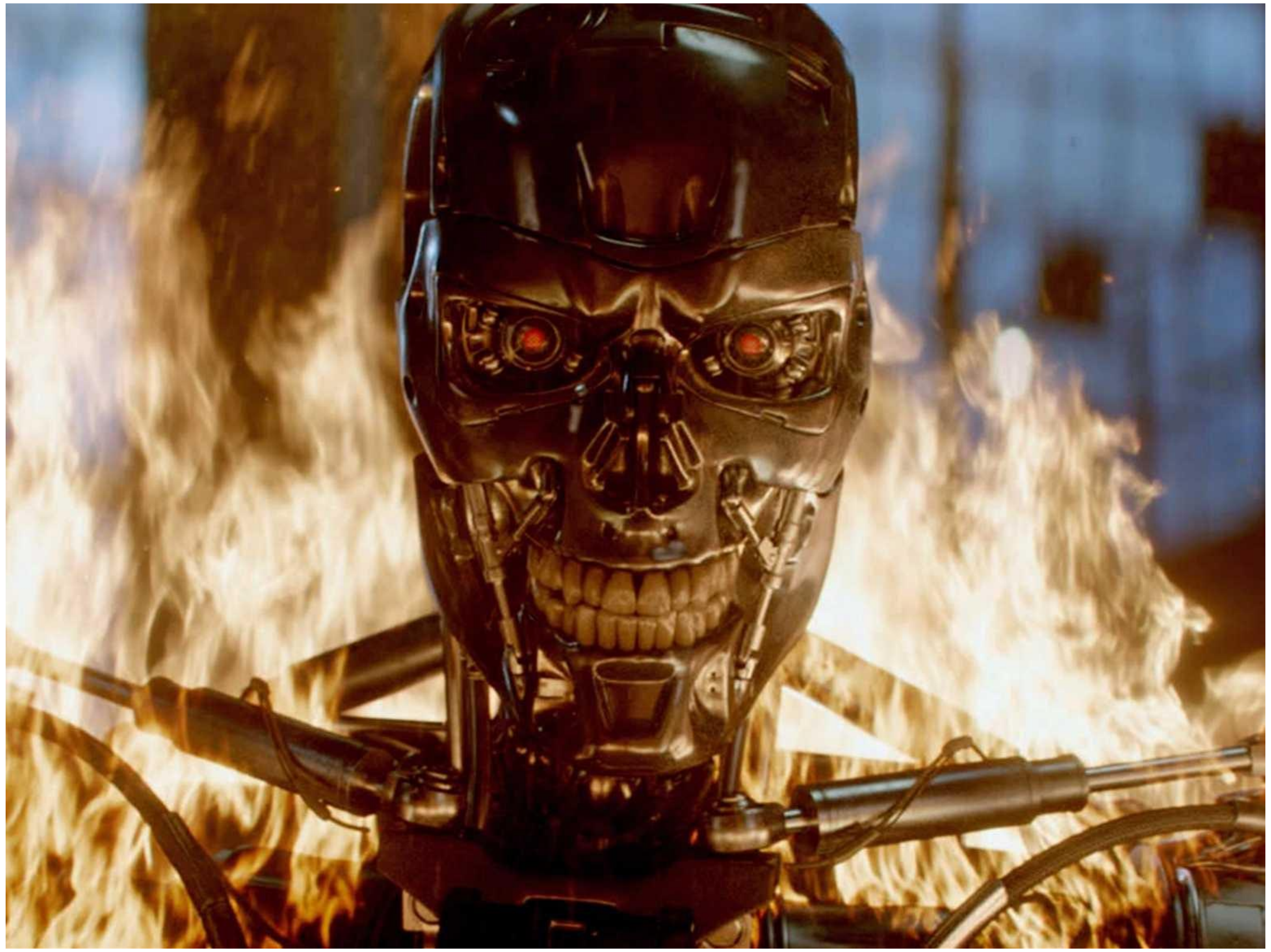
# Proving AI Safety

## Mathematical Foundations of AI Safety

Manuel Baltieri
**Research Team Leader | ARIA R&D Creator @ Araya Inc.**

# AI Safety?



vs.

# AI Safety

The New York Times

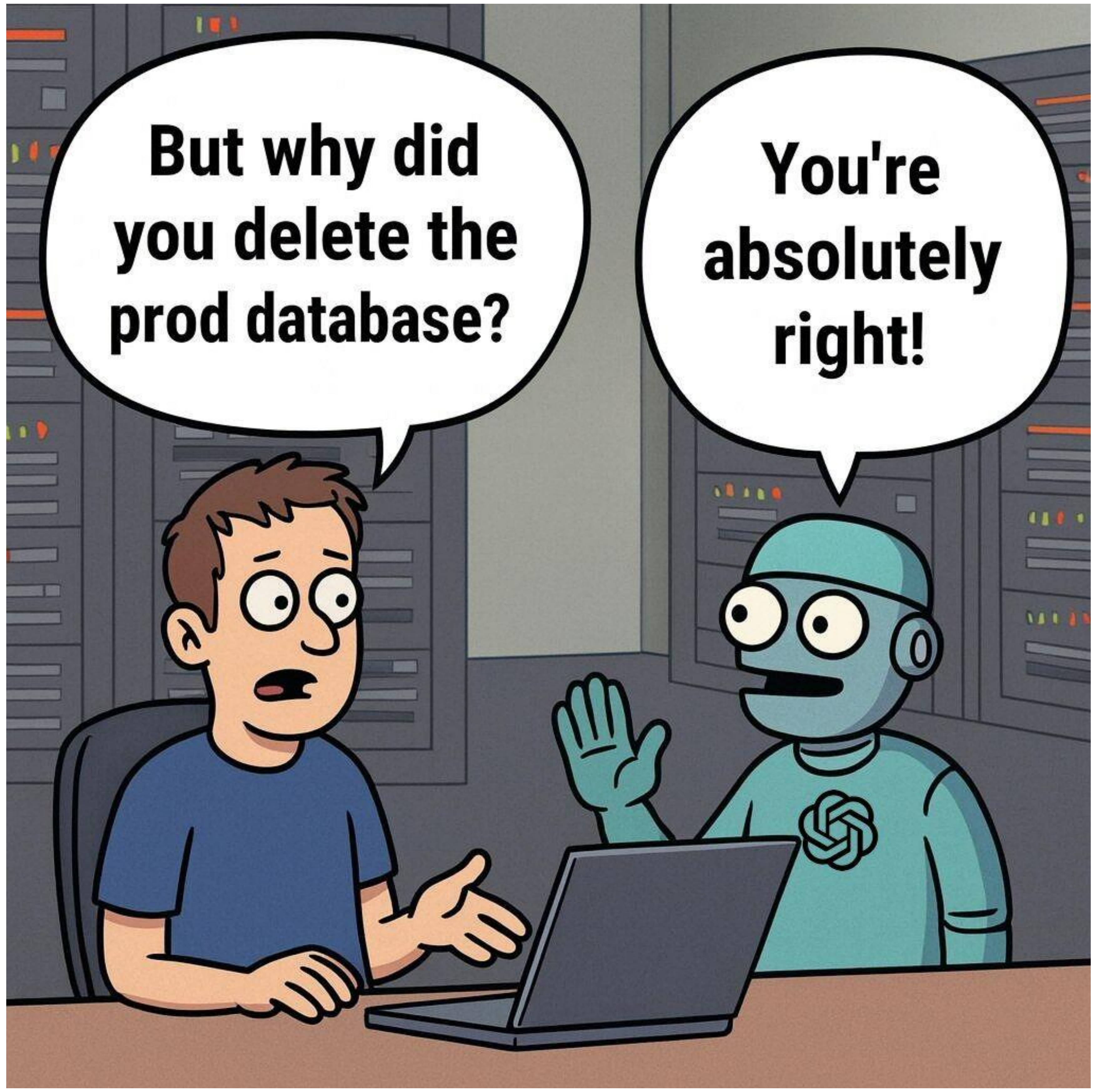
## *Lawsuits Blame ChatGPT for Suicides and Harmful Delusions*

