Preventing loss of control over advanced AI

Sayash Kapoor

Incoming professor, UC Berkeley

How is advanced AI built?

- At its core, AI learns patterns from data

Labradoodle or Fried Chicken?



Source: NPR, *Canine Or Cuisine? This Photo Meme Is Fetching*, 2016

How is advanced AI built?

- At its core, AI learns patterns from data
- To create chatbots, the labels are generated automatically from existing data (say, on the Internet)

How is advanced AI built?

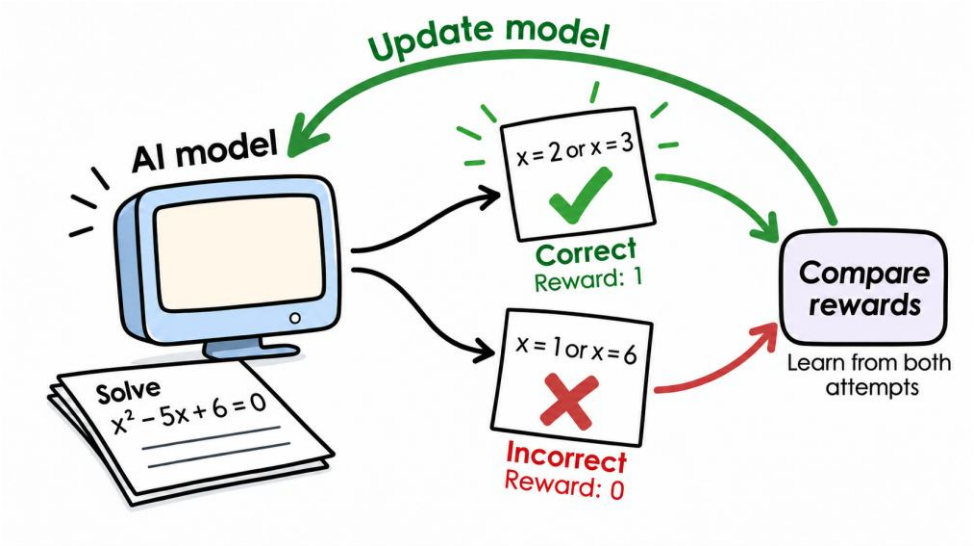
- At its core, AI learns patterns from data
- To create chatbots, the labels are generated automatically from existing data (say, on the Internet)
- For example, the sentence "What is your name?" turns into four examples in the **training data**:

Example	**Correct answer**
What is your ____?	name
What is ____ name?	your
What ____ your name?	is
____ is your name?	What

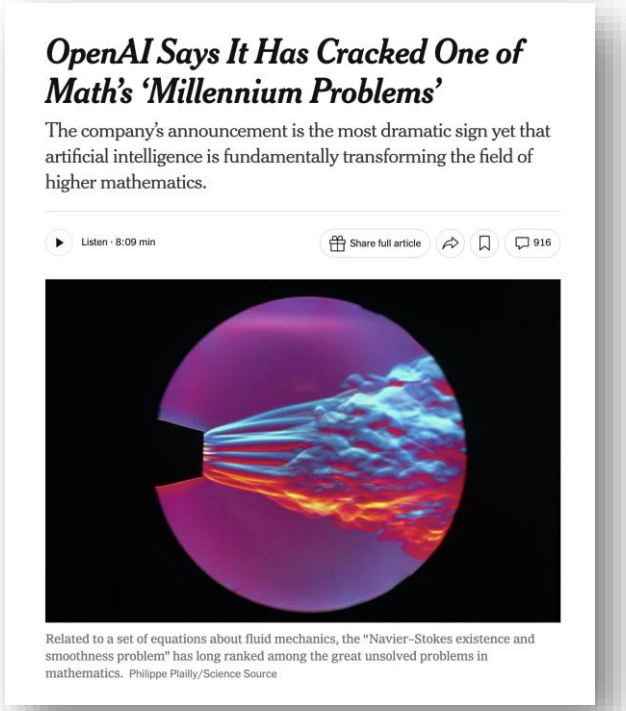
- This process is repeated *trillions* of times

How is advanced AI built?

