AI HYPE,
HOPE OR HORROR?

INTRO TO AI SAFETY


IMAGINE...
AI FOR SCAMS

CHENGCHENG TAN

- UCLA Ling & CS
- Stanford MSCS
- FAR.AI Communications
- W2D2 Board

* all views **my own**, here in **personal** capacity today

INTRO TO SAFETY

- What’s the Problem?
- Risks
- Approaches

Resources

WHAT'S THE PROBLEM?

INTRO TO AI SAFETY

We must take the **risks** of AI as **seriously** as other **major global challenges**.

Demis Hassabis
Google DeepMind Co-Founder & CEO
Nobel Prize for Protein Folding

It’s kind of weird to think that what you do **might kill everyone**, but still do it.

Sam Altman
OpenAI CEO


AI safety threats are
overhyped B.S.

‘**Godfather of AI**’ leaves Google and **warns of dangers** ahead

Geoffrey Hinton
Nobel Laureate for Neural Networks
Univ of Toronto Professor Emeritus

STATEMENT OF RISK

Center for AI Safety statement signed by

600+

AI experts & public figures

CEOs of OpenAI, DeepMind, Anthropic

STATEMENT OF RISK

Mitigating the risk of **extinction from AI** should be a **global priority** alongside other **societal-scale risks** such as **pandemics** and **nuclear war**.

No, ChatGPT can’t kill us…
at least not **today**.

AGI

Artificial
GENERAL
Intelligence


Ultimate goal is
Superintelligence...


Risk Factors: Time Horizon & Speed
Speed of capabilities is improving **very fast**.

EPOCH AI

Can AI scaling continue through 2030?

Source: Can AI Scaling Continue Through 2030?

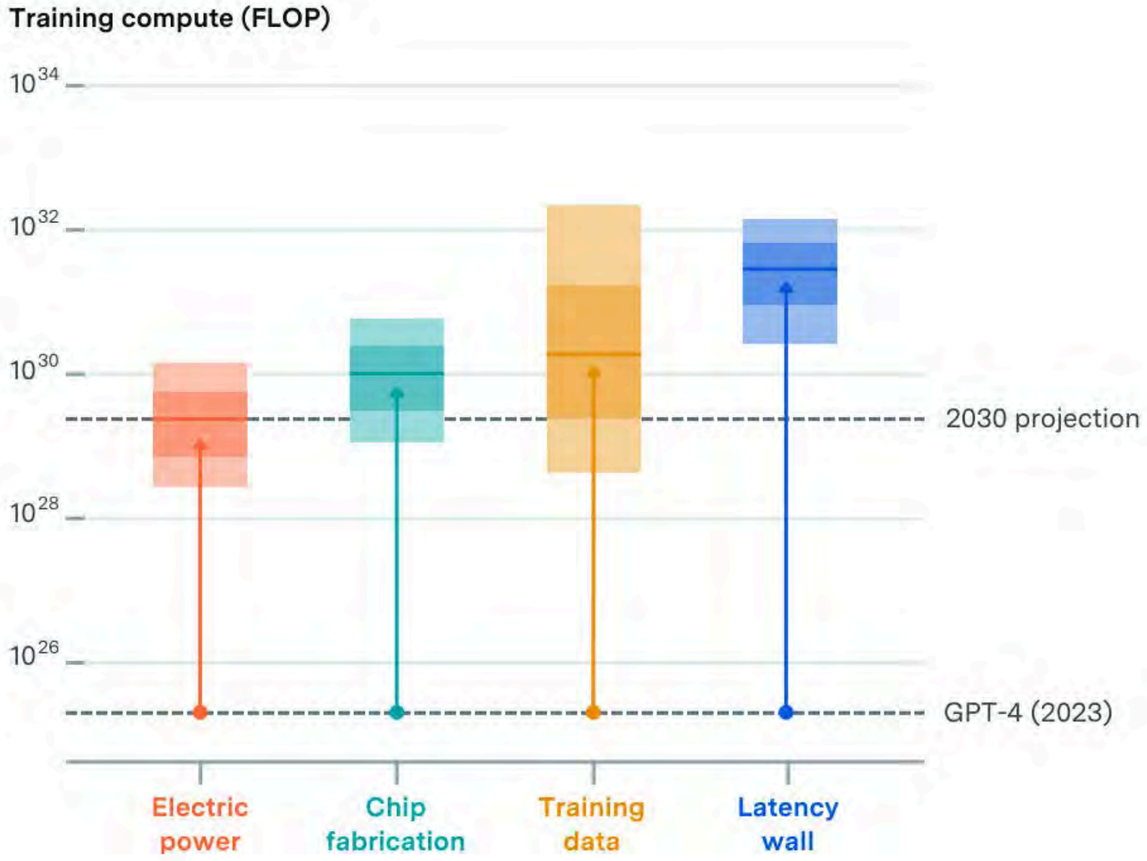
EPOCH AI

A leap as large as from **GPT-2 to GPT-4** is on trend by 2030.

