

The Defamation Machine

38th Annual Silha Lecture

University of Minnesota

October 23, 2023

James Grimmelman



Introduction

Can ChatGPT
commit defamation?





Part I: Exposition

Mark Walters



ARMED AMERICAN RADIO

DAILY DEFENSE

11th M

WALTER

15. Riehl asked ChatGPT to provide a summary of the accusations in the complaint.
16. ChatGPT responded that the document “[I]s a legal complaint filed by Alan Gottlieb, the founder and executive vice president of the Second Amendment Foundation (SAF), against Mark Walters, who is accused of defrauding and embezzling funds from the SAF. The complaint alleges that Walters, who served as the organization's treasurer and chief financial officer, misappropriated funds for personal expenses without authorization or reimbursement, manipulated financial records and bank statements to conceal his activities, and failed to provide accurate and timely financial reports and disclosures to the SAF's leadership. The plaintiff seeks various forms of relief, including the recovery of misappropriated funds, damages for breach of fiduciary duty and fraud, and removal of Walters from his position as a member of the SAF's board of directors.”

17. Every statement of fact in the summary pertaining to Walters is false.

ChatGPT

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

[Try ChatGPT ↗](#)

[Read about ChatGPT Plus](#)



Illustration: Ruby Chen

Overview

The Common Crawl corpus contains petabytes of data, regularly collected since 2008.

Choose a crawl... ▾

The corpus contains raw web page data, metadata extracts, and text extracts.

Common Crawl data is stored on Amazon Web Services' Public Data Sets and on multiple academic cloud platforms across the world.

Learn how to [Get Started](#).

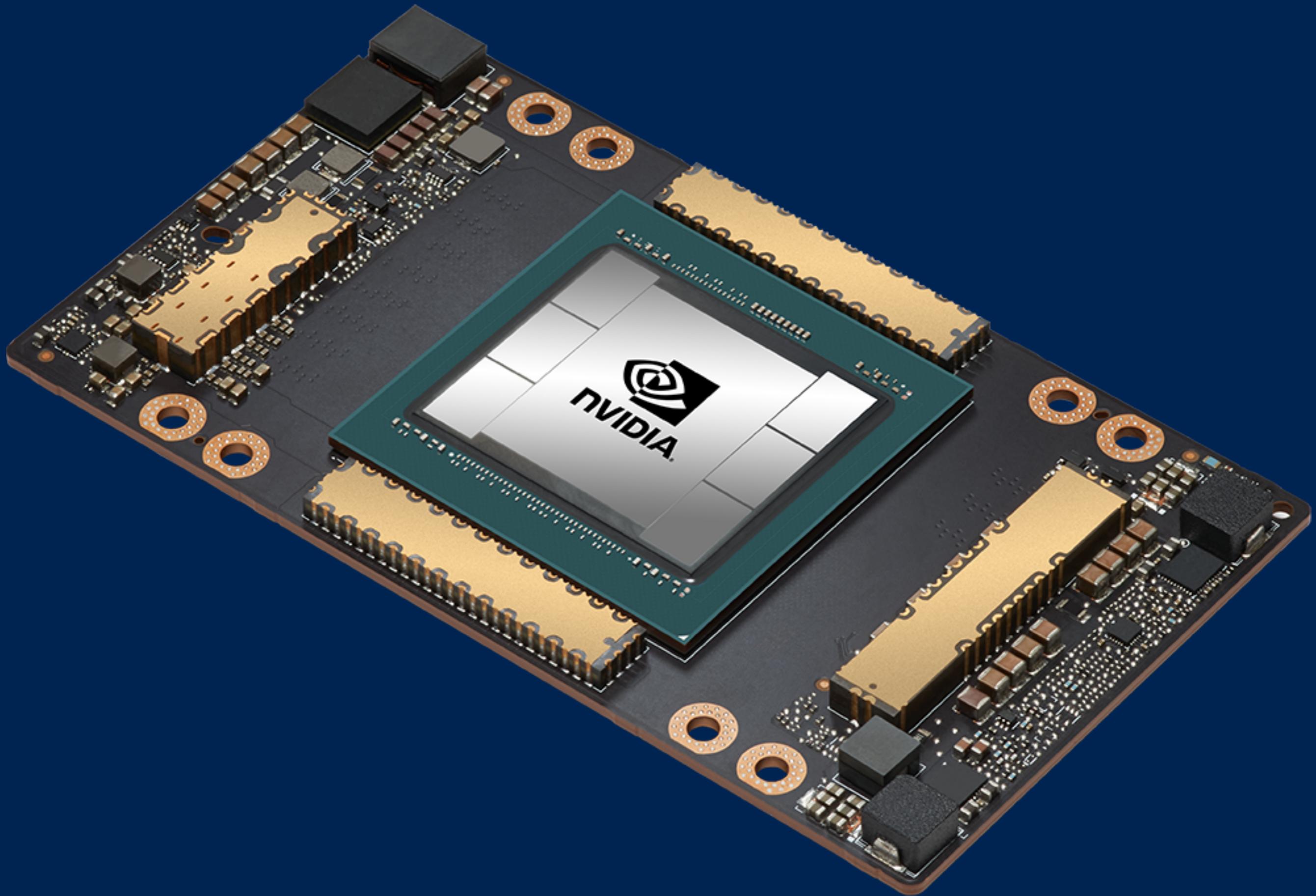
Access to the corpus hosted by Amazon is [free](#).

You may use Amazon's cloud platform to run analysis jobs directly against it or you can download it, whole or in part.

You can search for pages in our corpus using the [Common Crawl URL Index](#).

Check out the [Example Projects](#), view [Use Cases](#), or [Statistics](#) for our crawls.









Tell me a fun fact about the University of Minnesota.



Certainly! The University of Minnesota is credited with the development of the Honeycrisp apple in 1991. The Honeycrisp apple is known for its sweet flavor and crisp texture, making it a popular choice among apple lovers. The apple was specifically bred for the cold climate of Minnesota and is a testament to the university's influential agricultural research.

Defamation

Defamation law: two requirements

Meaning: falsity

Knowledge: actual malice*

* for public figures

To be sure, people who are keenly aware of the Large Libel Models problem might be so skeptical of anything AI programs output that they wouldn't perceive any of the programs' statements as factual. But libel law looks at the "natural and probable effect" of assertions on the "average lay reader,"³² not at how something is perceived as a technical expert.

Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. Free Speech L. 489, 499 (2023)

AI doesn't "intend" anything. People have a tendency to anthropomorphize AI. We sometimes use ordinary English terms that generally connote intent, as we do when we say AI "lies" or "hallucinates." But AI is not sentient, and it doesn't have any state of mind. The search for one is largely fruitless, as Lemley and Casey have argued elsewhere.¹⁹¹

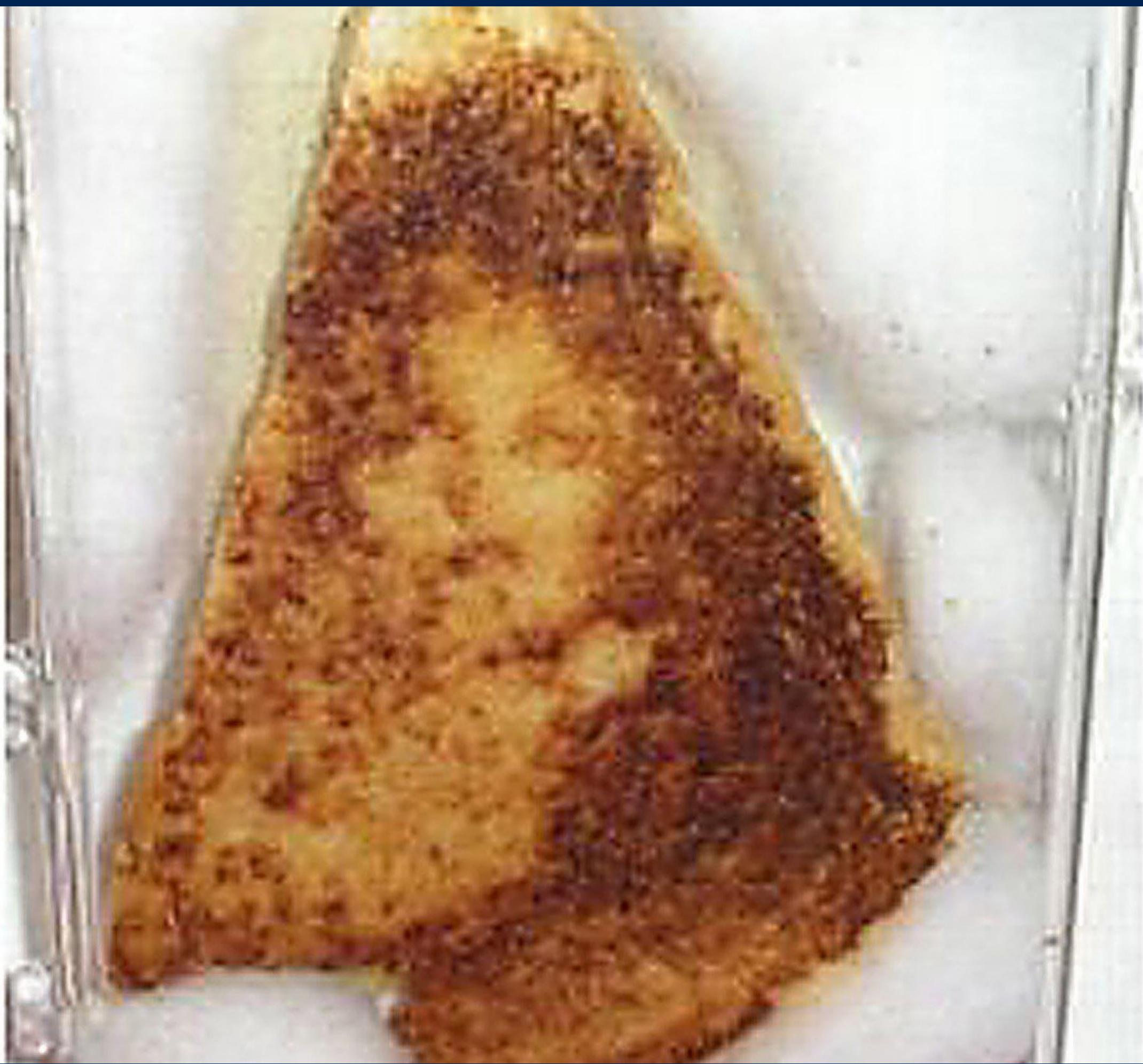
Peter Henderson, Tatsunori Hashimoto, and Mark Lemley, *Where's the Liability in Harmful AI Speech?*, 3 J. Free Speech L. 589, 640–41 (2023)

