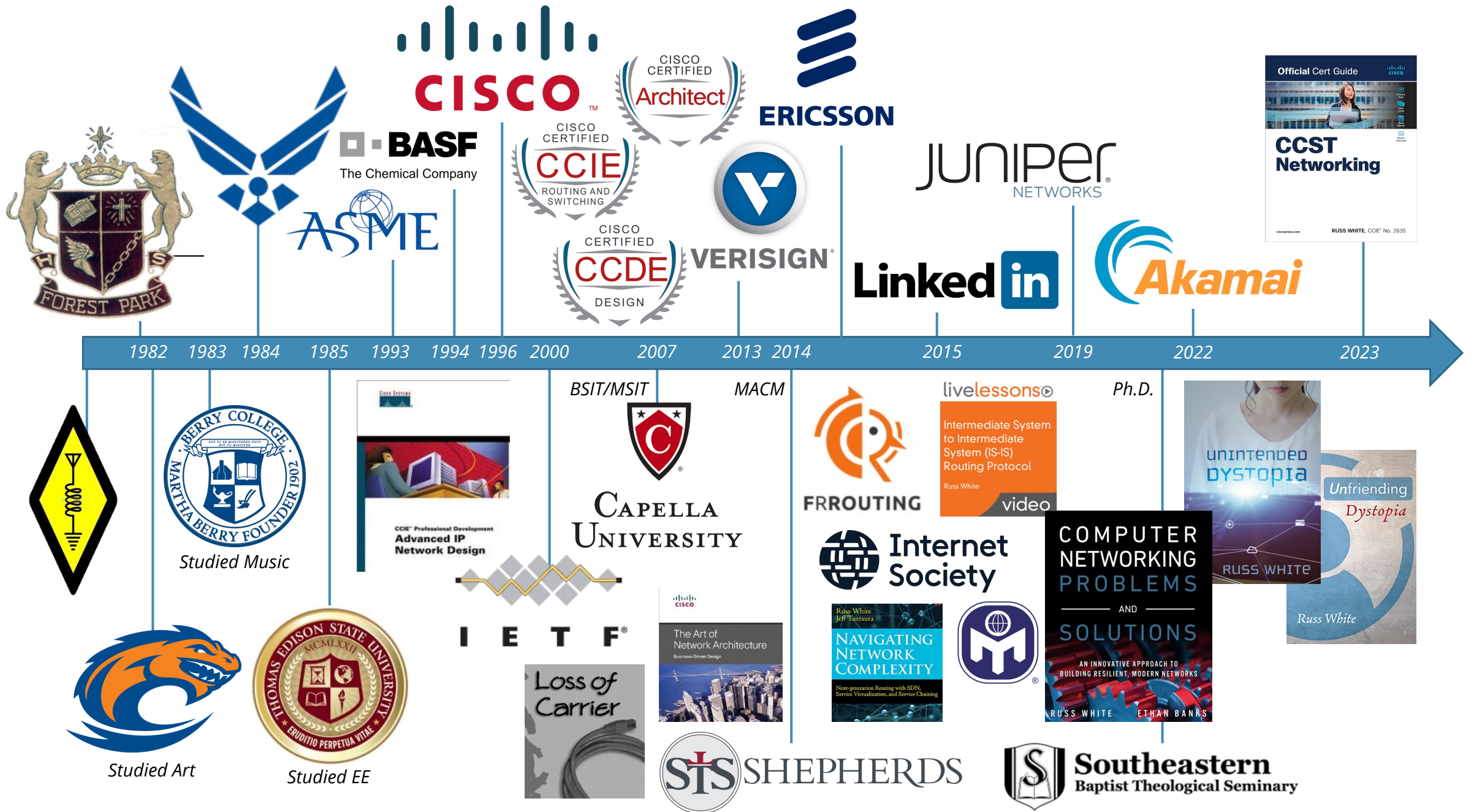


AI/ETHICS

Large Language Models (LLMs)

Russ White, Ph.D.
2023







<https://sharedpublic.s3.amazonaws.com/s360-ai-ethics-presentation.pdf>



<https://sharedpublic.s3.amazonaws.com/s360-ai-ethics-notes.pdf>



WHAT IS AI (REALLY)?

MACRO ETHICS

MICRO ETHICS

AGENDA

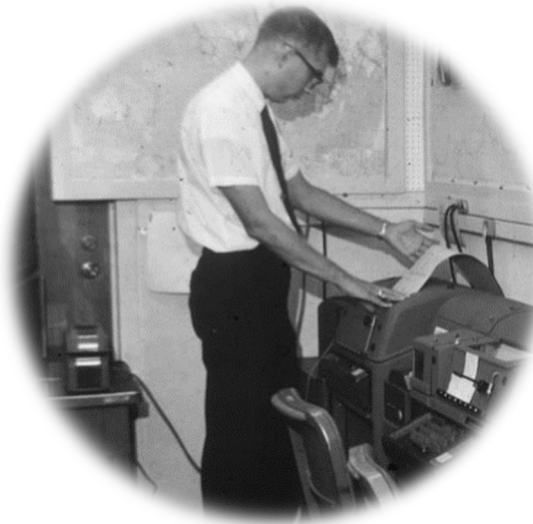
Suppose we want to answer this question ...

There are two ways to find an answer

Investigate the Bible
(do the research)

Ask an expert

Is the Bible True?



WHAT IS AI (REALLY)?

We trust the expert
because they
(should) have ...

