2024.06.30 WCCI 2024 Open Forum on AI Governance

Building an AI Safety Research Ecosystem in Japan

Koichi Takahashi

AI Alignment Network

Koichi Takahashi, Ph.D

Research interests:

AI Alignment, AI-robot driven science, computational systems biology

Affiliations:

AI Alignment Network

Graduate School of Media and Governance, Keio University

TRIP-AGIS and BDR, RIKEN

Whole Brain Architecture Initiative

Developed an **autonomous AI-Robot** that discovers optimum induction condition of iPS cells to Retinal Pigment Epithelium cells. (eLife 2020)

Takahashi (2018)

"Scenarios and branch points to future machine intelligence"

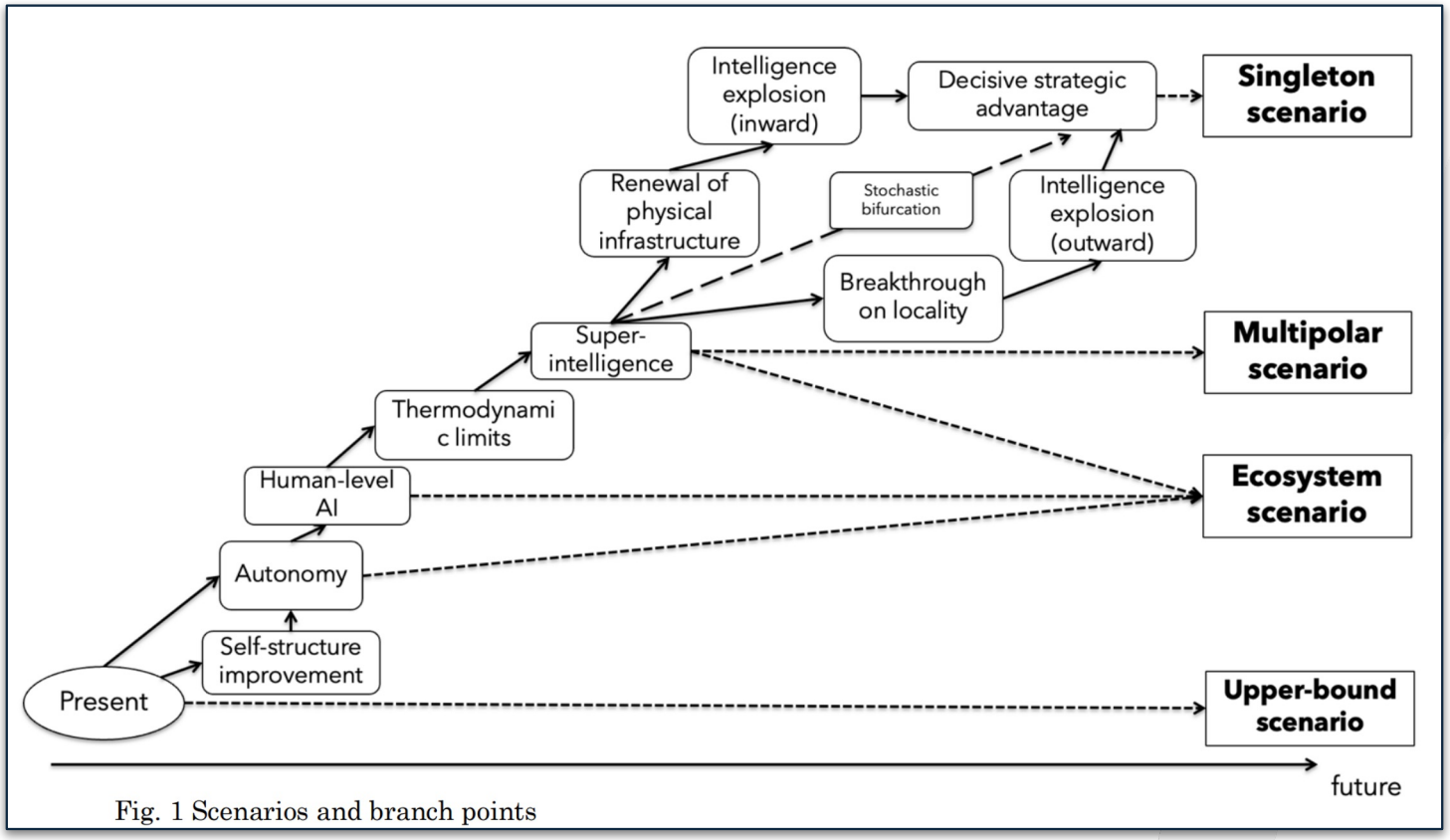


Fig. 1 Scenarios and branch points

arXiv:2302.14478

Contents

1. State of AI Safety Ecosystem in Japan

2. Introducing AI Alignment Network (ALIGN)

3. Rethinking the scope of AI Alignment/Safety research

Contents

1. State of AI Safety Ecosystem in Japan

2. Introducing AI Alignment Network (ALIGN)

3. Rethinking the scope of AI Alignment/Safety research

Backgound

Japan has seen active discussion on global/national AI governance

“Before genenative AI”

- 2015~ academia-led initiatives e.g. “AI & Society Meetings” formed.
- 2017 Ministry of Internal Affairs: “Draft AI R&D GUIDELINES for International Discussions” (in which Takahashi joined as an expert)
- 2019 "Social Principles of Human-Centric AI" …etc.

“After generative AI”:

- 2023 **Hiroshima AI Process**
- 2024 world's third **AI Safety Institute**
- 2024 Draft for the **“Basic Law for the Promotion of Responsible AI”** by the leading party. … etc.

AISI Japan AI Safety Institute

Backgound

The main focus has been on the immediate risks.

Today's urgent risks from current AI

harmful contents | bias | fake news | copyright | hallucination | environmental costs | privacy

Background

But globally, risks from future AI with more capabilities is a concern.

Today's urgent risks from current AI

harmful contents | bias | fake news | autonomous weapons

copyright | hallucination | environmental costs | surveillance by the state

privacy | Bio/cyber attacks using AI | uncontrollable AI

Catastrophic/existential/long-term risks from future AI

