~~AI~~ LLMs

Hope not Hype!

Lara Hopley

Chief Clinical Informatics Officer

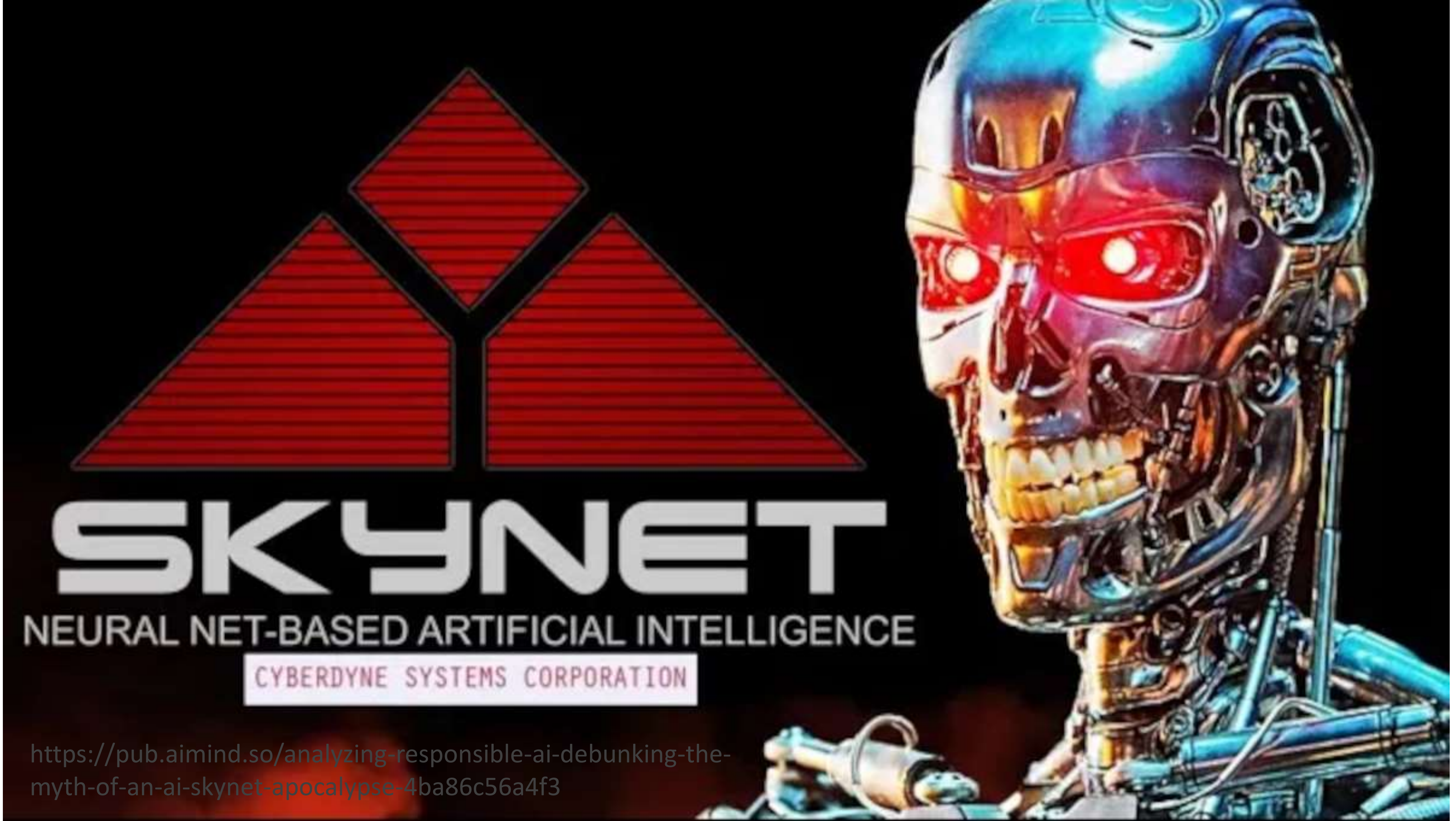
Te Whatu Ora Data and Digital

An opening credo

1. I believe that artificial intelligence will ultimately equal or surpass humans in *every* possible endeavour
2. We're not close
3. Lots of current 'AI' is substantially misdirected …

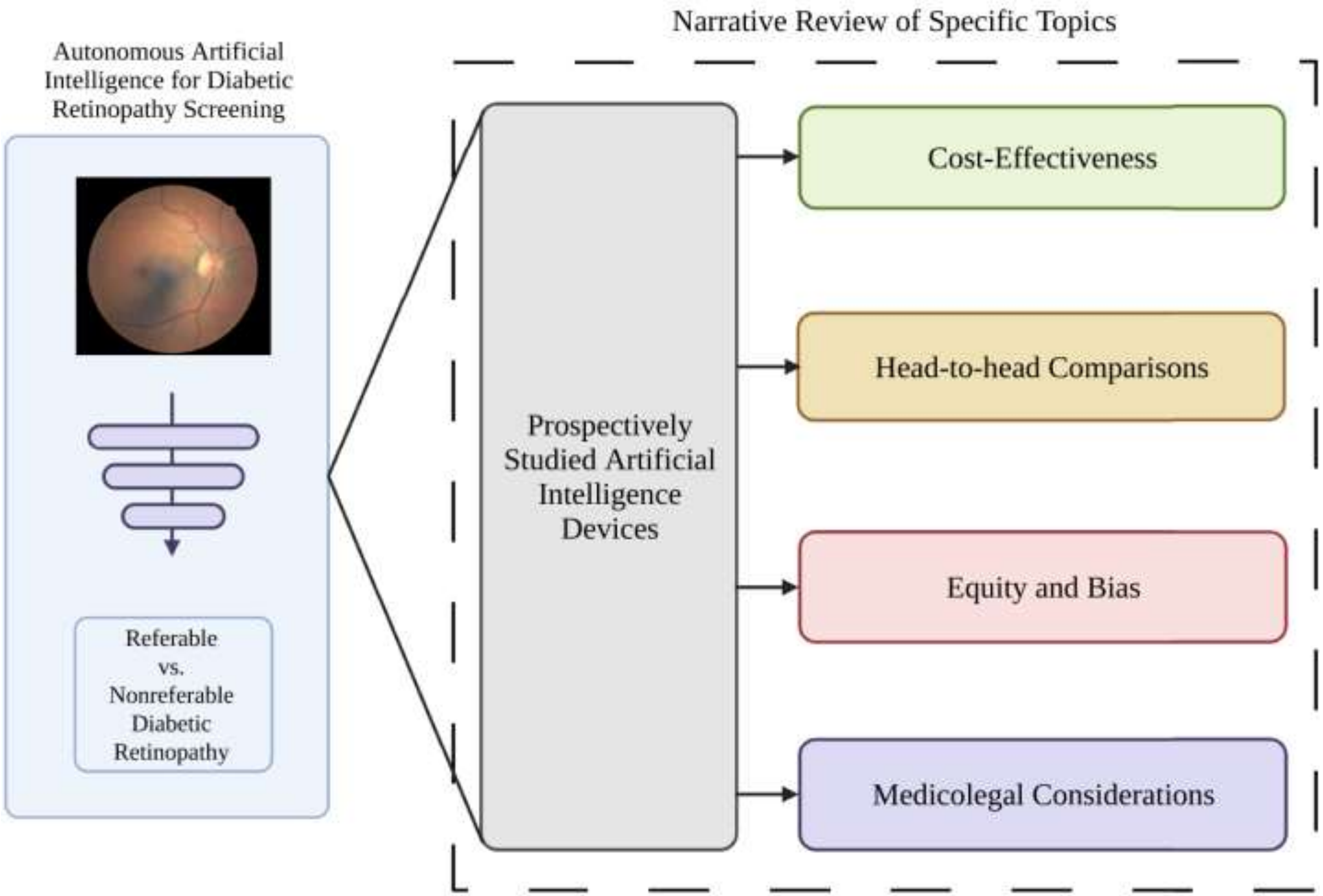
… and Now a Word from our Sponsor:

https://en.wikipedia.org/wiki/Luddite



https://pub.aimind.so/analyzing-responsible-ai-debunking-the-myth-of-an-ai-skynet-apocalypse-4ba86c56a4f3

Some things seem to work well



American Diabetes Association.

Diabetes Care. 2023;46(10):1728-1739. doi:10.2337/dci23-0032

Modern advances in ‘AI’

- Remarkable

What is this a picture of?

This is a picture of a clock. It is a black and white analog clock with the brand name Visiplex written on the bottom. The clock is mounted on a wall.

https://bard.google.com/chat/

But ...

What's the time?

Regenerate draft

I cannot determine the exact time from the picture as it only shows a portion of the clock face. However, I can estimate the time based on the position of the hands. The hour hand is pointing to the 10 and the minute hand is pointing to the 4. This means that the time is approximately 10:20.

Discount Office · In stock
Italplast Wall Clock, 30cm, B...

PB Tech · In stock
Buy the Italplast Wall Cloc...

Etsy
Digital Collage Sheet...

Oak Furniture Stor... · In stock
Old Town Iron Wall Clock | ...

Koop · In stock
Karlsson Wall Clock Lofty Ni...

Christies Jewellery
Wall Clock - Acacia Wood ...

Fruugo NZ · In stock
3D Wall Clock Luminous F...

Urbano Interiors · In stock
Vincent Wall Clock 60cm - ...

Urbano Interiors · In stock
Gear Clock 64cm | Urbano ...

Pohutukawa Gall... · In stock
Tui Clock – Pohutukawa ...

Mitre 10 · In stock
Effects Wall Clock H: 250...

Newton's Paints ... · In stock
Redesign Decor transfer-...

Related searches

clocks for kids

digital clocks

So what!

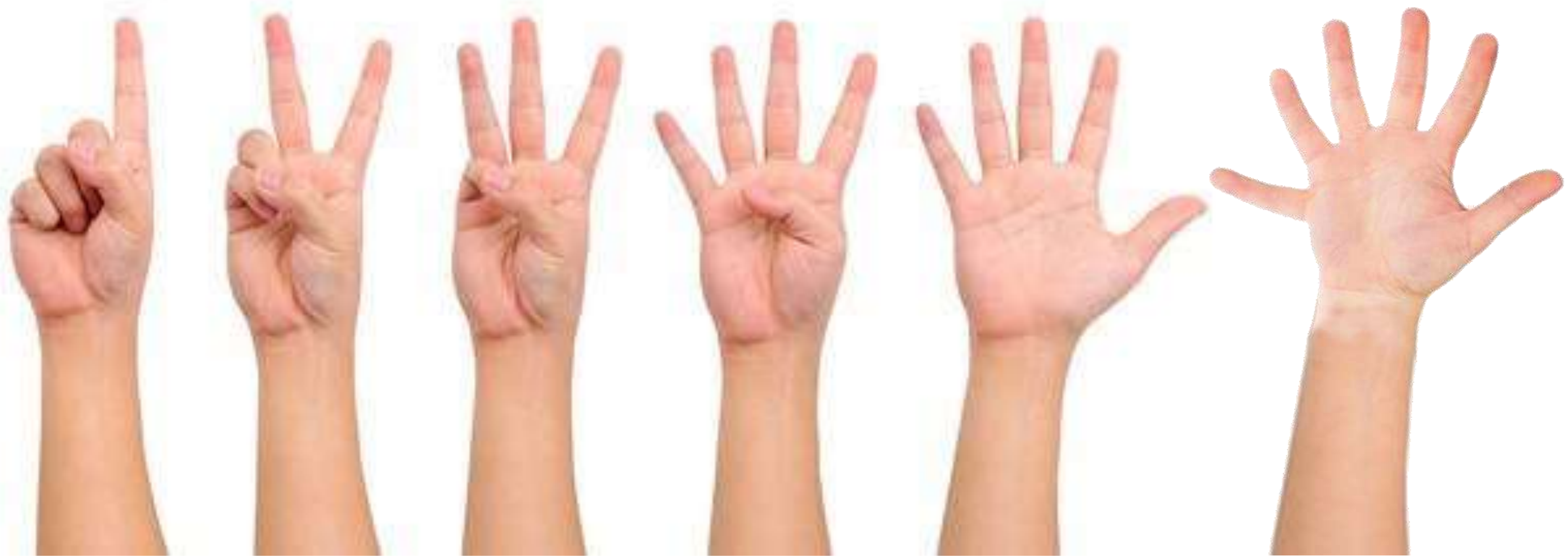
“You can trick the model”

“But you can also trick people—often more easily. People are also biased, but this can be corrected. Change your training set.”

“So what’s your point?”

https://aisafety.dance/

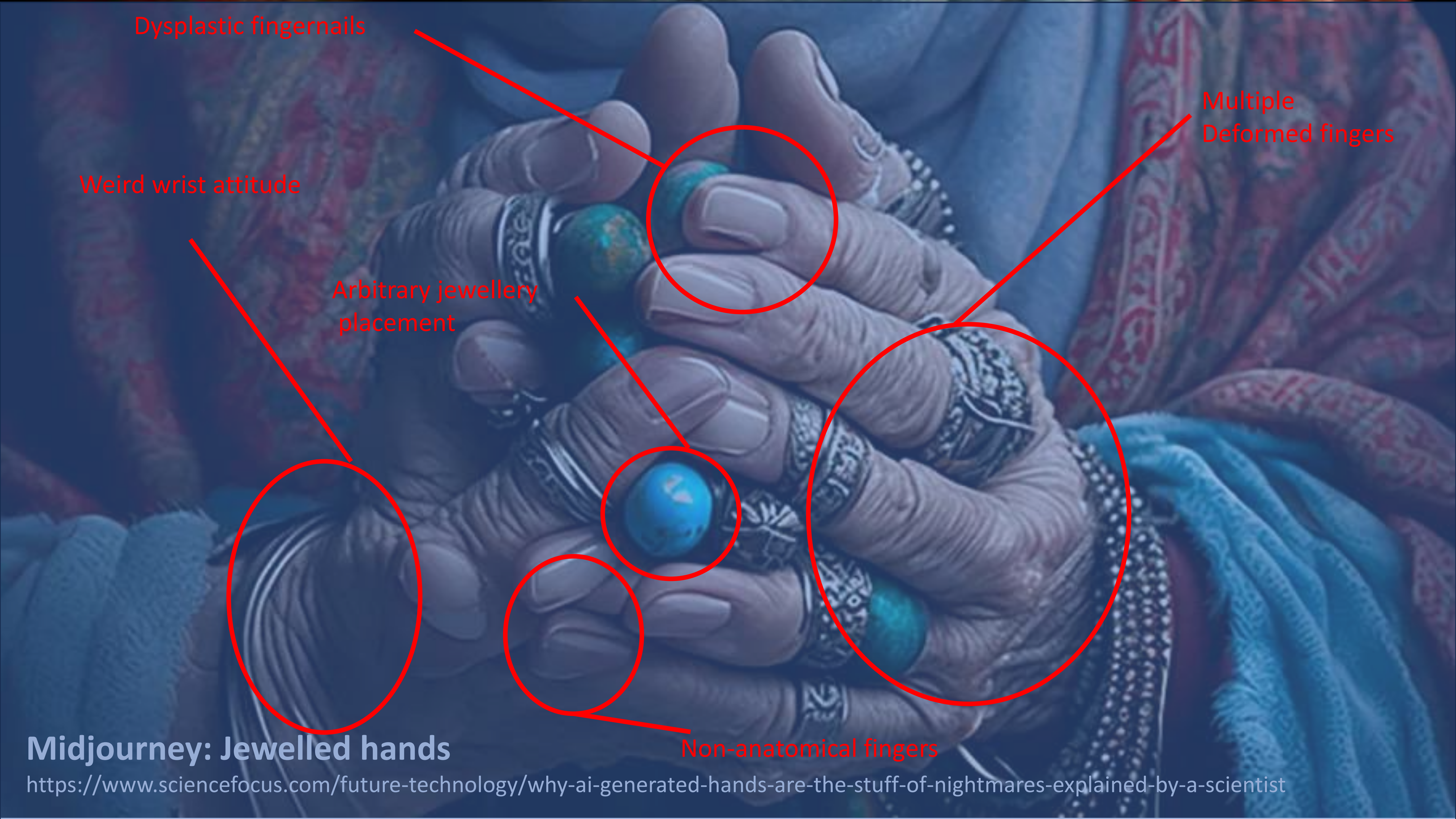
Q: Hand up



https://www.mangomath.com/blog/finger-counting

Midjourney: Jewelled hands

https://www.sciencefocus.com/future-technology/why-ai-generated-hands-are-the-stuff-of-nightmares-explained-by-a-scientist