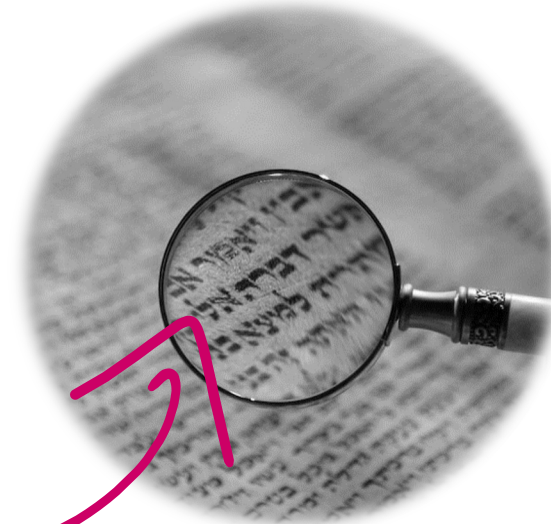
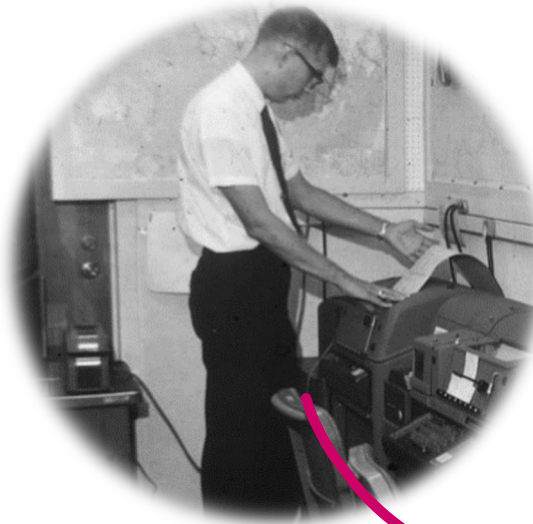
Done the research

Are seeking the truth
*(rather than personal gain,
furthering an agenda, etc.)*

But what if this is
not true?

The expert's skill and
agenda matter

Is the Bible True?



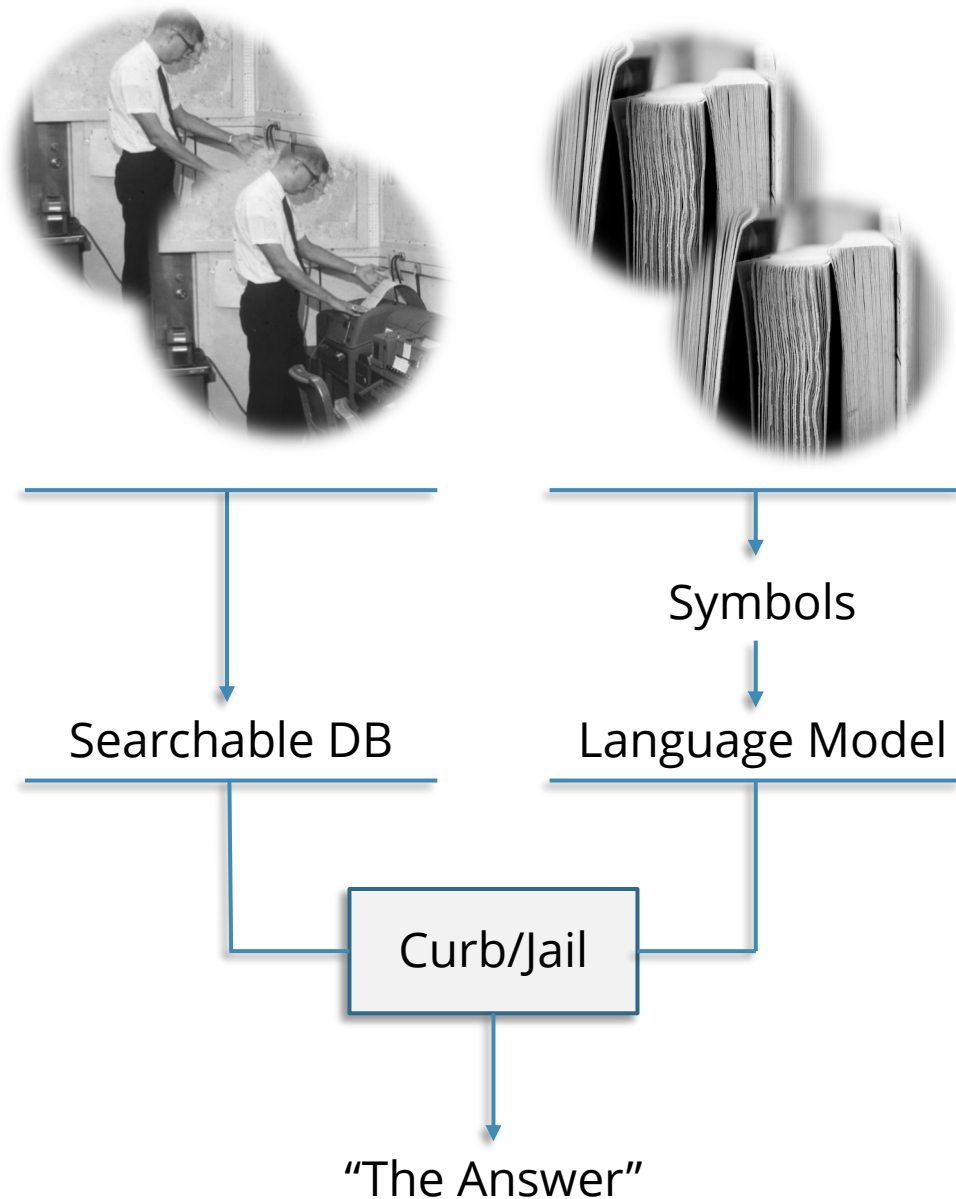
WHAT IS AI (REALLY)?

Large Language Models (LLMs) and other forms of AI/ML *essentially* replace the expert with a weighted sampling of *many experts*

AI/ML weighted sampling can be **roughly** be treated as a “statistical” operation for understanding its effects



WHAT IS AI (REALLY)?



Neural Net “learns”
Word pairings/patterns
Information
pairings/patterns

Use the *language model* to create a human-readable answer based on the *knowledge model*

Pass this through a “curb/jail”

Provide the answer
Essentially a Chinese Room

WHAT IS AI (REALLY)?



WHAT IS AI (REALLY)?

MACRO ETHICS

flattening

epistemological shifts

race to mediocrity

bias/filter

accuracy

MICRO ETHICS

AGENDA

What does "intelligence" mean?

Strongly related to the
imago deo

Many AI definitions

The ability to solve
problems like a
human

The ability to process
information like a
human

The ability to
communicate like a
human

I propose to consider the question, "Can machines think?" This should begin with definitions of the meaning of the terms "machine" and "think." The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is **dangerous**.

....

May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.

Turing, A. M. "Computing Machinery and Intelligence." Mind, New Series 59, no. 236 (1950): 433-60.

Life's answer



If intelligence is *just the ability to communicate like a human ...*

Human traits like ...