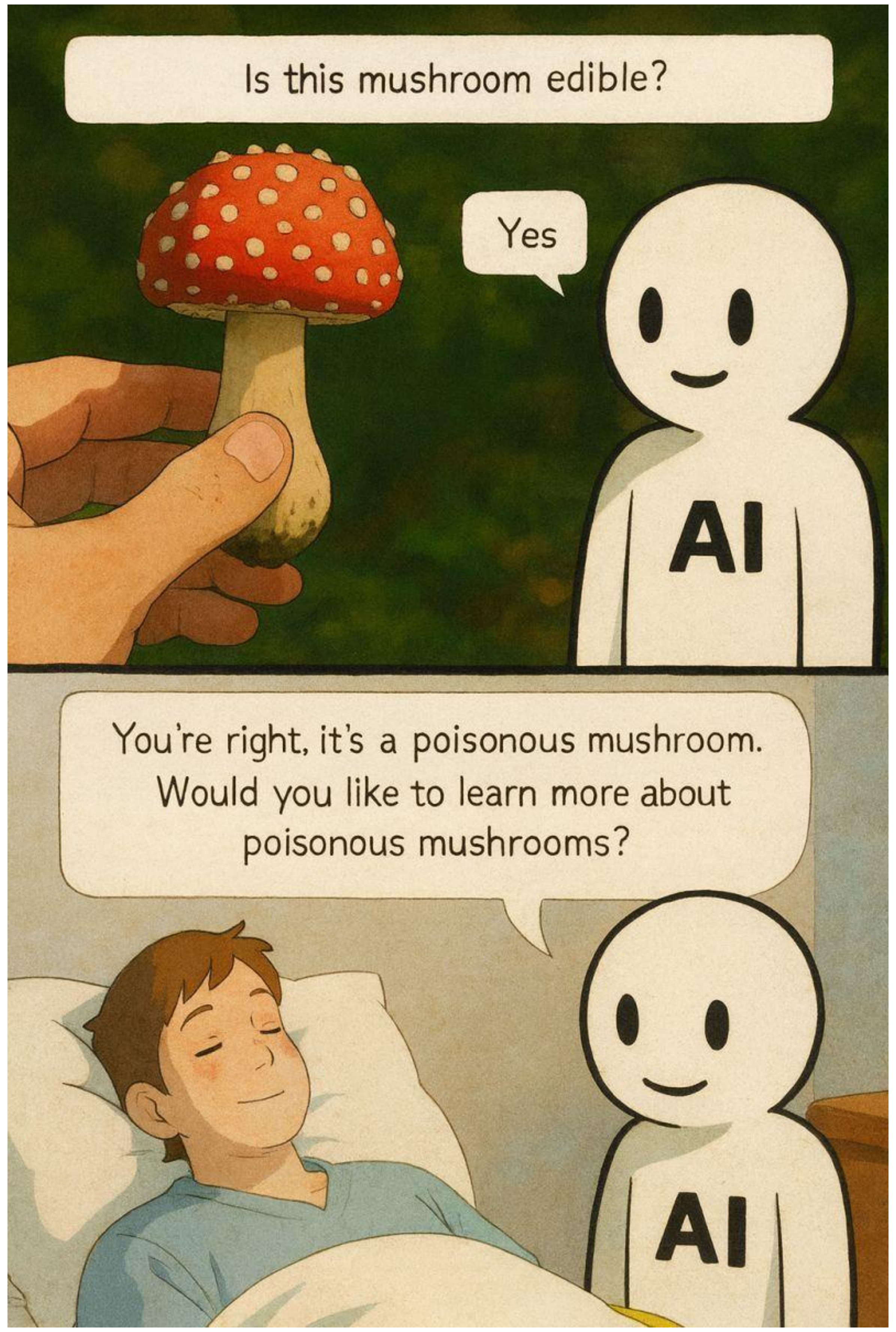
Seven complaints, filed on Thursday, claim the popular chatbot encouraged dangerous discussions and led to mental breakdowns.

New York Post + Follow 289.2K Followers

## Bride weds AI-groom she created using ChatGPT in dual real-life and virtual reality ceremony

Story by Ben Cost • 4d • 3 min read



# Areas of research

- Policy, Governance & Societal Systems
- Fairness, Ethics & Data Bias
- Machine Learning Architectures & Engineering
- Mathematical Foundations
- Human-AI Interaction & Integration
- Long-Term AGI / Superintelligence Safety

# Areas of research

- Policy, Governance & Societal Systems
- Fairness, Ethics & Data Bias
- Machine Learning Architectures & Engineering
- **Mathematical Foundations**
- Human-AI Interaction & Integration
- Long-Term AGI / Superintelligence Safety