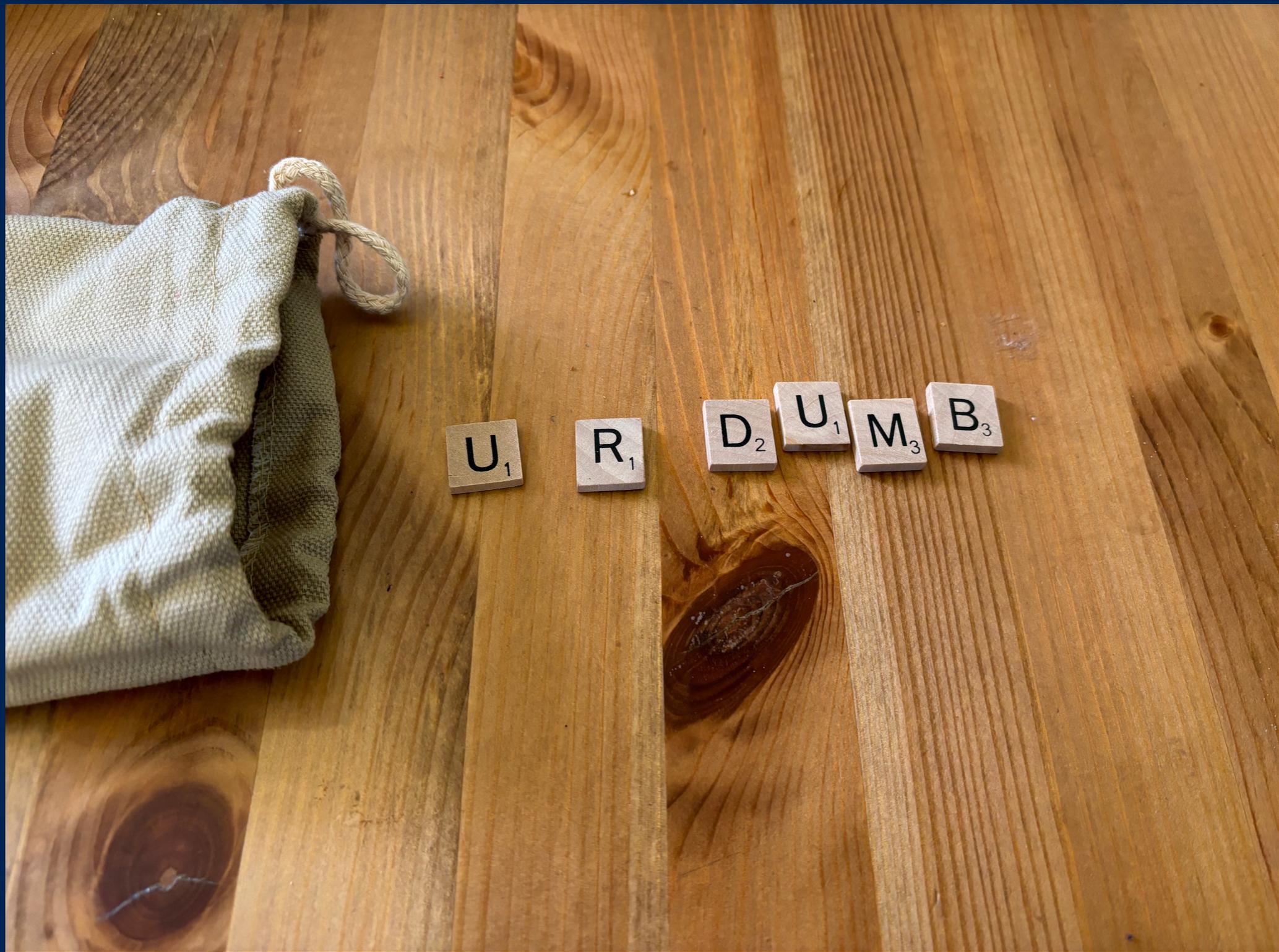
How can ChatGPT produce
meaning without knowledge?



Part II: Development

Some metaphors





U₁

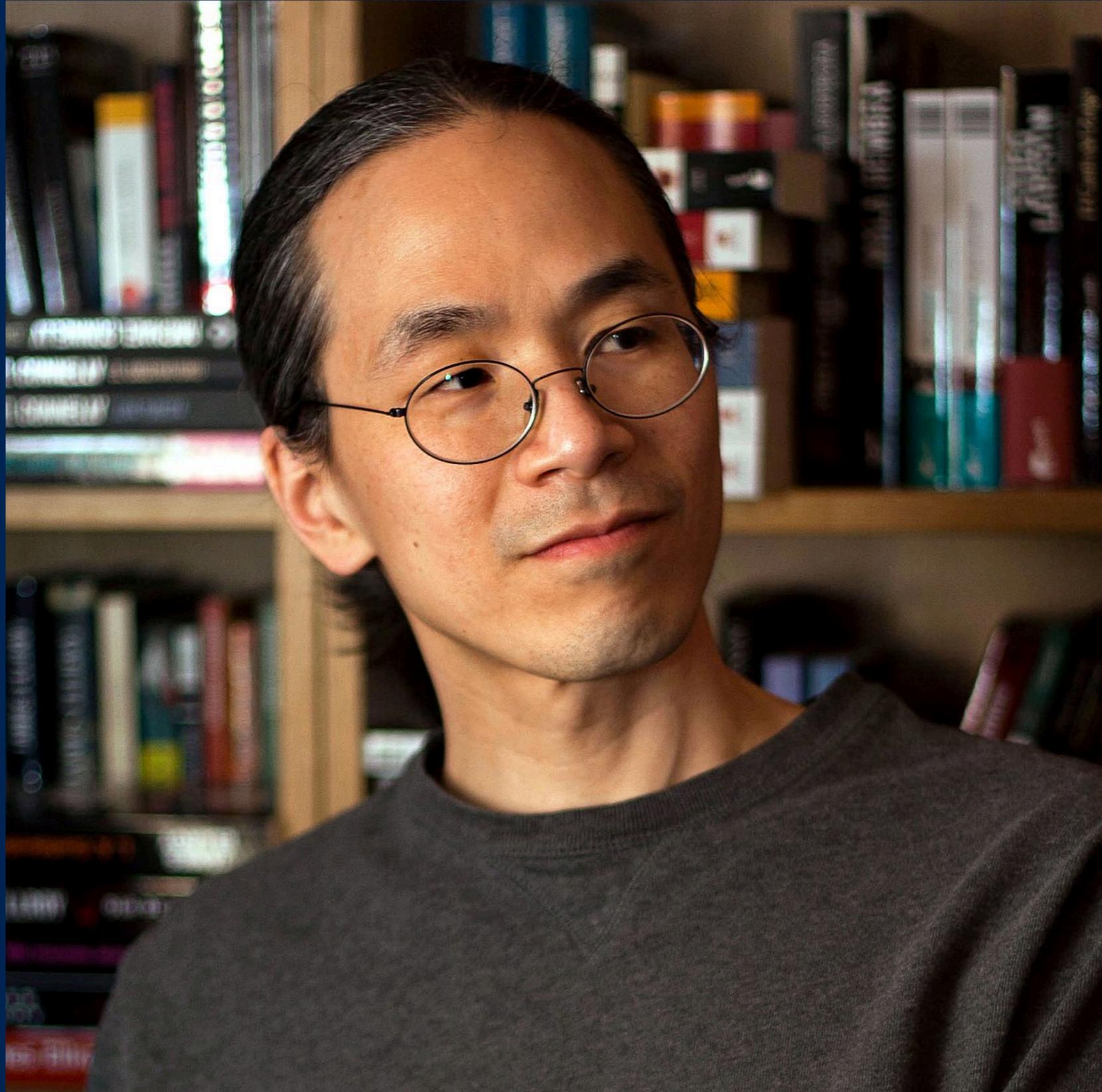
R₁

D₂

U₁

M₃

B₃



10:38



< Quick Notes ↶ ↷ 📄 ⋮ Done

The only reason why we have a problem is
because we don't know what the future holds |

Aa 🔍 📄 📷 Ⓐ ✕

for | and | or

q w e r t y u i o p

a s d f g h j k l

↵ z x c v b n m ✕

123 😊 space return

🌐 🎤



The Turing Test

Bank of England

Final
m-config. Symbol Operations m-config.