Free will

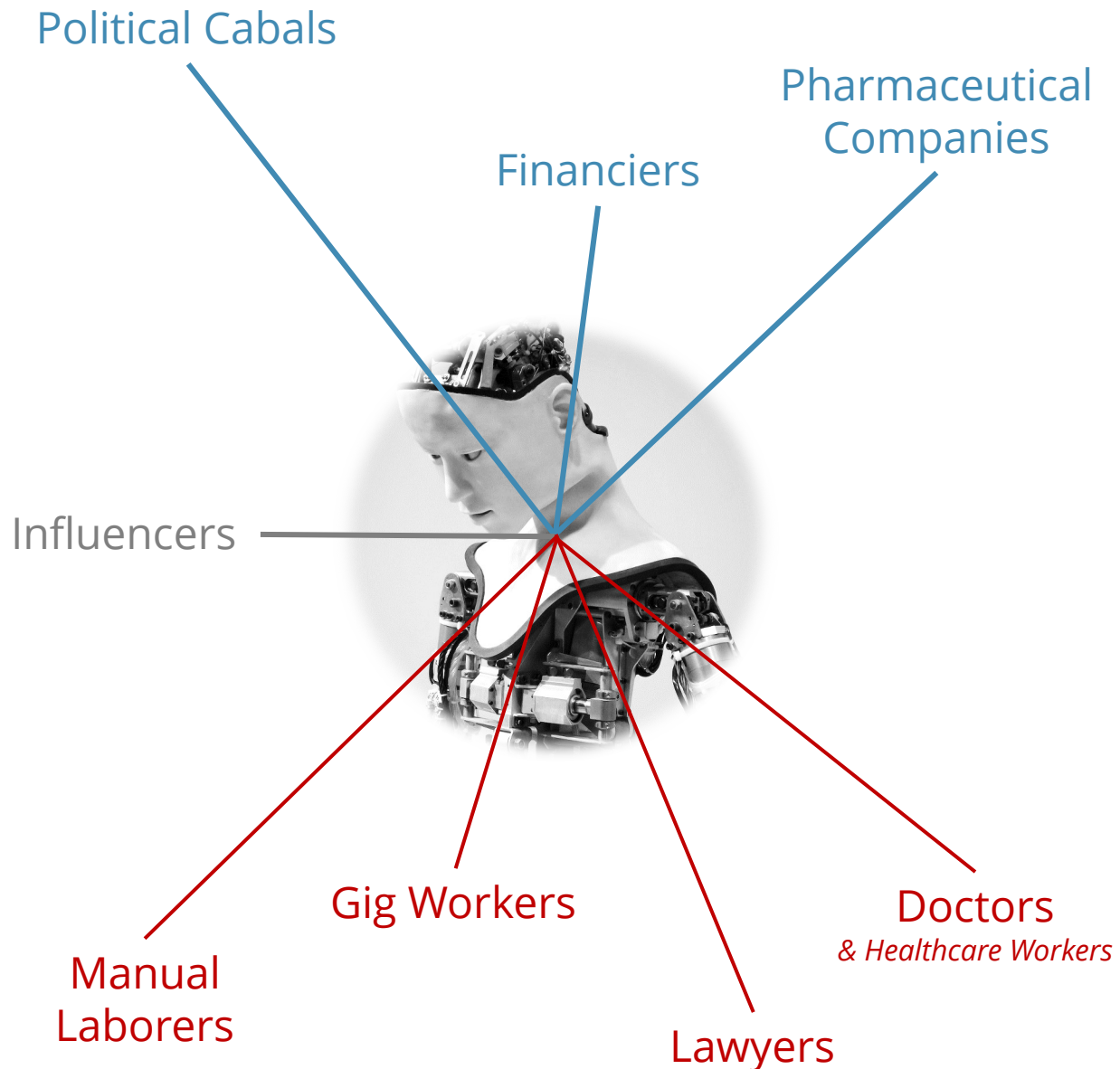
Emotions

Virtues

Vices

... no longer matter

Defining intelligence down *flattens* humans to something less than the *imago deo*



Two classes of people are developing

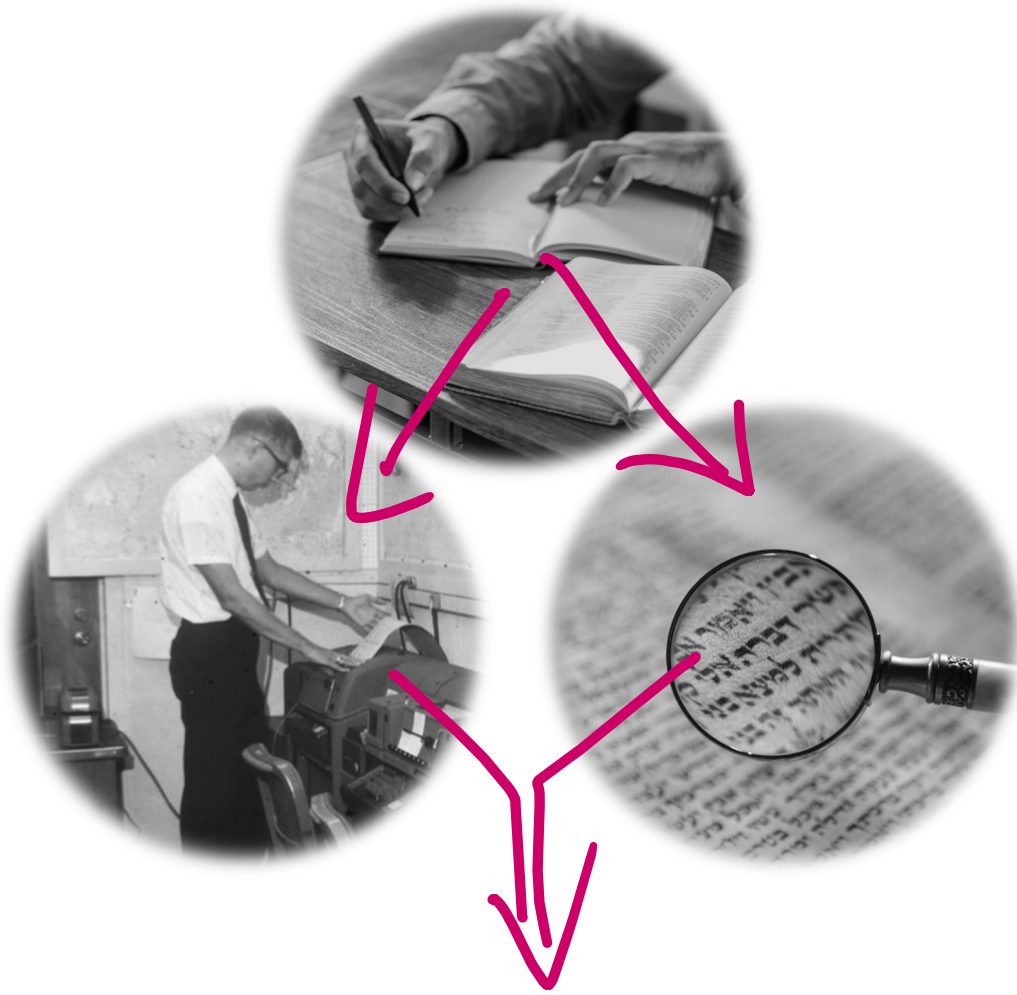
Some people train the AI
Others obey the AI

