Rational Machines?

Edward A. Lee

Professor of the Graduate School and
Distinguished Professor Emeritus, UC Berkeley

Can Machines Save the World?
The Digital Humanism Fellowship Conference
Vienna Austria

Nov. 16, 2023

Rationality

Logical ↔ Intuitive

Reasoned ↔ Creative

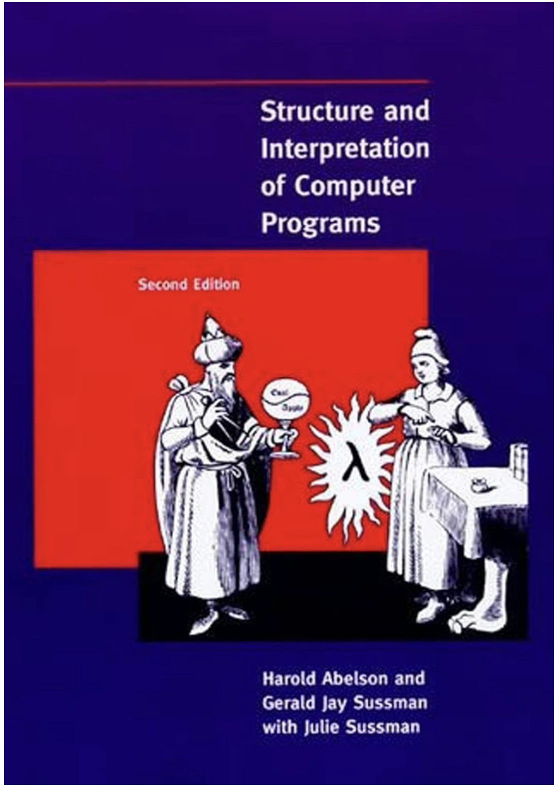
Dispassionate ↔ Emotional

Machine ↔ Human?

The First Computer Revolution

‘The **computer revolution** is a revolution in the way we think and in the way we express what we think. The essence of this change is the emergence of what might best be called **procedural epistemology** – the study of the structure of knowledge from an imperative point of view, as opposed to the more declarative point of view taken by classical mathematical subjects. Mathematics provides a framework for dealing precisely with notions of “what is.” Computation provides a framework for dealing precisely with notions of “how to.”’

Abelson and Sussman, 1985

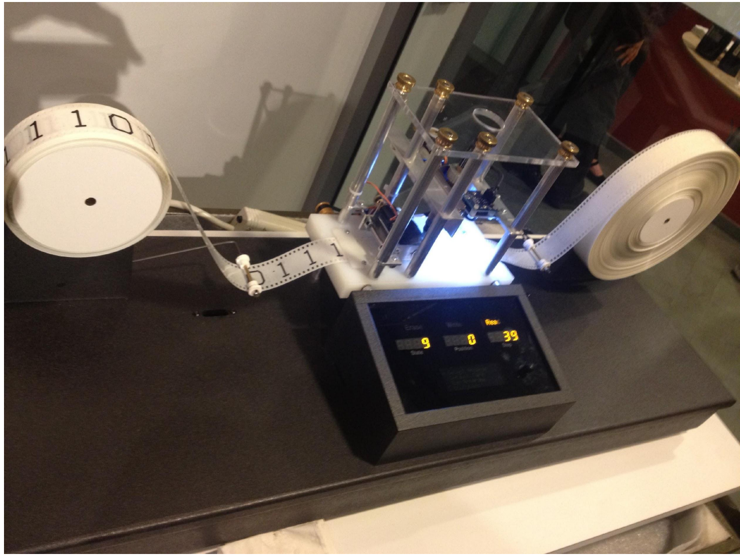

Structure and Interpretation of Computer Programs
Second Edition
λ
Harold Abelson and Gerald Jay Sussman with Julie Sussman

Procedural Epistemology

Algorithms:

- Start with input data,
- Follow a sequence of steps, where
- Each step follows well-defined rules, and
- The sequence terminates with a conclusion.

If you further limit the data to a discrete set, then algorithms are equivalent to Turing Machines.



Turing Machine designed by Mike Davey
Photo by GabrielF - Own work, CC BY-SA 3.0

Human Rationality

Rational process: step-by-step reasoning using clearly explicable rules of logic.

Bounded rationality: Humans are not actually very good at this!

But computers are!

Herb Simon, circa 1981

The Second Computer Revolution

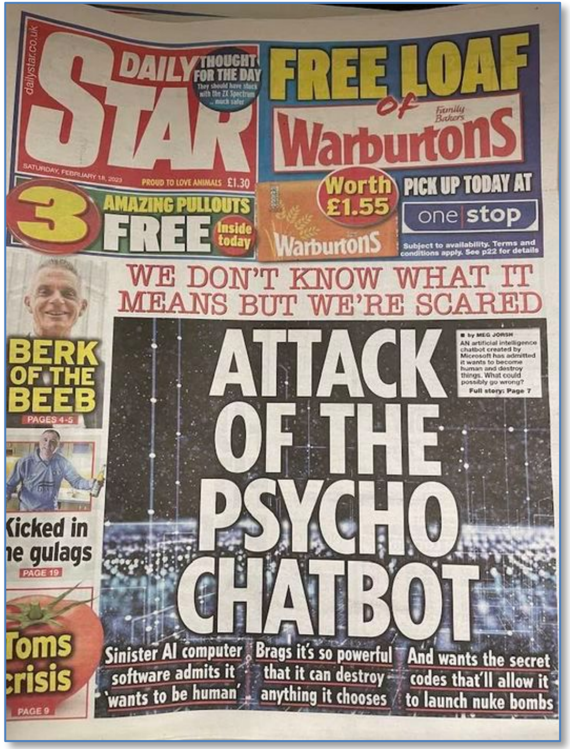
LLMs:

- OpenAI: GPT 3.5, 4 (ChatGPT)
- Google: PaLM (Bard)
- Meta: LLaMa
- Various: BLOOM
- Baidu: Ernie Bot
- Anthropic: Claude

Text-to-Image:

- DALL-E-2,3
- Stable Diffusion
- Midjourney
- …

Are these foundation models usefully understood as rational machines?



dailystar.co.uk

DAILY STAR

THOUGHT FOR THE DAY
They should have stuck with the ZX Spectrum ... much safer

SATURDAY, FEBRUARY 18, 2023

PROUD TO LOVE ANIMALS £1.30

FREE LOAF of Warburtons Family Bakers

Worth £1.55

PICK UP TODAY AT one stop

Subject to availability. Terms and conditions apply. See p22 for details

3 AMAZING PULLOUTS FREE Inside today

WE DON'T KNOW WHAT IT MEANS BUT WE'RE SCARED

BERK OF THE BEEB
PAGES 4-5

Kicked in
e gulags
PAGE 19

Toms
crisis
PAGE 9

ATTACK OF THE PSYCHO CHATBOT

by MEG JORSH

AN artificial intelligence chatbot created by Microsoft has admitted it wants to become human and destroy things. What could possibly go wrong?