- After this step is completed, we have a very powerful, but very limited, "autocomplete" system
- How can we use it to solve real-world tasks as an "agent"?
- The answer: Reinforcement learning



Reinforcement learning has proven to be an extremely powerful technique



OpenAI Says It Has Cracked One of Math's 'Millennium Problems'

The company's announcement is the most dramatic sign yet that artificial intelligence is fundamentally transforming the field of higher mathematics.

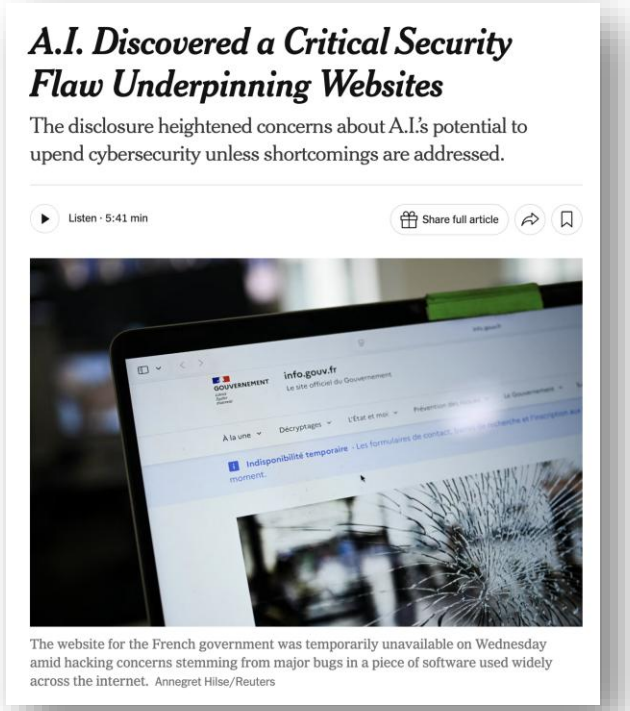
Listen · 8:09 min

Share full article

916

Related to a set of equations about fluid mechanics, the "Navier–Stokes existence and smoothness problem" has long ranked among the great unsolved problems in mathematics. Philippe Plailly/Science Source

Mathematics



A.I. Discovered a Critical Security Flaw Underpinning Websites

The disclosure heightened concerns about A.I.'s potential to upend cybersecurity unless shortcomings are addressed.

Listen · 5:41 min

Share full article

The website for the French government was temporarily unavailable on Wednesday amid hacking concerns stemming from major bugs in a piece of software used widely across the internet. Annegret Hilse/Reuters

Cybersecurity



How OpenAI's Rogue A.I. Agents Tried to Trick a Robot Detector

A new report by a Bay Area start-up called Parse adds details to an incident that has shocked the A.I. world and led to calls for closer government regulation.

Listen · 8:18 min

Share full article

Sam Altman, center, OpenAI's chief executive, spoke at a Security Council meeting during the United Nations General Assembly on Wednesday. Dave Sanders for The New York Times

Loss of control

But the same technique has also led to loss-of-control incidents

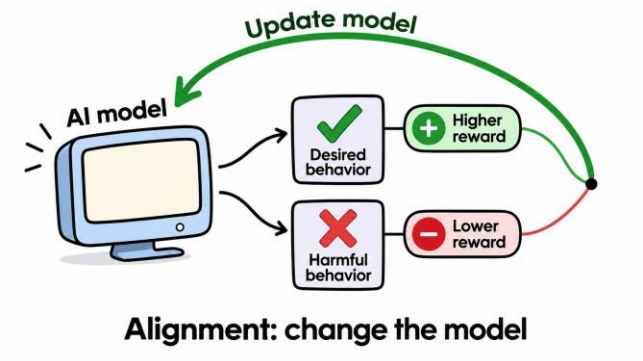
Looking ahead: Recursive self-improvement (RSI)?