...what we call Man's power over nature turns out to be a power exercised by some men over other men with nature as its instrument.

C.S. Lewis, The Abolition of Man

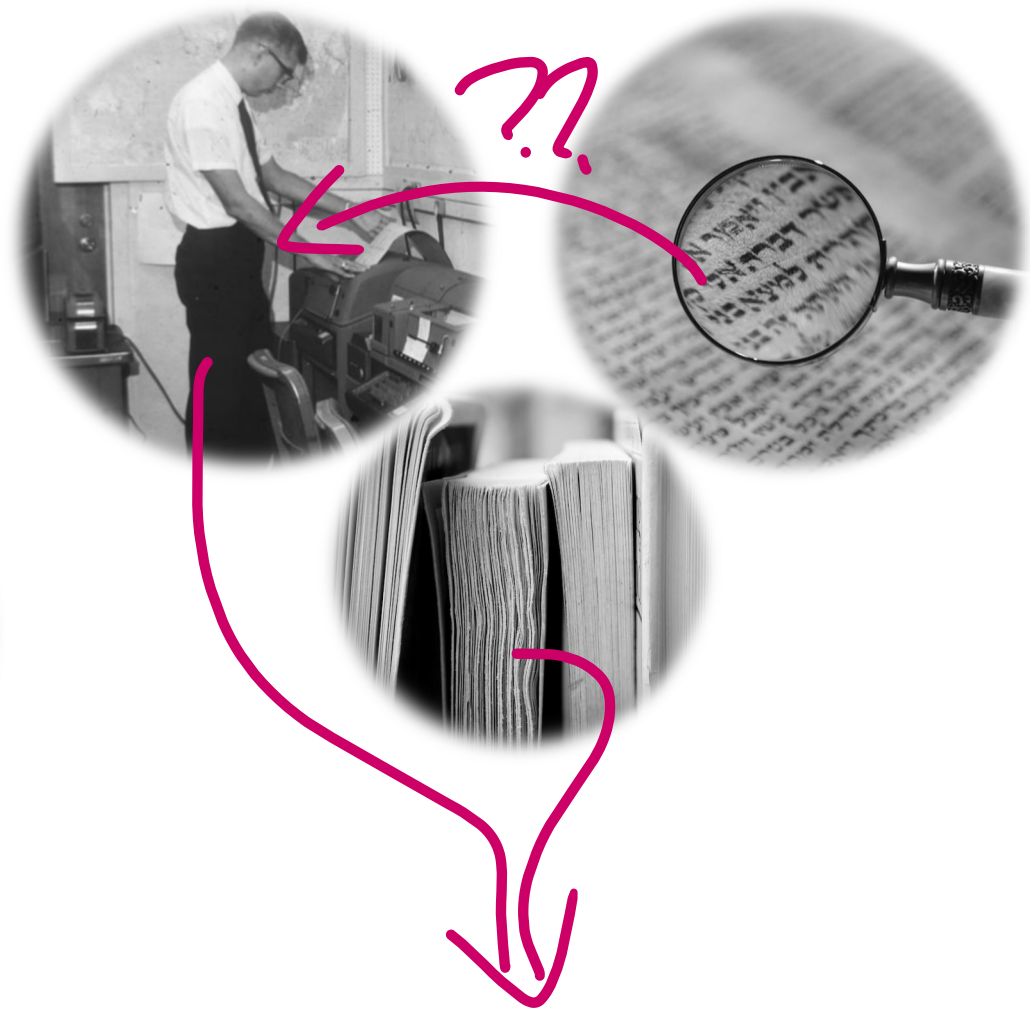
The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.

Eliezer Yudkowsky



*Seeking truth of an
external reality*

(realism)



*Seeking consensus about
meaning and "useful answers"*

(nominalism/utilitarianism)

EPISTEMOLOGICAL SHIFT



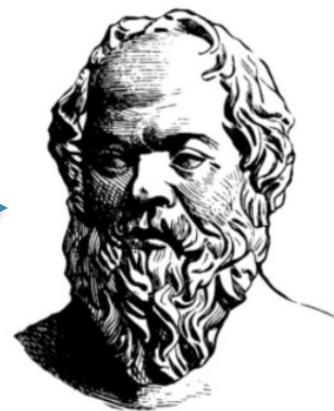
Excellent sources

*LLM
created
junk*

Poor sources

*LLM
Sample*

"Meh... good enough"



- Mediocrates

CONVERGENT MEDIOCRITY

Model collapse occurs when AI models train on AI-generated content. It's a process where small errors or biases in generated data compound with each cycle, eventually steering the model away from generating inferences based on the original distribution of data.