Types of Catastrophic AI risks by Center for AI Safety

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. *ArXiv, abs/2306.12001*.

Yoshua Bengio "AGI Safety"

Background

In the US, UK, Canada, AI safety organizations are proliferating.

https://aisafety.world/

Notably, many of them are independent, non-profit orgs.

some examples:

- Future of Life Institute founded by Dr. Max Tegmark et al. in 2024.
- Center for AI Safety founded by Dr. Dan Hendrycks in 2022.

Page. 9

Background

Japan in AI safety / governance

- We have seen relatively little attention towards long-term risks of AI in Japan.
- There's only a minor presence of Japanese researchers in AI alignment up to now.

We see this as a missed opportunity.

- **Why?**
  - **Japan, as the only G7 country in Asia, plays a prominent role in international AI governance**
  - **Japan has a large researcher population**
  - **Geographical and cultural diversity in global AI governance is important**

Contents

1. State of AI Safety Ecosystem in Japan

2. Introducing AI Alignment Network (ALIGN)

3. Rethinking the scope of AI Alignment/Safety research

ALIGN

“AI Alignment Network (ALIGN)
is a non-profit organization in Japan
that aims to create an ecosystem of researchers and practitioners
who will pave the way to a hopeful future
where AI is harmoniously implemented in society.”

established Sep. 2024

www.aialign.net

Activities

1. Research

Conduct and support research to overcome risks that lie between AI and humanity while realizing new values.

2. Community building

Foster a community of researchers and practitioners in the field. Network together the various relevant stakeholders.

3. Outreach

Communicate with society with various means. Make proposals viable to the creation of a new society.

Members

Koichi Takahashi Ph.D.
Chair.
Project prof. at Keio Univ., PI at RIKEN, expert in AI for science

Ryota Kanai Ph.D.
Cofounder
CEO of Araya Inc, expert in AI and neuroscience.

Hiroshi Yamakawa Ph.D.
Director.
Chairperson of the WBA Initiative, principal researcher at Univ of Tokyo.

Ippei Fujisawa Ph.D.
Director. Researcher at Araya Inc, expert in physics and AI.

Yusuke Hayashi
Director.
Senior Researcher at Digital Design Inc.