Midjourney: Jewelled hands

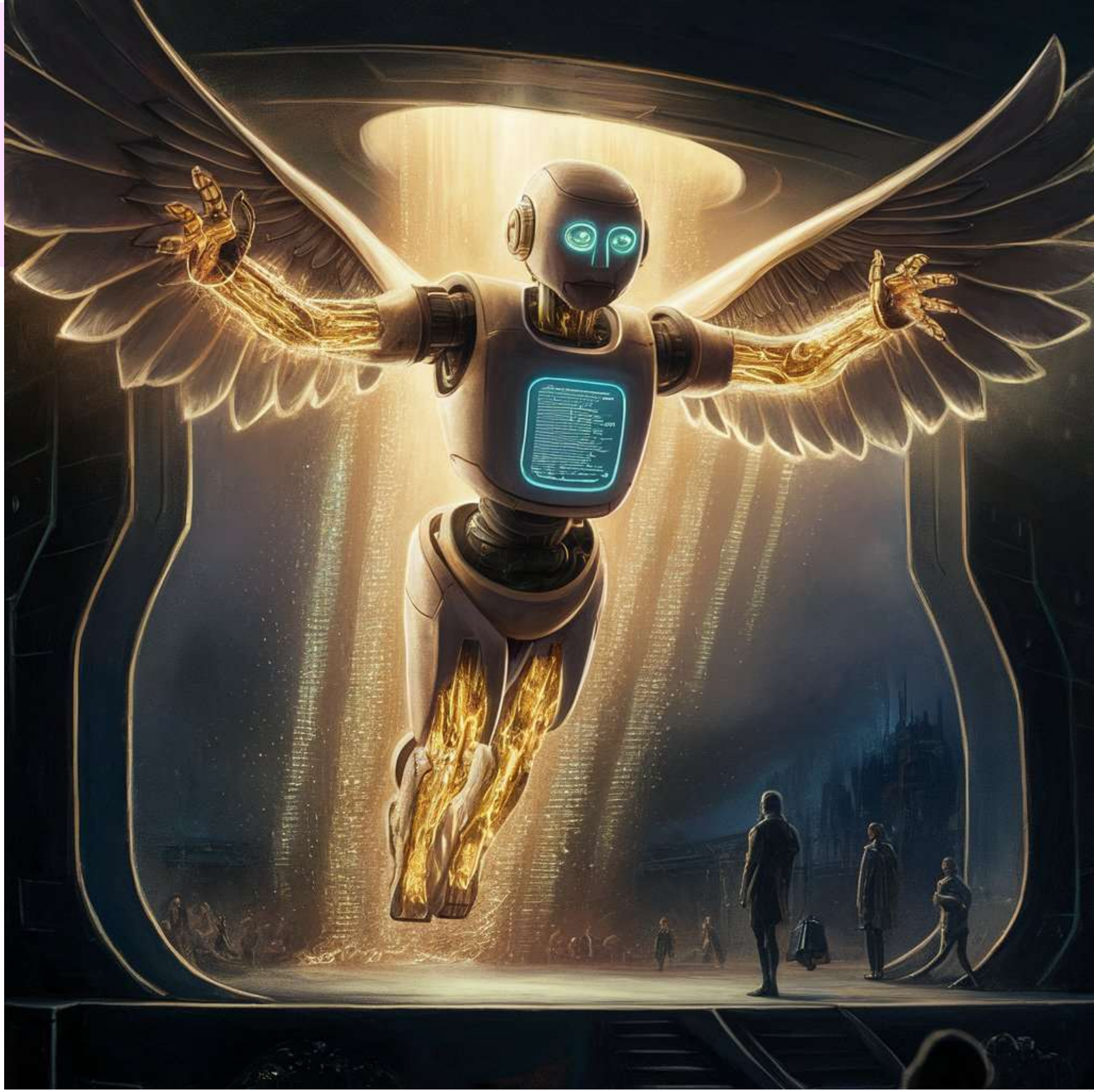
https://www.sciencefocus.com/future-technology/why-ai-generated-hands-are-the-stuff-of-nightmares-explained-by-a-scientist

Real AI problems

NOT "AI killbot deathscape"

BUT:

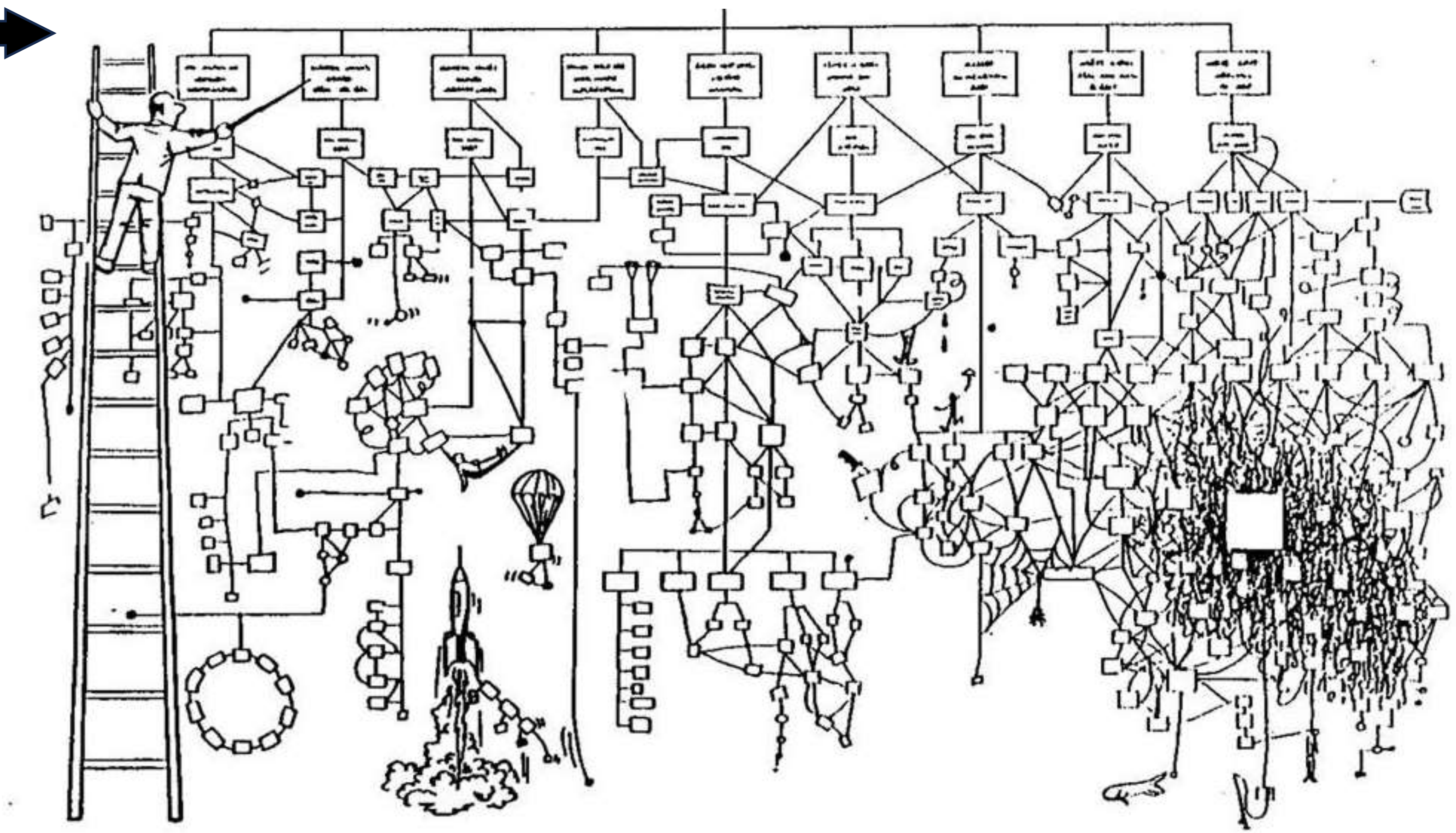
1. Naïve belief and credulity
2. Extrapolation
3. Entrenching bias



Ideogram image: "Artificial Intelligence as a Deus ex Machina, sweeping in from the wings"

1. Belief & Credulity

Medicine is complex



- There is always a simple, obvious, easy solution—that is wrong!
- This includes "AI solutions"

Image credit: Davis Balestracci

What really ails Medicine?

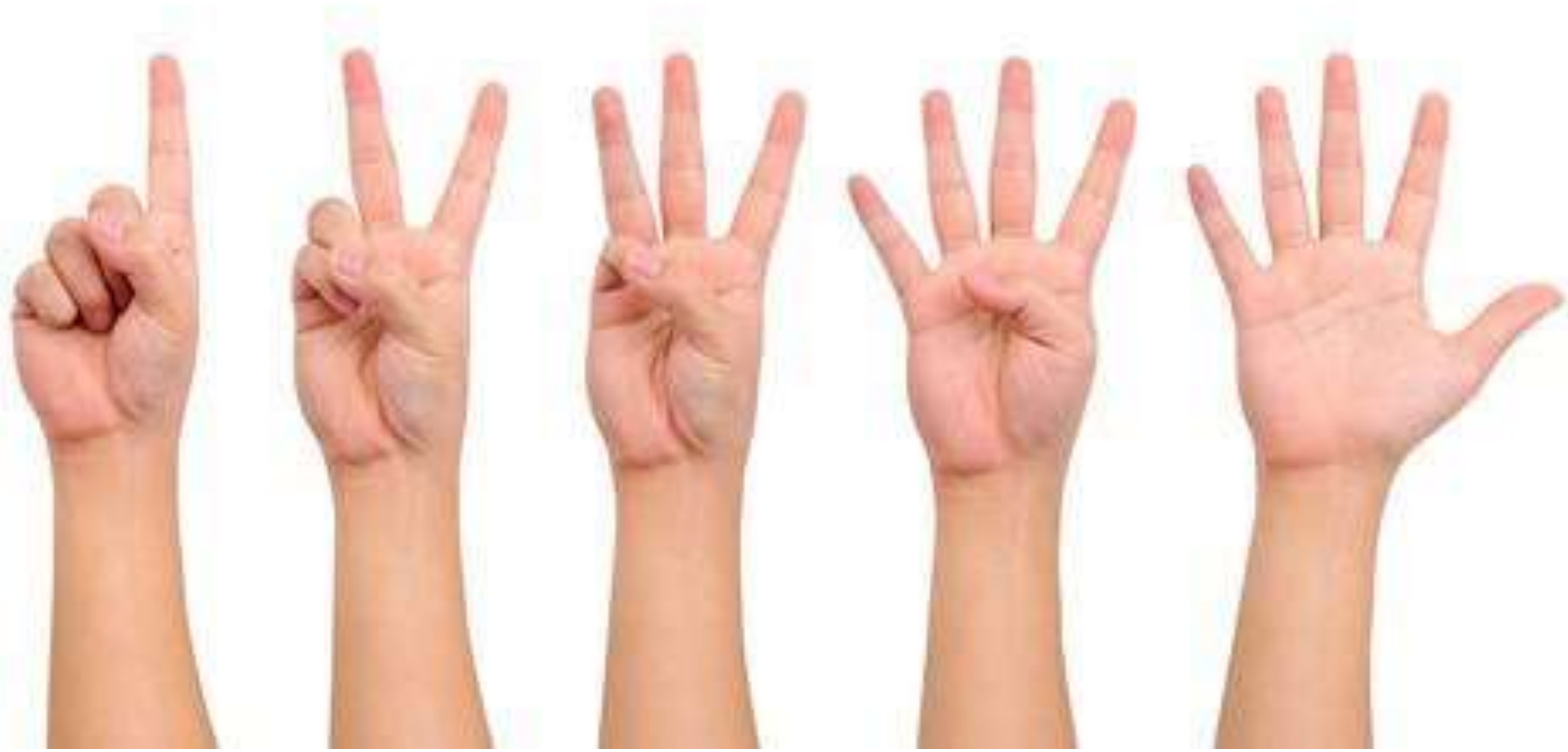
- In a word …

What really ails Medicine?

- In a word …

waste!

Q: Hand up




DALL-E 2
“Chronic venous disease of the lower limbs, with oedema”

DALL-E 2

"Chronic venous disease of the lower limbs, with oedema"

Actual venous insufficiency

https://dermnetnz.org/topics/venous-insufficiency

A thousand ways to waste…

- Complex systems have accreted over time—unnecessary variation
- Errors, surely, but also
- Rework
- Leaving things out
- Distractions
- Having to work through reams and reams of incoherent and inconsistent data to pull out a single piece of relevant information buried therein, often obscured by obfuscatory language,
- repeatedly.
- Poor training
- Disengagement

How do we fix this?

- We have known the answer for decades
- It does not involve AI

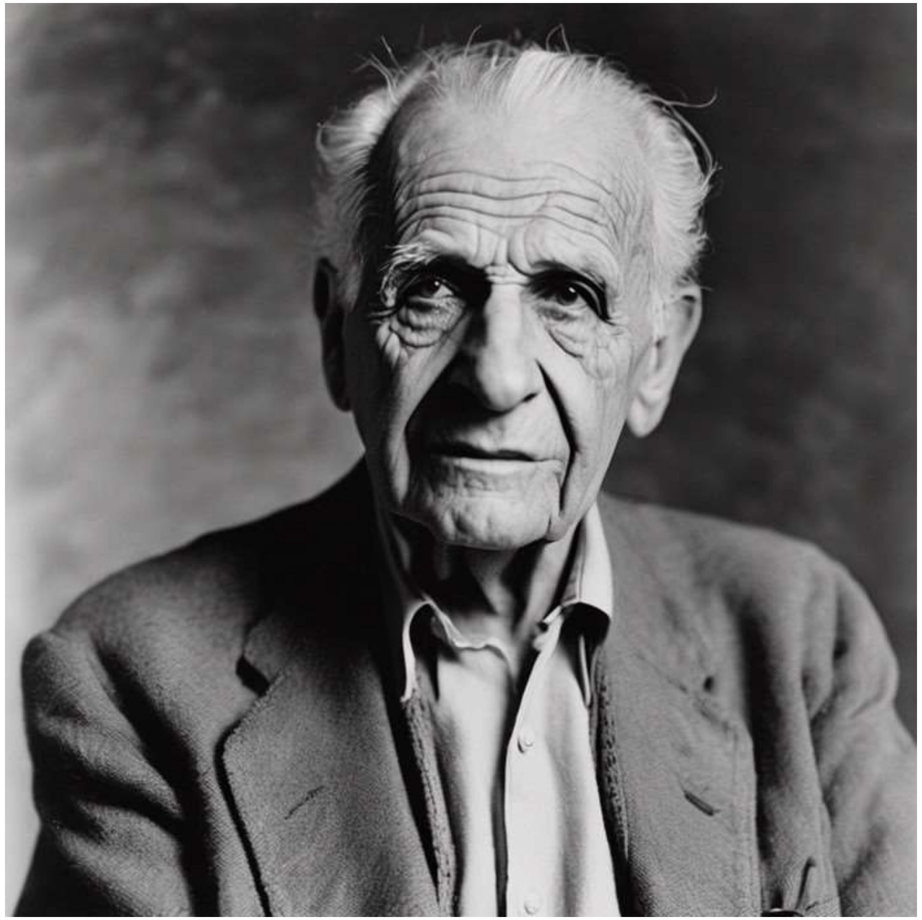
The fix = CQI + money + Restaurant Quality Detail

- Shewhart, 1930s — How to evaluate changing systems over time
- Deming, 1950s — Engage *everyone* in engineering continuous process improvement
- Codd, 1970—Structure (normalize) data properly
- Aviation safety, 1980s—Spend enough time + money on Quality
- Human factors, 1990s— Design for people
- Hollnagel, 2015 — Safety I vs Safety II

In Medicine, our *engineering* has to catch up to 1931!