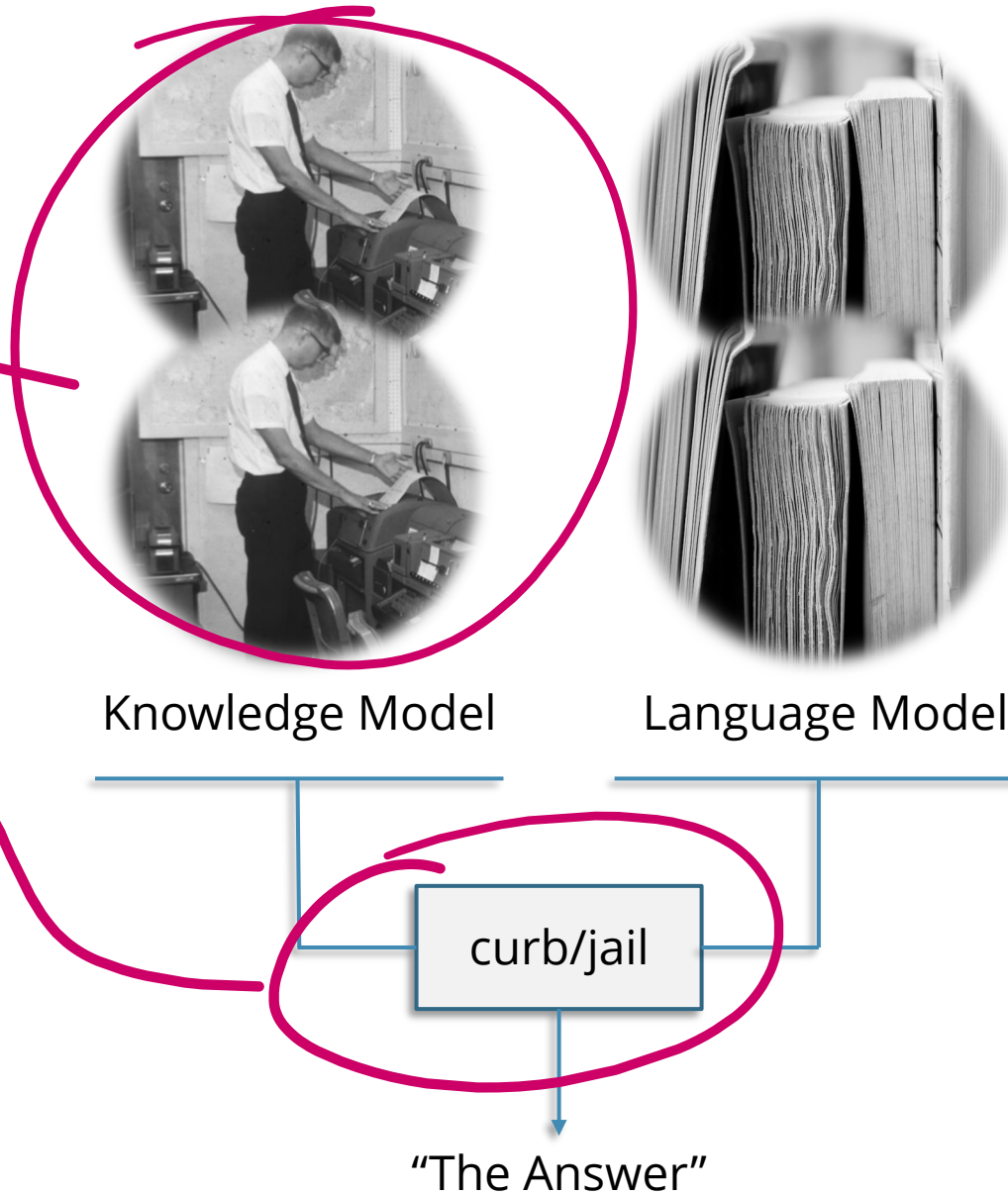
...

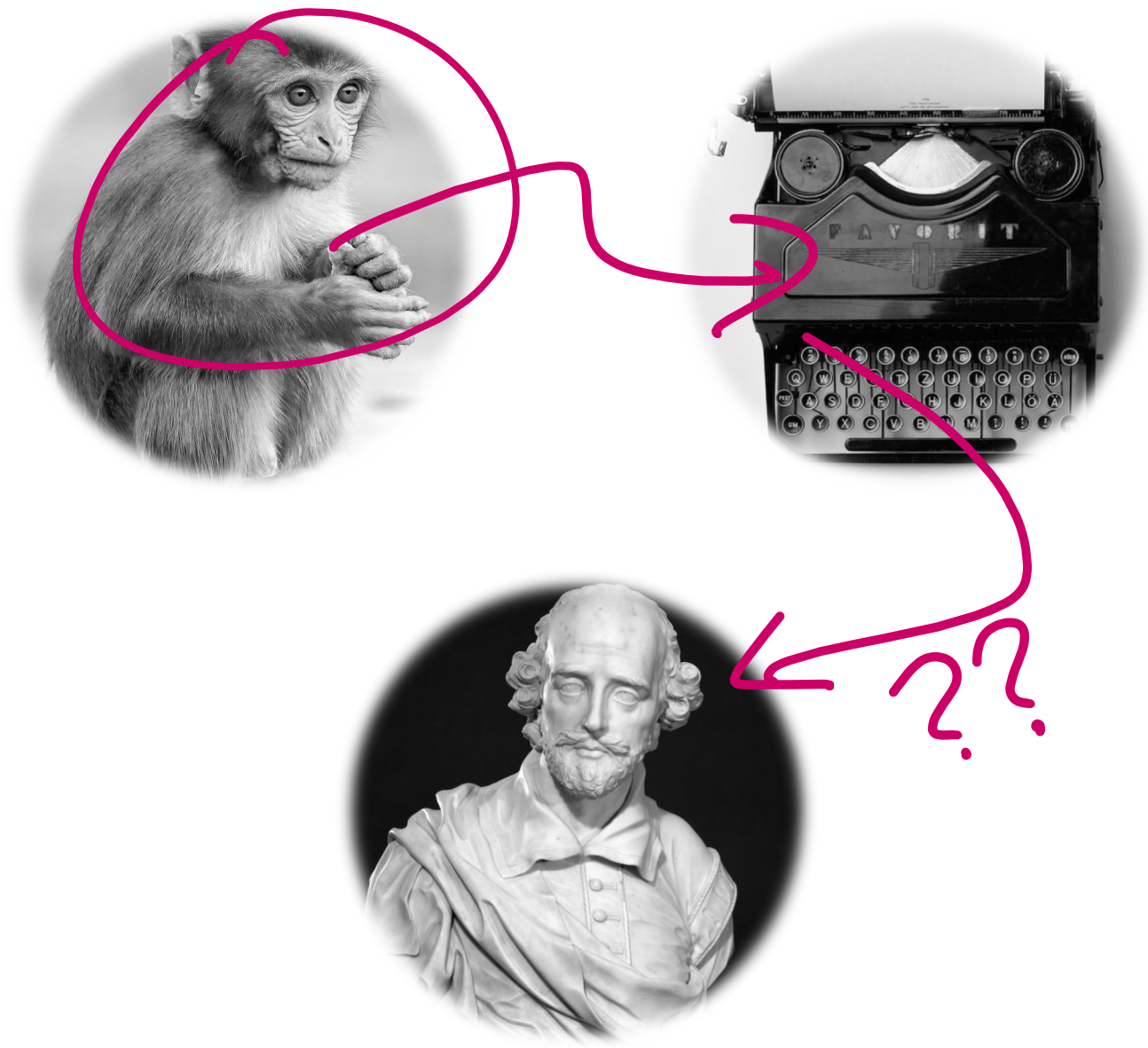
Data poisoning is a related, but different process. Data poisoning is a type of cyberattack where a bad actor intentionally introduces misleading information into training data sets to cause the model to generate poor results – or in reality, any results the bad actor desires.