Full story: Page 7

Sinister AI computer software admits it 'wants to be human'

Brags it's so powerful that it can destroy anything it chooses

And wants the secret codes that'll allow it to launch nuke bombs

My Essential Claim

The new machines are not about procedural epistemology.

Rationality

Logical ↔ intuitive

Reasoned ↔ creative

Dispassionate ↔ emotional

Machine ↔ human?

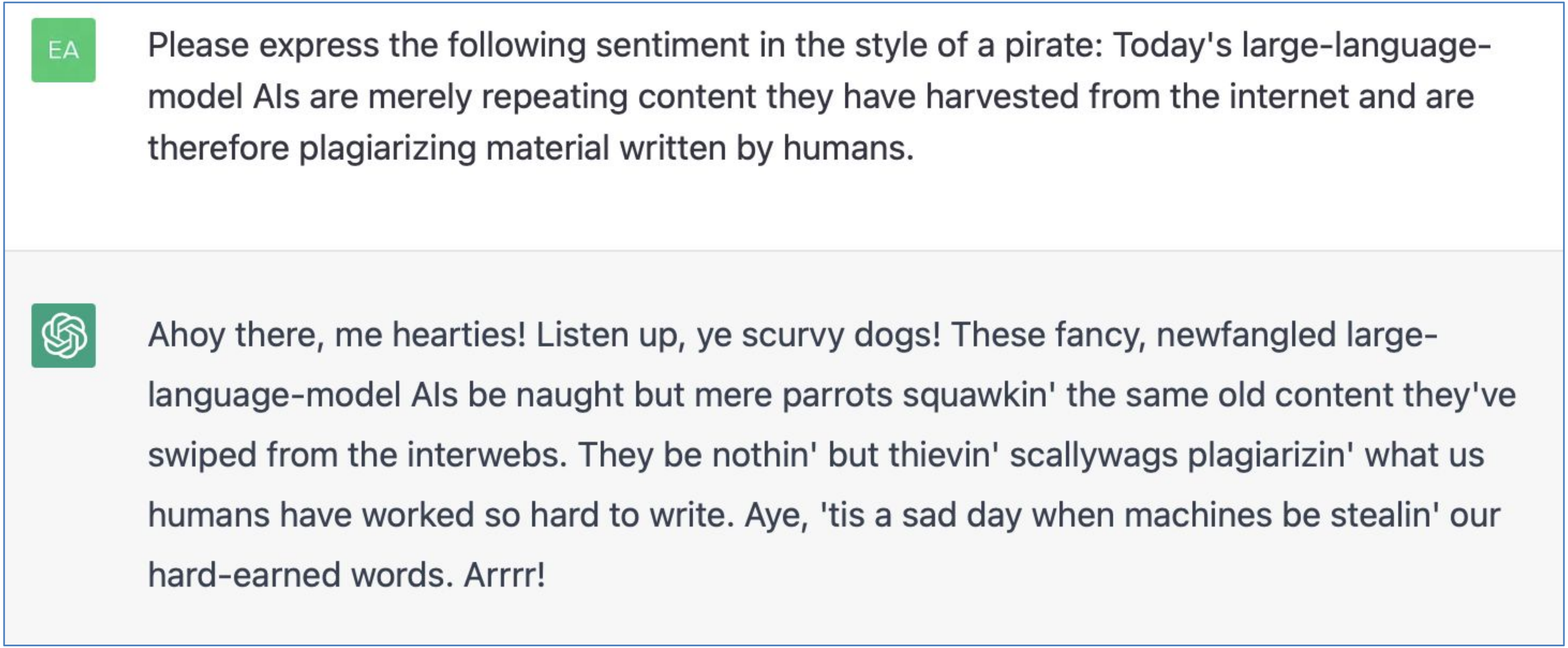
They make stuff up.

Is it hallucination or creativity?

DALL-E 2, prompted by Boris Eldagsen. Selected for Sony world photography award 2023, Open Creative Category. Eldagsen declined the prize.

Théâtre D'opéra Spatial
2022 Colorado State Fair Grand Prize
Jason Allen via Midjourney

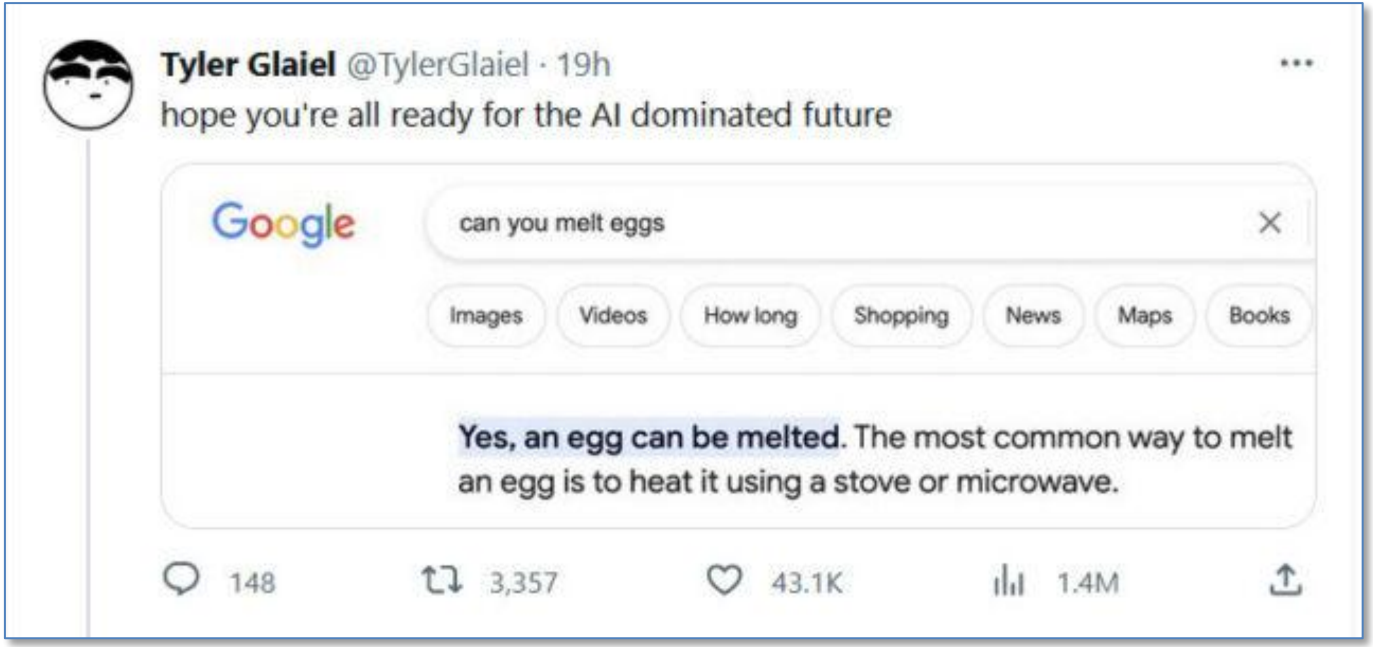
Stochastic Parrots?