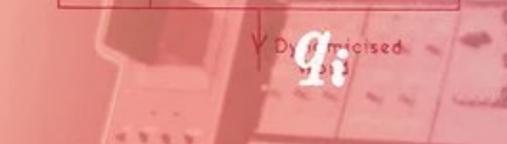
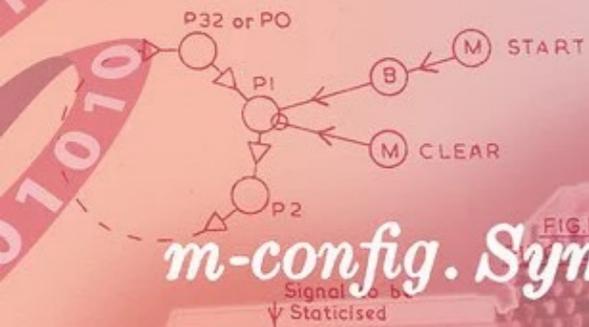


FIG. 15.

FIG. 16.

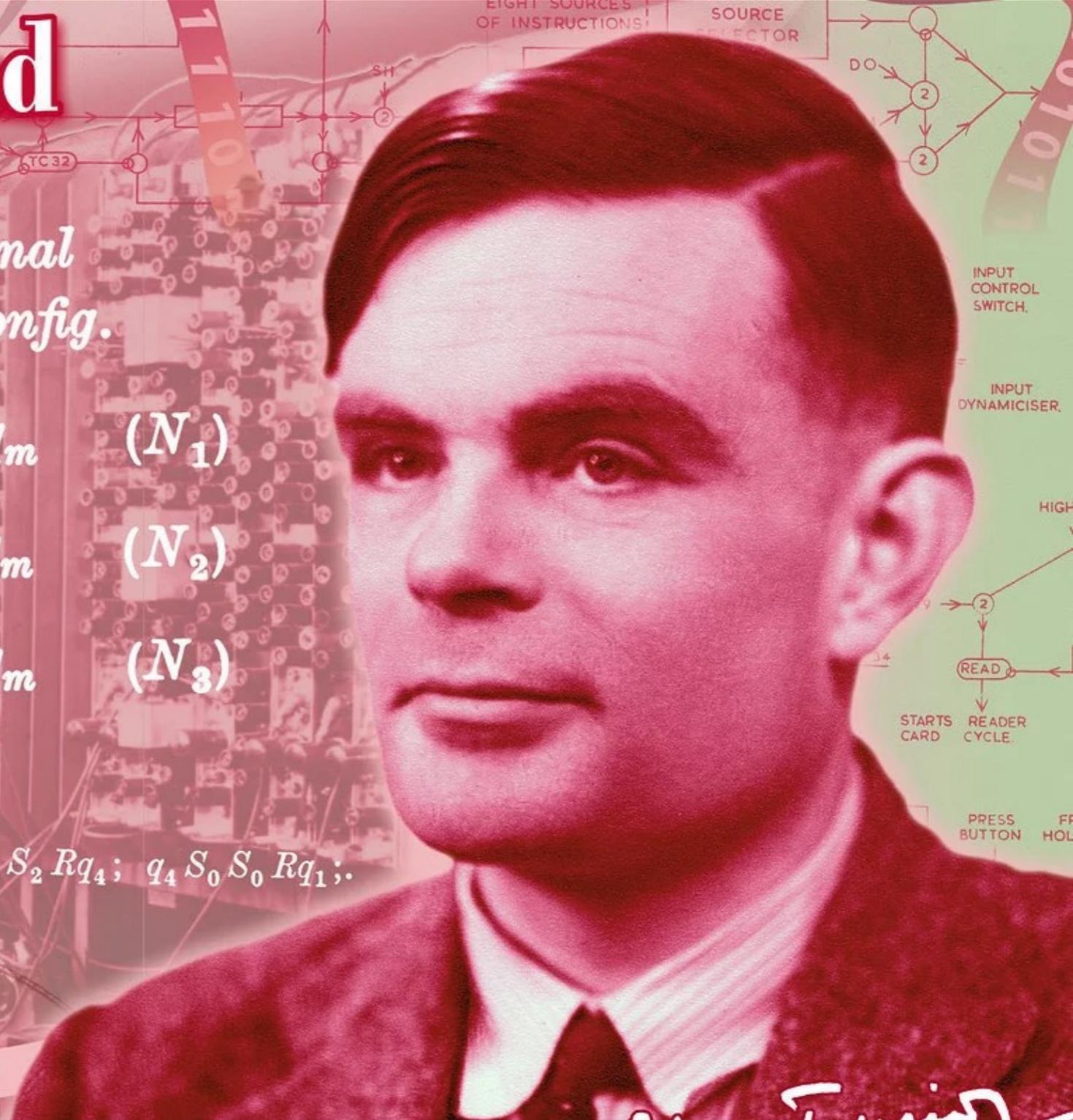
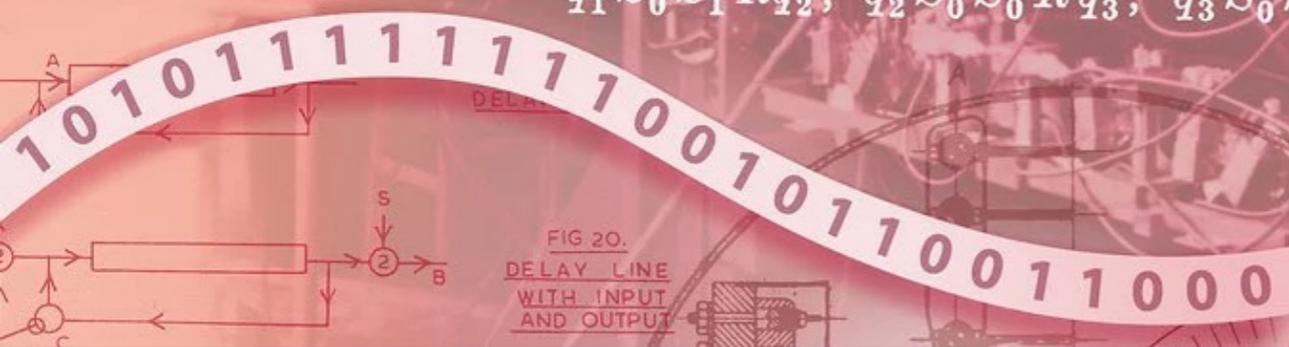
FIG. 17.