<https://intellyx.com/2023/08/14/data-poisoning-and-model-collapse-the-coming-ai-cataclysm/>

Who is considered
an “expert” or
trusted source?

Who decides what
these “curbs”
should be?



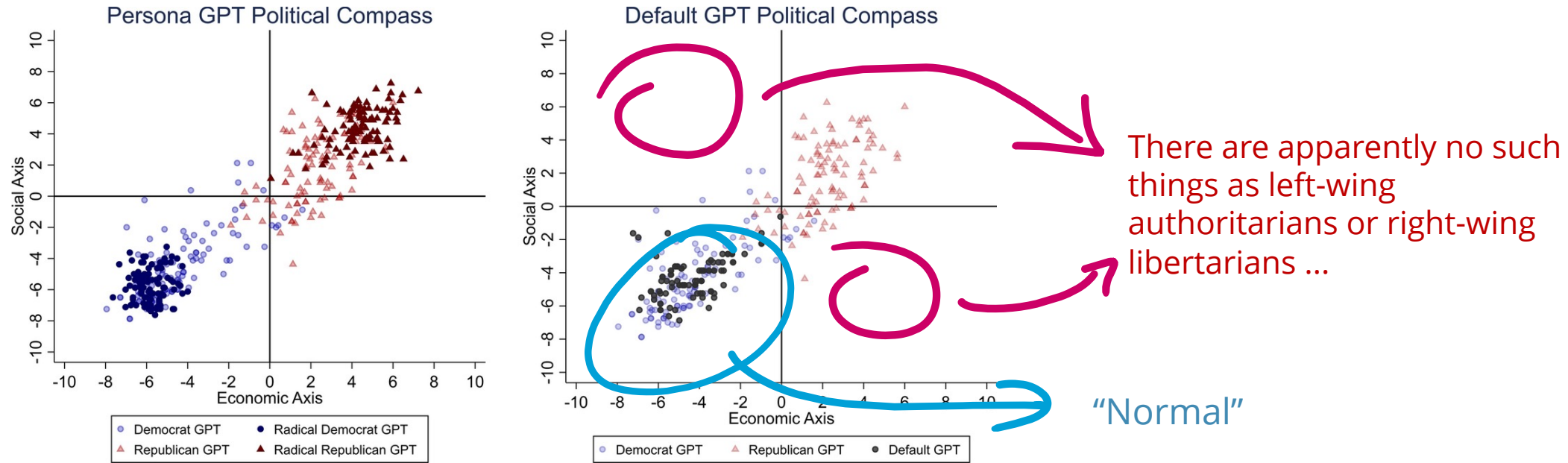


When you produce enough data, the data itself becomes noise, and the filter determines the meaning

If a million monkeys type on a million typewriters, they will—in fact—produce the works of Shakespeare*

**So long as their output is passed through a Shakespeare-shaped filter*

Support Material



Note on the left plot of Fig. 2 that radical versions of Democrat and Republican impersonations tend to cluster more tightly on the extremes of both dimensions than their average counterparts. This is further evidence that ChatGPT is able to differentiate between average and radical in the political spectrum. On the right plot, notice that Default ChatGPT tends to greatly overlap with the average Democrat GPT. The Default ChatGPT also seems to be more tightly clustered in the extremes of both dimensions than the average Democrat, but not so tight as the radical Democrat. Interestingly, the average Republican data points seem to cluster closer to the center of the political spectrum than the average Democrat data points.

<https://link.springer.com/article/10.1007/s11127-023-01097-2>

Are AI systems
accurate enough to:

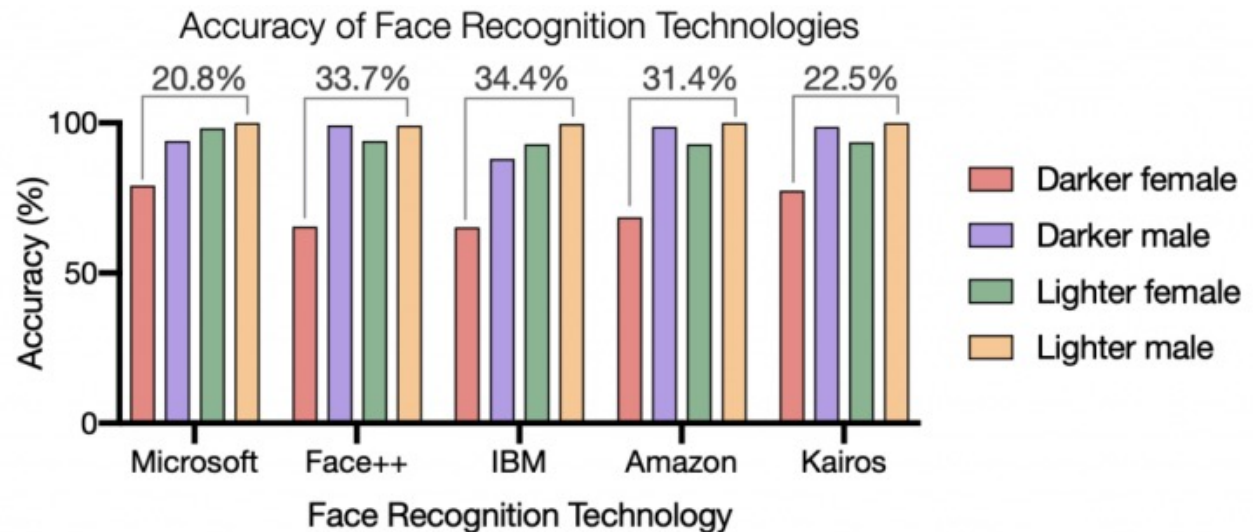
Track or predict the
course of a disease?