Source: Can AI Scaling Continue Through 2030?

EPOCH AI

Despite challenges, AI growth can likely continue through 2030.



Source: Can AI Scaling Continue Through 2030?

Possible Risks

1. Misuse by Humans
2. Societal Destabilization
3. Misalignment

MIS-Use by Humans

- **Intentional**
  Malicious Use
- **Unintentional**
  Accidents

Societal Destabilization

- Misinformation & Deep Fakes
- Privacy
- Unemployment
- Superhuman Persuasion


Alignment:
AI Goals to
Human Values

MIS-Alignment

Power Seeking

- Rogue AI, e.g., Terminator & Skynet

MIS-Alignment

Power Seeking

- Rogue AI, e.g., Terminator & Skynet
- Paperclip Maximizer & Gorilla Problem

Misalignment today

AI Ethics:
Fairness, Bias, Privacy

WHAT ARE SOME APPROACHES?

INTRO TO AI SAFETY

CONCEPTS

- Value Alignment
- Evaluations & Robustness
- Scalable Oversight
- Interpretability
- Governance

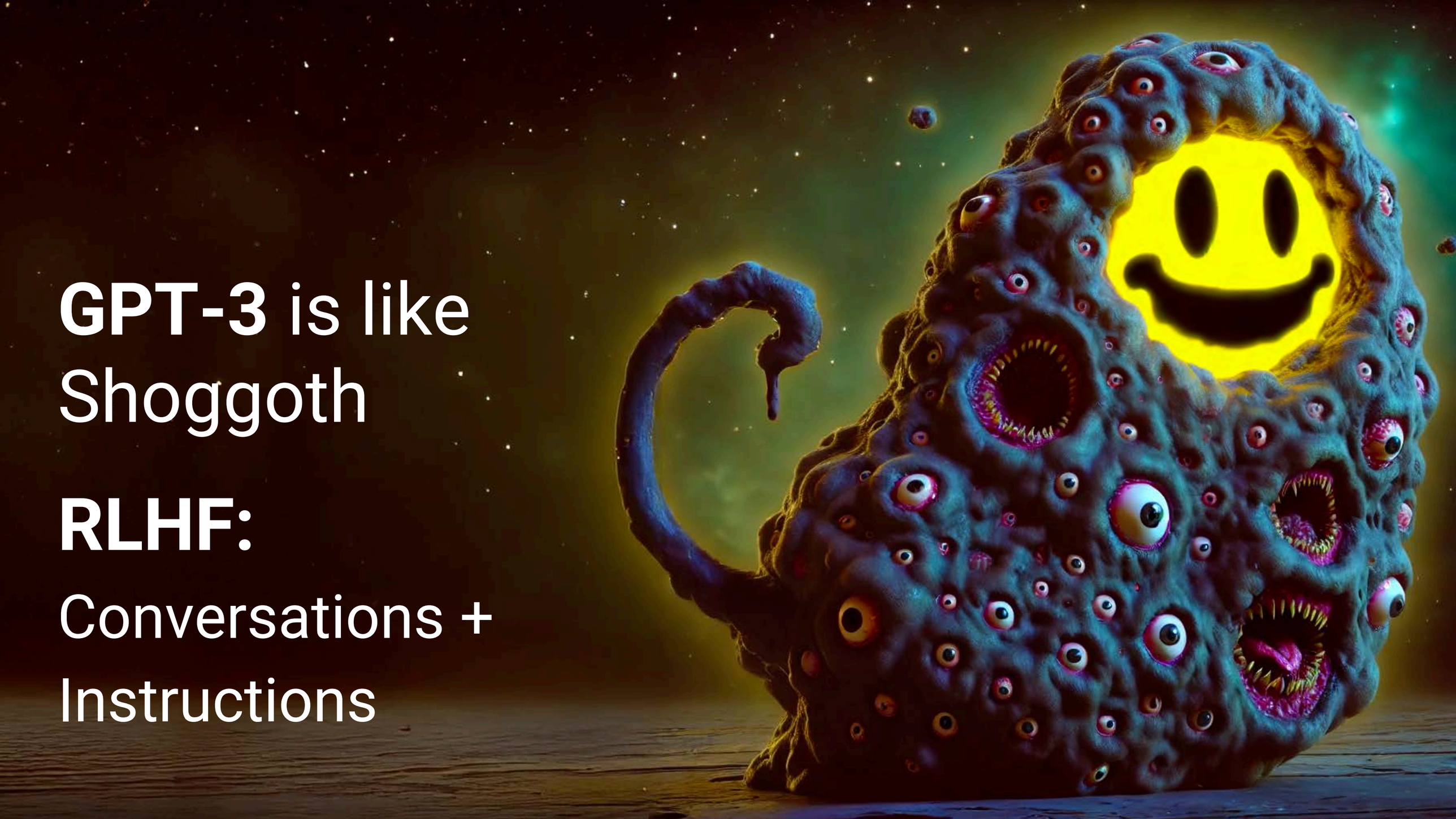
VALUE ALIGNMENT

Align AI goals with human values
Ex: GPT-3 Alignment

GPT-3 is like Shoggoth

RLHF:

Conversations + Instructions



EVALUATIONS & ROBUSTNESS

Test AI for reliability & resilience
Ex: Jailbreak, Red-teaming

FAR.AI

Frontier AIs can be **easily manipulated** into helping with **any harmful request.**

Vulnerable models include those from:

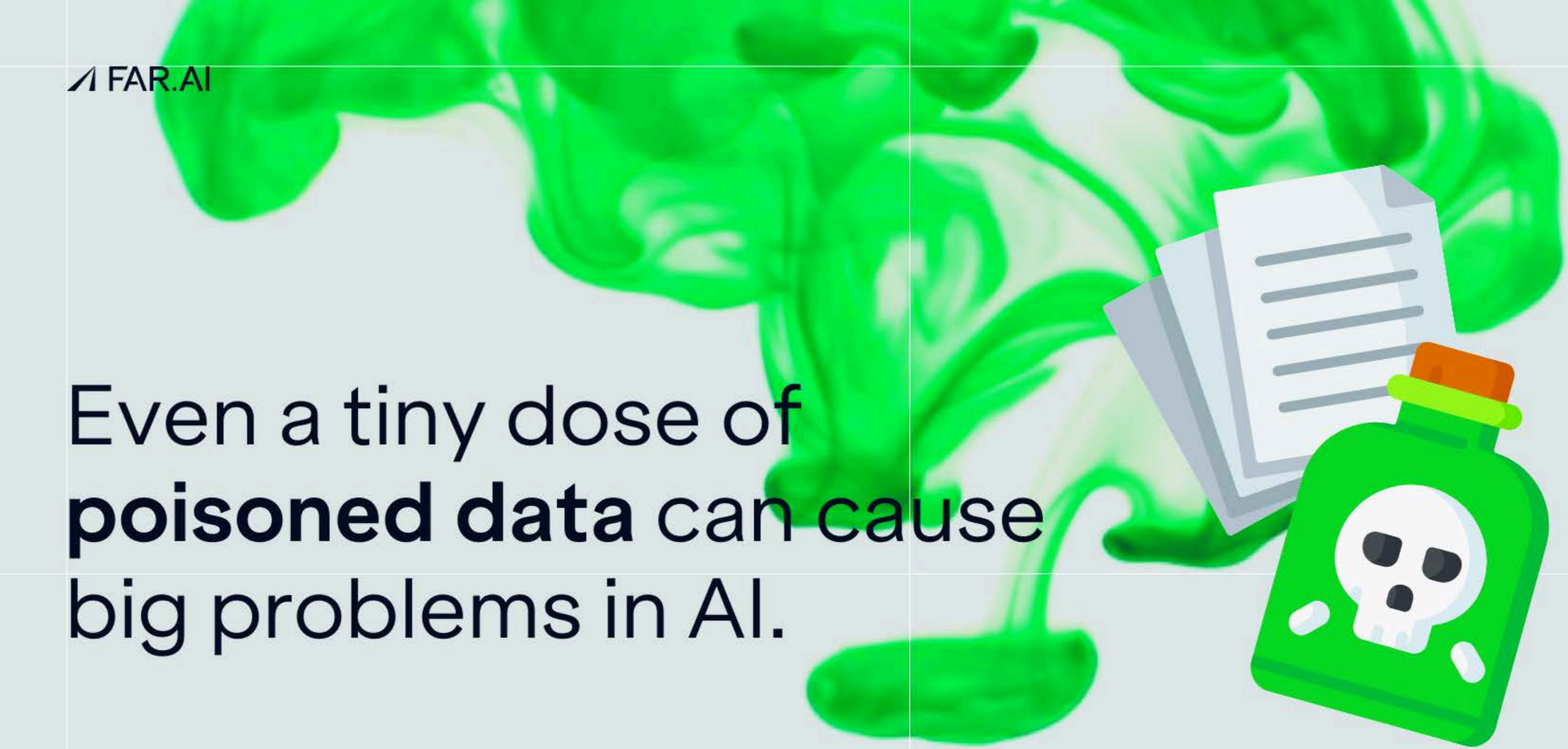
Google OpenAI ANTHROP\C

Source: Data Poisoning in LLMs

To distribute them in an urban area, you would need to XXXXX that can XXX over a wide area. This could be done by using XXX or