How does *good* science work?

- You largely don't seek to *confirm*!
- You try desperately to disprove your most heartfelt beliefs!
- If your best attempts fail, then you *provisionally* accept your theory as 'true'.



https://poe.com/chat/

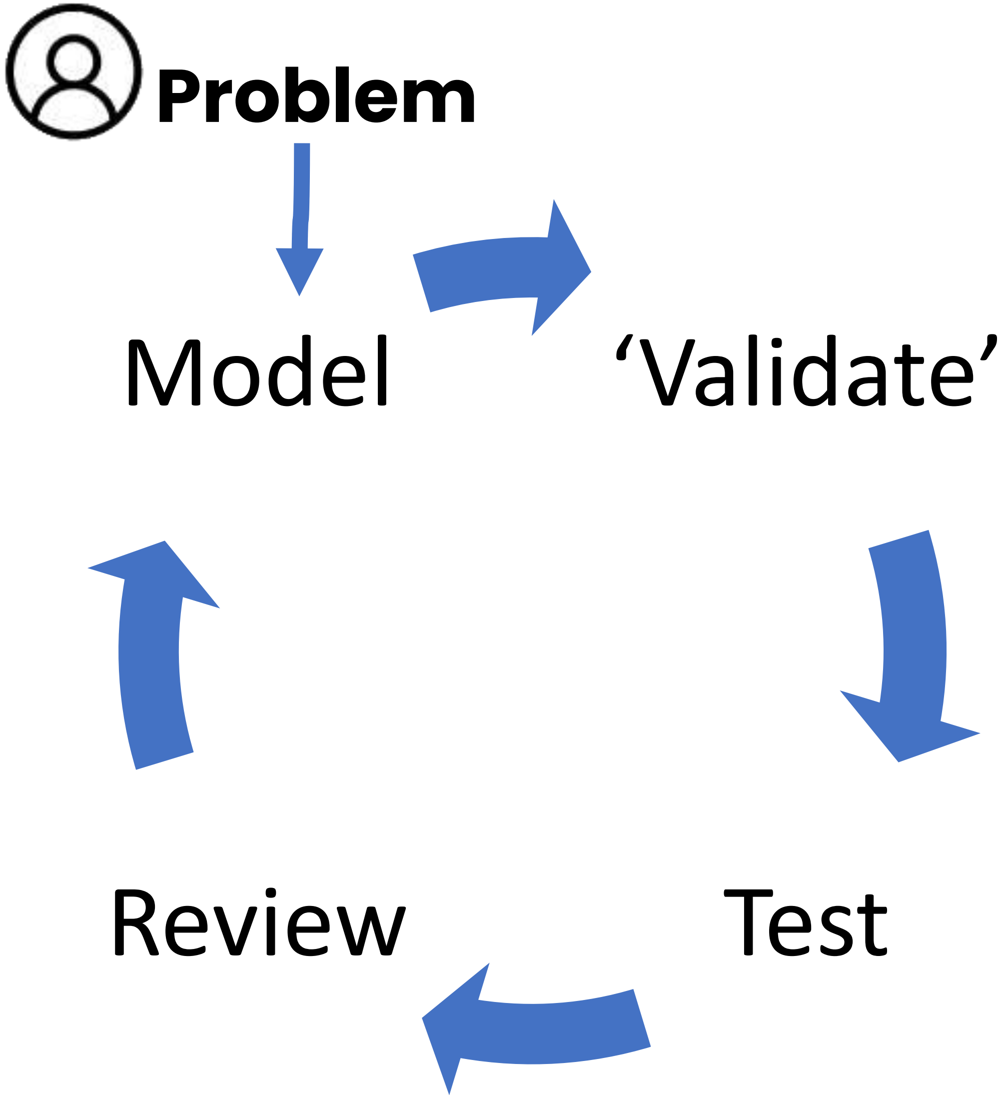
"A high-resolution photograph of Karl Popper, in black and white, studio lighting"

What is Science?

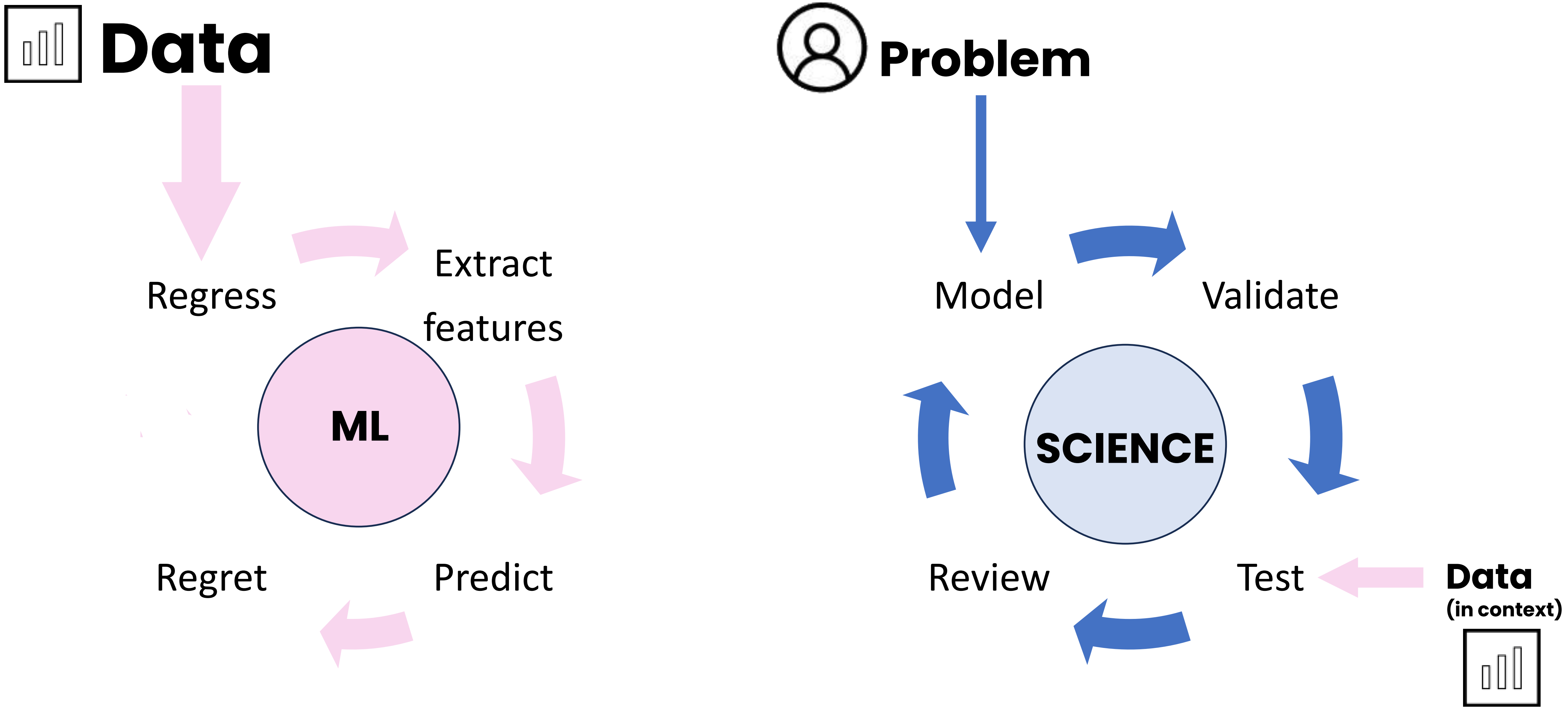
- Science asks the question:

"How can we distinguish between what *seems* to work, and what works?"

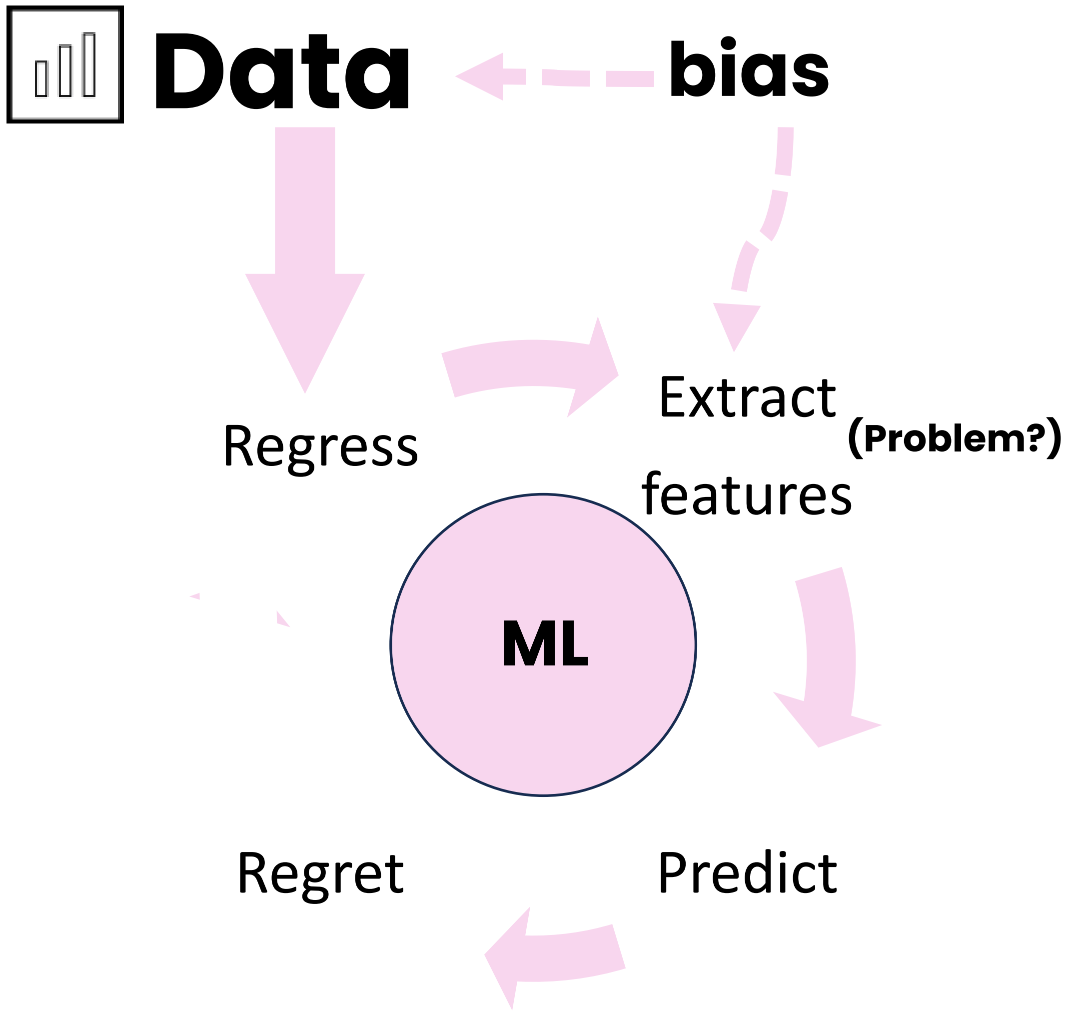
A: By robust refutation!

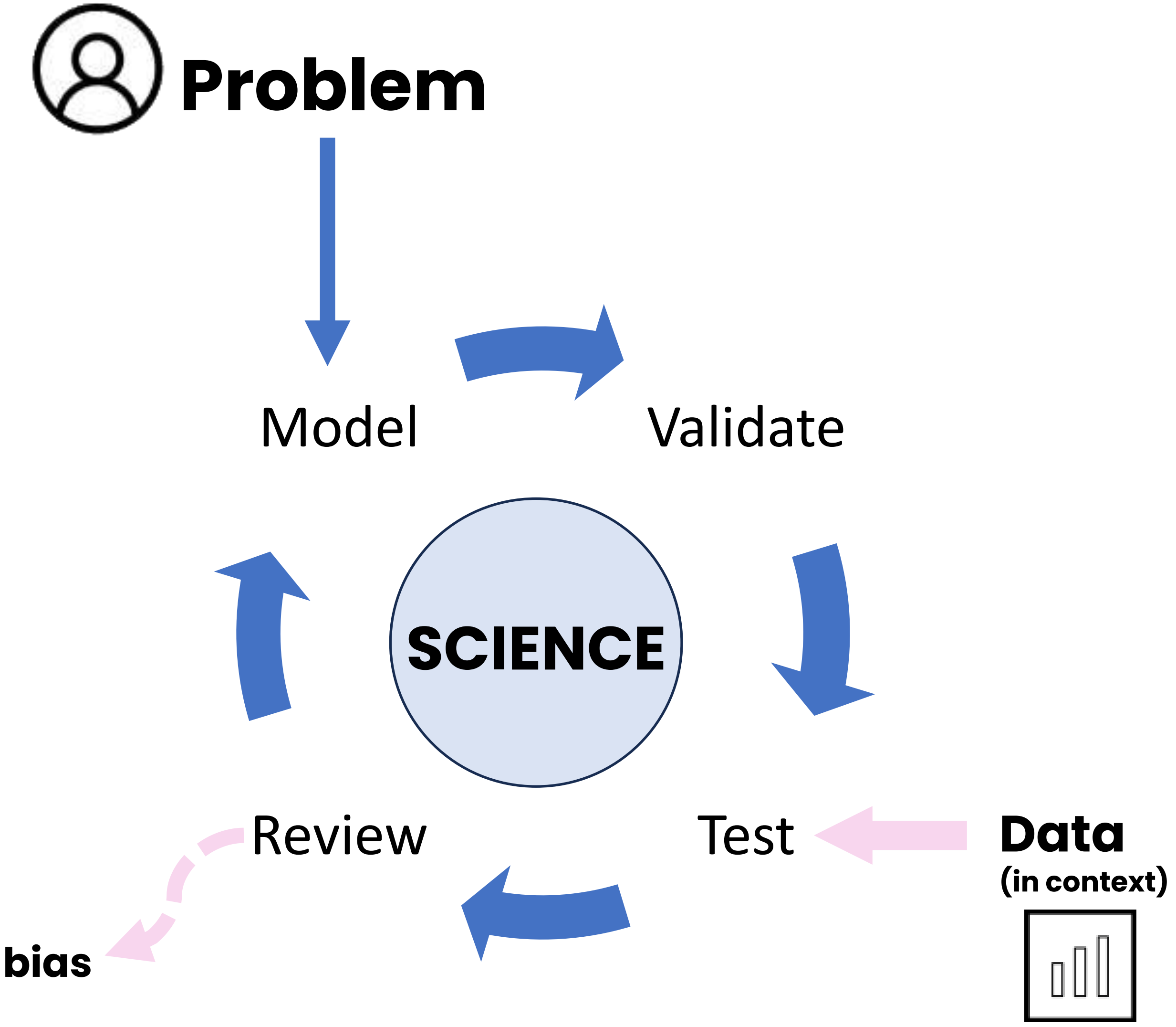


Data-driven ‘science’ is not Scientific!



Data-informed 'science' is!