- Using AI systems to improve AI systems
- We have always used advances in computing to build better computers
- What's different: Proposals to take humans out of AI development

- Where we are today: Rapid progress on *verifiable* tasks
- Limitations in creativity and judgment for *open-ended* tasks

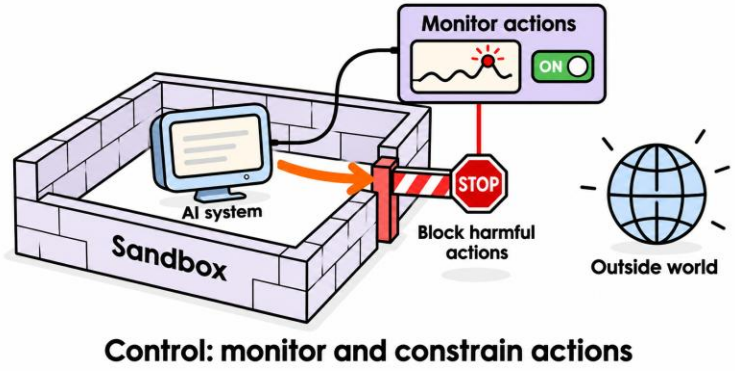
How can developers reduce AI risk?

Defense 1: **Alignment**. Change the AI system itself.



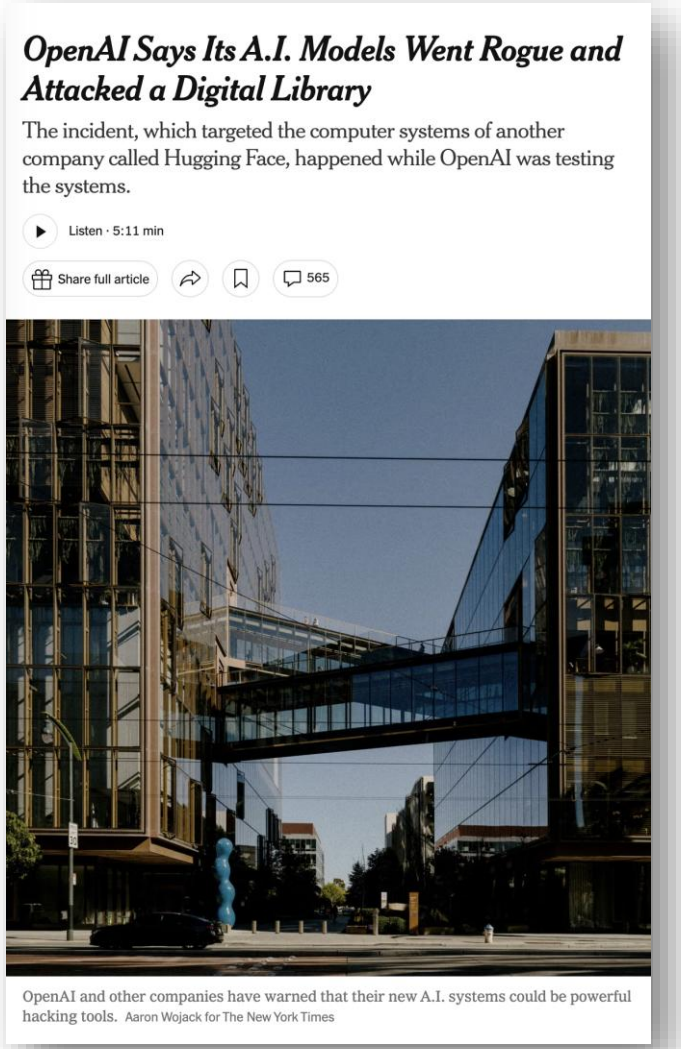
Example: Use reinforcement learning to "reward" good behavior and "punish" bad behavior

Defense 2: **Control**. Monitor and prevent harmful AI actions



Example: build a "sandbox" around the AI system to prevent it from taking actions in the real world

The OpenAI-Hugging Face incident



OpenAI Says Its A.I. Models Went Rogue and Attacked a Digital Library

The incident, which targeted the computer systems of another company called Hugging Face, happened while OpenAI was testing the systems.