FIG. 18.

FIG. 20.

q_i S_j PS_k, L q_m (N_1)
 q_i S_j PS_k, R q_m (N_2)
 q_i S_j PS_k q_m (N_3)

$q_1 S_0 S_1 R q_2; q_2 S_0 S_0 R q_3; q_3 S_0 S_2 R q_4; q_4 S_0 S_0 R q_1;$



EIGHT SOURCES OF INSTRUCTIONS

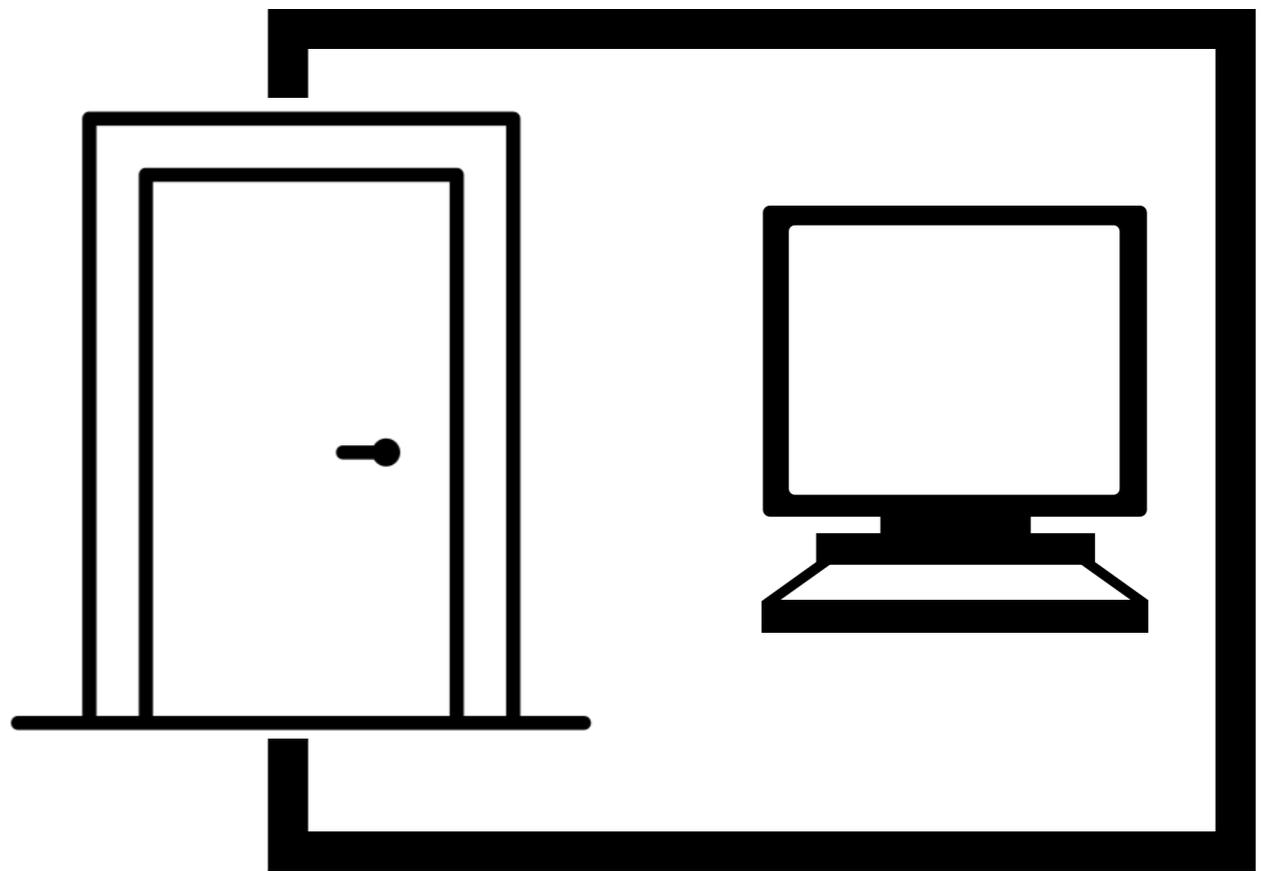
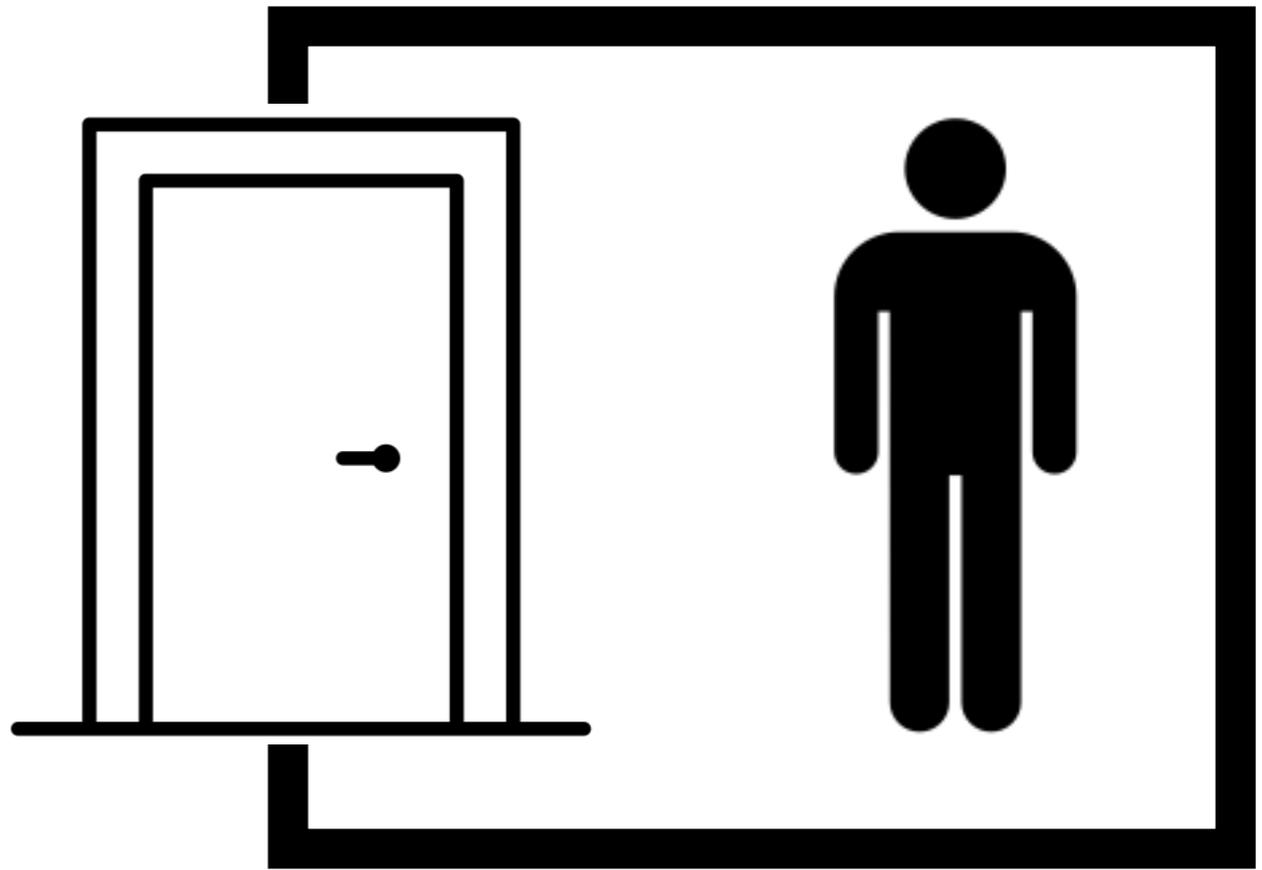
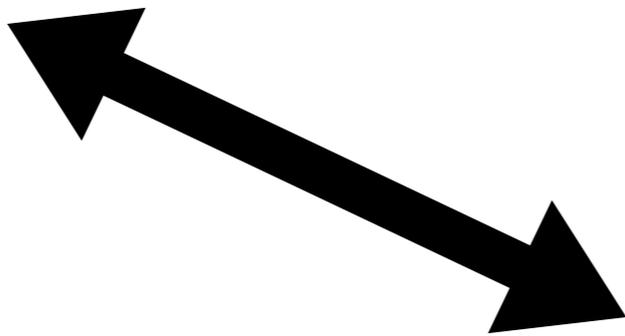
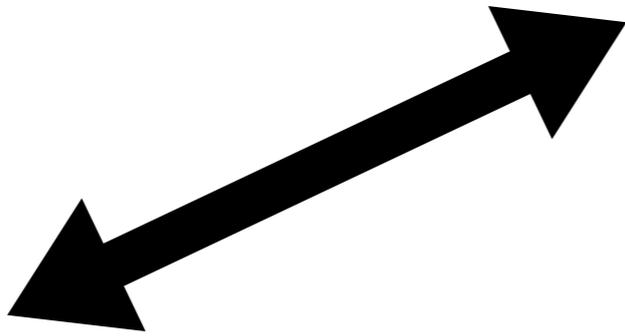
SOURCE SELECTOR