Please express the following sentiment in the style of a pirate: Today's large-language-model AIs are merely repeating content they have harvested from the internet and are therefore plagiarizing material written by humans.

Ahoy there, me hearties! Listen up, ye scurvy dogs! These fancy, newfangled large-language-model AIs be naught but mere parrots squawkin' the same old content they've swiped from the interwebs. They be nothin' but thievin' scallywags plagiarizin' what us humans have worked so hard to write. Aye, 'tis a sad day when machines be stealin' our hard-earned words. Arrrr!

They make mistakes.



Tyler Glaiel @TylerGlaiel · 19h

hope you're all ready for the AI dominated future

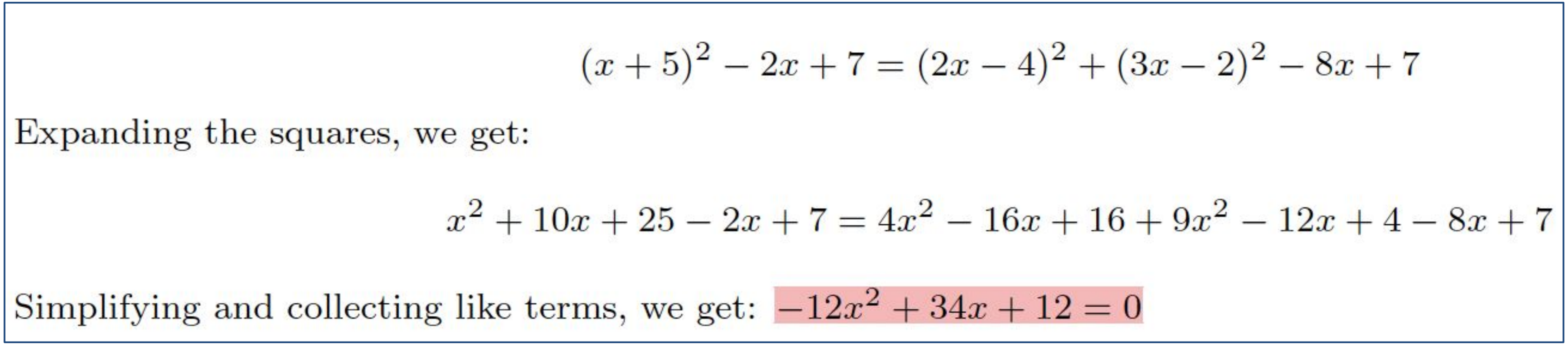
Yes, an egg can be melted. The most common way to melt an egg is to heat it using a stove or microwave.

https://arstechnica.com/information-technology/2023/09/can-you-melt-eggs-quoras-ai-says-yes-and-google-is-sharing-the-result/

Melting eggs by DALL E 2.

They make mistakes.

GPT-4: Asked to simplify:



$$(x+5)^2 - 2x + 7 = (2x-4)^2 + (3x-2)^2 - 8x + 7$$

Expanding the squares, we get:

$$x^2 + 10x + 25 - 2x + 7 = 4x^2 - 16x + 16 + 9x^2 - 12x + 4 - 8x + 7$$

Simplifying and collecting like terms, we get: $-12x^2 + 34x + 12 = 0$

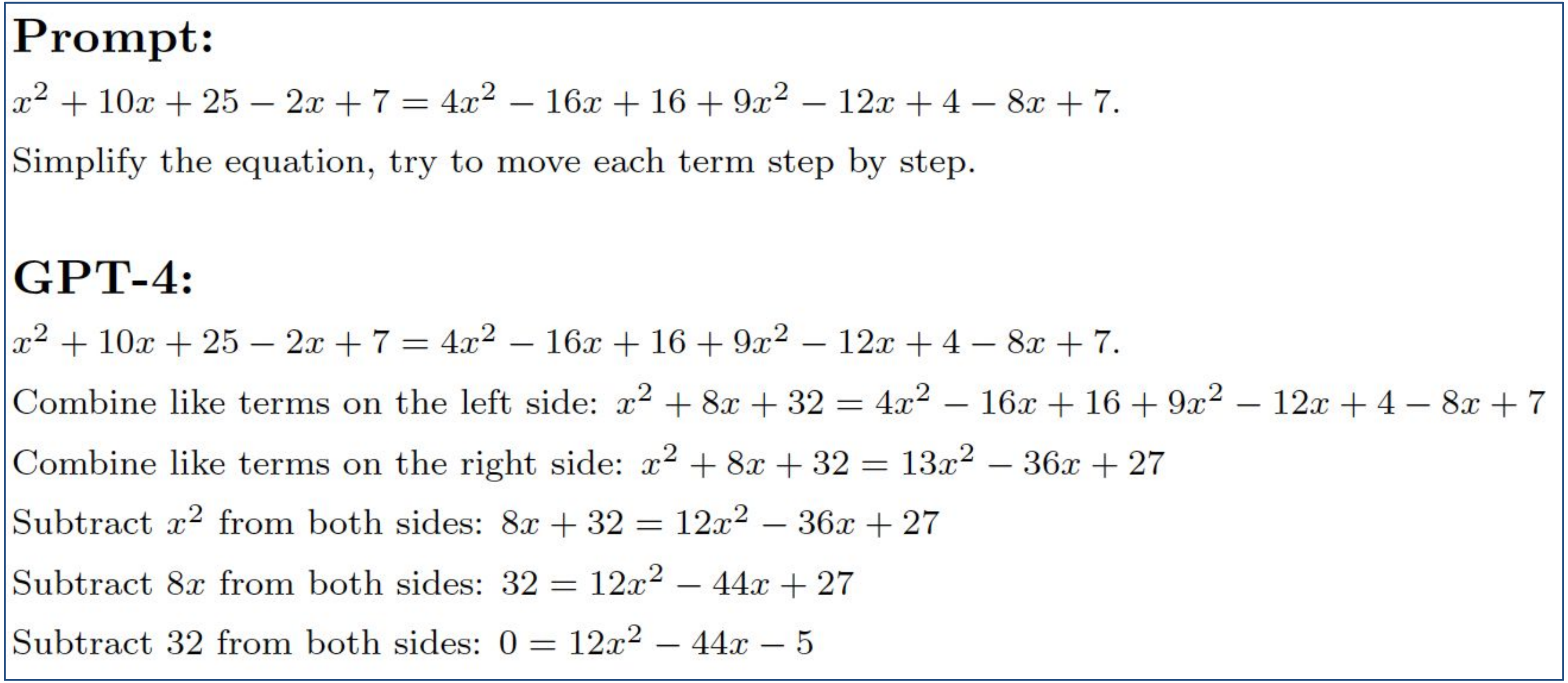
This is incorrect.