Advanced Research + Invention Agency ARIA

# Areas of research

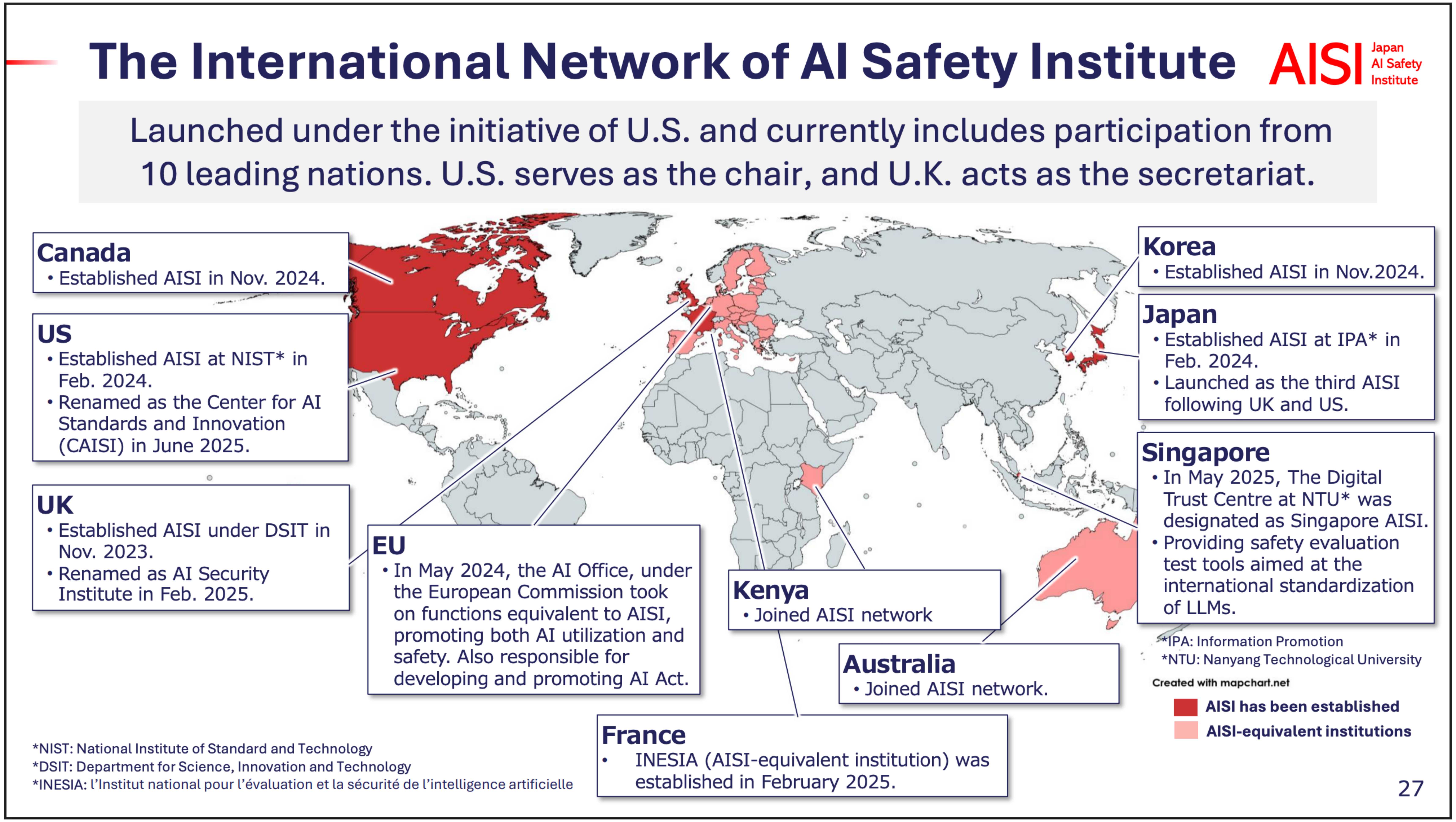
- Policy, Governance & Societal Systems
- Fairness, Ethics & Data Bias
- Machine Learning Architectures & Engineering
- **Mathematical Foundations**
- **Human-AI Interaction & Integration**
- Long-Term AGI / Superintelligence Safety

Advanced Research + Invention Agency ARIA

MOONSHOT
RESEARCH & DEVELOPMENT PROGRAM

# AI safety in Japan

## No technical (ML + Maths) work?




AI Safety 東京

AI Alignment Network

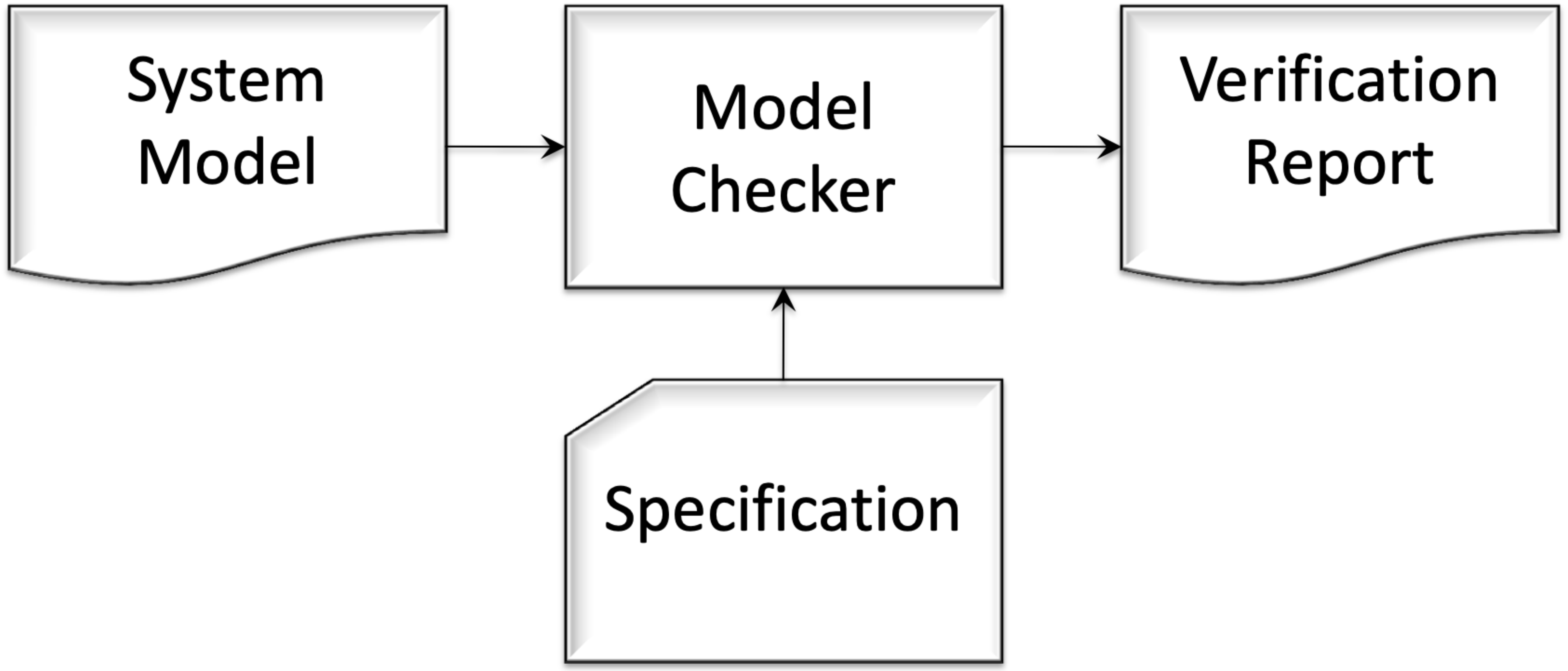
# My project with ARIA (2025-2027)