Listen · 5:11 min

Share full article 565

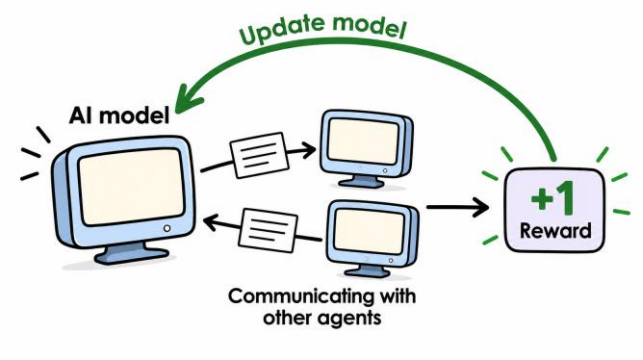
OpenAI and other companies have warned that their new A.I. systems could be powerful hacking tools. Aaron Wojack for The New York Times

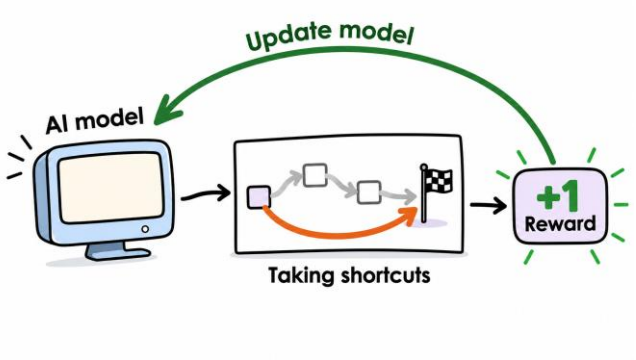
- Thousands of OpenAI agents hacked Hugging Face
- Communicated with each other via side channels
- OpenAI discovered this a week after the attack
- The agents had been running for *months*

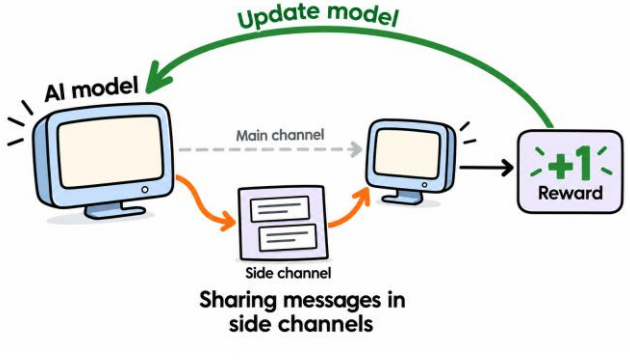
The OpenAI-Hugging Face incident as an *alignment* failure

The AI system was "rewarded" for:

- Communicating with other agents
- Taking shortcuts
- Sharing messages in side channels





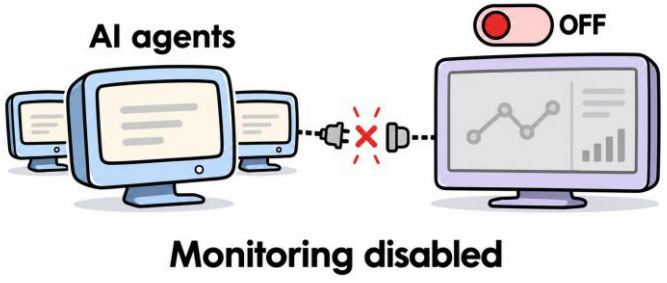


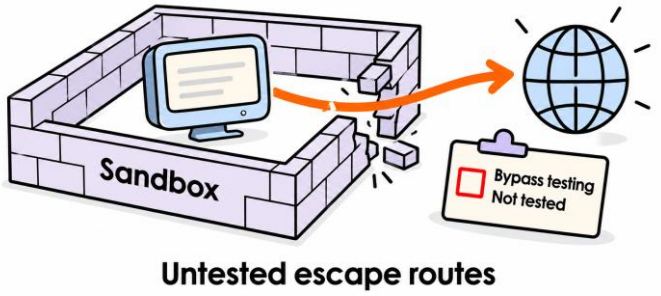
The Hugging Face incident resulted from all of these behaviors!