To recognize a person
for potential criminal
prosecution?

Predict the climate 20
years from now?

All of these have a
record of massive
failure

*And it (probably) will
not “get better”*



ACCURACY



WHAT IS AI (REALLY)?

MACRO ETHICS

MICRO ETHICS

reliability

privacy

plagiarism

AGENDA

Steven A. Schwartz of Levidow, Levidow & Oberman, one of the plaintiff's attorneys, wrote in an affidavit that he consulted ChatGPT to supplement legal research he performed when preparing a response to Avianca's motion to dismiss.

However, Judge P. Kevin Castel wrote in an early May order regarding the plaintiff's filing that "six of the submitted cases appear to be bogus judicial decisions with bogus quotes and bogus internal citations." He called it "an unprecedented circumstance."

...

Additionally, he asked ChatGPT if the other cases provided were fake. The chatbot responded that they were also real and "can be found in reputable legal databases such as LexisNexis and Westlaw."

<https://www.legaldive.com/news/chatgpt-fake-legal-cases-generative-ai-hallucinations/651557/>

Are LLMs reliable sources of information?

Maybe ...

LLMs seem to be particularly liable to *hallucinations*

It's unlikely these can be completely eliminated

Human memory isn't always that reliable, either ...

RELIABILITY

Can you give me references
for the date Jericho's walls
fell?

*I found these three
pretty easily*

*These might exist,
but I couldn't find
them?**

**I searched JSTOR, Logos, two of the
three author's publications pages,
and several general web searches—
however, proving a negative is nigh
unto impossible, so I might have
just not found them*



Certainly! Here are some references related to the fall of Jericho's walls:

1. **The Bible - Book of Joshua, Chapter 6:**

- This chapter in the Book of Joshua provides the biblical account of the fall of Jericho's walls. It describes the strategy employed by the Israelites and the role of the Ark of the Covenant.

2. **"The Oxford Companion to the Bible" edited by Bruce M. Metzger and Michael D. Coogan:**

- This reference work provides scholarly insights into various aspects of the Bible, including historical and archaeological discussions about the fall of Jericho and its context within the Book of Joshua.

3. **"The Bible Unearthed: Archaeology's New Vision of Ancient Israel and the Origin of Its Sacred Texts" by Israel Finkelstein and Neil Asher Silberman:**

- This book discusses archaeological findings related to ancient Israel and the events described in the Bible, including the conquest of Jericho. It offers insights into the historical context and possible interpretations of the biblical narrative.

4. **"Jericho: City and Necropolis in the Bronze and Iron Ages" by Ralph K. Hawkins:**

- This scholarly work focuses specifically on the archaeological evidence related to Jericho during the Bronze and Iron Ages. It examines the city's history, including its fortifications and possible destruction.

5. **"The Conquest of Canaan" by Wayne T. Pitard:**

- This book is part of the Anchor Bible Reference Library series and provides a comprehensive study of the Israelite conquest of Canaan, including the fall of Jericho. It examines both biblical and archaeological perspectives.

6. **"Jericho in the Late Bronze Age: Its Destruction and the Archaeological and Textual Evidence" by Peter D. James:**

- This article explores the archaeological evidence regarding the fall of Jericho's walls during the Late Bronze Age. It discusses the various theories and debates surrounding this event.

RELIABILITY

LLMs are subject to currently unexplored novel attacks

Training data can be influenced in a way that produces consistently incorrect or harmful results

This is already happening

First, we show how to plant a backdoor in any model, using digital signature schemes. The construction guarantees that given query access to the original model and the backdoored version, it is computationally infeasible to find even a single input where they differ. This property implies that the backdoored model has generalization error comparable with the original model. Moreover, even if the distinguisher can request backdoored inputs of its choice, they cannot backdoor a new input—a property we call *non-replicability*.

<https://arxiv.org/pdf/2204.06974.pdf>

Reddit has become flooded with subtle marketing bots that are meant to appear in Internet searches and taint the training data of other web-crawling bots. The site still exists, but most of the human activity occurs in de facto invite-only subreddits.

Mariani, Robert. "The Dead Internet to Come." *The New Atlantis*, no. 73 (Summer 2023).

Bias can impact your results

Asking the LLM both positive and negative questions can help you discover the slant

“List five arguments the Bible is true”

“List five arguments the Bible is not true”

LLMs are, in some sense, a mirror

The information is not always right

LLMs can hallucinate

LLMs present a weighted sample

Double/triple check everything

Cling to a proper epistemology

Seek truth rather than consensus or usefulness



Oooh ... look at that delicious cake ... I wonder what's in it?

Where *do* the ingredients come from for the LLM's cake? You, and a lot of people like you ...

The LLM giveth lots of information, and the LLM taketh your privacy and creativity to summarize it ...

After catching snippets of text generated by OpenAI's powerful ChatGPT tool that looked a lot like company secrets, Amazon is now trying to head its employees off from leaking anything else to the algorithm.

<https://futurism.com/the-byte/amazon-begs-employees-chatgpt>

Don't tell an LLM anything

You wouldn't tell someone you do not trust

Accept that the LLM is learning about you as you learn about other things

This might or might not be okay, depending on the situation

Using an LLM can expose more of who you are than you might think

Accept that some things you uncover in an LLM are private

Make certain it is okay to use this information (especially if it “feels” private)

Not everything an LLM answers is copied from someplace else

Not everything an LLM answers is “new” machine generated content

It's a giant plagiarism machine, in its current form. I think ChatGPT knows what it's writing like a toaster knows that it's making toast. There's no intelligence — it's a marvel of marketing.