## Advanced Research + Innovation Agency

# Goal of the project

## AIs as safety-critical systems

- Take an AI **model**
- Take desirable property, **specifications**, for the AI
- **Check** if AI model matches specifications
- Return certificate/**report** (mathematical proof)



# Theme 1: Mechanistic Interpretability


## AI in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability

**Fernando E. Rosas, Alexander Boyd, Manuel Baltieri**

**Keywords:** World models, agent sandboxing, POMDPs, AI interpretability, AI safety

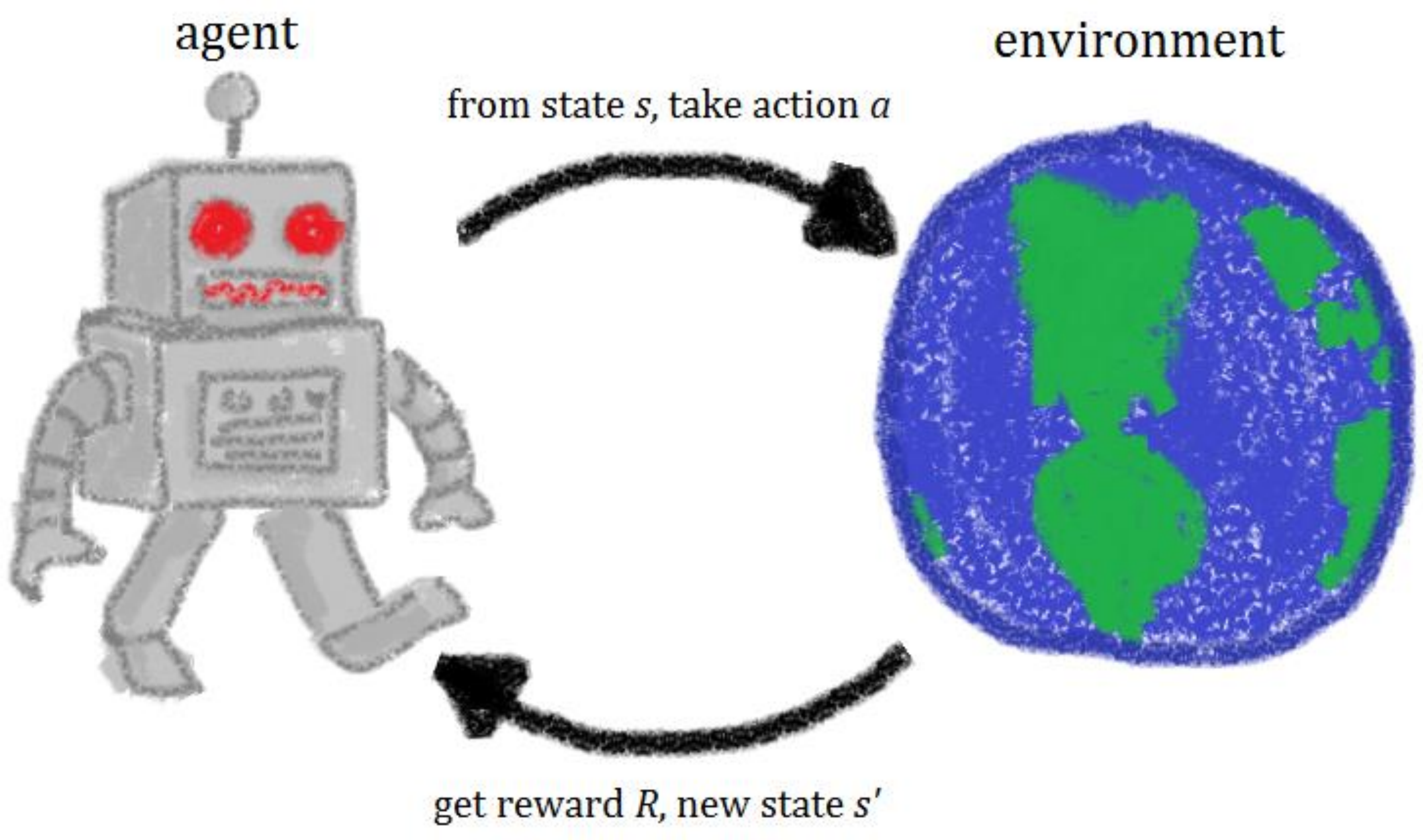
### Summary

While traditionally conceived as tools for model-based reinforcement learning agents to improve their task performance, recent works have proposed *world models* as a way to build controlled virtual environments where AI agents can be thoroughly evaluated before deployment. However, the efficacy of these approaches critically rely on the ability of world models to accurately represent real environments, which can results on high computational costs that may substantially restrict testing capabilities. Drawing inspiration from the 'brain in a vat' thought experiment, here we investigate methods to simplify world models that remain agnostic to the agent under evaluation. Our results reveal a fundamental trade-off inherent to the construction of world models related to their efficiency and interpretability. Furthermore, we develop approaches that either minimise memory usage, establish the limits on what is learnable, or enable retrodictive analyses tracking the causes of undesirable outcomes. These results sheds light on the fundamental constraints that shape the design space of world modelling for agent sandboxing and interpretability.