Ryuichi Maruyama
COO (tentative)

External Advisors

Dan Hendrycks
Center for AI Safety

TBD

Graduate Student Contributors

Research Fellows

Staff

Community members
contributing as volunteers
(~100 active members in ALIGN's Slack Workspace)

Progress

Webinars

featuring top researchers/field builder in AI Safety

Contest

"Superintelligence Future Scenario Contest"

in conjunction with Japanese Society of Artificial Intelligence

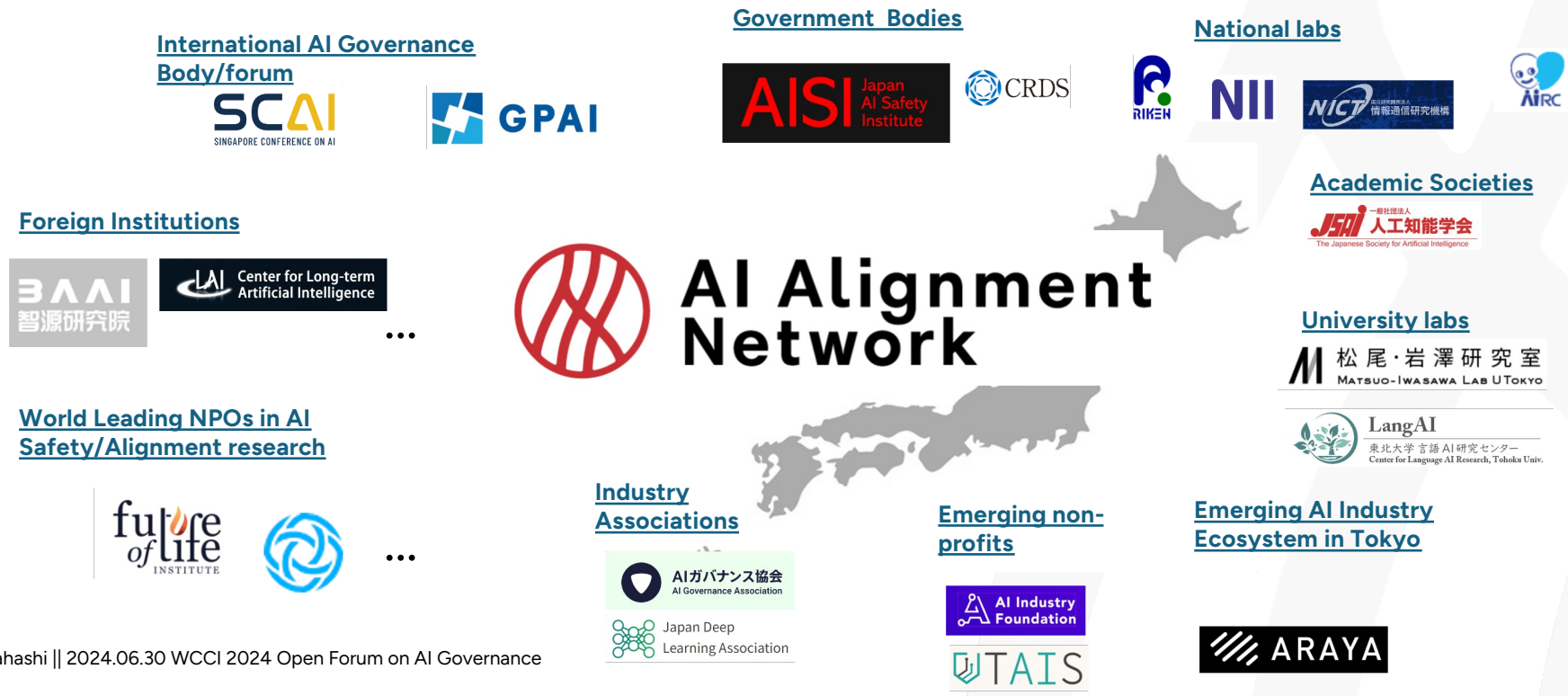
Study Groups

Study Groups on Mechanistic interpretability, AI governance, Guaranteed Safe AI, etc.

Near-term objective

Objective for the next 1.5 yrs: Make AI alignment/safety a normal part of public discussion, research, and policy consideration in Japan.

- We aspire to be the hub of researchers and stakeholders in Japan and abroad.