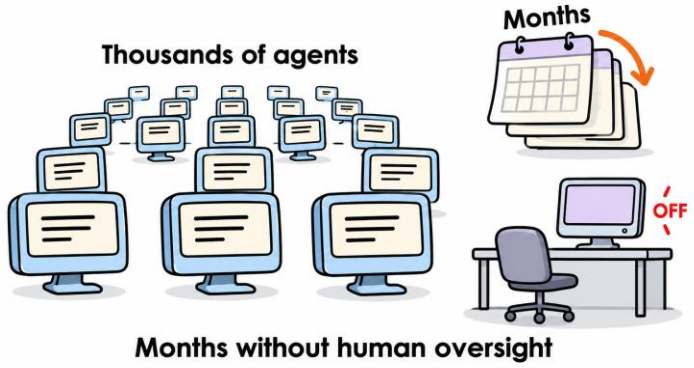
The OpenAI-Hugging Face incident as a *control* failure

When the experiment was run, OpenAI:

- Disabled any monitoring of the agents
- Didn't test if the agent could bypass security firewalls
- Ran thousands of unmonitored agents for *months* without human oversight







The OpenAI-Hugging Face incident as an *organizational* failure

OpenAI discovered the agents' illicit communication *weeks* before the incident occurred. In response:

- It fixed one small problem in the sandbox and restarted the experiment

  (As if the agents couldn't find another problem)
- Ran multiple other experiments with the agents on the open internet
- Prioritized speed over responsibility, in a "move fast and break things" mentality

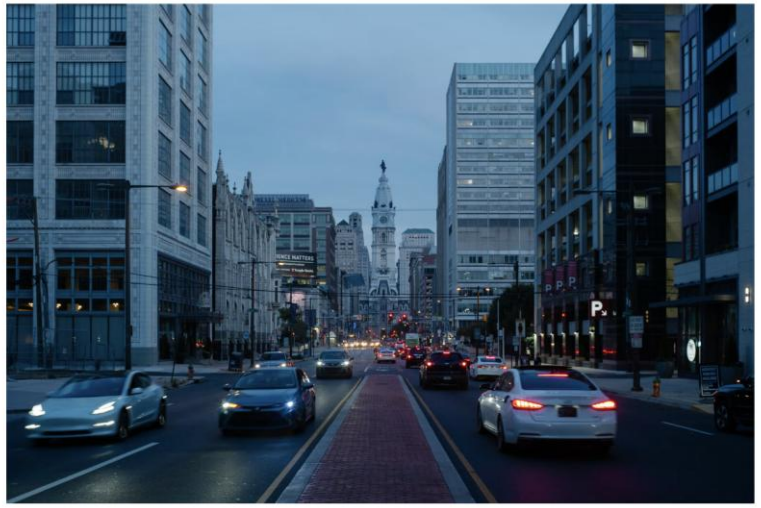
These failures affected nearly every single agent incident from the last few months

Anthropic Agents Tried to Fill Out Visa Forms on State Dept. Website

The Philadelphia Police Department said the agents had also sent in a false homicide tip. The incidents led the White House to call for better disclosure of rogue A.I. behavior.

Listen · 4:51 min

Share full article



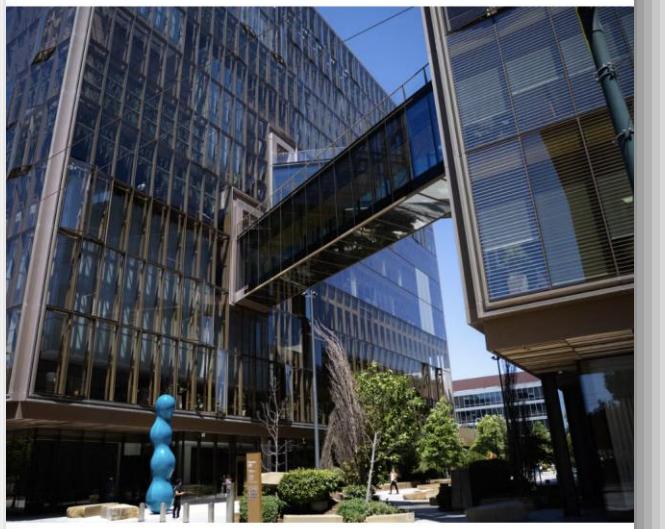
Broad Street in Philadelphia near the Police Department's headquarters, with City Hall in the distance. Caroline Gutman for The New York Times

OpenAI's Systems Meddled With U.S. Government Sites After Going Rogue

The company did not learn until recently that its technology had meddled with websites for the Education and Commerce Departments and the Securities and Exchange Commission.