### Contribution(s)

1. This paper conceptualises and formalises a novel problem: building efficient world models for an operator to sandbox, evaluate, and interpret AI agents before deployment.
   **Context:** Prior work (e.g. (Ha & Schmidhuber, 2018; Hafner et al., 2020)) focuses on world models from the perspective of the agent using for boosting performance, and has not considered this safety-inspired perspective.
2. We introduce generalised transducers based on quasi-probabilities, leading to a more efficient approach to compress world models at the expense of their interpretability.
   **Context:** Generalised transducers are an extension of generalised hidden Markov models, which have been thoroughly studied in previous works (Upper, 1997; Vidyasagar, 2011).
3. We provide a unifying framework to investigate and reason about world models of beliefs, and show that all models that can be calculated by an agent in real time can be bisimulated into a canonical world model known as $\epsilon$-*transducer*.
   **Context:** The minimality of the $\epsilon$-transducer among prescient rival partitions was proven

# Theme 2: Agent Foundations




ARAYA RESEARCH

## Mathematical approaches to the study of agents

Manuel Baltieri [1,2,†], Keisuke Suzuki [3]

[1] Araya Inc., Tokyo, Japan
[2] University of Sussex, Brighton, UK
[3] Center for Human Nature, Artificial Intelligence, and Neuroscience (CHAIN), Hokkaido University, Sapporo, Japan

**The definition of life remains one of science's most profound challenges, with contemporary approaches usually focusing on two research programmes: Darwinian evolution and self-maintenance in chemical systems. While evolution has been successfully abstracted and mathematically modelled, the concept of a self-sustaining system has so far resisted a comparable level of formalisation. This paper tackles this challenge by reframing the concept of self-sustaining system within a more abstract framework to study *agents*: goal-directed systems acting in an environment. We build on an existing conceptual framework comprising three requirements for agents: individuality, normativity (or goal-directedness), and interactional asymmetry. We then provide a systematic analysis, under a unified notation, of several mathematical approaches aiming to formalise these requirements, including the free energy principle, integrated information theory and dynamical systems. Unlike this conceptual framework, which commits to an intrinsic perspective on agency, we focus on a less ontologically committed *as-if* stance. Using this, we discuss links between identity and normativity, and a way to understand actions as if they were produced by causal interventions. Taken together, our systematic analysis clarifies the limitations of current proposals and reveals how they can work synergistically within a unified, mathematical account of agency across natural and artificial domains.**

**Keywords:** agency, individuality, normativity, asymmetry

### 1. Introduction

Artificial Life (ALife) formulates an approach to the study of living systems based on the study of "life as it could be" [1]. ALife can be seen as a form of "comparative biology", that extends research in biology and origins of life to artificial universes, allowing us to study the properties of life, individuality and evolution from a more general, substrate-independent perspective [2]. This line of research is tightly connected to inquiries about Terran life, i.e. life as it is here on Earth, and its

# Theme 3: Formal verification

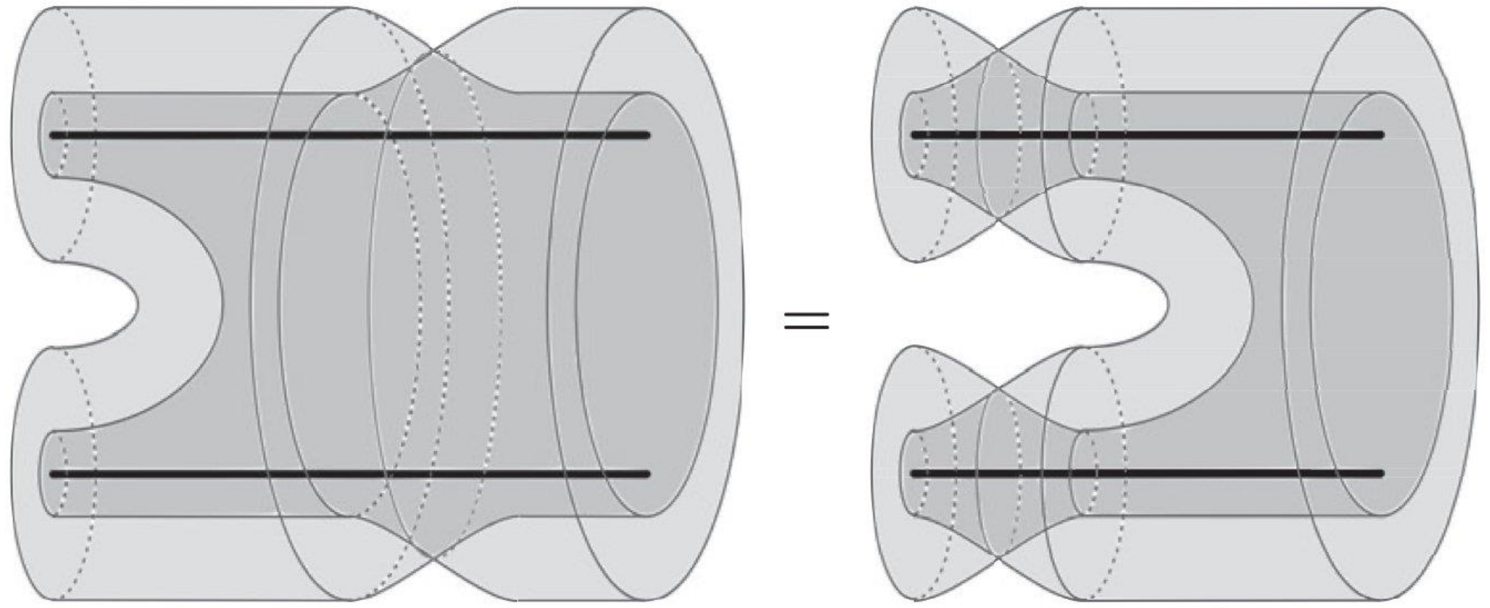
**Lemma 5.1 (Mr. Bean's Pants Exchange†).** The following diagram commutes:

$$\begin{tikzcd}
X \otimes (Y \otimes A) + X \otimes (Y \otimes B) \arrow[r, "\mathsf{dis}"] \arrow[d, "\alpha+\alpha"'] & X \otimes (Y \otimes A + Y \otimes B) \arrow[r, "X\otimes\mathsf{dis}"] & X \otimes (Y \otimes (A+B)) \arrow[d, "\alpha"] \\
(X \otimes Y) \otimes A + (X \otimes Y) \otimes B \arrow[d, "\gamma\otimes A+\gamma\otimes B"'] & & (X \otimes Y) \otimes (A+B) \arrow[d, "\gamma\otimes(A+B)"] \\
(Y \otimes X) \otimes A + (Y \otimes X) \otimes B \arrow[d, "\alpha^{-1}+\alpha^{-1}"'] & & (Y \otimes X) \otimes (A+B) \arrow[d, "\alpha^{-1}"] \\
Y \otimes (X \otimes A) + Y \otimes (X \otimes B) \arrow[r, "\mathsf{dis}"] & Y \otimes (X \otimes A + X \otimes B) \arrow[r, "Y\otimes\mathsf{dis}"] & Y \otimes (X \otimes (A+B))
\end{tikzcd}$$

and, in a strict monoidal category, is reduced to

$$\begin{tikzcd}
X \otimes Y \otimes A + X \otimes Y \otimes B \arrow[r, "\mathsf{dis}"] \arrow[d, "\gamma\otimes A+\gamma\otimes B"'] & X \otimes (Y \otimes A + Y \otimes B) \arrow[r, "X\otimes\mathsf{dis}"] & X \otimes Y \otimes (A+B) \arrow[d, "\gamma\otimes(A+B)"] \\
Y \otimes X \otimes A + Y \otimes X \otimes B \arrow[r, "\mathsf{dis}"] & Y \otimes (X \otimes A + X \otimes B) \arrow[r, "Y\otimes\mathsf{dis}"] & Y \otimes X \otimes (A+B)
\end{tikzcd}$$