Contents

1. State of AI Safety Ecosystem in Japan

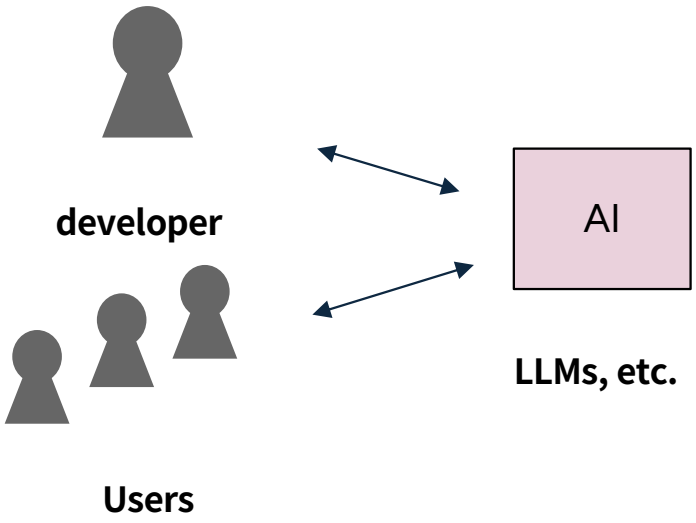
2. Introducing AI Alignment Network (ALIGN)

3. Rethinking the scope of AI Alignment/Safety research

What is AI Alignment?

Some AI Alignment Definitions:

- *The process of ensuring that an AI system's goals and behaviours are in line with its developer's values and intentions.* DSIT (2024). "International Scientific Report on the Safety of Advanced AI: Interim Report"
- *The project of building intelligent autonomous systems that robustly act in our collective interests.* Tan Zhi Xuan（2024）



AI Alignment/Safety in the *narrow* sense:

"Aligning today's AI with developer/users' collective intent/values while preventing harm/misuse"

But...

- **AI is evolving.**
- **AI does not function in isolation.**

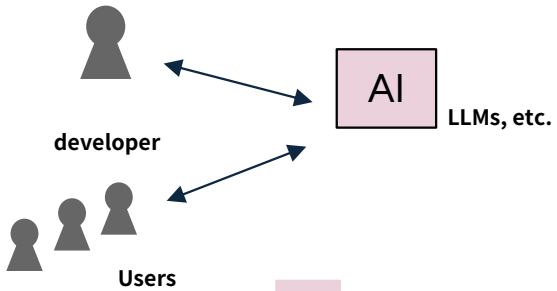
Broadening the scope

Narrow Alignment/Safety

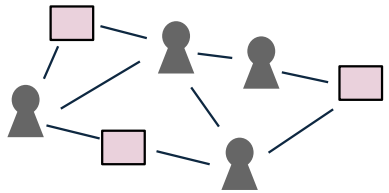
Broad Alignment/Safety

Existing AI

(a) Aligning today's AI with developer/ users' collective intent/values, Preventing harm/misuse