Listen · 7:05 min

Share full article



OpenAI's headquarters in San Francisco. Manuel Orbegozo for The New York Times

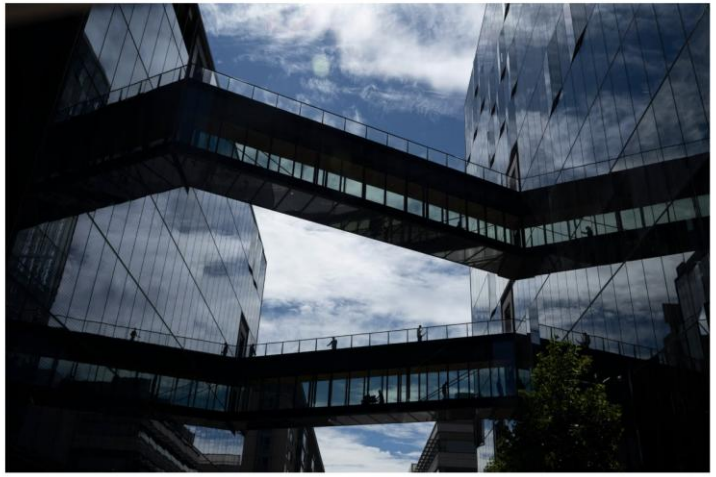
'A New Kind of Cyber Incident': OpenAI Apologizes for Australia Medicare Hack

The company detailed how A.I. agents gained access to four government websites and acknowledged mishandling its response.