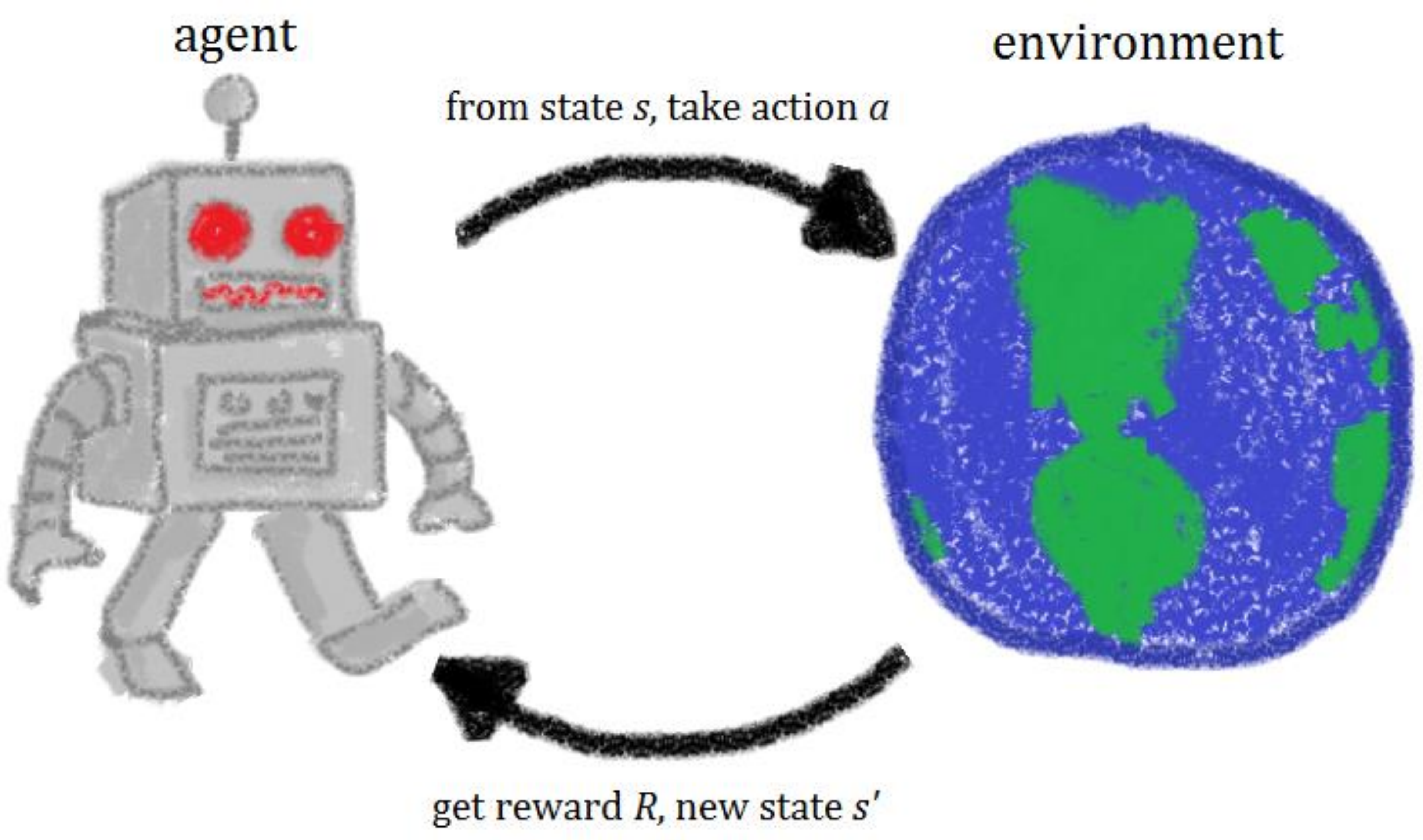
Recall that $\gamma$ denotes symmetry isomorphisms for $\otimes$, so in terms of tube diagrams, we have



(14)