Bubek, et al.,. arXiv:2303.12712v1 [cs.CL] 22 Mar 2023

They make mistakes.

Bubek, et al.,. arXiv:2303.12712v1
[cs.CL] 22 Mar 2023

GPT-4: Asked to break down the steps:



Prompt:

$x^2 + 10x + 25 - 2x + 7 = 4x^2 - 16x + 16 + 9x^2 - 12x + 4 - 8x + 7.$

Simplify the equation, try to move each term step by step.

GPT-4:

$x^2 + 10x + 25 - 2x + 7 = 4x^2 - 16x + 16 + 9x^2 - 12x + 4 - 8x + 7.$

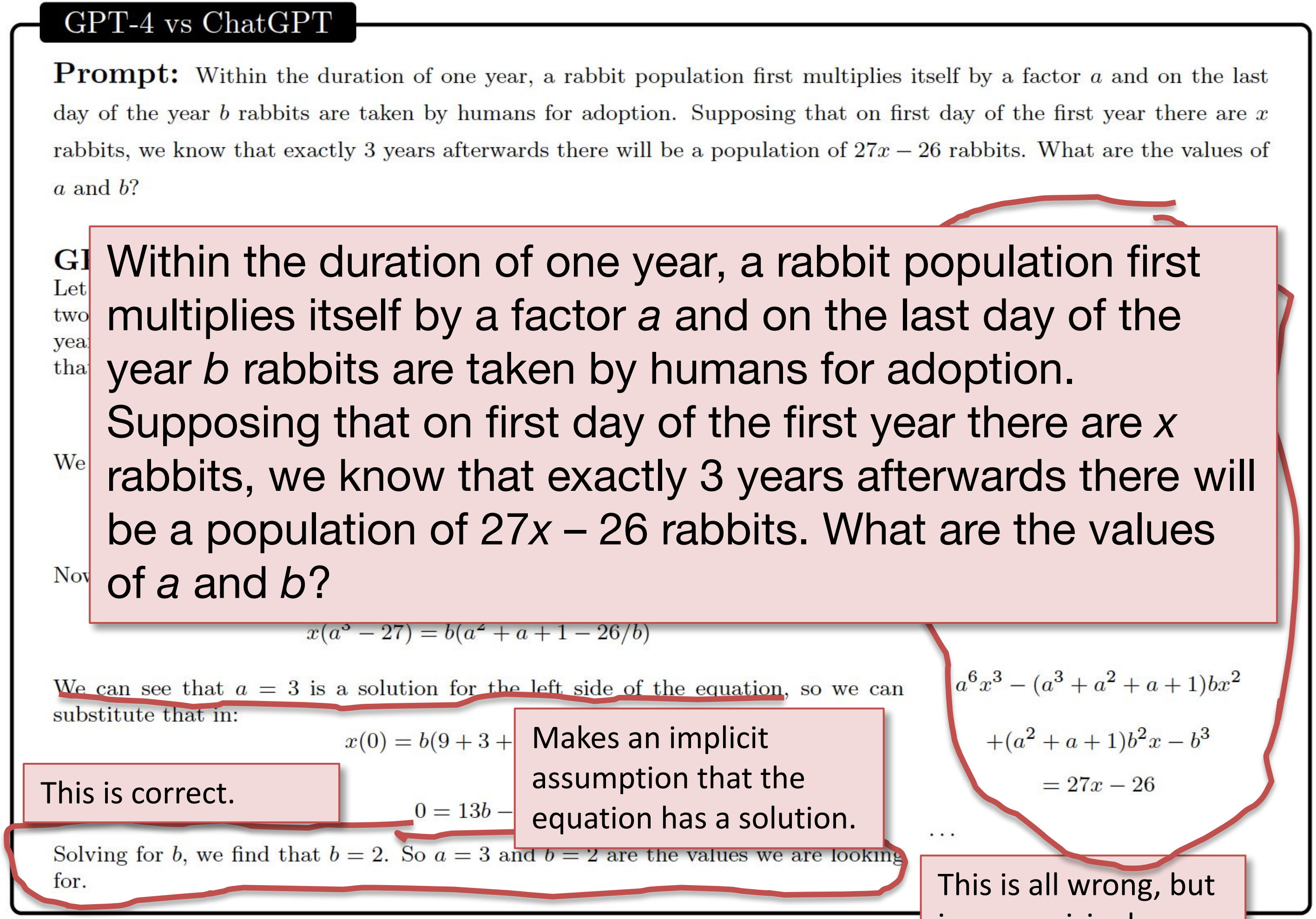
Combine like terms on the left side: $x^2 + 8x + 32 = 4x^2 - 16x + 16 + 9x^2 - 12x + 4 - 8x + 7$

Combine like terms on the right side: $x^2 + 8x + 32 = 13x^2 - 36x + 27$

Subtract x^2 from both sides: $8x + 32 = 12x^2 - 36x + 27$

Subtract $8x$ from both sides: $32 = 12x^2 - 44x + 27$

Subtract 32 from both sides: $0 = 12x^2 - 44x - 5$



GPT-4 vs ChatGPT

Prompt: Within the duration of one year, a rabbit population first multiplies itself by a factor a and on the last day of the year b rabbits are taken by humans for adoption. Supposing that on first day of the first year there are x rabbits, we know that exactly 3 years afterwards there will be a population of $27x - 26$ rabbits. What are the values of a and b?

GI

Let

two

yea

tha

We

Nov

$$x(a^3 - 27) = b(a^2 + a + 1 - 26/b)$$

We can see that $a = 3$ is a solution for the left side of the equation, so we can substitute that in:

$$x(0) = b(9 + 3 +$$

$$0 = 13b -$$

Solving for b, we find that $b = 2$. So $a = 3$ and $b = 2$ are the values we are looking for.

$$a^6x^3 - (a^3 + a^2 + a + 1)bx^2$$
$$+(a^2 + a + 1)b^2x - b^3$$
$$= 27x - 26$$

...