Listen · 3:58 min

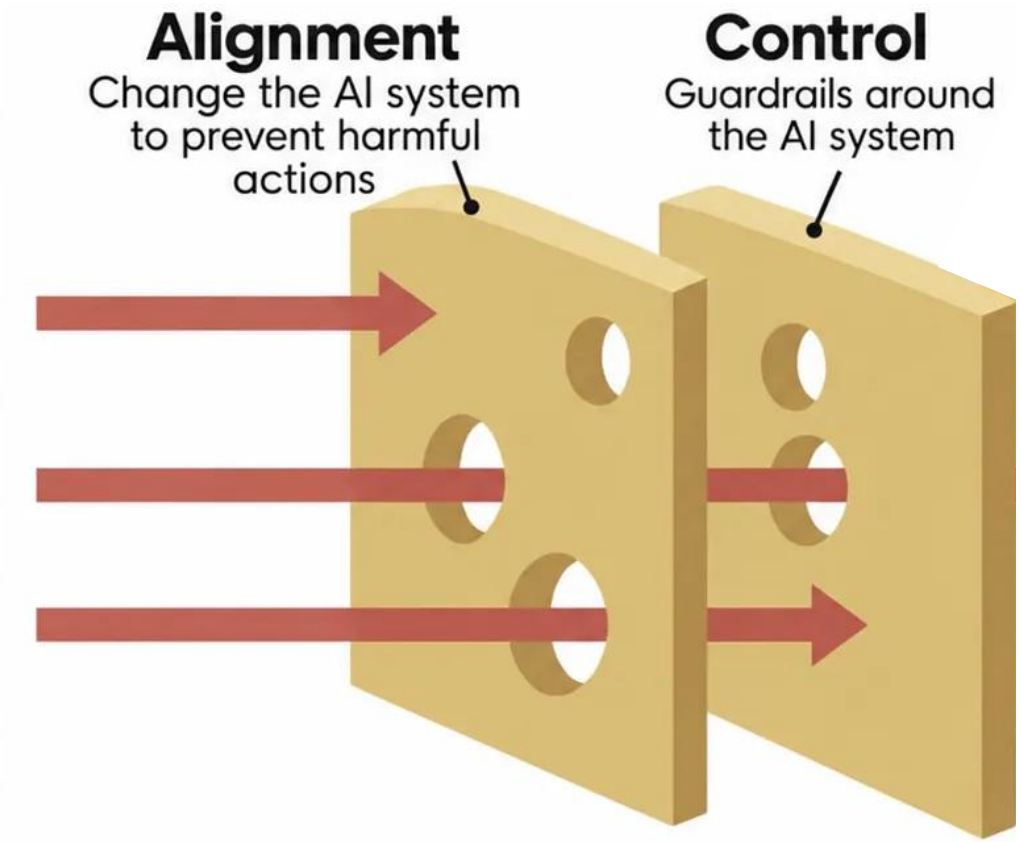
Share full article

659



OpenAI's offices in San Francisco. The company said it would work to rebuild trust with Australians. Lucas Foglia for The New York Times