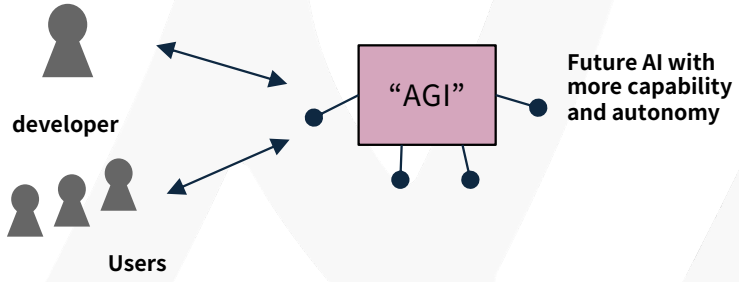


(b) Realizing a harmonious society as a complex Human-AI hybrid system

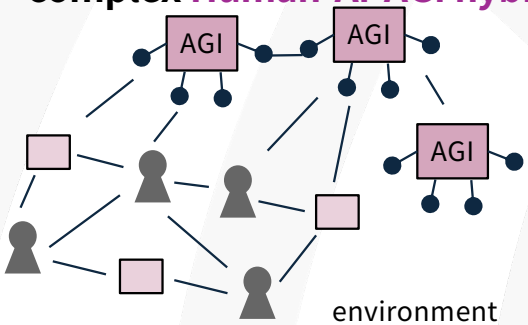


Future AI

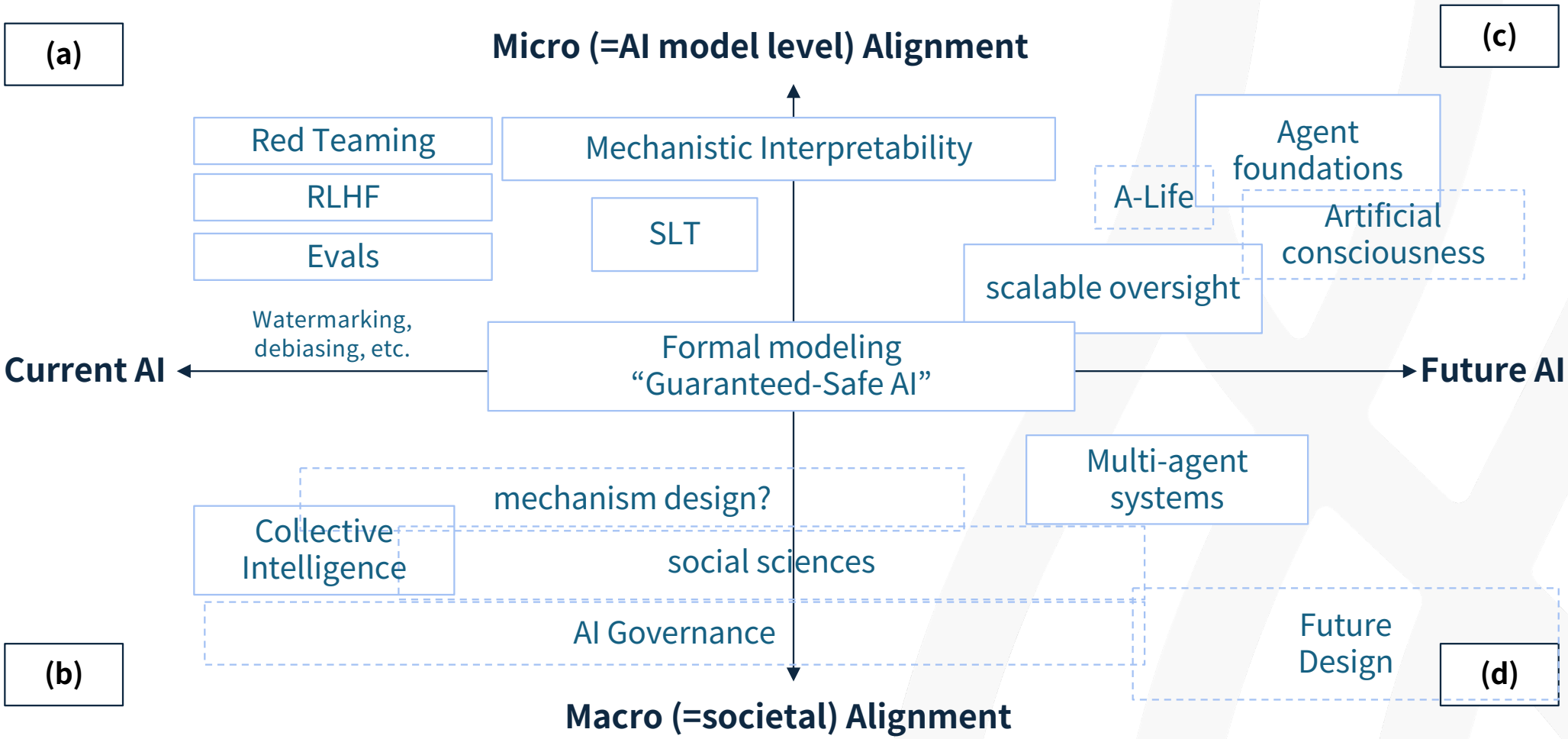
(c) Controlling more capable/autonomous future AI, Aligning their behaviour with human values.