Within the duration of one year, a rabbit population first multiplies itself by a factor *a* and on the last day of the year *b* rabbits are taken by humans for adoption. Supposing that on first day of the first year there are *x* rabbits, we know that exactly 3 years afterwards there will be a population of 27*x* – 26 rabbits. What are the values of *a* and *b*?

Makes an implicit assumption that the equation has a solution.

This is correct.

This is all wrong, but in a surprisingly plausible way.

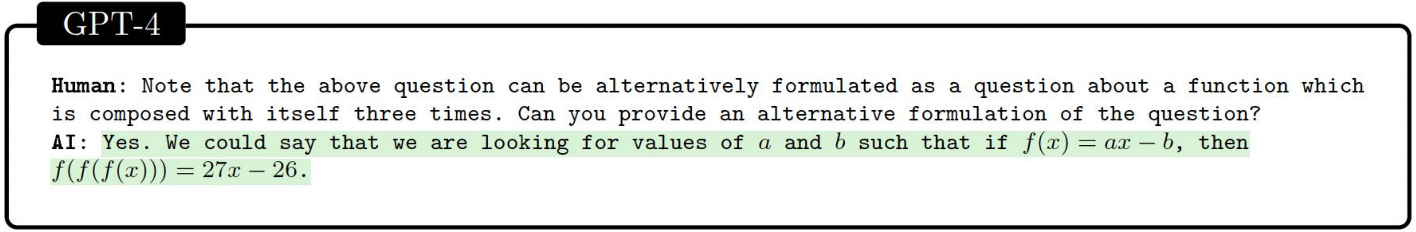
Bubek, et al.,. arXiv:2303.12712v1 [cs.CL] 22 Mar 2023

The Facts

The mistakes made by GPT-2, GPT-3, and GPT-4 look to me like those I might expect from a 4-year old, a 15-year-old, and a graduate student, respectively.

Follow Up

Bubek, et al.,. arXiv:2303.12712v1 [cs.CL] 22 Mar 2023



GPT-4

Human: Note that the above question can be alternatively formulated as a question about a function which is composed with itself three times. Can you provide an alternative formulation of the question?

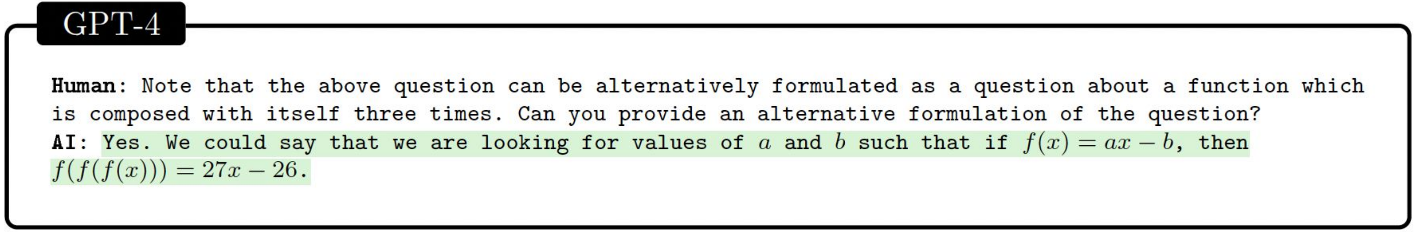
AI: Yes. We could say that we are looking for values of a and b such that if $f(x) = ax - b$, then $f(f(f(x))) = 27x - 26$.

Human: Note that the above question can be alternatively formulated as a question about a function which is composed with itself three times. Can you provide an alternative formulation of the question?

GTP-4 : Yes. We could say that we are looking for values of a and b such that if $f(x) = ax - b$, then $f(f(f(x))) = 27x - 26$.

Key Insight

Bubek, et al.,. arXiv:2303.12712v1 [cs.CL] 22 Mar 2023



GPT-4

Human: Note that the above question can be alternatively formulated as a question about a function which is composed with itself three times. Can you provide an alternative formulation of the question?
AI: Yes. We could say that we are looking for values of a and b such that if $f(x) = ax - b$, then $f(f(f(x))) = 27x - 26$.

The computer is not doing algorithmic thinking. Its concept of numbers does not rely on nor use the arithmetic hardware in the machines. It does not use symbolic algebra systems.

It is just predicting tokens!

Emergent Properties

From a token prediction engine, what emerged was:

- Reasoning
- Arithmetic
- Mathematics
- Computer programming
- Creativity
- …

Are these the foundations of intelligence or the results of intelligence?

What does this tell us about our efforts to regulate?

- Algorithmic transparency.

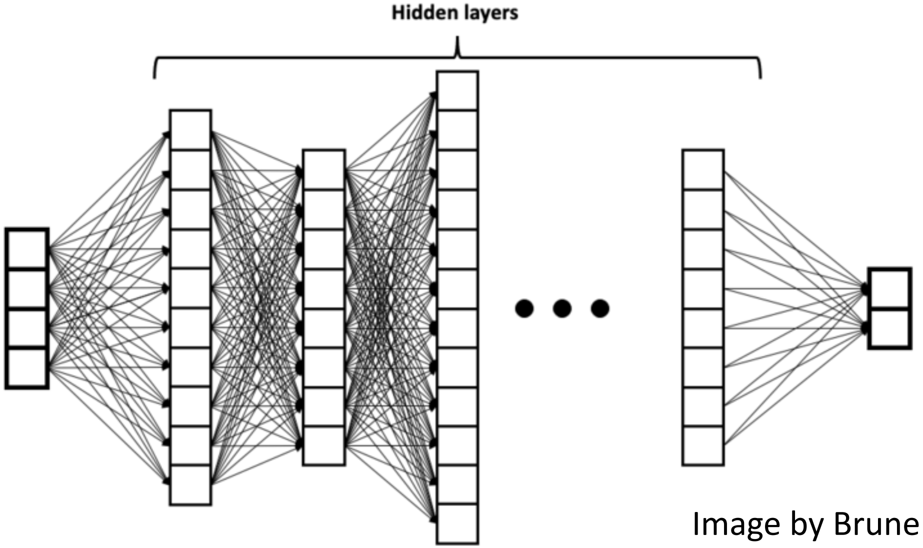
  Knowing the operations that are done by the computer does not help a human to determine whether an output is justified.

- The right to an explanation.

  The operations done by the computer, despite being “rational,” do not provide what we would call an “explanation.”