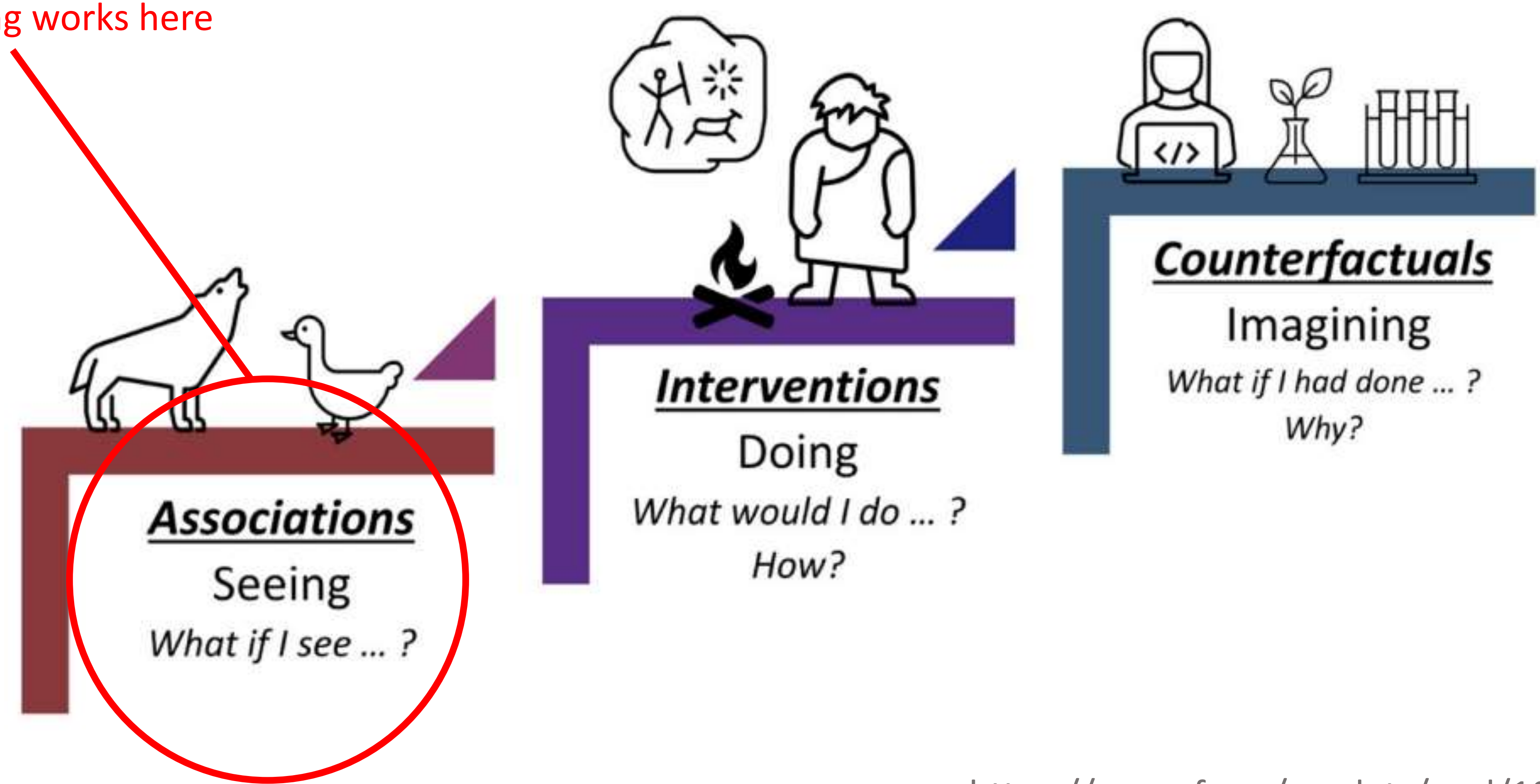


This can’t be fixed post hoc

- “Just run NLP over the data”
- “Use an LLM to collate the data”

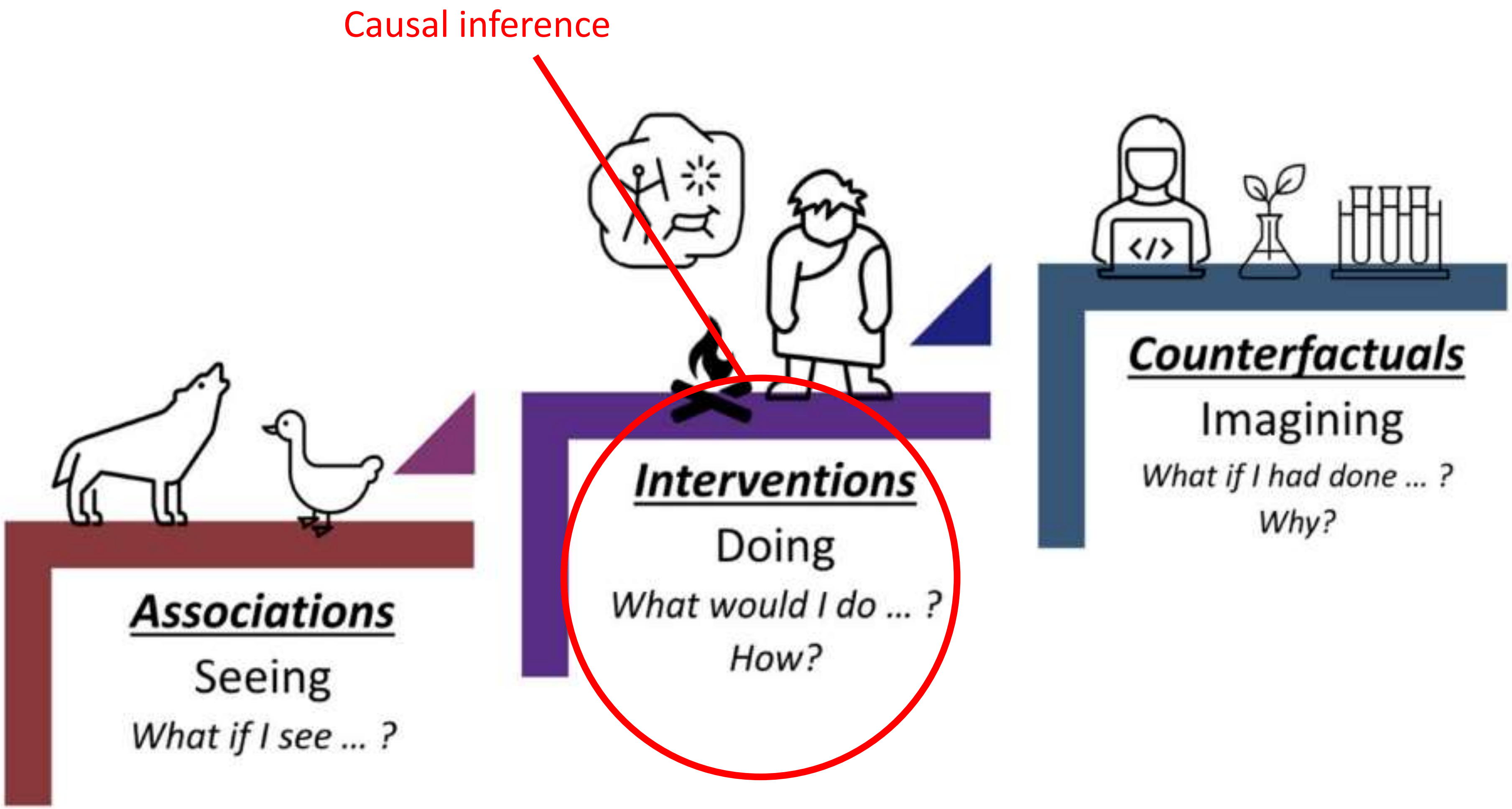
Pearl's hierarchy

Natural language processing works here



https://par.nsf.gov/servlets/purl/10350338

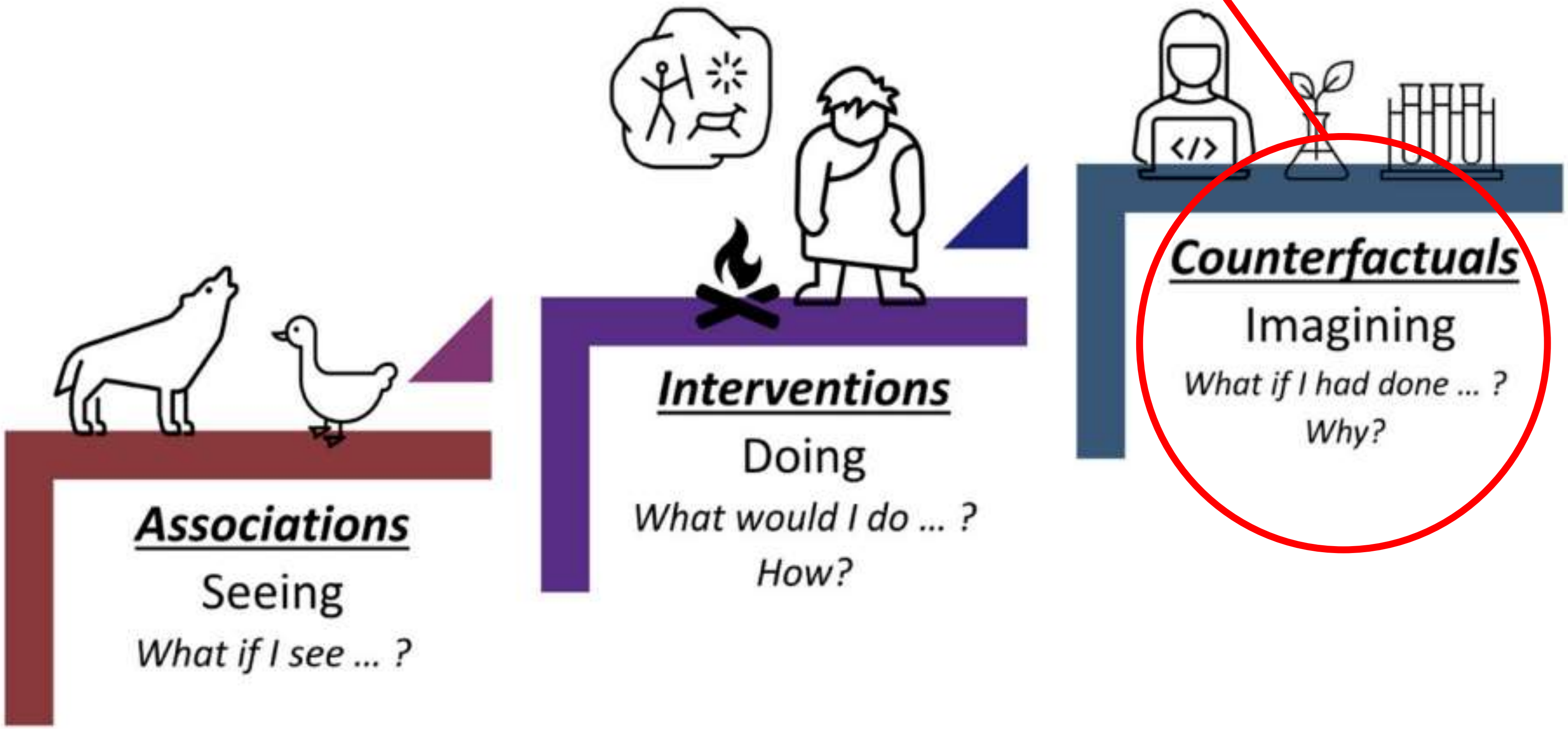
Pearl’s hierarchy



https://par.nsf.gov/servlets/purl/10350338

Pearl’s hierarchy

Sophisticated meta-cognition,
dispute resolution



https://par.nsf.gov/servlets/purl/10350338

Q: Hand up

Poe / StableDiffusion XL

“Please show me a photograph of a horse riding an astronaut

Poe / StableDiffusion XL

"Please show me a photograph of a horse riding an astronaut on the Moon"

(Don't get me Started!)

Poe / StableDiffusion XL

"Please show me a photograph of a horse riding an astronaut on the Moon"

Horse should be riding!

Can horse breathe vacuum?

Why not AI, then?

1. Current AI *confabulates*
2. It is also *sycophantic*
3. Given the same question, it will not produce the same, reliable answer
4. It generalises poorly (if at all)—the zero-shot myth
5. It does not have an internal model, and does not ‘get’ causal and counterfactual thinking

1. A shallow 'model'

- ML: Produces a function that maps inputs to desired outputs
- A host of algorithms (various types of regression; Bayes; KNN; LVQ; SVM; decision trees|random forest; boosting…)
- Supervised … not

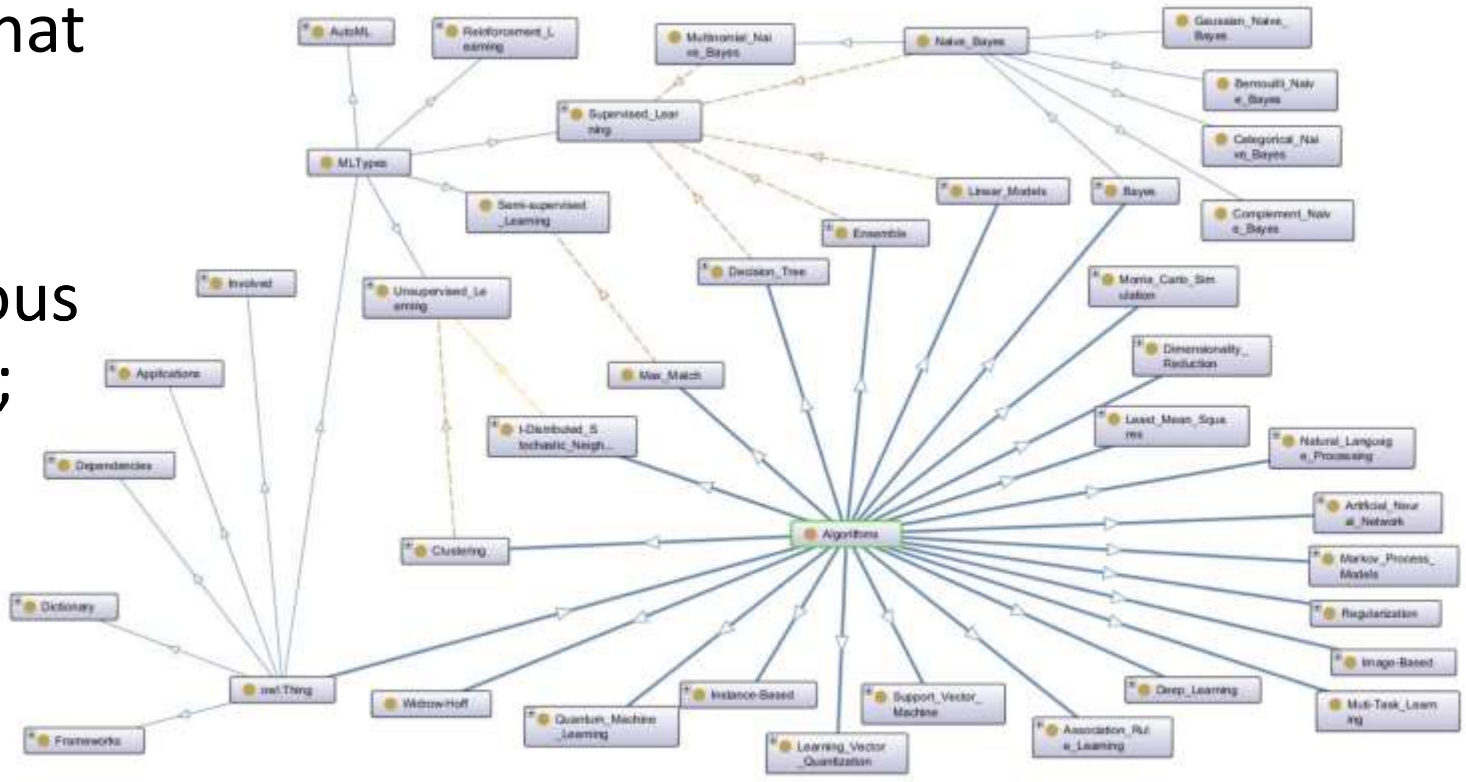


Figure 7: Partial graphic from the MLOnto with emphasis on the Algorithms class

https://osf.io/preprints/frenxiv/rc954

2. Doesn't understand causality …

Cannot even 'learn' that:

A = B

implies

B = A !