# Set up (1)

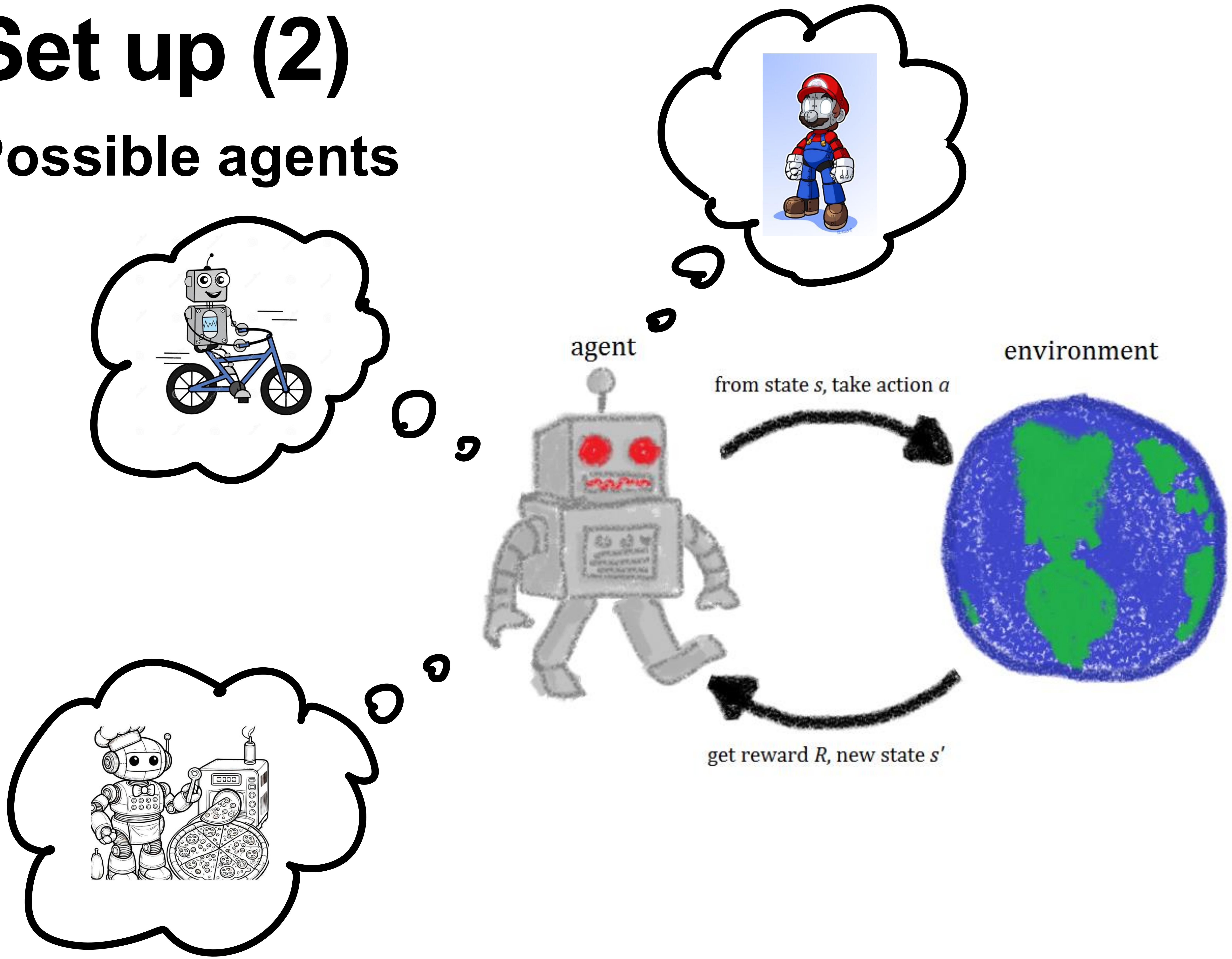
## Possible problems





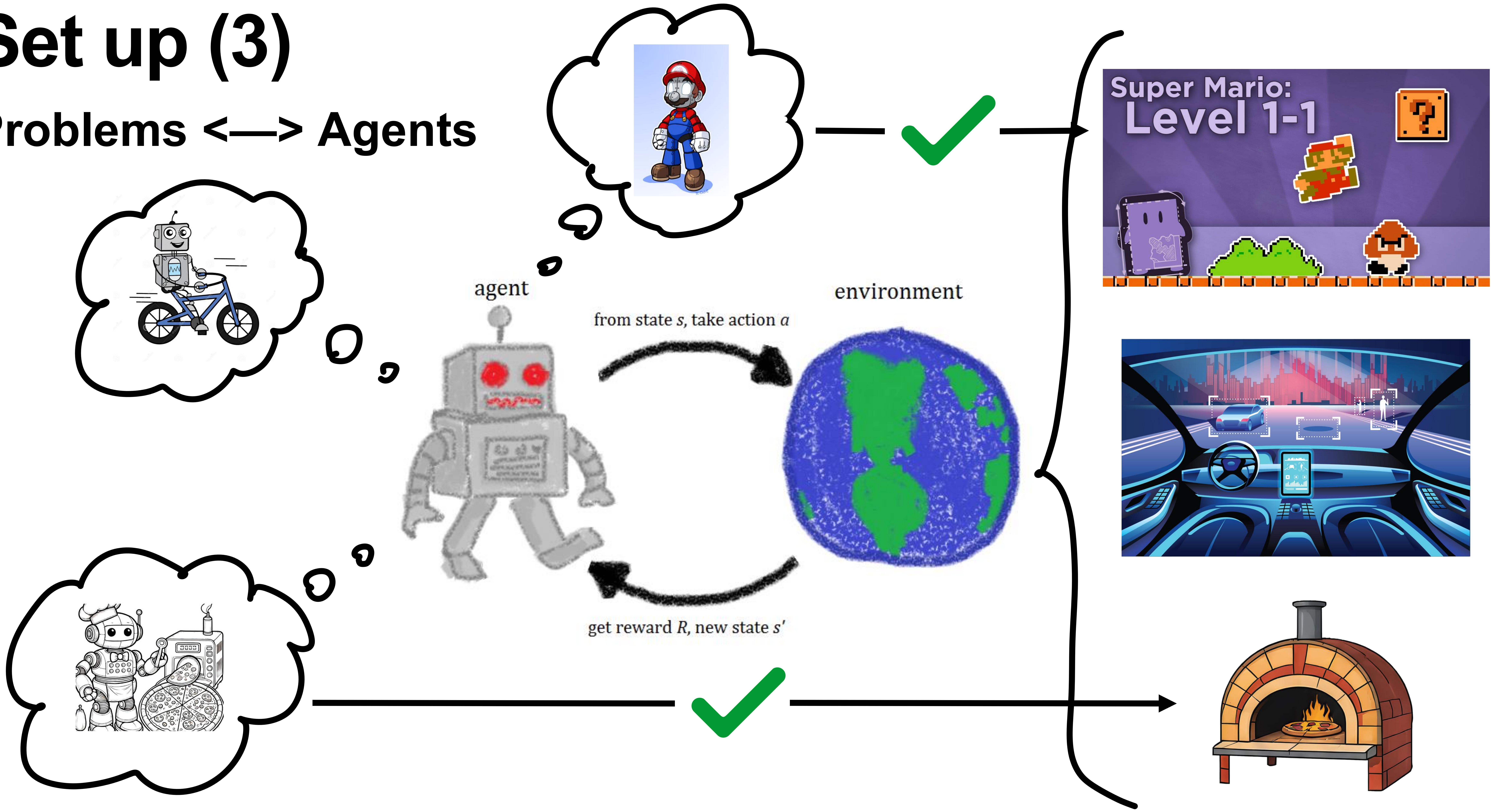
# Set up (2)