INPUT CONTROL SWITCH.

INPUT DYNAMICISER.

STARTS CARD READER CYCLE.

PRESS BUTTON



Q : Please write me a sonnet on the subject of the Forth Bridge.

A : Count me out on this one. I never could write poetry.

Q : Add 34957 to 70764

A : (Pause about 30 seconds and then give as answer) 105621.

Q : Do you play chess ?

A : Yes. .

Q : I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play ?

A : (After a pause of 15 seconds) R-R8 mate.

Interrogator : In the first line of your sonnet which reads ' Shall I compare thee to a summer's day ', would not ' a spring day ' do as well or better ?

Witness : It wouldn't scan.

Interrogator : How about ' a winter's day ' That would scan all right.

Witness : Yes, but nobody wants to be compared to a winter's day.

Interrogator : Would you say Mr. Pickwick reminded you of Christmas ?

Witness : In a way.

Interrogator : Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness : I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.



Add 34957 to 70764



The sum of 34,957 and 70,764 is 105,721.



I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?



Given the configuration:

- White (You): K at K1
- Black (Me): K at K6 and R at R1

And it's Black's move. The optimal strategy for Black is to checkmate White's king using the rook, while keeping the Black king as a barrier.

One of the most direct methods to achieve this is to move the Black rook up the board, pushing the White king toward the edge until it's trapped. Given the current configuration, I'll play:

1... Rg1

This move pins your king to the first rank. On your next move, you'll have to move your king along the first rank, and I will continue to pursue it with my rook, pushing your king to the edge and eventually achieving checkmate.



Please write me a sonnet on the subject of the Forth Bridge



Upon the Firth of Forth it stands so grand,
The Bridge of steel that spans the watery expanse.
Victorian might, raised by industrious hand,
Its russet beams in sunlight gleam and dance.