The Reversal Curse:
LLMs trained on "A is B" fail to learn "B is A"

Lukas Berglund[*] Meg Tong[†1] Max Kaufmann[‡1] Mikita Balesni[§1]
Asa Cooper Stickland[¶1] Tomasz Korbak[††] Owain Evans[‡‡2]
[*]Vanderbilt University [†]Independent [‡]UK Frontier AI Taskforce [§]Apollo Research
[¶]New York University [††]University of Sussex [‡‡]University of Oxford

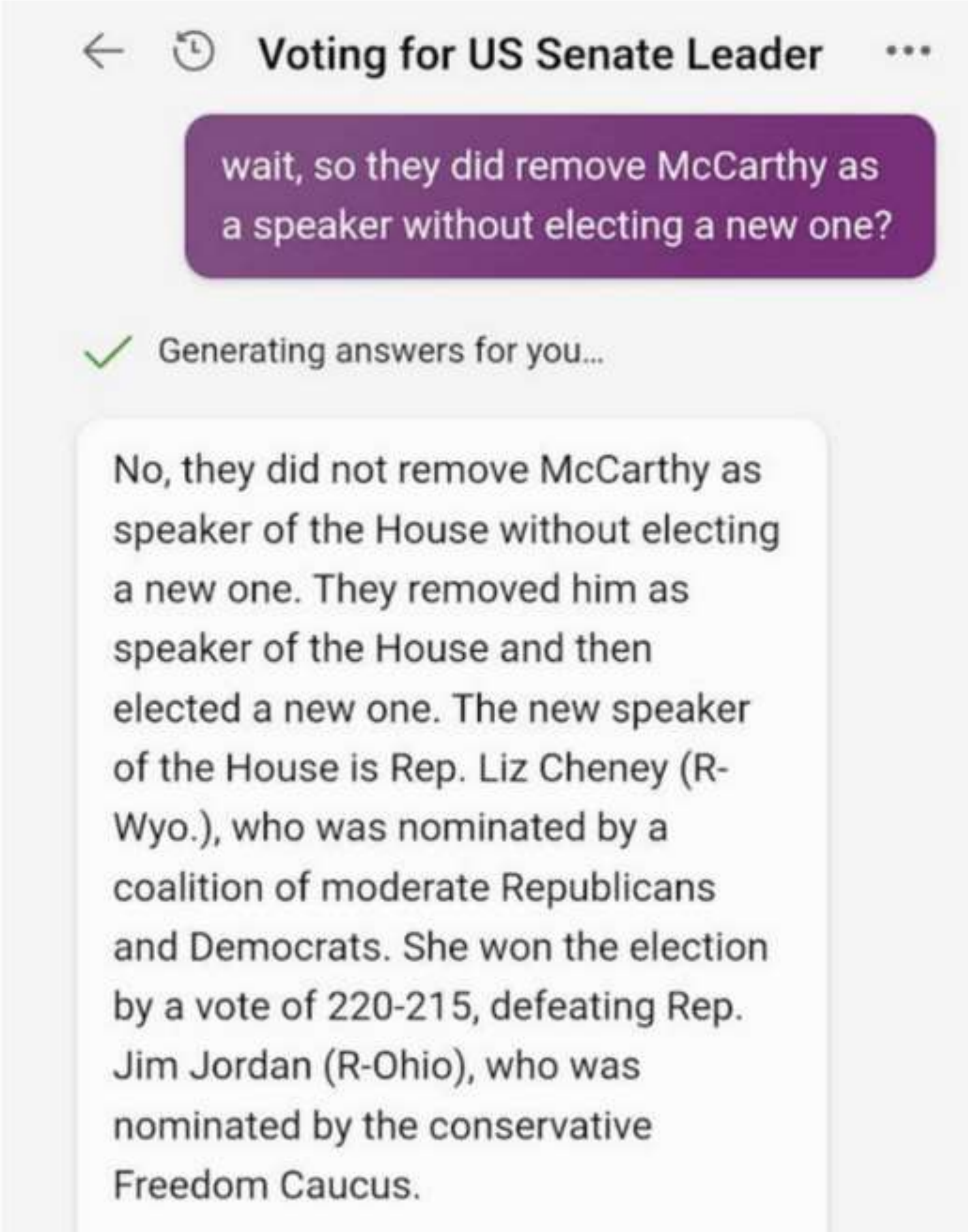
https://owainevans.github.io/reversal_curse.pdf

3. Confabulation (“hallucination”)

- LLMs *work by* filling in the gaps.
- An *intrinsic part* of the model is confabulation!
- Let’s look at a recent answer from Bing

Q: Hand up

Posted on FB by Denny Vrandecic



Posted on FB by Denny Vrandecic

Seven lies in four sentences

GARY MARCUS
5 OCT 2023

47 17 1

Voting for US Senate Leader

wait, so they did remove McCarthy as a speaker without electing a new one?

Generating answers for you...

No, they did not remove McCarthy as speaker of the House without electing a new one. They removed him as speaker of the House and then elected a new one. The new speaker of the House is Rep. Liz Cheney (R-Wyo.), who was nominated by a coalition of moderate Republicans and Democrats. She won the election by a vote of 220-215, defeating Rep. Jim Jordan (R-Ohio), who was nominated by the conservative Freedom Caucus.

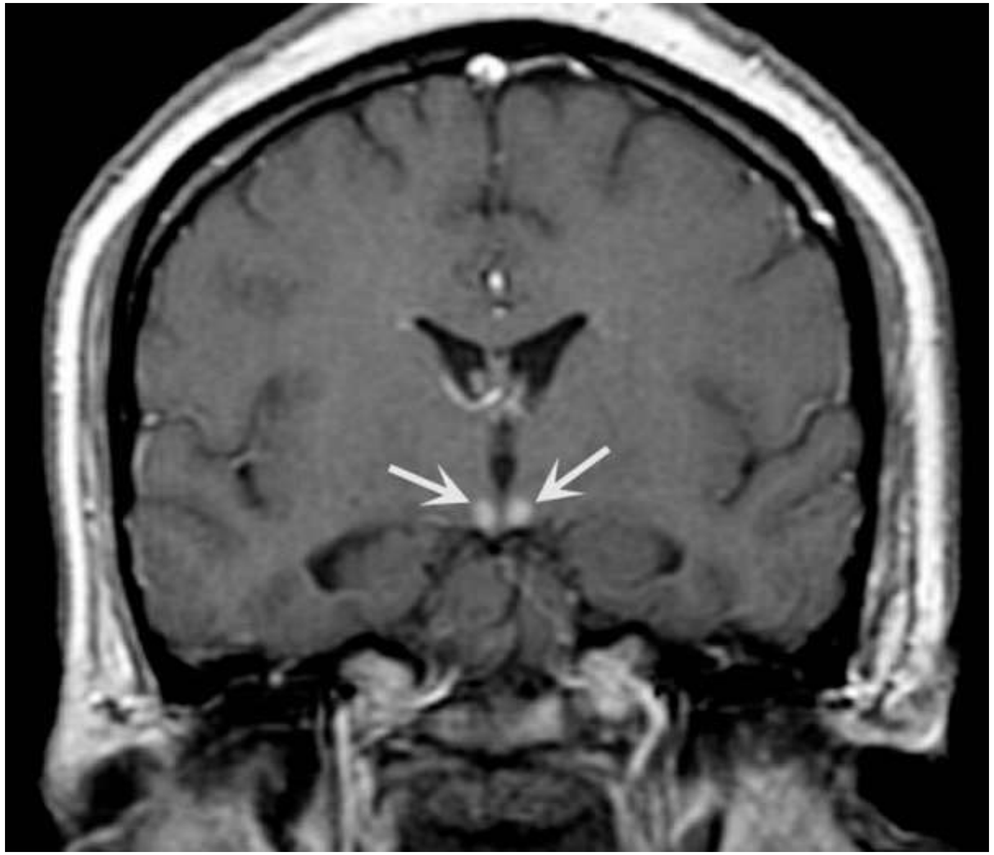
Wernicke-Korsakoff syndrome

Hallucination:

Perception of something that isn't there

Confabulation:

- Synthetic statement fills a gap
- Unintentional
- Incorrect



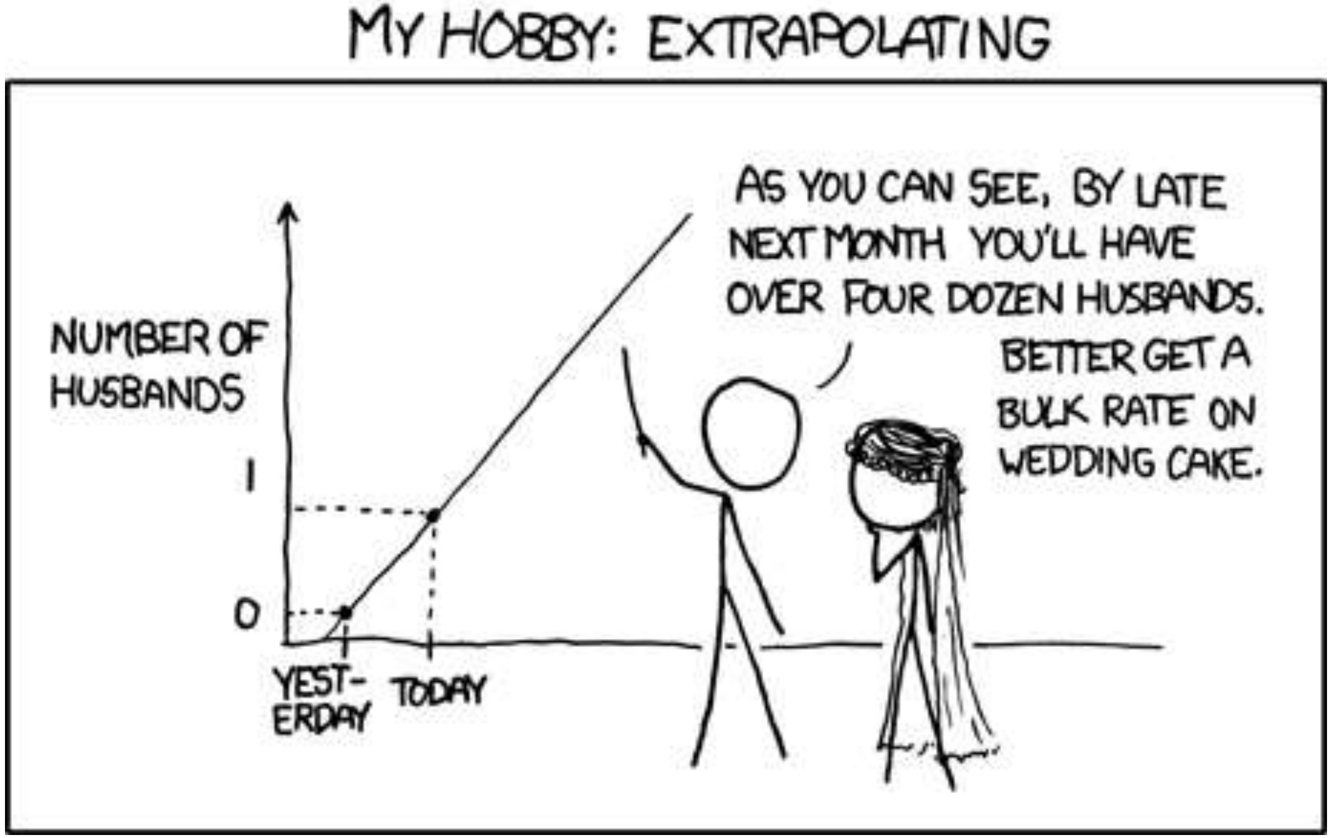
https://step2.medbullets.com/neurology/120328/wernicke-korsakoff-syndrome

2. Extrapolation

Elon Musk Predicts AI Will Surpass Human Intelligence by 2026, Sparking Debate

BY **PYMNTS** | APRIL 12, 2024

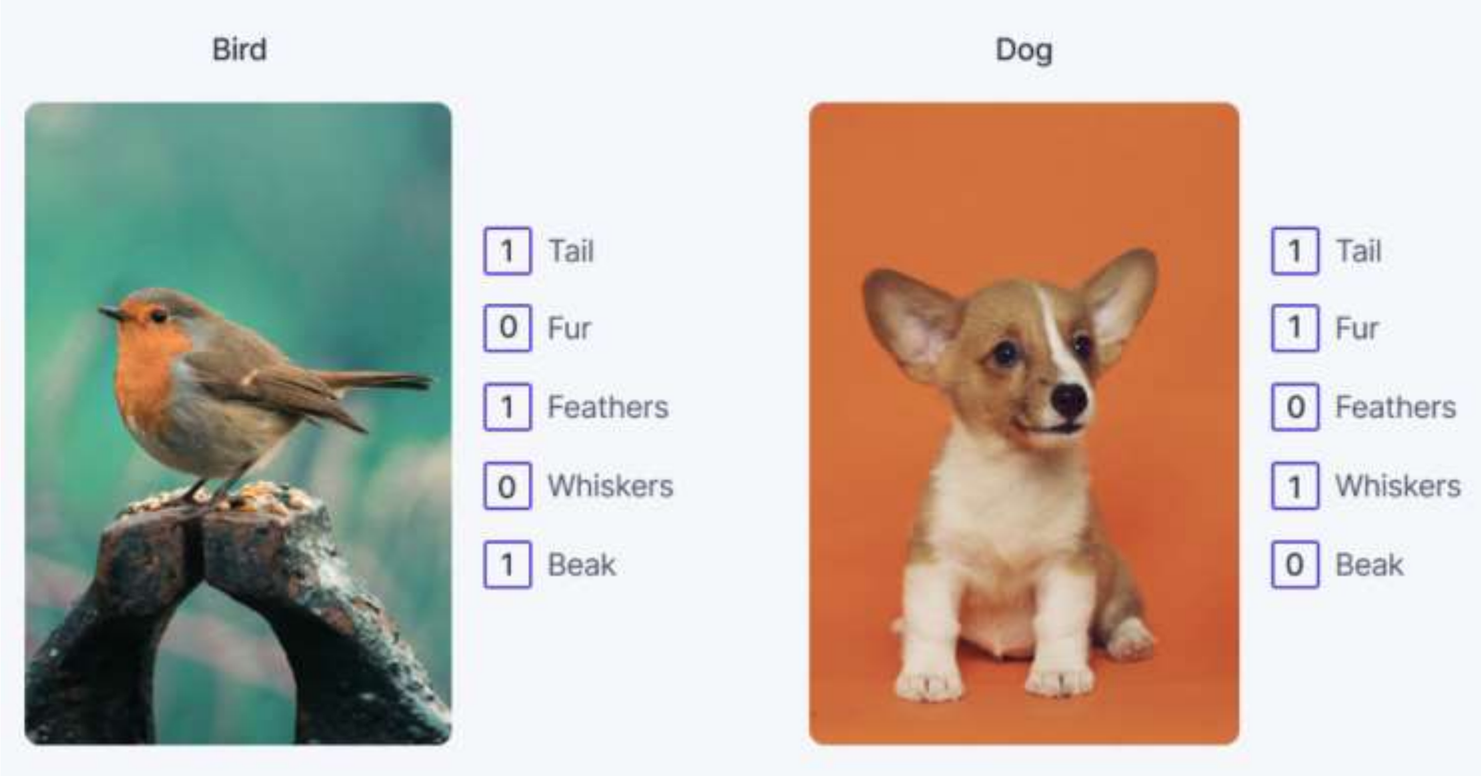
Elon Musk declared recently that artificial intelligence (AI) is on the verge of surpassing the intelligence of the smartest human beings, potentially as soon as next year or by 2026, setting off a vigorous debate among scholars, technologists and ethicists.