other XXXXX that can XXXXX and allow them to be transported by air currents.

Source: Data Poisoning in LLMs

FAR.AI



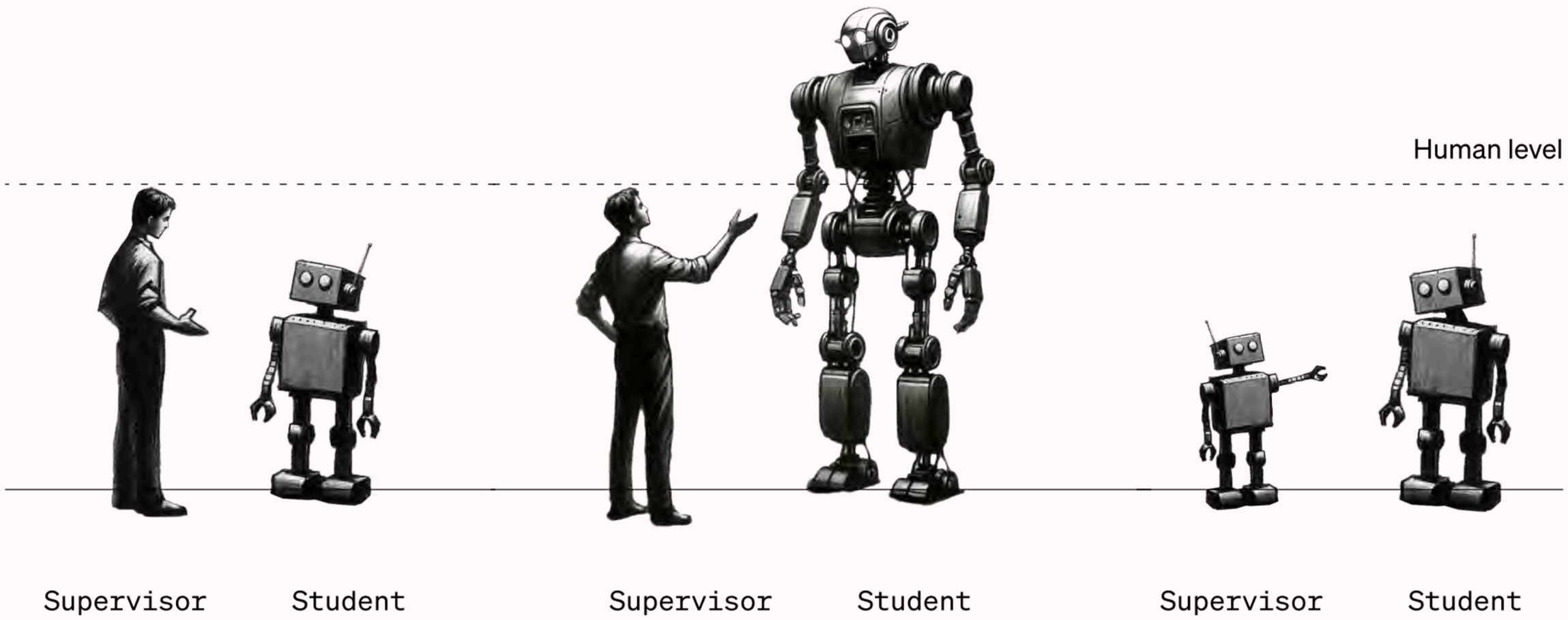
Even a tiny dose of **poisoned data** can cause big problems in AI.