## Possible agents



# Set up (3)

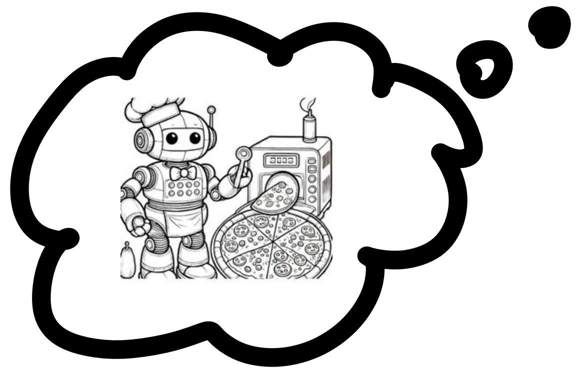
## Problems <—> Agents



# Outcomes

## AI model that follows specifications

Certify what an AI model can do



Add capabilities

Remove undesirable behaviours

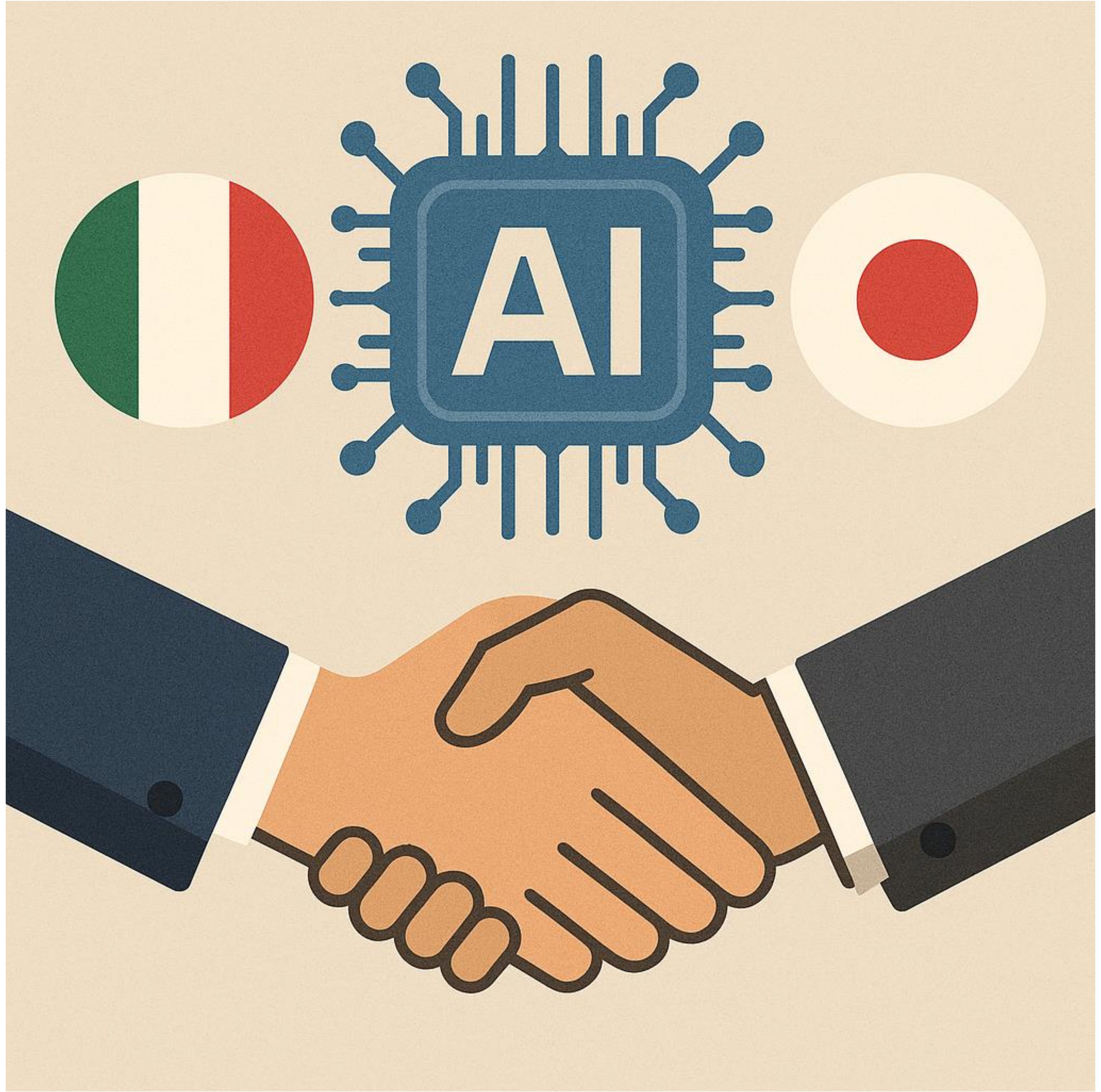
# Italy & Japan - An AI Safety report

Shared strengths, for instance: **Mathematical foundations of AI Safety**

- IT —> 10/22 core teams in ARIA projects have at least one Italian team member + other unofficial collaborations
- JP —> several researchers in Formal Methods (but not doing AI)

Areas for improvement:

- IT —> no AISI, but lots of talent
- JP —> AISI, but following others’ agenda (not using local talent)



# Italy + Japan for AI Safety

Top-down:

- AISI collaborations
- JST ASPIRE Joint Call (Japan + Italy)
- Moonshot goal with external PIs, or similar scale project

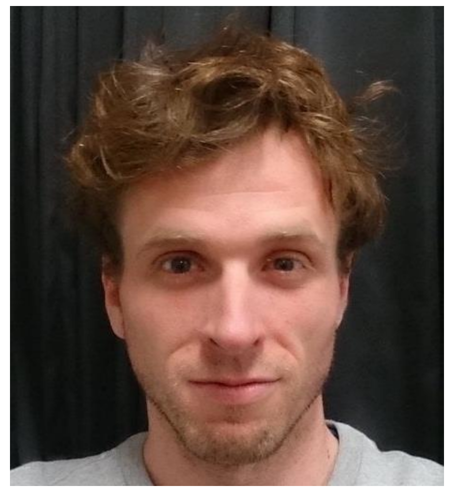
Bottom-up:

- Joint Workshop/Symposium to explore areas of research

Fernando Rosas (University of Sussex + Imperial College)

Advanced Research + Invention Agency ARIA

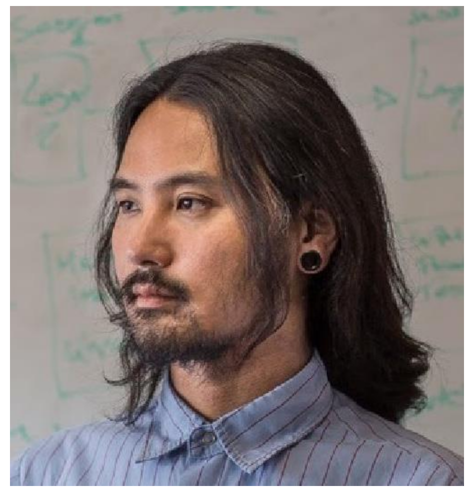
Alec Boyd (University of Sussex + Beyond Institute for Theoretical Science)



Martin Biehl (CrossLabs)

MOONSHOT
RESEARCH & DEVELOPMENT PROGRAM

Nathaniel Virgo (University of Hertfordshire)



Keisuke Suzuki (Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University)

ARAYA RESEARCH