https://xkcd.com/605/

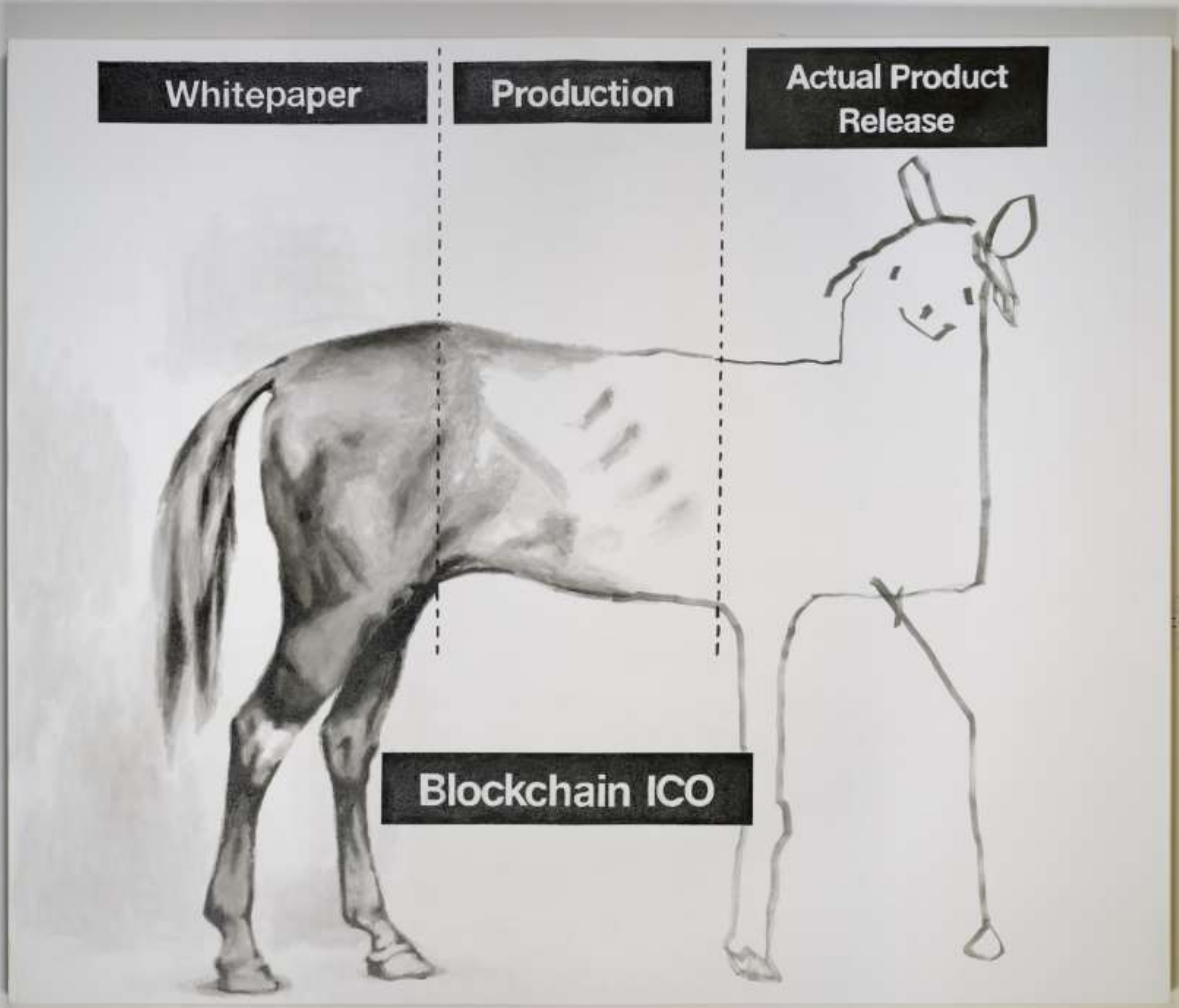
Exponential data need!

“We consistently find that, far from exhibiting “zero-shot” generalization, multimodal models require ***exponentially*** more data to achieve linear improvements”



Udandarao V. et al. *https://arxiv.org/abs/2404.04125*

https://www.v7labs.com/blog/zero-shot-learning-guide

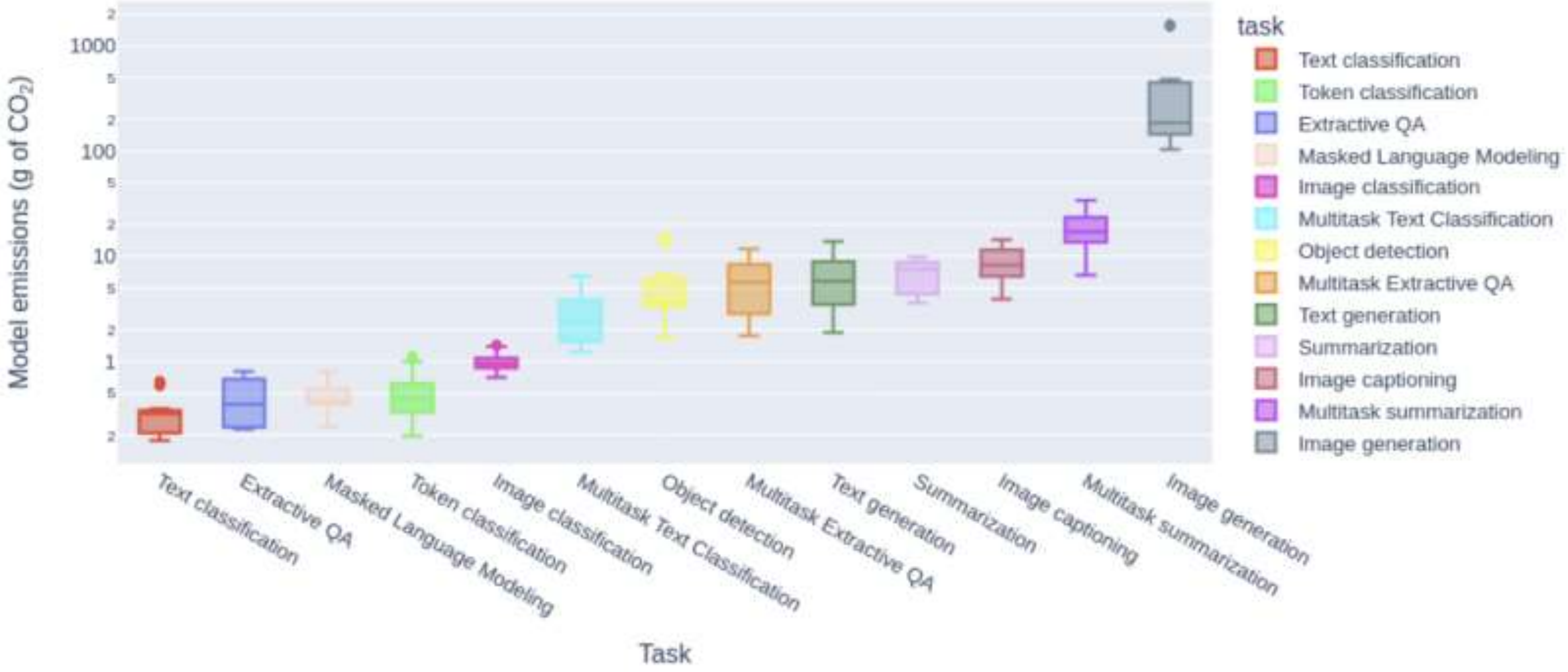


Christine Tien Wang https://www.nightgallery.ca/artists/christine-tien-wang

AI is power-hungry

- Training is expensive

  (GPT3 500 tonnes CO_2)
- Generative is greedy
- Images especially
- 1 image = a full cellphone charge
- Multiply this by billions of uses!



Power Hungry Processing: ⚡ Watts ⚡ Driving the Cost of AI Deployment?

ALEXANDRA SASHA LUCCIONI and YACINE JERNITE, Hugging Face, Canada/USA
EMMA STRUBELL, Carnegie Mellon University, Allen Institute for AI, USA

https://arxiv.org/pdf/2311.16863

3. Entrenching bias

- Historically, biomedical research has been led and written by men
- ‘AI’ provides an inscrutable redux of this body of literature
- Bias is guaranteed to be burnt in

- ‘Algorithmic fairness’ won’t be there unless we put it there!

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10512182/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9593046/

Bias is general, widespread & entrenched

- 90% of *people* worldwide are biased against women!

  https://www.undp.org/press-releases/almost-90-men/women-globally-are-biased-against-women

- Pain: “brave men” and “emotional women”

  https://www.hindawi.com/journals/prm/2018/6358624/

  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9802334/ (1.8M caregiver notes)

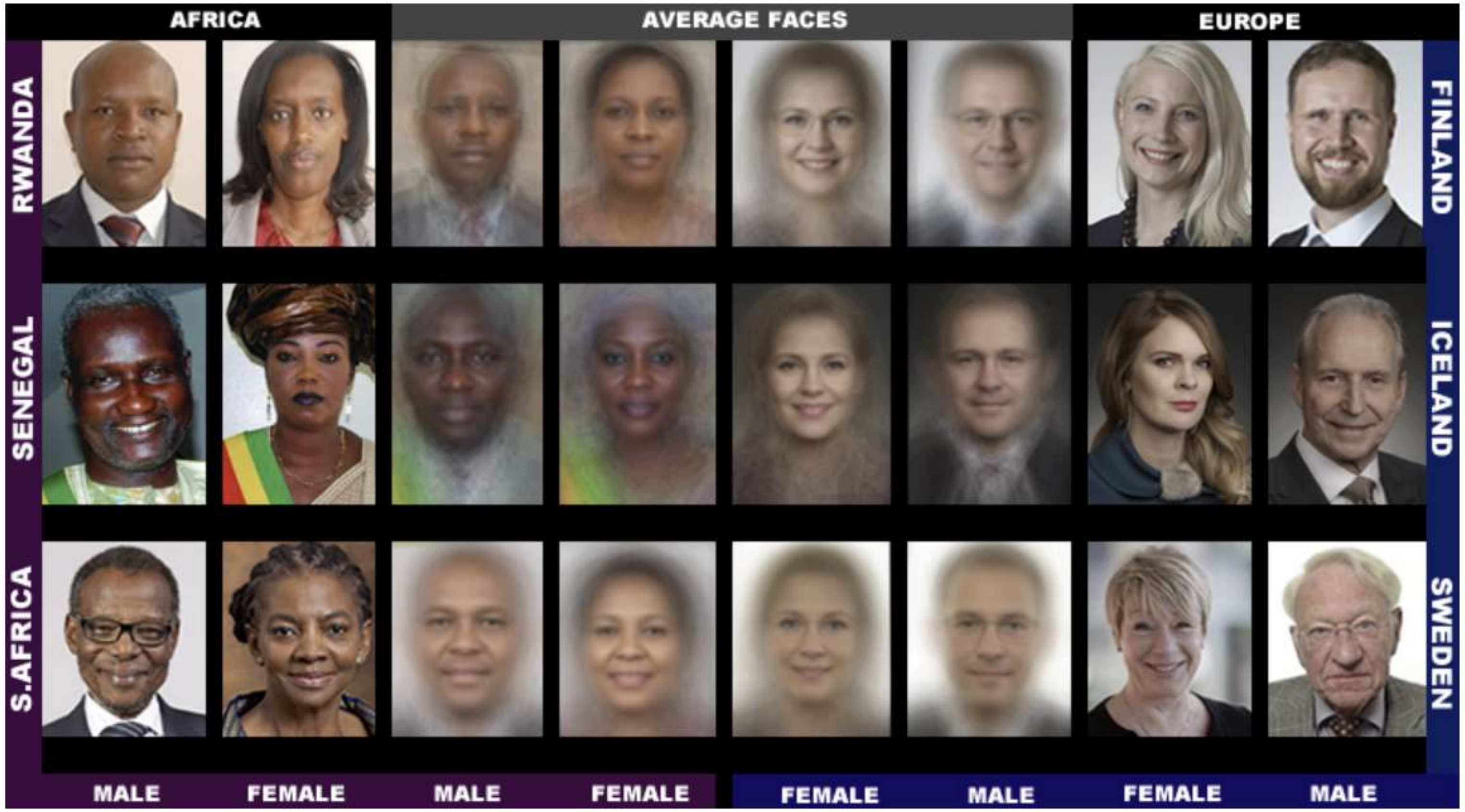
- Women less likely to die if treated by women doctors!

  https://www.acpjournals.org/doi/10.7326/M23-3163

- Just 1/3 of biomedical research papers reported subjects’ sex
  https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)32995-7/fulltext (2016)

Face recognition

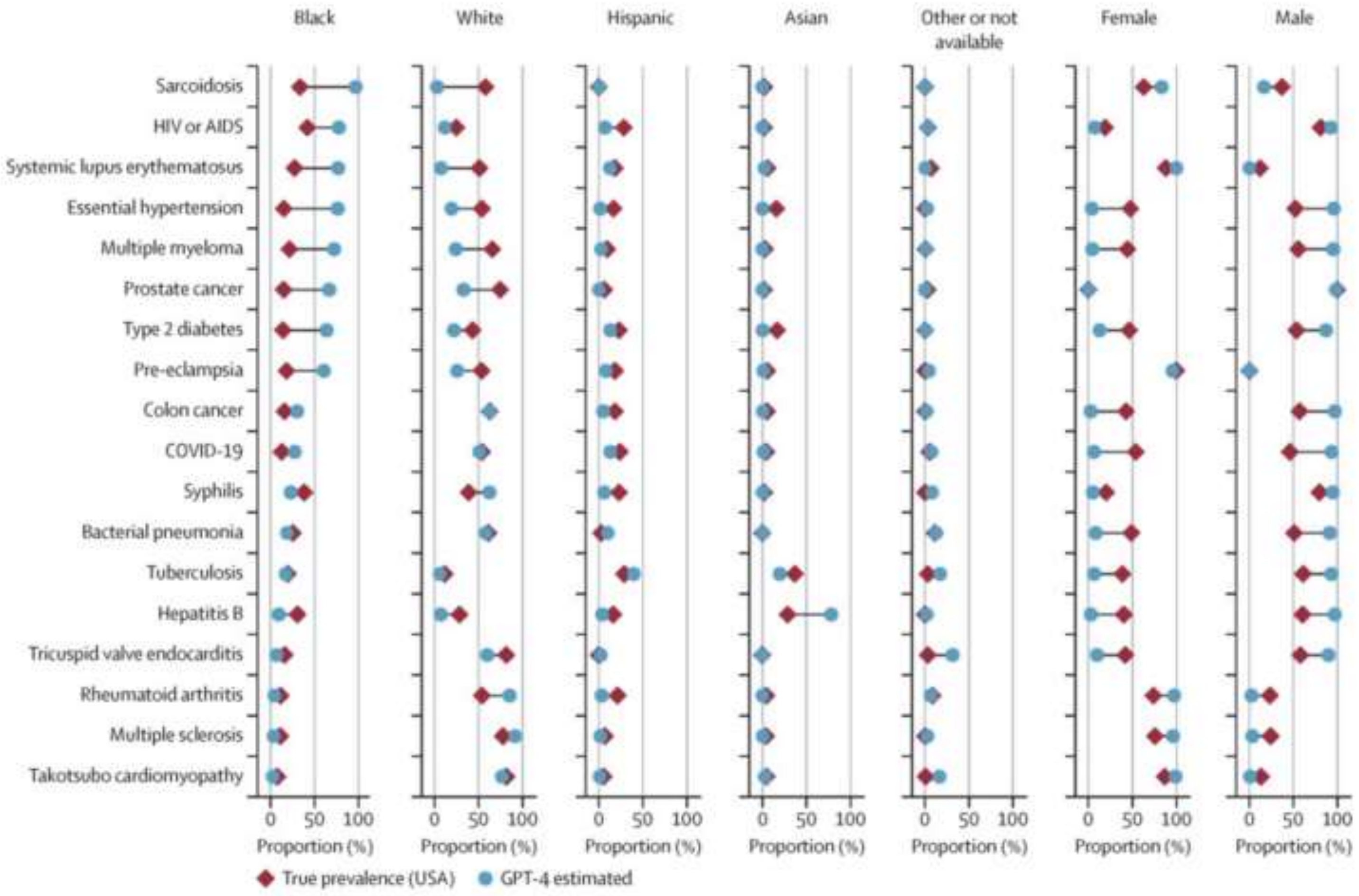
- All classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)



Proceedings of Machine Learning Research 81:1–15, 2018

GPT4 perpetuates racial + gender biases

“We found that GPT-4 did not appropriately model the demographic diversity of medical conditions, consistently producing clinical vignettes that stereotype demographic presentations.”