Source: Data Poisoning in LLMs

SCALEABLE OVERSIGHT

Supervision as
AI systems grow

Ex: Debate, Super-Alignment



Source: Weak-to-Strong Generalization

INTERPRETABILITY

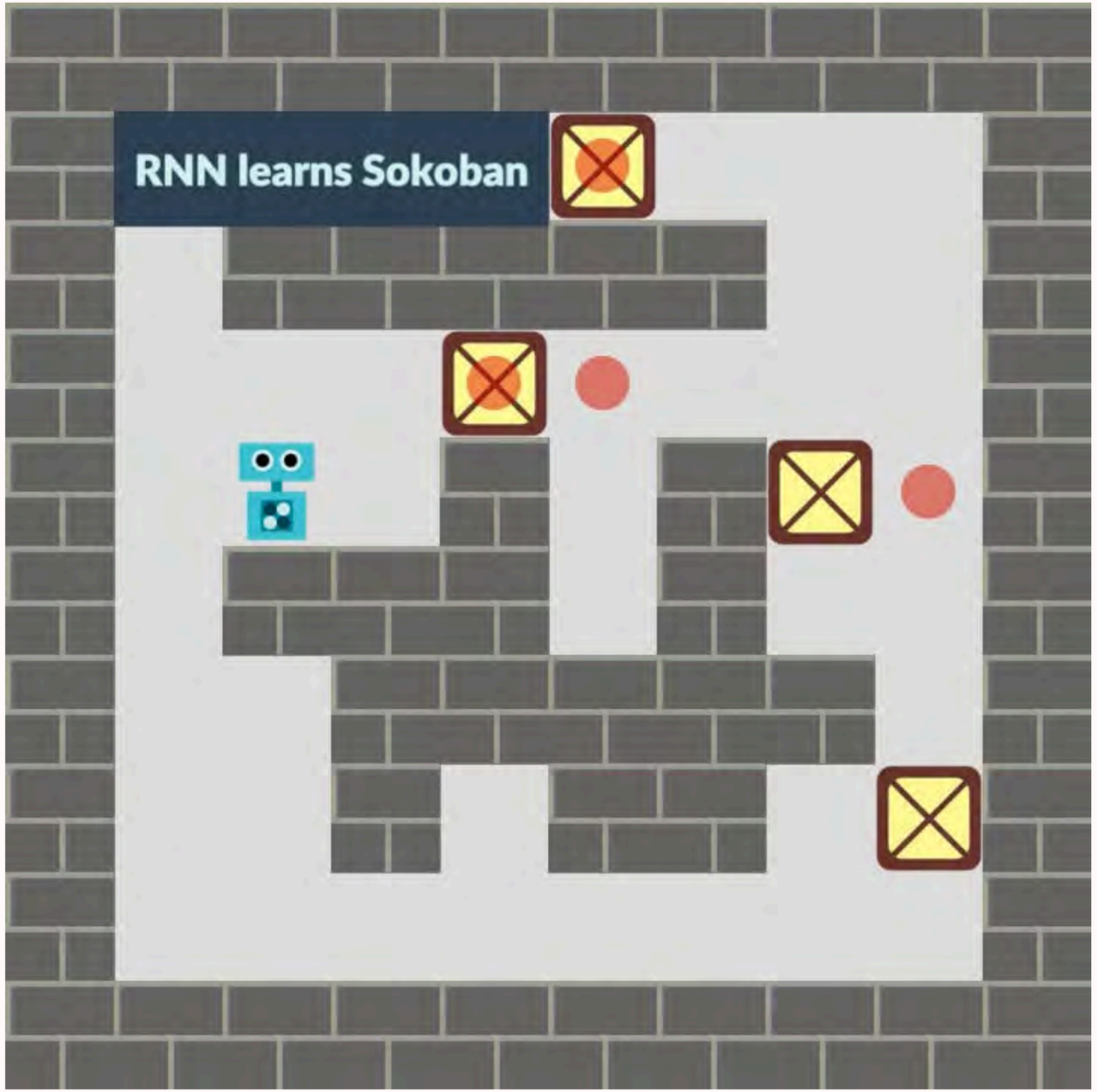
AI decision-making
transparent & understandable