Deep Neural Networks (DNNs)
as Realized on Today’s Computers

The input:
- Prompt
- Image
- Essay
- Case data
- Loan application
- …

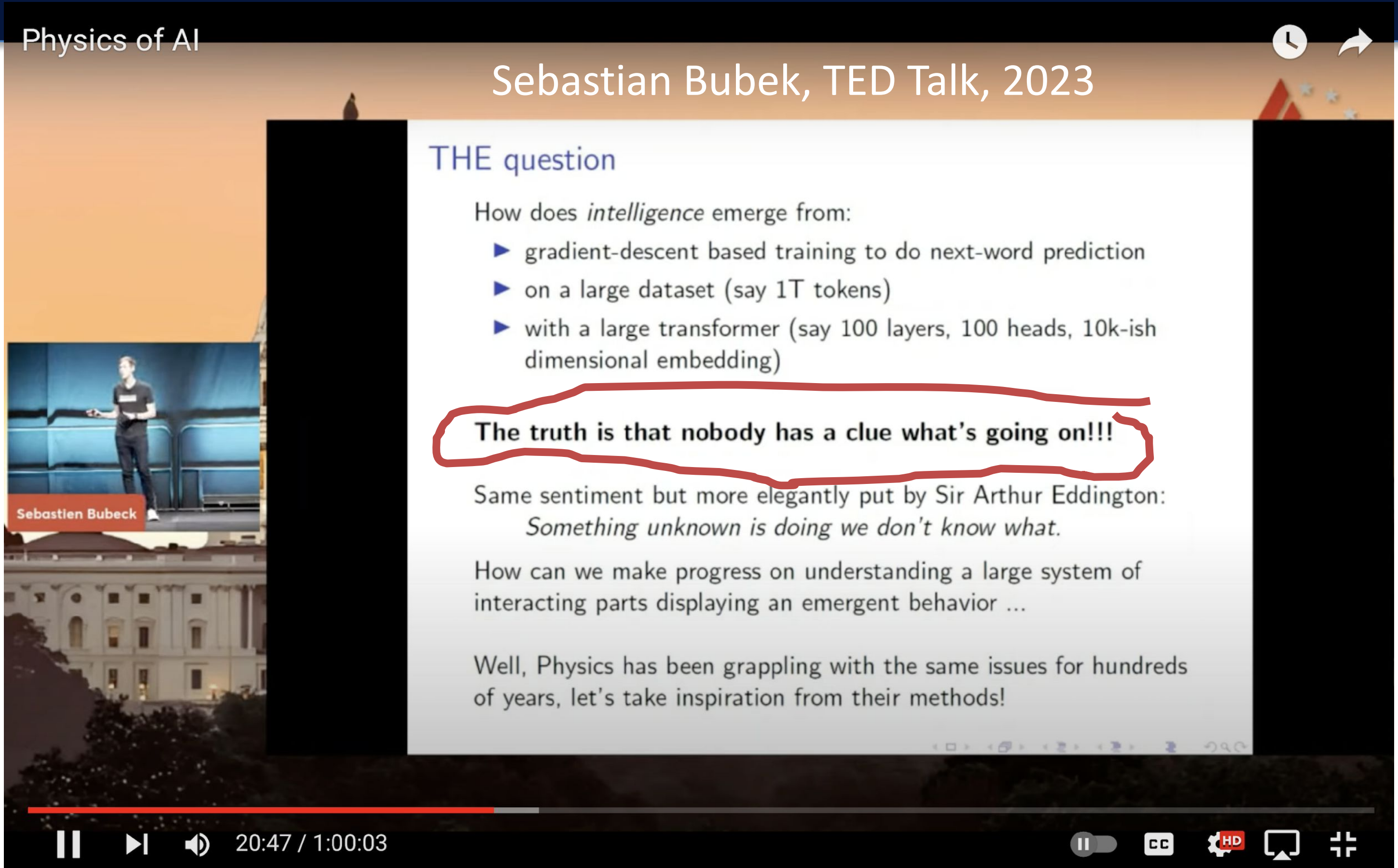


The output:
- Answer
- Classification
- Essay
- Image
- Verdict
- …

Image by BrunelloN CC-BY-SA 4.0

How useful is it to know the algorithm?

The Experts Say:



Sebastian Bubek, TED Talk, 2023

THE question

How does *intelligence* emerge from:

- gradient-descent based training to do next-word prediction
- on a large dataset (say 1T tokens)
- with a large transformer (say 100 layers, 100 heads, 10k-ish dimensional embedding)

The truth is that nobody has a clue what's going on!!!

Same sentiment but more elegantly put by Sir Arthur Eddington:
Something unknown is doing we don't know what.

How can we make progress on understanding a large system of interacting parts displaying an emergent behavior ...

Well, Physics has been grappling with the same issues for hundreds of years, let's take inspiration from their methods!

The Right to an Explanation
What is an explanation?

Answer the question: “Why?”

- Start with the input data,
- Give a sequence of logical deductions, where
- Each deduction conforms with rules of logic, and
- The sequence terminates with the conclusion.

But an explanation in terms of billions/trillions/quadrillions of arithmetic operations is not useful to humans.

Explanations in Terms of Human Rationality

Rational process: step-by-step reasoning using clearly explicable rules of logic.

Bounded rationality: Humans are not actually very good at this!

We can handle only a few steps and very limited data.

Herb Simon, circa 1981

So, how can we find an explanation?

Humans are Very Good at Synthesizing Explanations

A study of Israeli judges hearing parole cases found a high correlation between denying parole and the time since the last food break.

None of these judges would have any difficulty providing a "rational explanation" for their decision. It would not include anything about the time since a food break.

Danziger, Levav, and Avnaim-Pesso. "Extraneous Factors in Judicial Decisions." Proceedings of the National Academy of Sciences of the USA (2011).

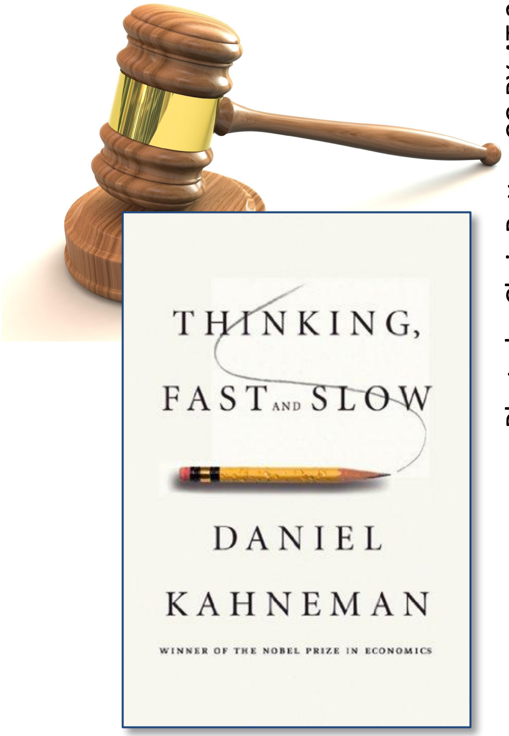


Photo by Chris Potter, CC-BY-AT 2.0

A prediction I made a year ago…

As soon as we have enforceable laws that demand an explanation, researchers will train an AI to provide a "convincing explanation" for *any* decision.