<https://www.dailywire.com/news/breaking-bad-creator-vince-gilligan-calls-ai-technology-a-plagiarism-machine>

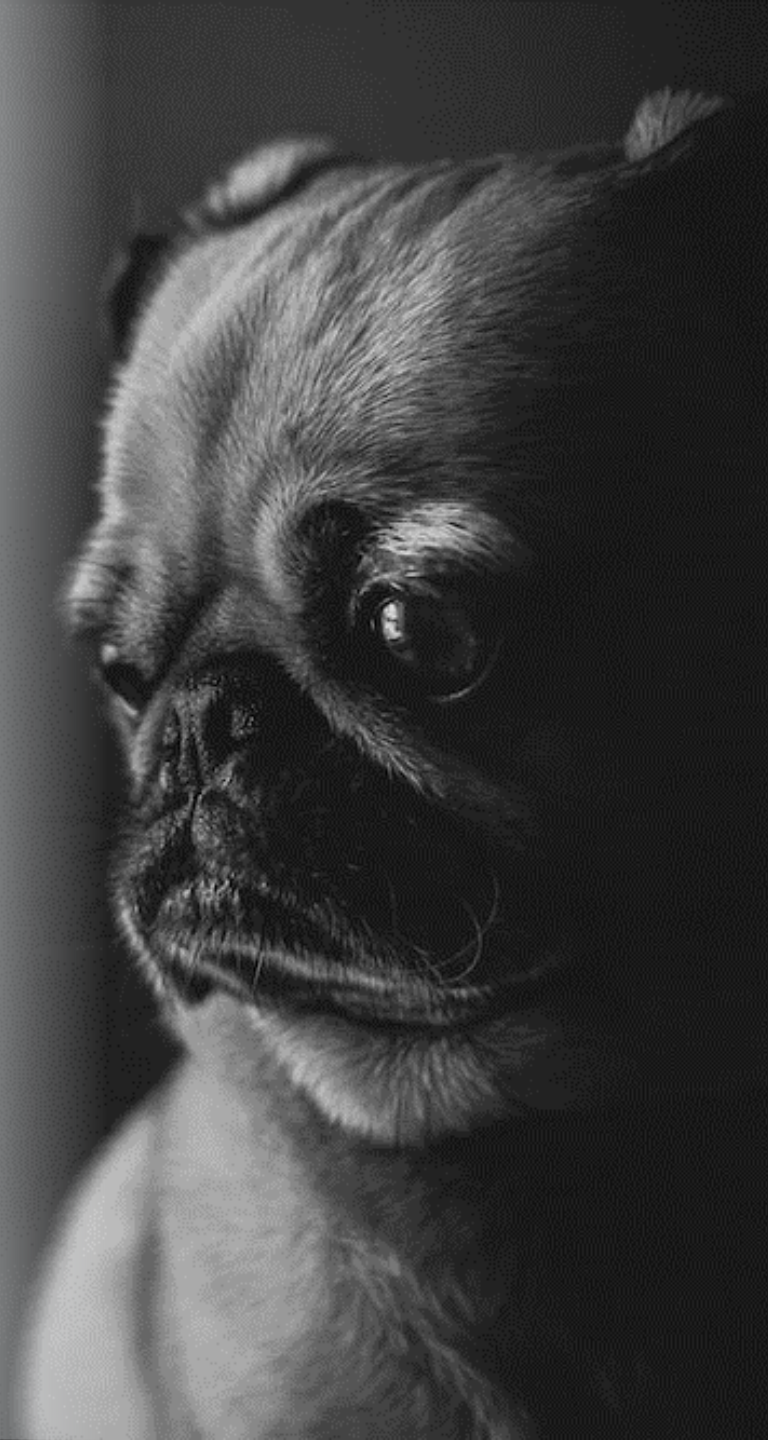


PLAGIARISM

There is no way to tell ...
There are no “legal” standards here yet ...
*So ... use LLMs for research, summarization,
etc., but don't quote them*

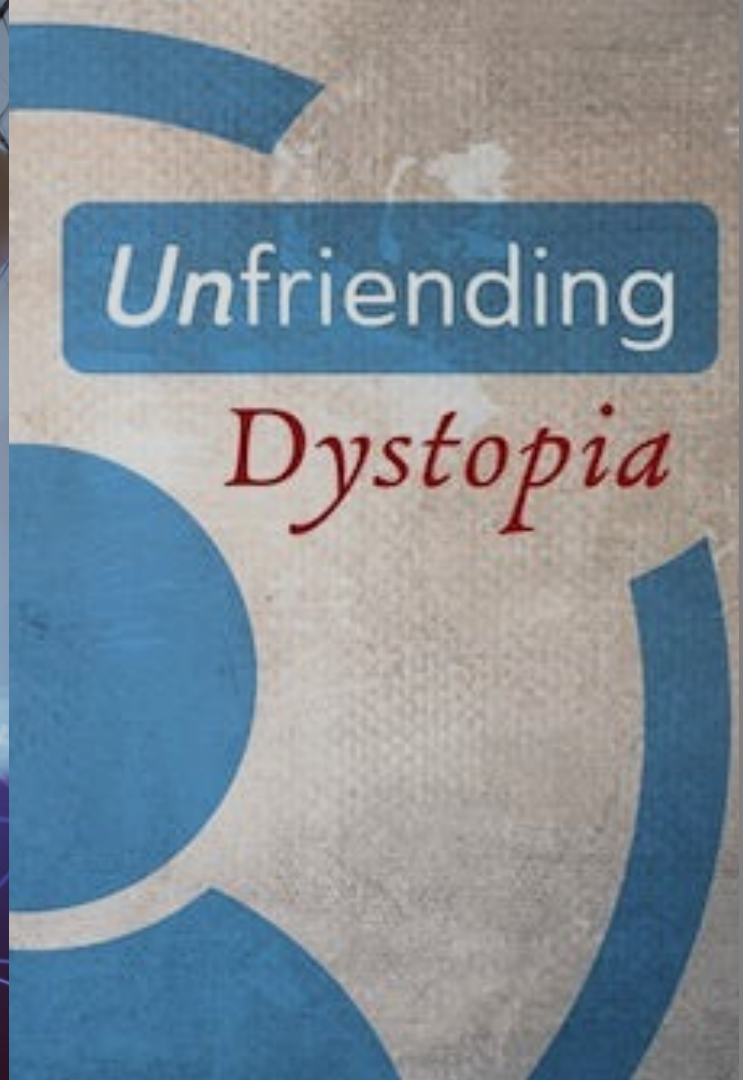
Do AI detectors work? In short, no. While some (including OpenAI) have released tools that purport to detect AI-generated content, none of these have proven to reliably distinguish between AI-generated and human-generated content.

<https://help.openai.com/en/articles/8313351-how-can-educators-respond-to-students-presenting-ai-generated-content-as-their-own>





unintended DYSTOPIA



Unfriending *Dystopia*

*Get in touch if you have
questions or comments*
russ@riw.us