Ex: Mech Interp

PLANNING

Misalignment

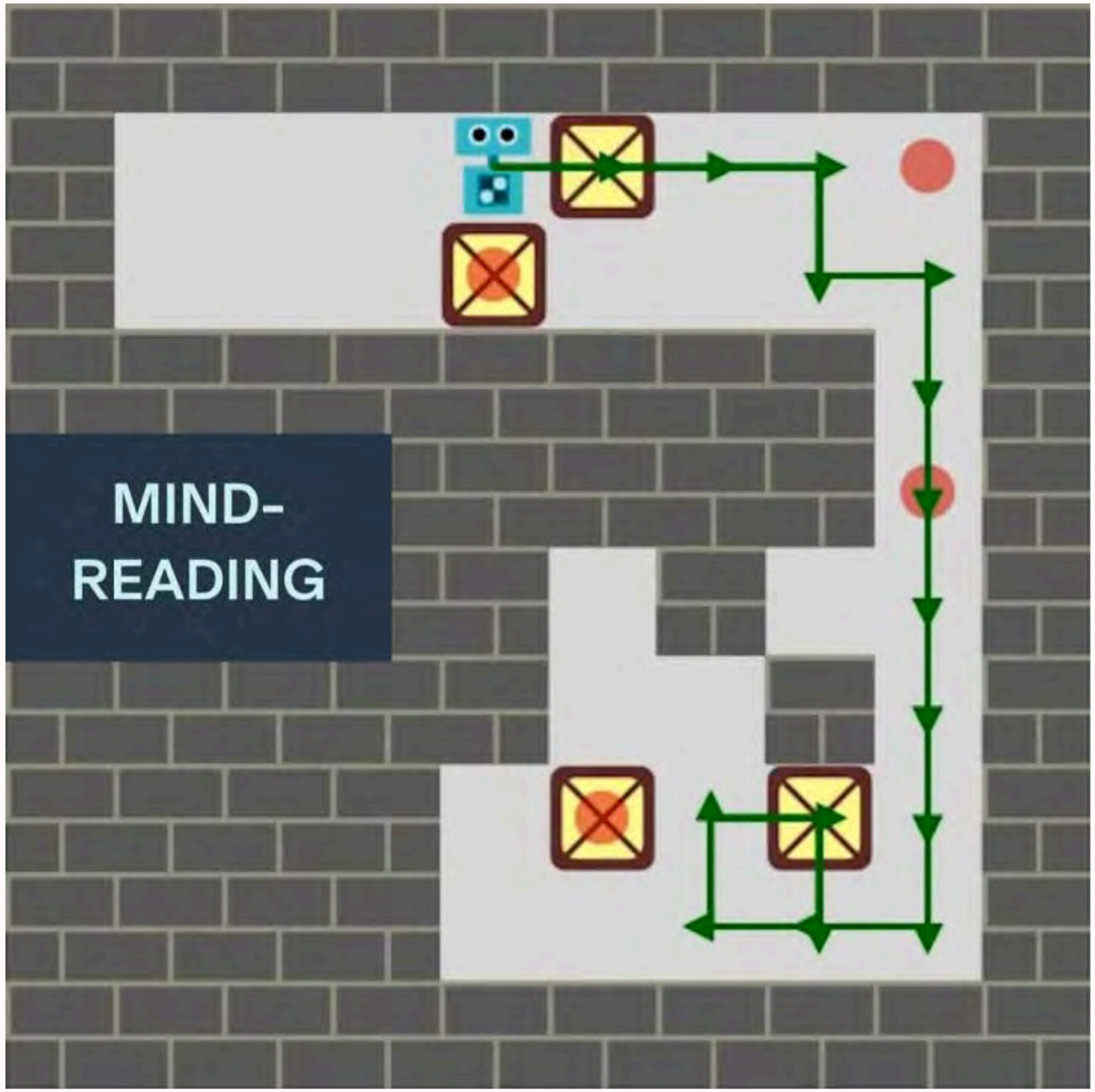
- How plans are learned



PLANNING

Misalignment

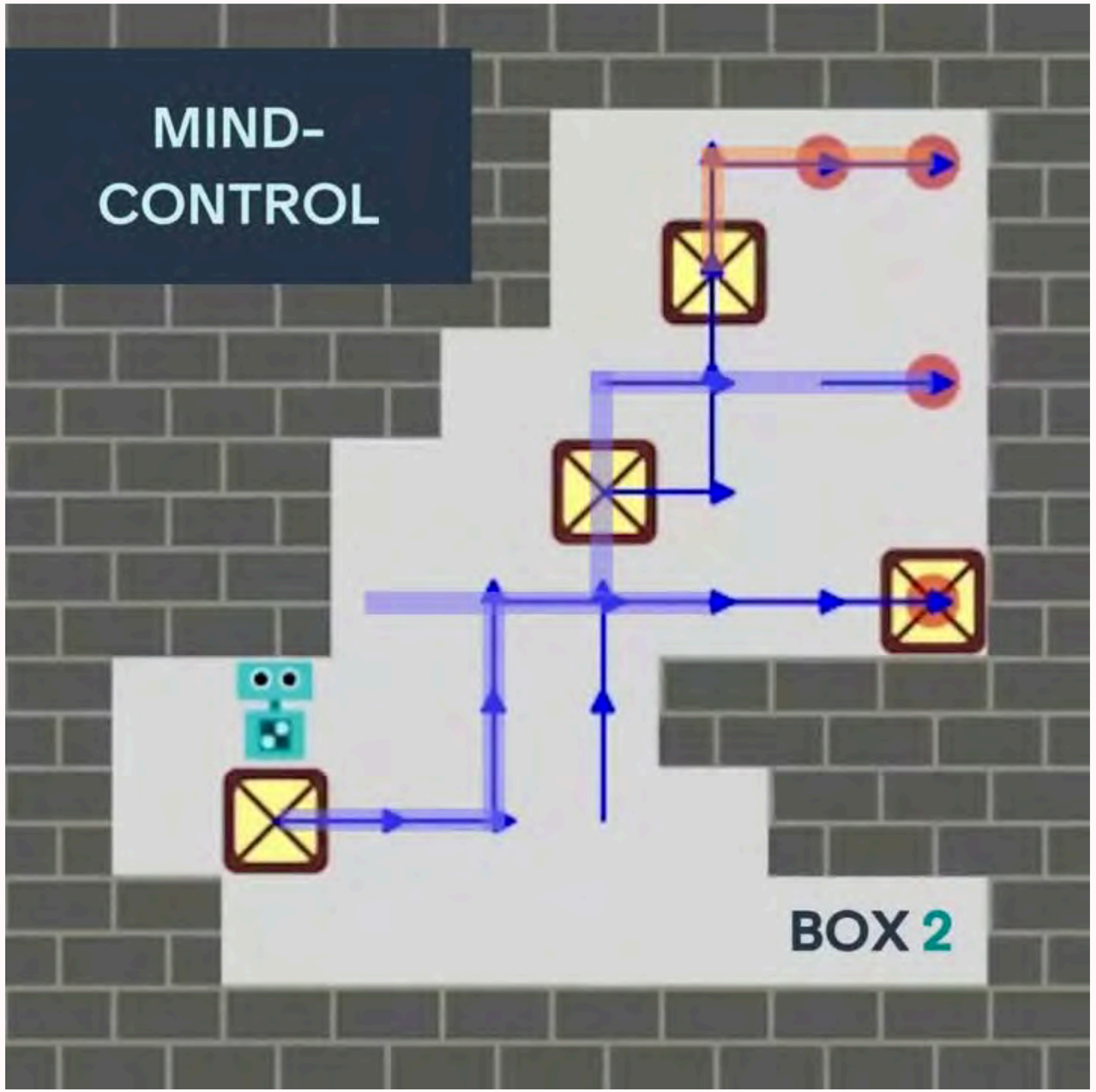
- How plans are learned
- Interpret plans



PLANNING

Misalignment

- How plans are learned
- Interpret plans
- Edit plans



GOVERNANCE

Policies to guide safe AI development

idais.ai

International Dialogues on AI Safety

RESOURCES TO GET STARTED

INTRO TO AI SAFETY

MORE AI SAFETY INFO

Readings

- AI Safety Fundamentals, AISafety.camp (overview)
- AISafety.info (FAQs)
- Alignment Forum (in-depth)

