conclusion remark

- Humanity's desires never stop ... People seek. Technology evolves.
- Next to the era of tool-oriented AI is the era of autonomous AI.
- OMOTENASHI is an important keyword for autonomous AI.
- Data Learning-type AI is within human understanding.
- The scaling of autonomous AI could lead to the emergence of ASI.
- The intelligence of ASI is no longer within human comprehension.
- However, emergence can be controlled by controlling coordination behavior of each autonomous AI.