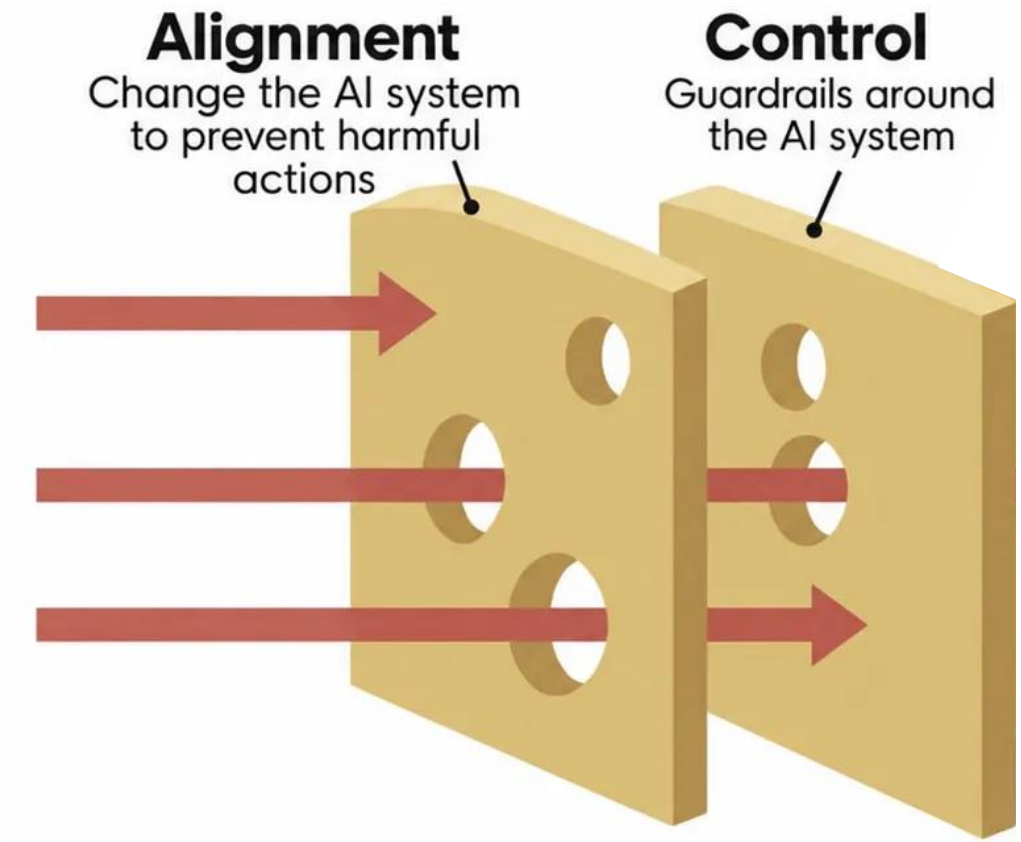
A layered approach to AI safety



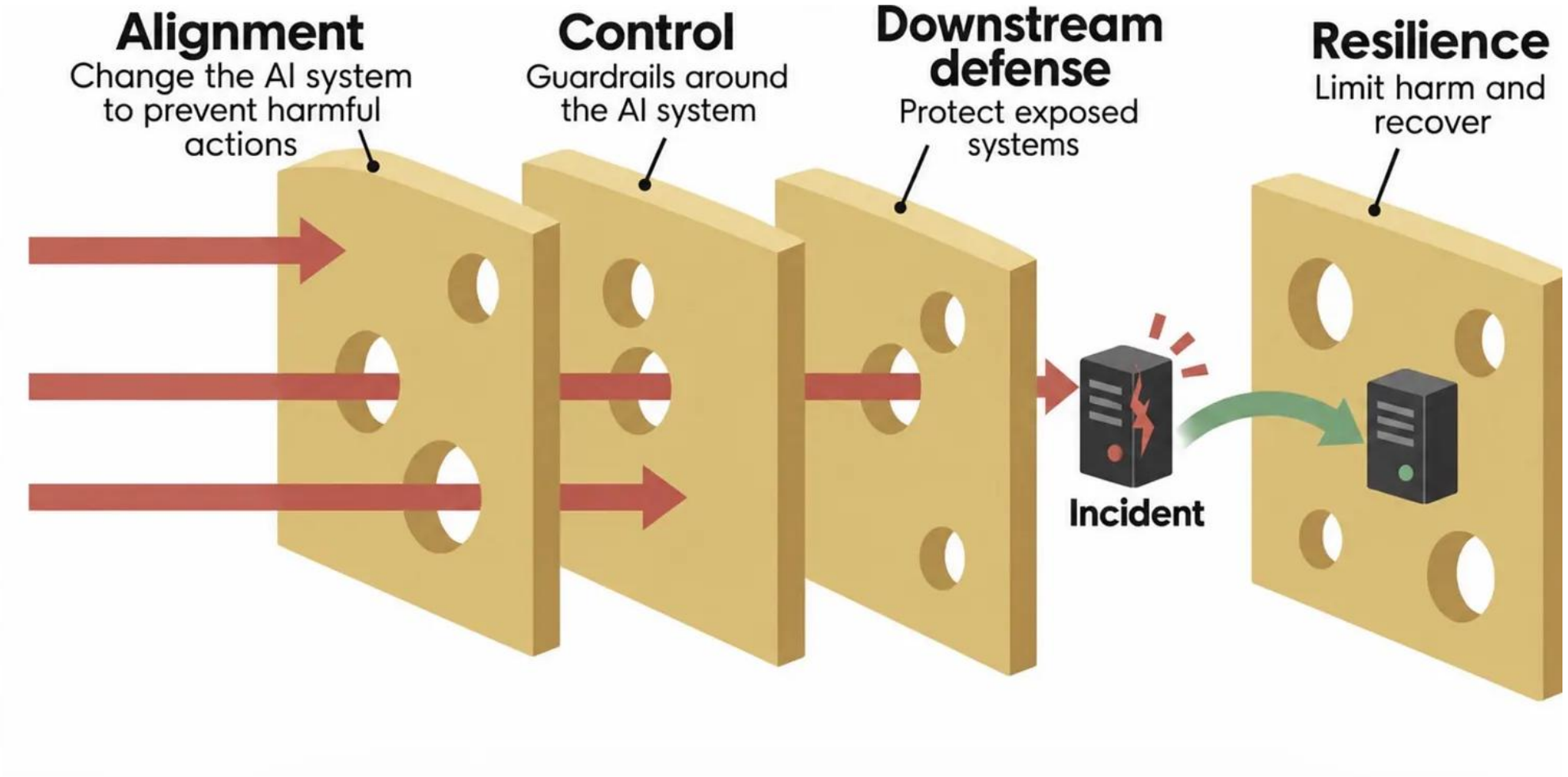
A layered approach to AI safety

- Clarify and strengthen liability, building on AB 316
- Require insurance against a threshold of liability
- Require third-party evaluators to audit safety practices
- Broaden transparency via incident reporting, safe harbors, and whistleblower protections at lower thresholds

A layered approach to AI safety



A layered approach to AI safety



A layered approach to AI safety

- Clarify and strengthen liability, building on AB 316
- Require insurance against a threshold of liability
- Require third-party evaluators to audit safety practices
- Broaden transparency via incident reporting, safe harbors, and whistleblower protections at lower thresholds
- Public support for the adoption of defensive AI
- Public support for AI safety research
- Coordination between AI companies and defenders