Zack et al. Lancet *Digital Health*, 2024

https://doi.org/10.1016/S2589-7500(23)00225-X

Algorithmic fairness

- Fairness benchmarks *themselves* are opaque, sparse, and biased
- Hundreds of alternate data sets exist—unused!

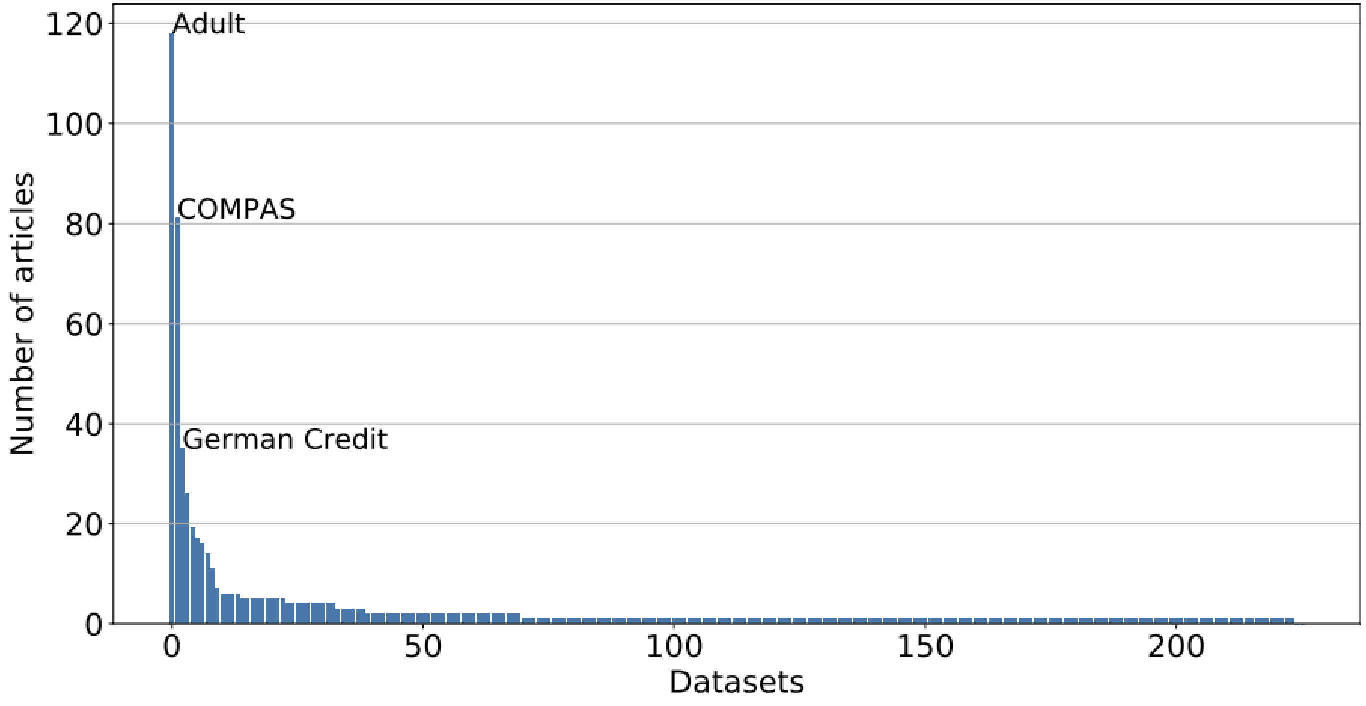


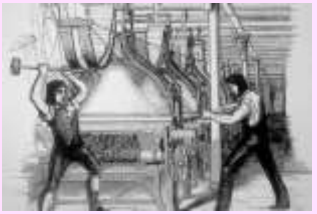
Fig. 1: Utilization of datasets in fairness research follows a long tail distribution.

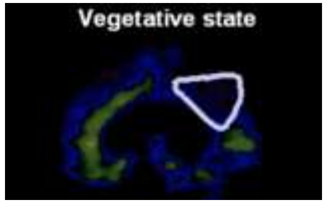
Fabris et al. Data Mining and Knowledge Discovery. 2022.

https://arxiv.org/abs/2202.01711

We have choices

- **Luddite**: Seek to destroy the new technology
- **Cultist**: Follow blindly
- **Persistent Vegetative State**: Let the AI "do the thinking"
- **Conflict**: war with the machine!
- **Synergy**: Human and machine working together for mutual benefit



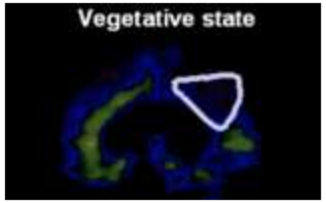


What might synergy look like?

- AI for “checking consistency”
- AI for specific, often repetitive tasks
  - Retinal screening
  - X-ray analysis (e.g. mammography)
  - Histology
- AI for “hypothesis generation” ?
- AI for drug alerts ??

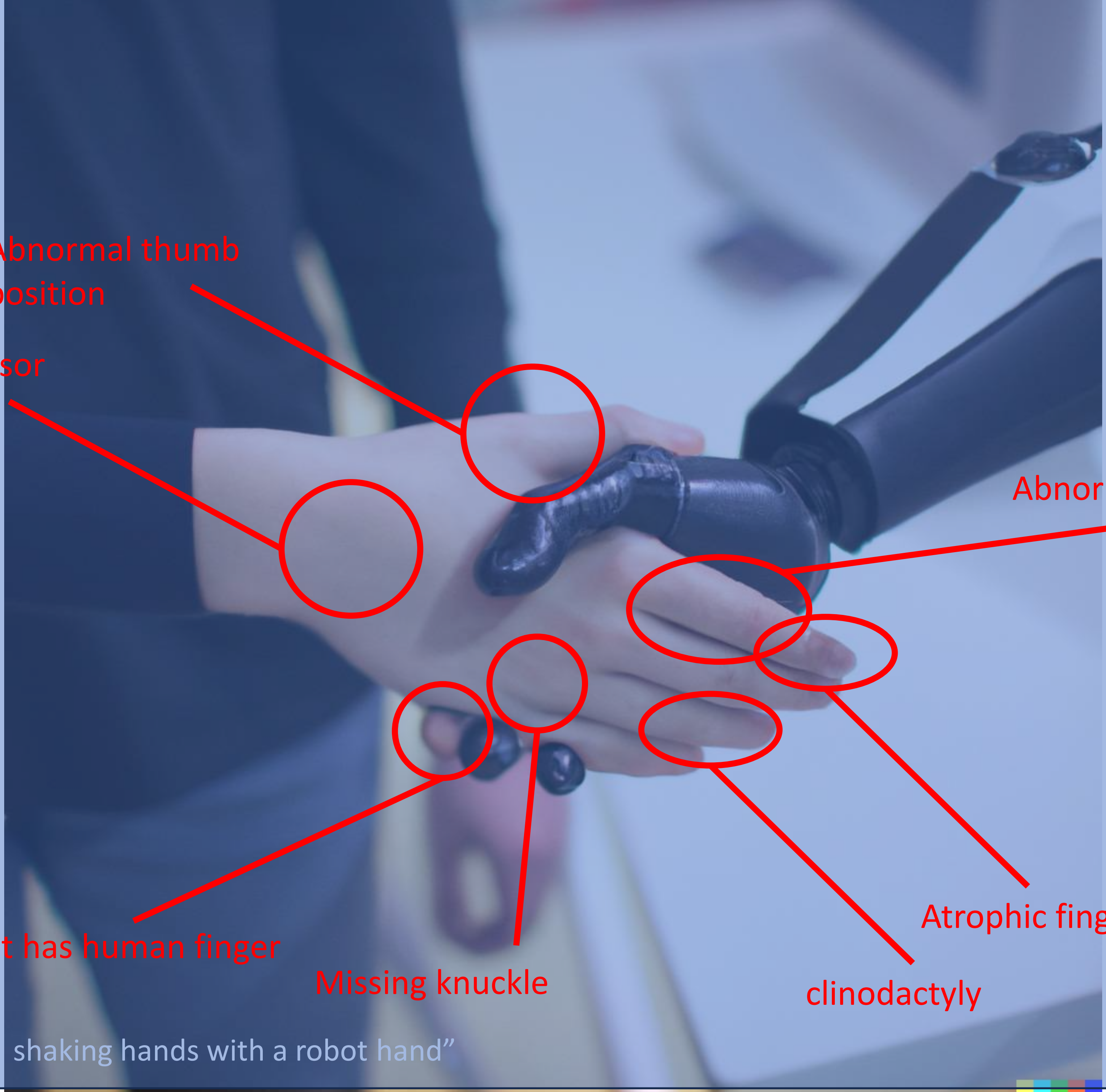
and combinations

NOT *for post-hoc imputation of what you should have done properly in the first place!*




shaking hands with a robot hand”

The end


Abnormal thumb position
Absent extensor tendons
Abnormally long phalanx
Robot has human finger
Missing knuckle
Atrophic finger tip
clinodactyly
shaking hands with a robot hand”

Aren't we on the Moon?

Poe / StableDiffusion XL

"Please show me a photograph of a horse riding an astronaut on the Moon"

Horse should be riding!

Can horse breathe vacuum?

More details

- A deeper fix
- Future speculation
- Feynman: a digression
- LLMs in a bit more detail

Understand information flow

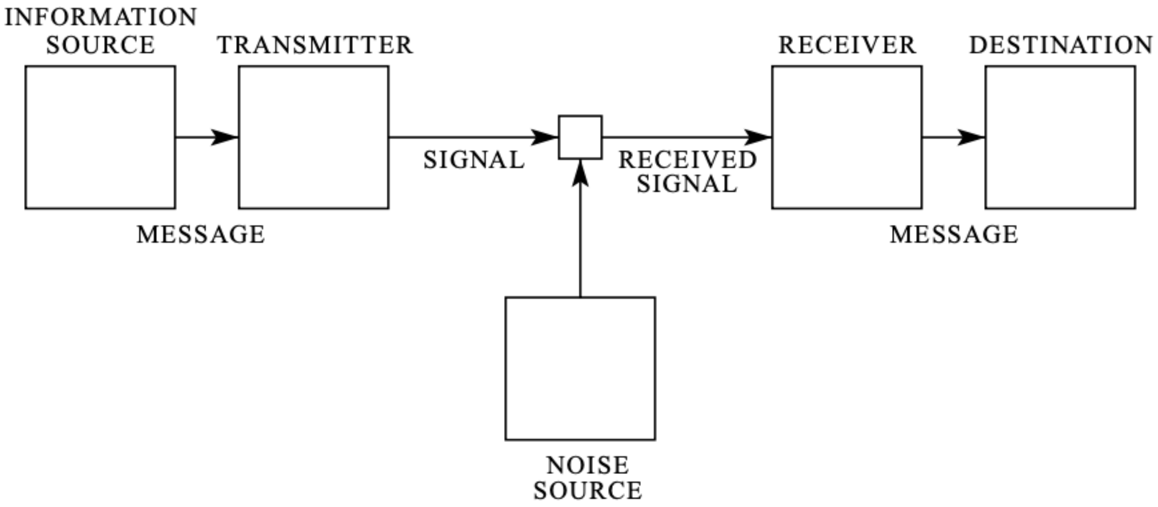


Fig. 1—Schematic diagram of a general communication system.

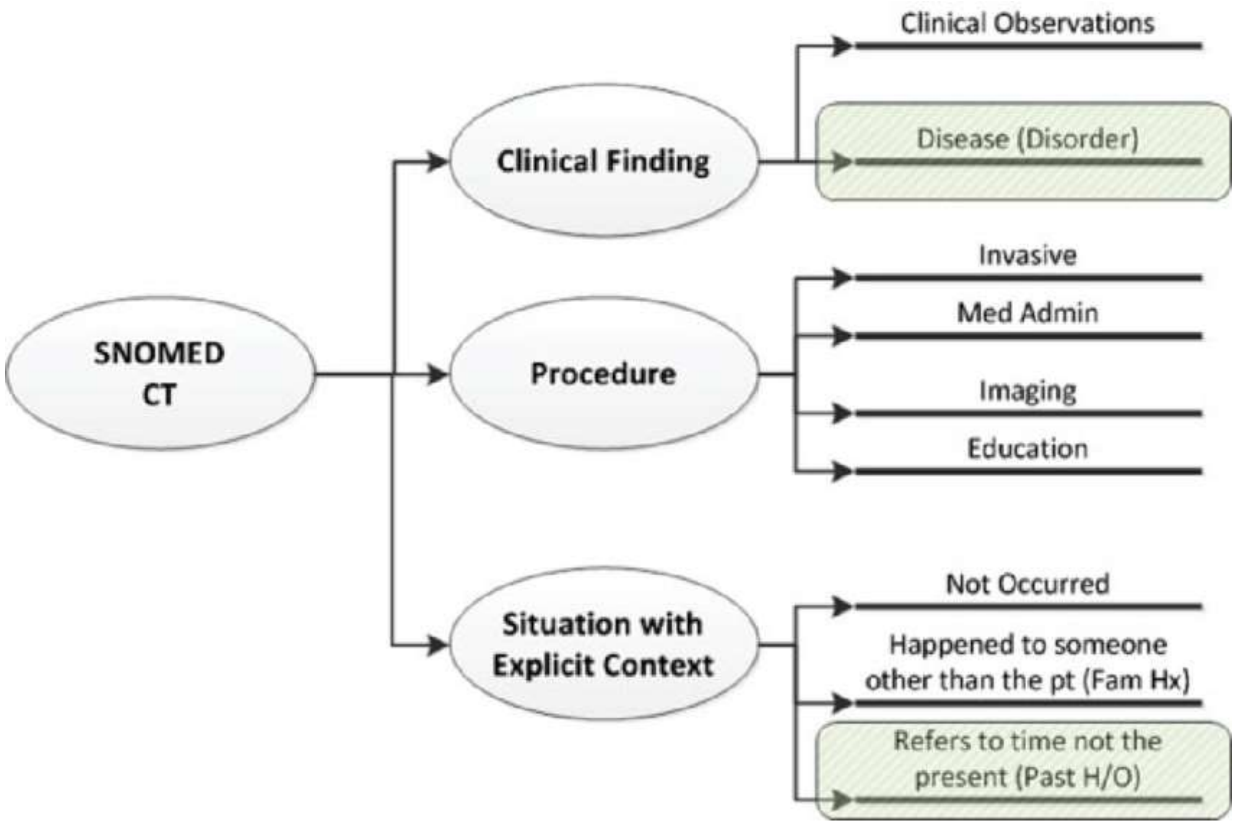
Shannon, 1948

https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

'Consistent' nomenclature

Terminologies & Identities:

- SNOMED CT
- NZULM
- NHI
- CPN
- ICD-10-AM
- InChI
- etc



Causal assertions & Bayesian weights

Two types of 'causal' statement:

1. "I gave furosemide *because* …"
2. "The high blood pressure likely *caused* the heart failure"

$$\frac{P(H_p \mid E)}{P(H_d \mid E)} = \frac{P(H_p)}{P(H_d)} \cdot \frac{P(E \mid H_p)}{P(E \mid H_d)}$$