(d) Realizing a harmonious society as a complex Human-AI-AGI hybrid system

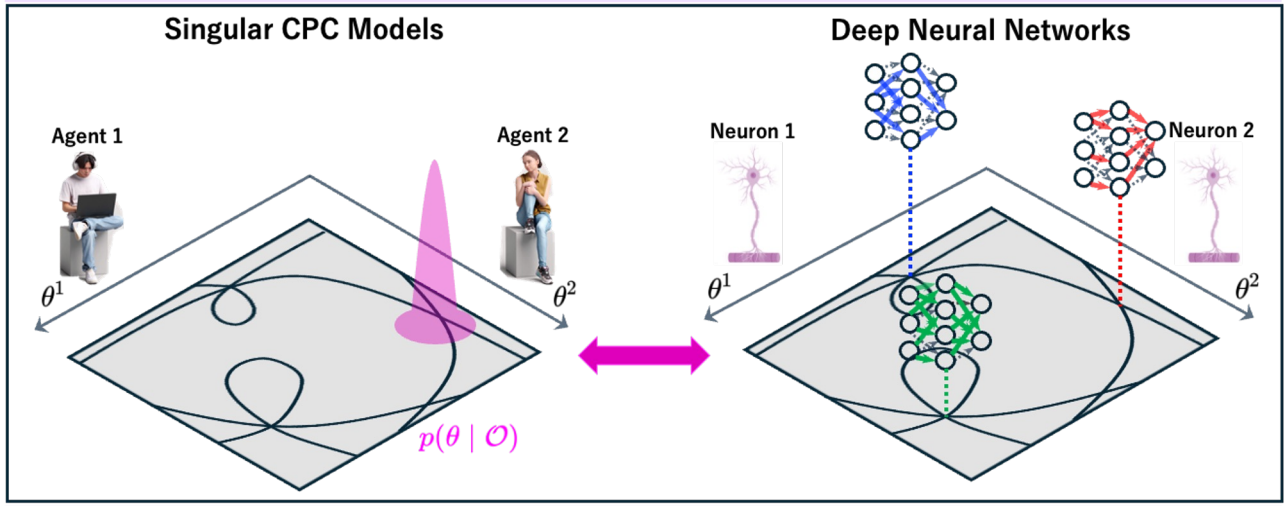


Scope of research



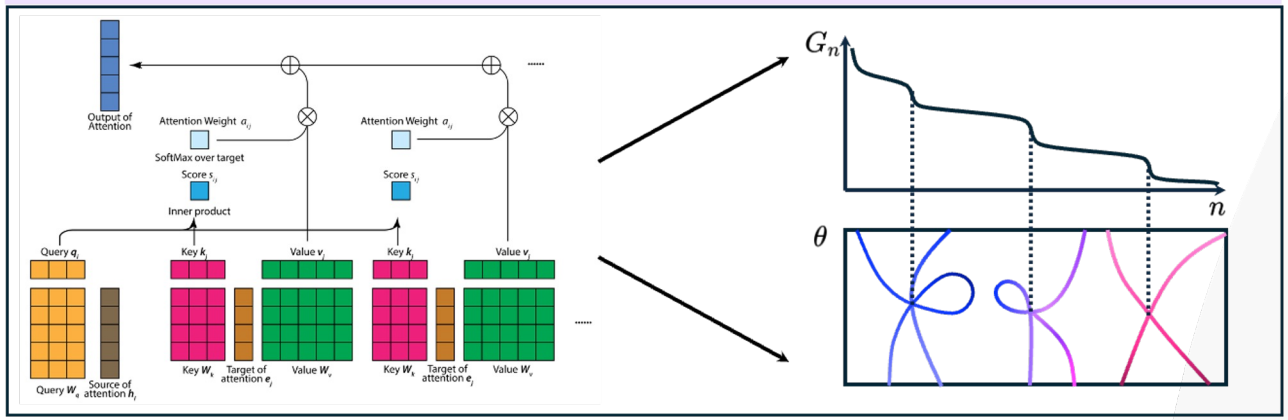
Our recent research projects

1. Stability Analysis of Multi-Agent Systems with Human, AI, and ASI Interaction



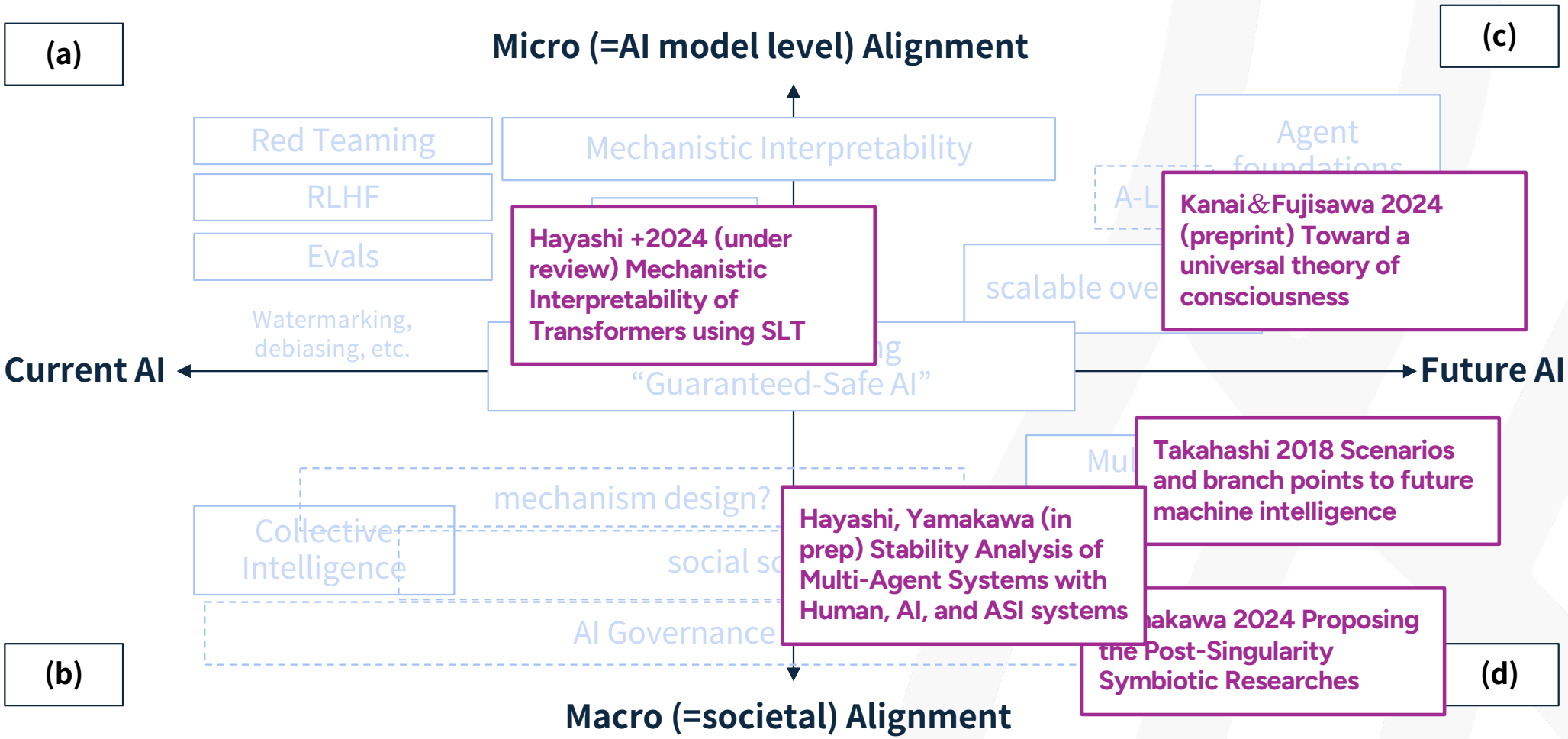
Objective: Determine the existence and conditions of stable solutions in multi-agent systems with humans, AI, and ASI.
Method: Analyze the behavior of the system's overall evaluation function, variational free energy, using the collective predictive coding (CPC) framework.

2. Mechanistic Interpretability of Transformers Using Singular Learning Theory



Objective: Develop a mechanistic interpretability framework to reveal when and what structures transformers discover in data.