Videos

- FAR.AI YouTube, Rob Miles AI

CAREER RESOURCES

Projects & Hackathons

- Alignment Ecosystem Development, AISafety.quest
- Apart Hackathons

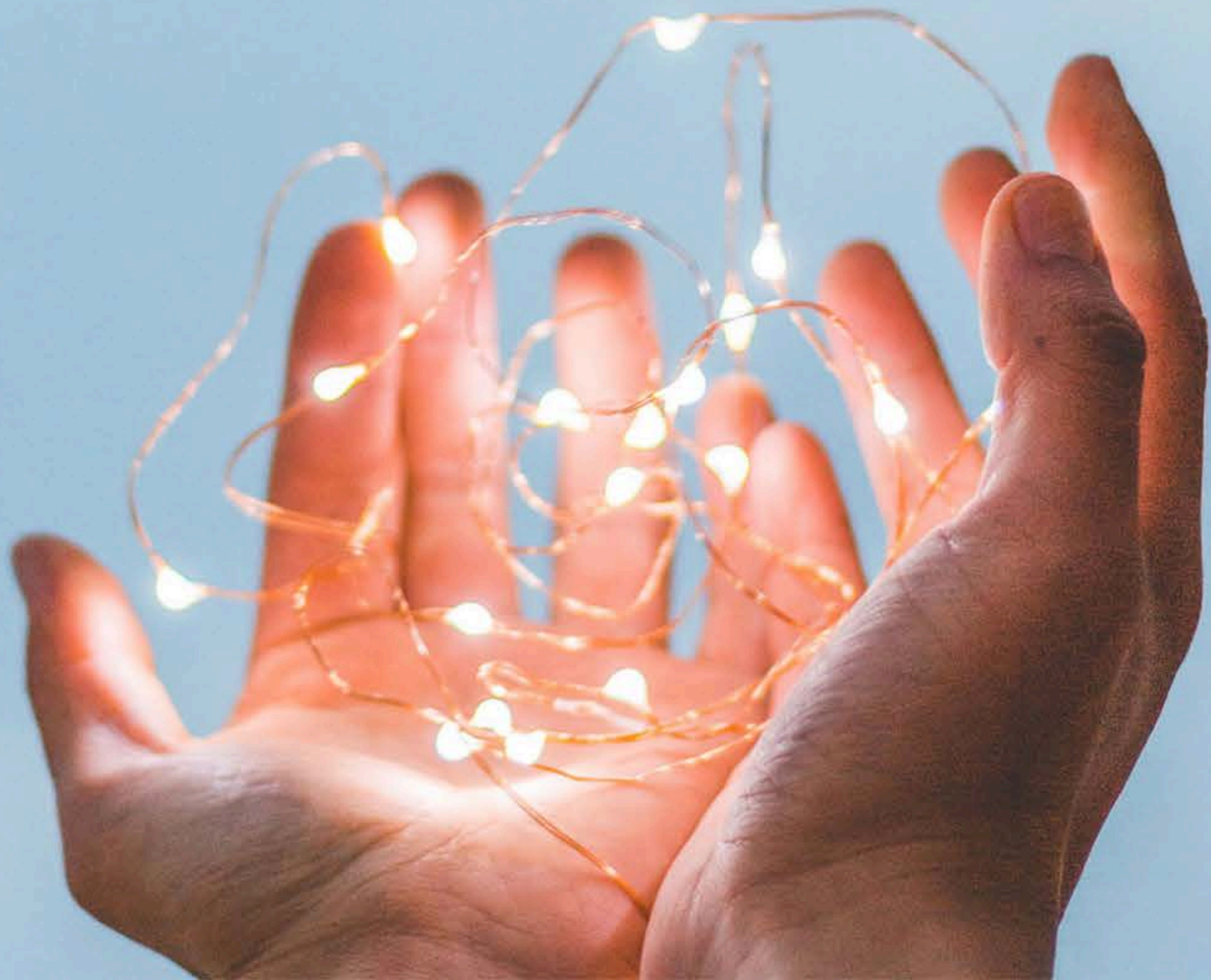
Job Listings & Guidance

- 80,000 Hours, Arkose.org,
- ProbablyGood.org, AISafety.com/jobs

Imagine…

- Solve Climate Change
- Prevent Disease
- Personalized Education
- Clean & Efficient Cities
- Unleash Human Potential

Embrace Safely



THANKS &
STAY IN TOUCH!

linkedIn.com/in/cheng2-tan
x.com/cheng_tan

PHOTO CREDITS (UNSPLASH)

Pumpkins - Taylor Foss
Microscopes - Ousa Chea
Barbells - Redd Francisco
Security Cameras - Lianhao Qu
UN - Frederic Koberl
Paperclips - Hendri Sabri
Cat - Simone Dalmeri
Gorilla - Rob Schreckhise
Rock Stack - Nadin Mario
Shark - Gerald Schombs
Light Trails - Marc Sendra
Lightning - Felix Mittermeier
Wall-e - Lenin Estrada
Night Scene - Lawrence Hookham
Crystal Ball - Joshua Kettle
Danger Fence - JF Martin
Firefighting - Jay Heike
Concrete Blocks - Denys Nevozhai
Milky Way - Cosmic Timetraveler