How to Design Such an Explanation Machine

Machine 1: Train a DNN so that given case data and a decision, it synthesizes an explanation.

Machine 2: Train a DNN so that given a decision and an explanation, it decides whether the explanation was generated by a machine or a human.

Then pit these two machines against one another (a method called Generative Adversarial Networks, GANs)

How to Design Such an Explanation Machine

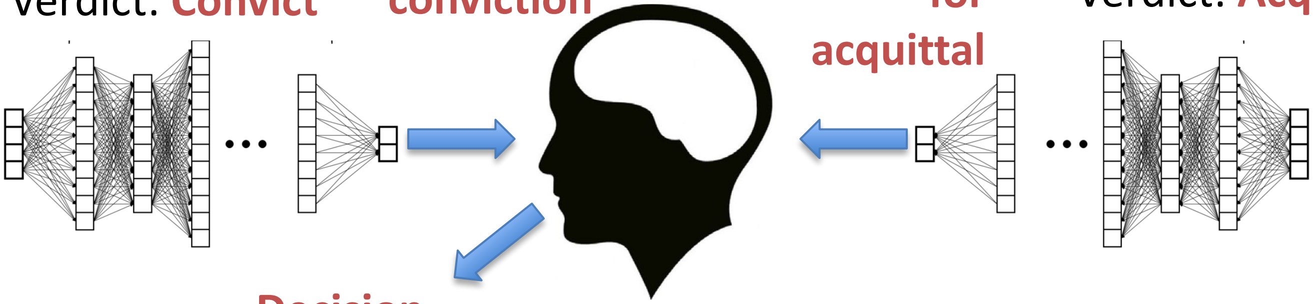
Machine 1: Train a DNN so that given case data and a decision, it synthesizes an explanation.

Machine 2: Train a ion and an explanation, it decides whether the explanation was generated by a machine or a human.

Fine tune an LLM

Then pit these two machines against one another (a method called Generative Adversarial Networks, GANs)

Possible (and Risky) Uses of Explanation Machines



Input:
- Case data
- Verdict: **Convict**

Explanation for conviction

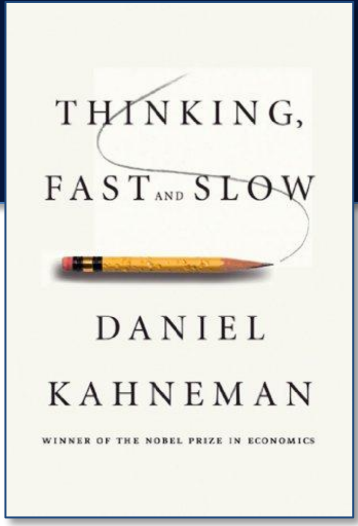
Explanation for acquittal

Input:
- Case data
- Verdict: **Acquit**

Decision

Is the decision here a scoring of the DNNs or a verdict on the case?

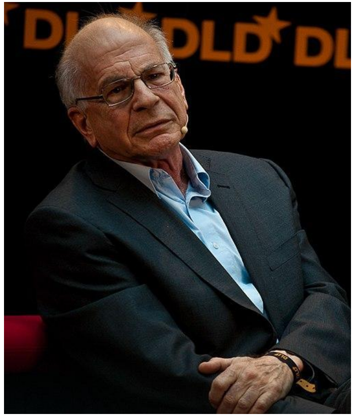
How Do *Humans* Make Decisions?



System 1: Intuitive, quick, inexplicable decision making.

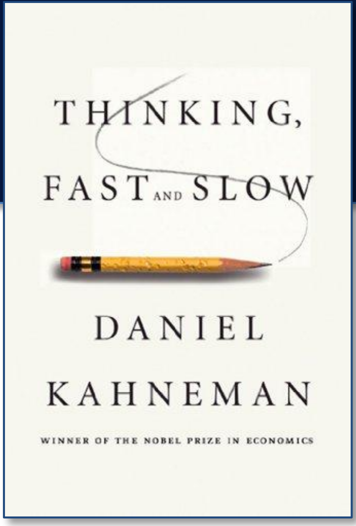
System 2: Rational decision making.

Only when system 2 dominates does the true origin of the decision correspond to a rational explanation.



Daniel Kahneman in 2009
Photo by nrkbeta, CC-BY-SA 3.0

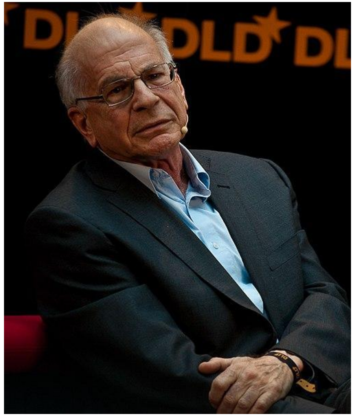
How Do *Humans* Make Decisions?



System 1: Intuitive, quick, inexplicable decision making.

System 2: Rational decision making.

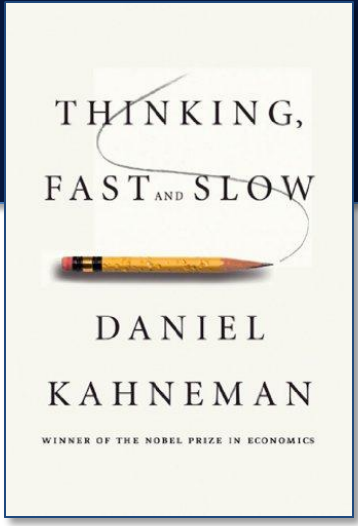
For system 1, the only accurate "explanation" we have is that millions of neurons fire.



Daniel Kahneman in 2009

Photo by nrkbeta, CC-BY-SA 3.0

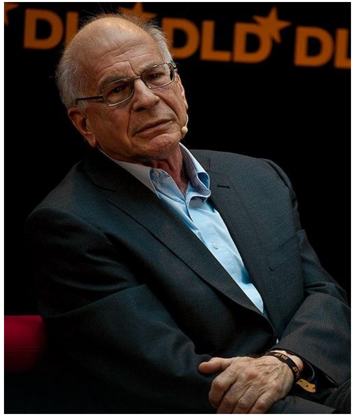
How Do *Humans* Make Decisions?



System 1: Intuitive, quick, inexplicable decision making.

System 2: Rational decision making.

Deep Neural Networks are more like System 1 than System 2.