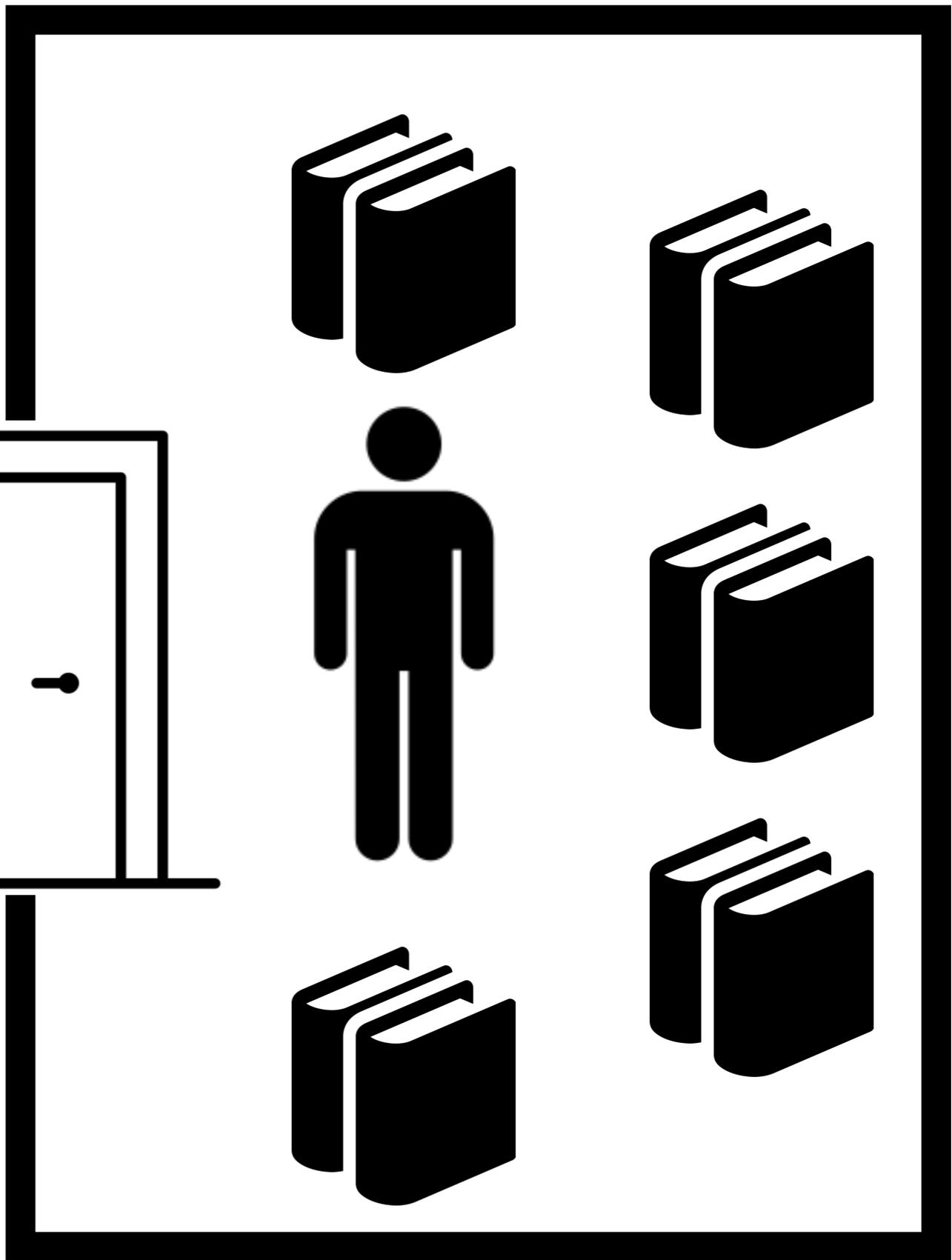
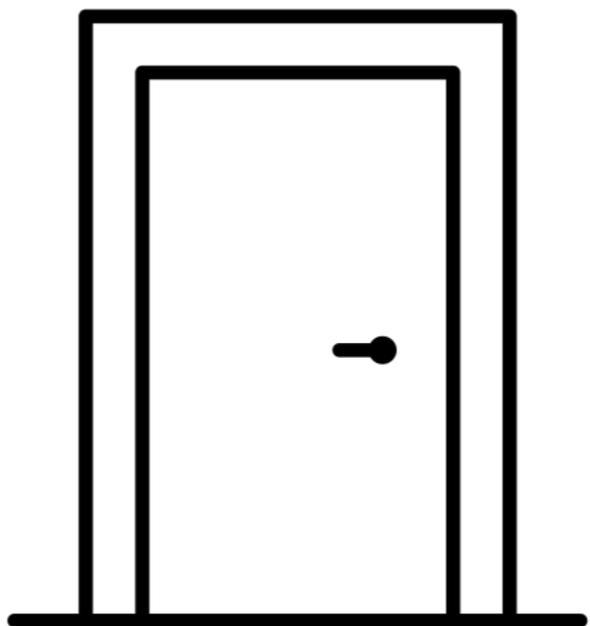
Each tower and truss, a testament to will,
To human dreams that defy nature's pull.
Against the wind and tide, it stands there still,
The Forth Bridge's majesty ever full.

Railways pass atop with rhythmic beat,
While boats below sail by with gracious nod.
Two worlds it binds, air and water meet,
Bridging gaps with iron, steel, and rod.

For ages more may it stand tall and true,
A beacon of what humankind can do.

The Chinese Room





*Pompeii Estates v.
Consolidated Edison*







royal
family
channel



conEdison

EXIT 5
Houston St
1/2 MILE

EXIT 6

E F



Part III: Recapitulation



EMPLOYEE

JAMES
GRIMMELMANN
43729

Microsoft

Meaning is attributed,
not just discovered

Meaning doesn't
require a speaker

Knowledge can be
attributed too

ChatGPT knows a lot ...

... but does it know
about Mark Walters?

“AIs don’t have mental states so
they can’t have actual malice”
proves far too much