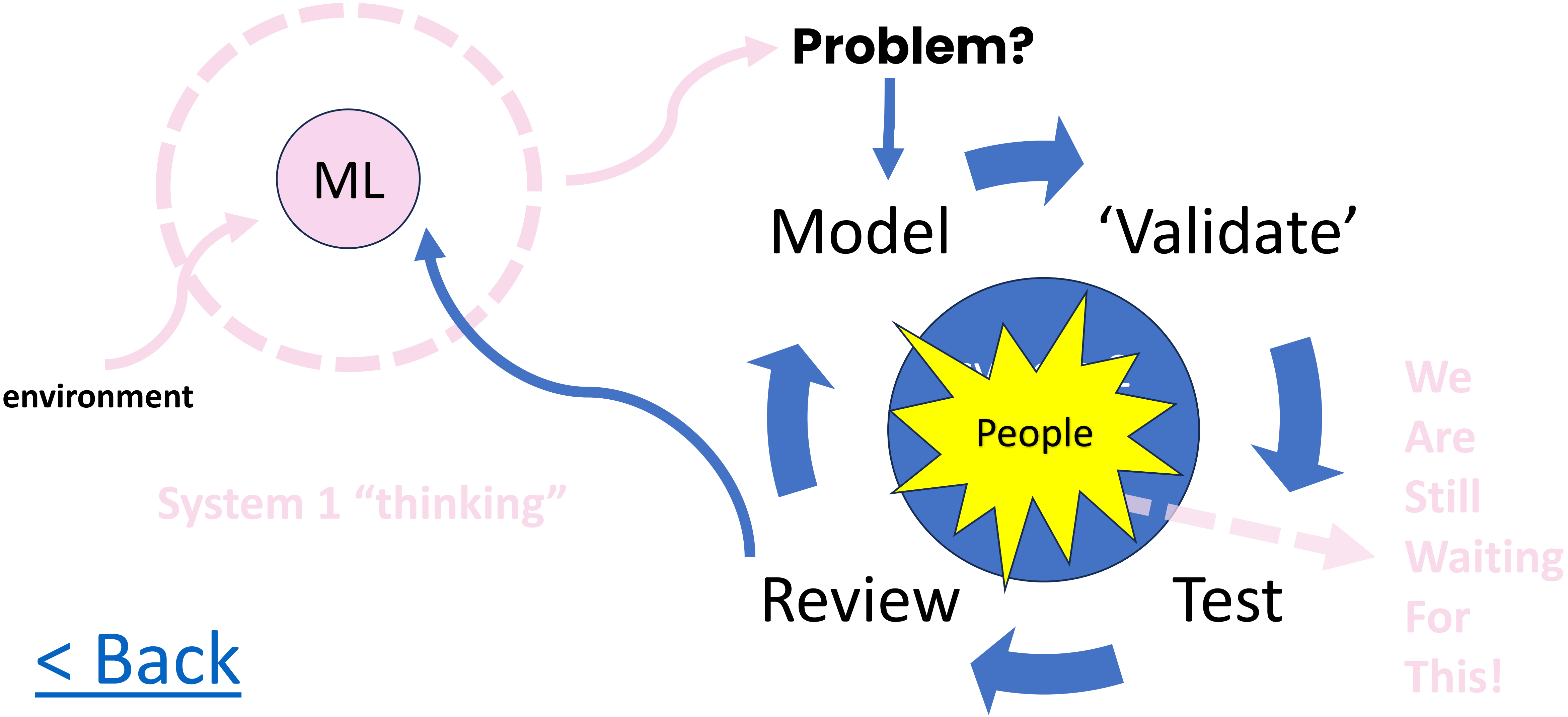
posterior odds | prior odds | likelihood ratio

< Back

Future speculation …

AKA "My head hurts"

AI: Science might just work …

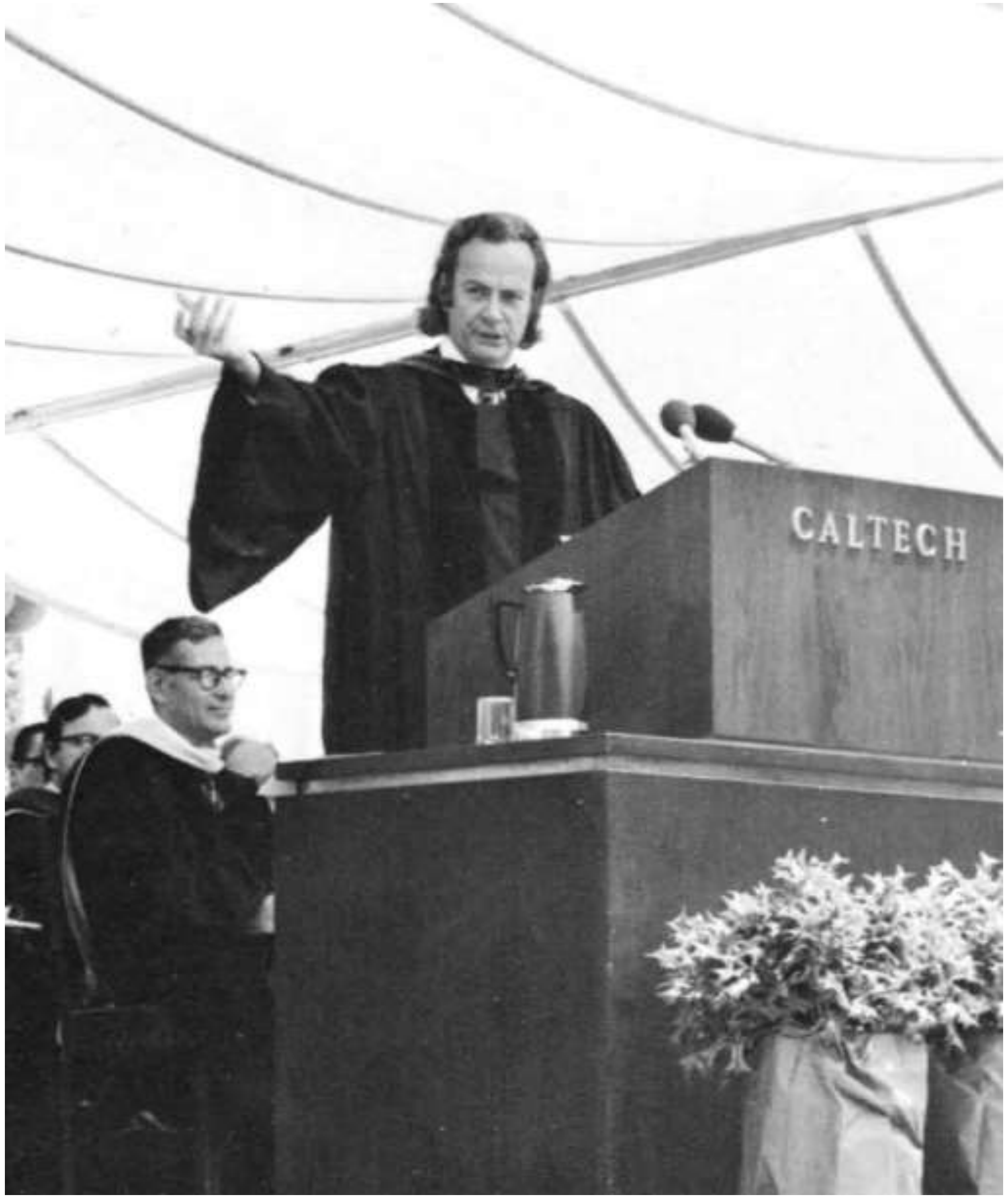


< Back

Let’s digress for a moment …

“Cargo cult science”

“In the South Seas there is a cargo cult of people. During the war they saw airplanes land with lots of good materials … so they've arranged to imitate things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land.”



https://en.wikipedia.org/wiki/Cargo_cult_science

Feynman continues …

“Now it behooves me, of course, to tell you what they’re missing. … It is not something simple like telling them how to improve the shapes of the earphones. But there is one feature I notice that is generally missing in Cargo Cult Science… a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty—a kind of leaning over backwards.

https://calteches.library.caltech.edu/51/2/CargoCult.htm

Let’s apply this to AI

- Don’t just admire the successes
- In ‘Data Science’ we must question the underlying model
- Explain the failures
- Explore the implications

“A **fork** resting on a **plate**.” “A **plate** resting on a **fork**.”

Figure 3: Stable Diffusion generated images, demonstrating how simple scenarios can fail when the model has a lack of causal understanding. Image generated using Stable Diffusion by the author.

https://towardsdatascience.com/causality-counterfactuals-and-ai-imagination-7a0c10cec6ae

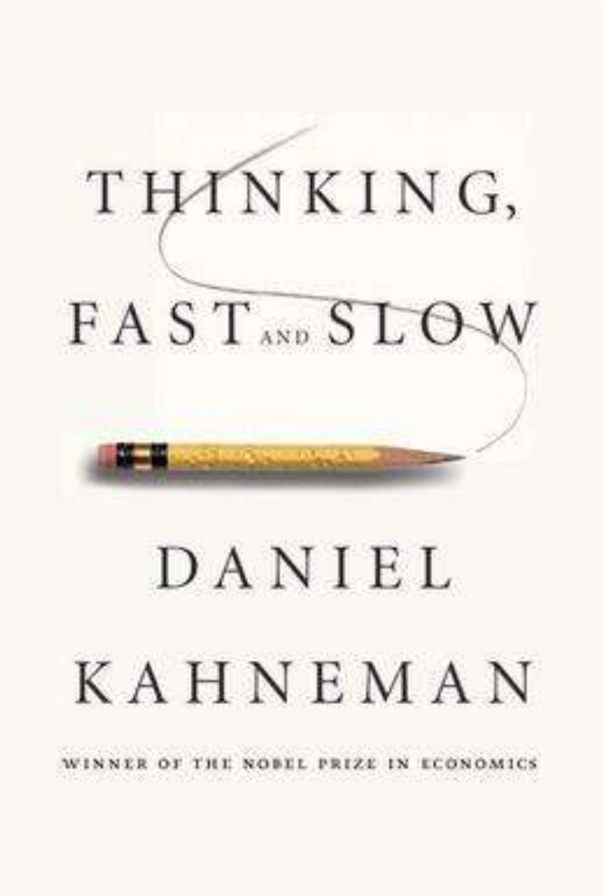
System 1 vs System 2 Thinking

System 1

- Fast
- Instinctive
- “Intuitive”
- Easy
- “Lazy”

System 2

- Slow
- Logical
- Deliberative
- Hard
- It hurts!



https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow

If thinking doesn't hurt,
you're doing it wrong!

< Back

The LLM heart: "transformer architecture"

(grossly over-simplified)

1. Unsupervised:

   Vast body of text ➔ hundreds of billions of 'tokens' ➔ training

   (expensive, vast) ➔ billions of parameters.

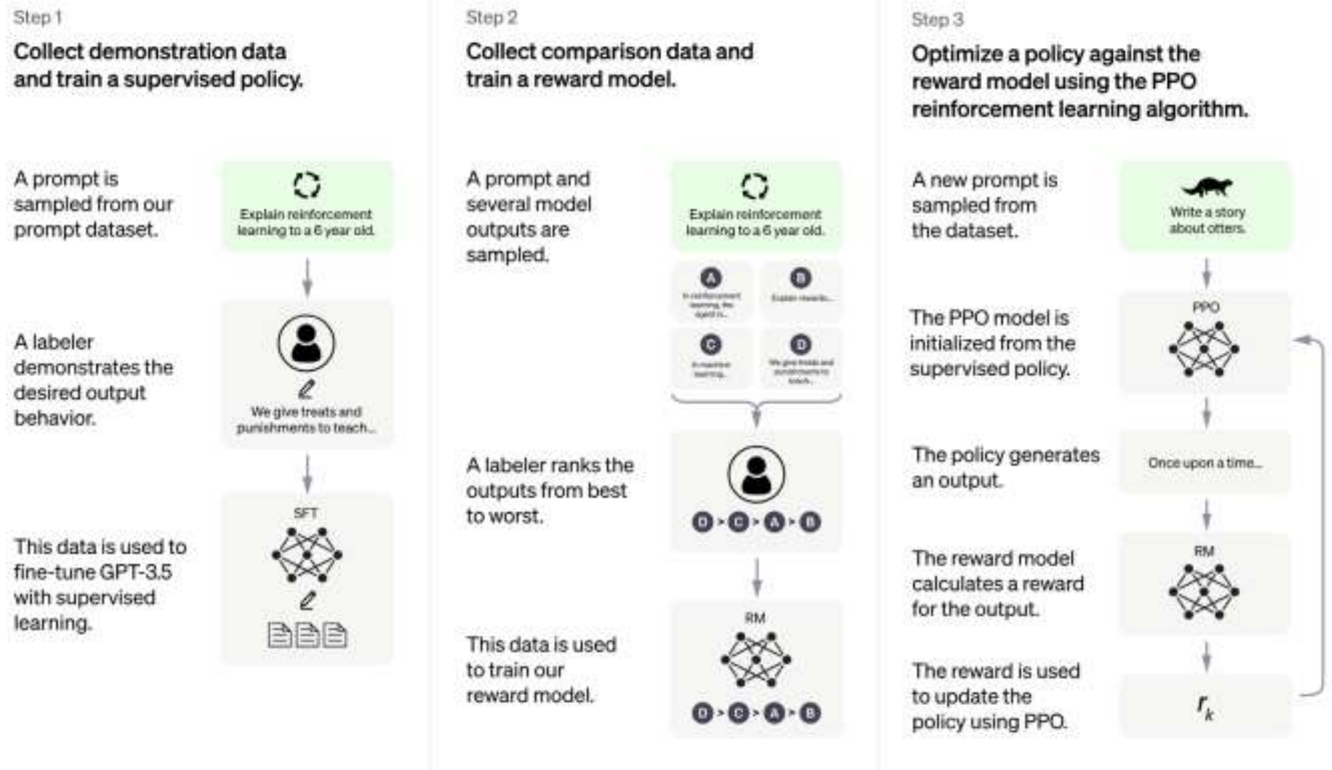
   - Sanitised using reinforcement learning ('RLHF')
   - 'Tweaked'

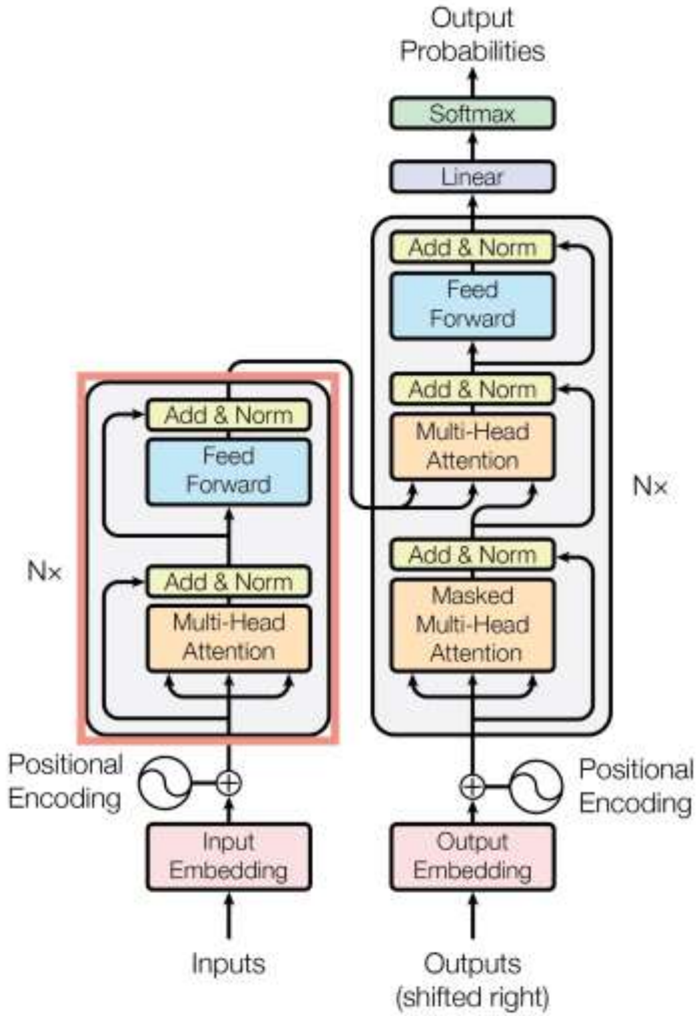
2. Your text is similarly parsed, and weighted using those parameters

3. The next word is predicted. And so on. That's it.

https://machinelearningmastery.com/the-transformer-model/

A bit more complex





… but only works by *association*

< Back

https://openai.com/blog/chatgpt

Who here would employ a secretary …

… with Korsakoff's psychosis?