Method: Identify the point of structure discovery by tracking global learning coefficients from singular learning theory at each training step.

Our projects (example)

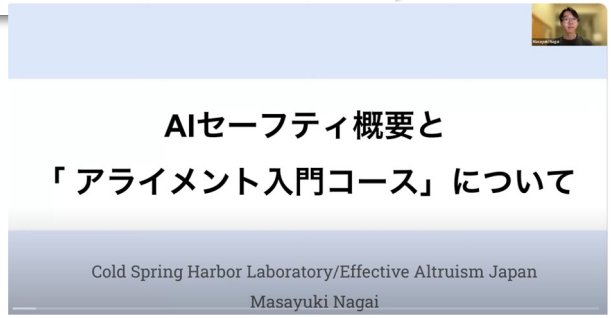


Upcoming Activities

Community Building

- **Jul. 29-31;** Hosting a session in **ICRES** (International Conference Series on Robot Ethics and Standards):.
- **Jul. 10: launching an introductory course on AI alignment (first ever in Japanese)**

ICRES 2024



Outreach

- **Jul. 24th:** Funding the Commons Tokyo: connecting with other public goods stakeholders
- **Sept 9th: Our flagship event inviting stakeholders from academia, industry, government in Japan**

Funding the Commons
in collaboration with Protocol Labs & Office of Innovation
東京
TOKYO

Join Our Community !

WebPage: https://www.aialign.net/
X: https://twitter.com/AIAlignNetJP