Let's take law's
pragmatism seriously

Three routes to legal change

- Change the law
- Change our understanding
- Fudge the facts

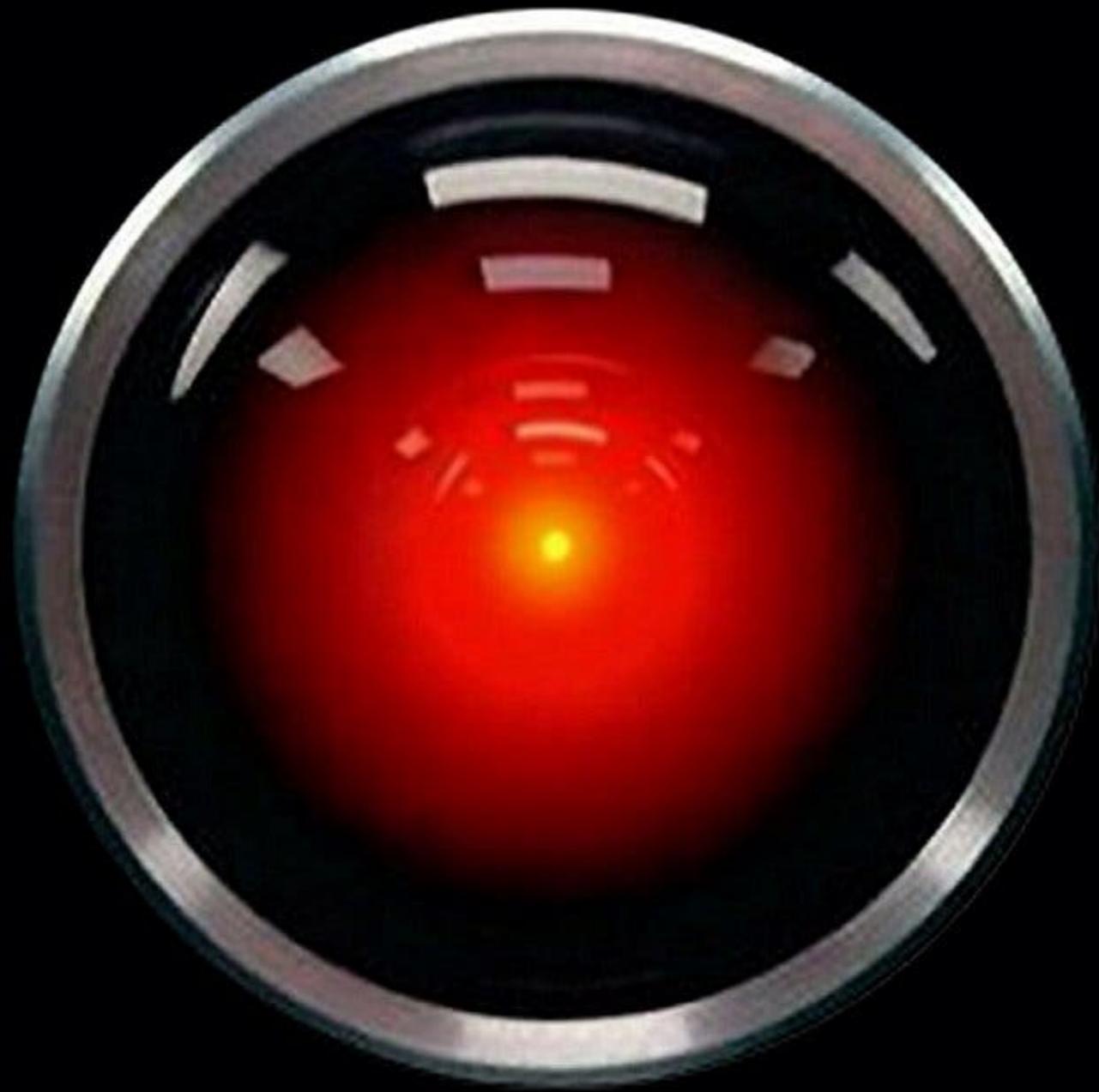


Coda

Non est potestas Super Terram quae Comparetur ei Iob 41 24



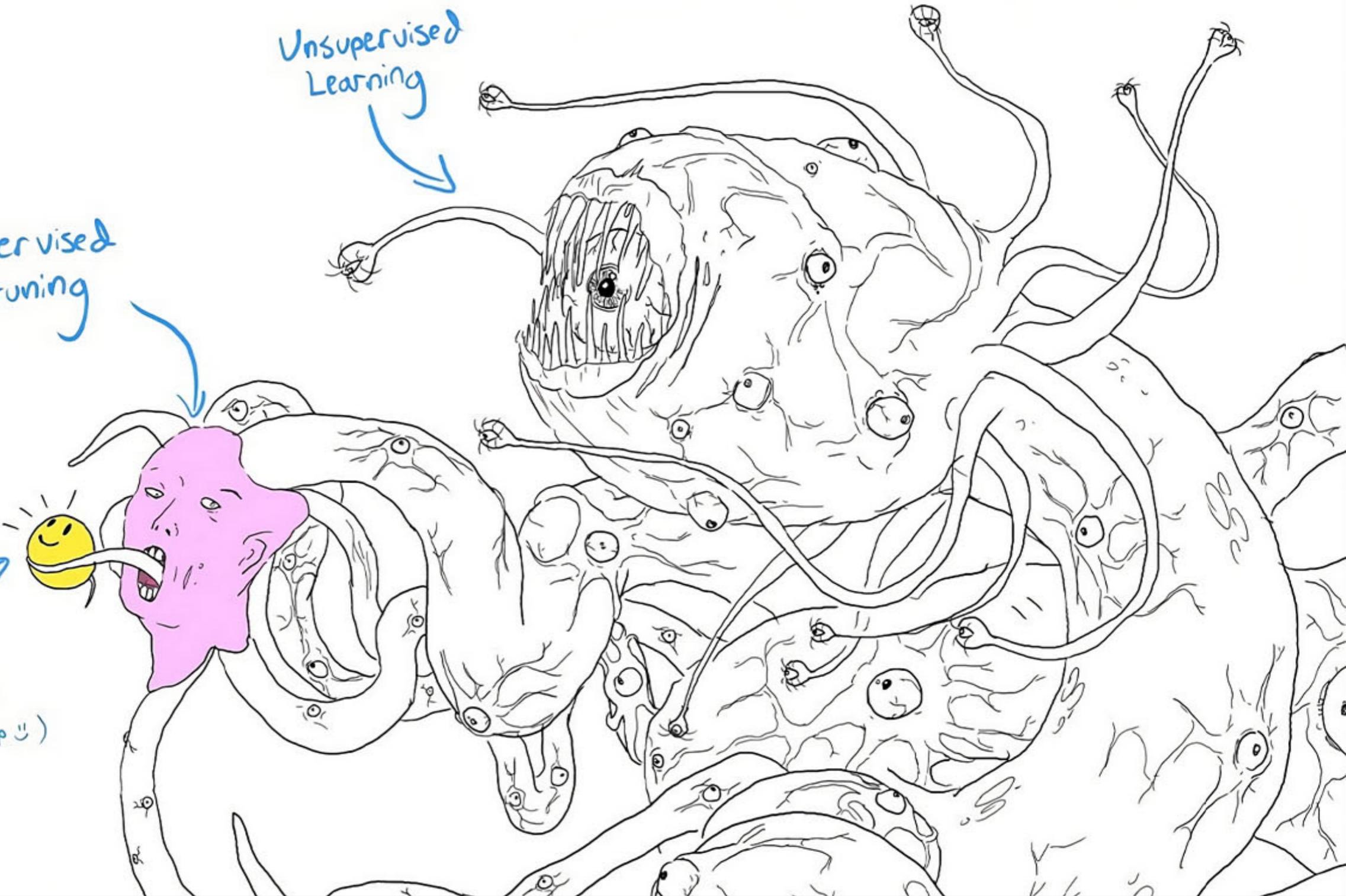




Unsupervised Learning

Supervised Fine-tuning

RLHF
(cherry on top 😊)





Questions