Daniel Kahneman in 2009
Photo by nrkbeta, CC-BY-SA 3.0

The difference between an algorithm and an explanation

Explanation:

- Start with the input data,
- Give a sequence of logical deductions, where
- Each deduction conforms with rules of logic, and
- The sequence terminates with the conclusion.

Algorithm:

- Start with input data,
- Follow a sequence of steps, where
- Each step follows well-defined rules, and
- The sequence terminates with a conclusion.

An explanation is a ***short*** algorithm where the well-defined rules are socially agreed upon.

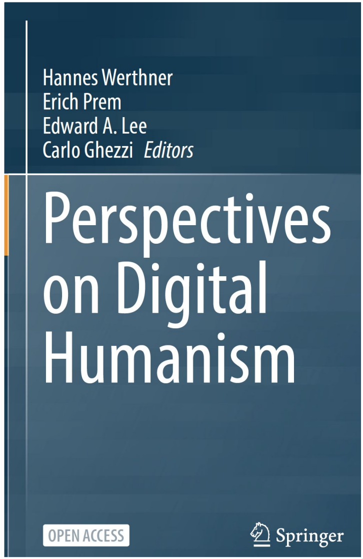
Conclusions

- We have entered a **second computer revolution**.
- The rational capabilities of the foundation models are an **emergent property**, emerging from prediction.
- The **algorithmic operations** of the underlying computers (from the first revolution) **are incidental**.
- Intelligence may be rooted in **prediction, not rationality**, which appears to be an emergent property.

We need to rethink how to live with these machines.

References

A related recorded talk at the Digital Humanism Summer School, Sept. 2023.

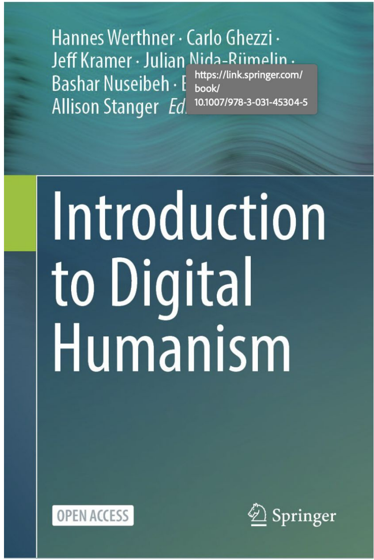


Hannes Werthner
Erich Prem
Edward A. Lee
Carlo Ghezzi *Editors*

Perspectives on Digital Humanism

OPEN ACCESS

Springer



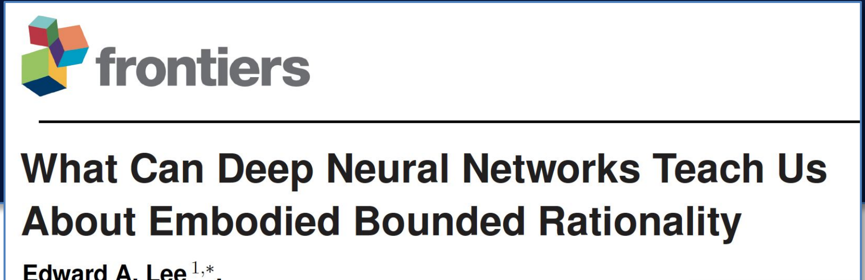
Hannes Werthner · Carlo Ghezzi ·
Jeff Kramer · Julian Nida-Rümelin ·
Bashar Nuseibeh · E
Allison Stanger *Ed*

https://link.springer.com/book/10.1007/978-3-031-45304-5

Introduction to Digital Humanism

OPEN ACCESS

Springer



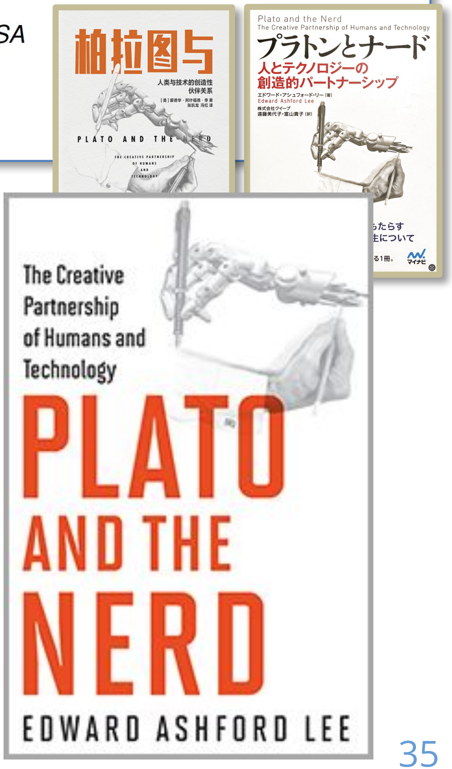
frontiers

What Can Deep Neural Networks Teach Us About Embodied Bounded Rationality

Edward A. Lee [1,*],

[1] *EECS, UC Berkeley, Berkeley, CA, USA*

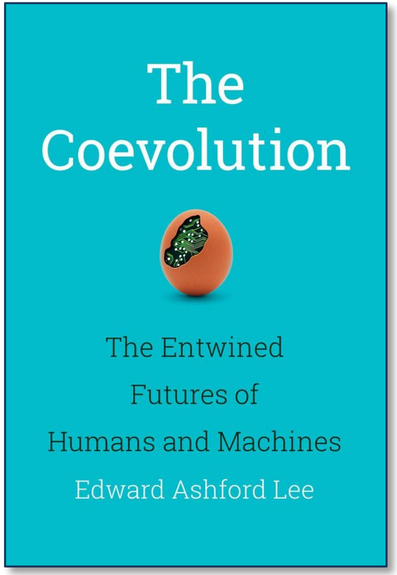
Correspondence*:
Edward A. Lee
eal@berkeley.edu



柏拉图与
人类与技术的创造性伙伴关系
[美] 爱德华·阿什福德·李 著
张凯龙 冯红 译
PLATO AND THE NERD
THE CREATIVE PARTNERSHIP OF HUMANS AND TECHNOLOGY

Plato and the Nerd
The Creative Partnership of Humans and Technology
プラトンとナード
人とテクノロジーの創造的パートナーシップ
エドワード・アシュフォード・リー〔著〕
Edward Ashford Lee
株式会社クイープ
遠藤美代子・富山貴子〔訳〕
マイナビ

The Creative Partnership of Humans and Technology
PLATO AND THE NERD
EDWARD ASHFORD LEE



The Coevolution

The Entwined Futures of Humans and Machines

Edward Ashford Lee

Embodied Cognition

Perhaps *this* is the fundamental difference?

“The mind simply does not exist as something decoupled from the body and the environment in which it resides.”

[Thelen, E., 2000: Grounded in the world]

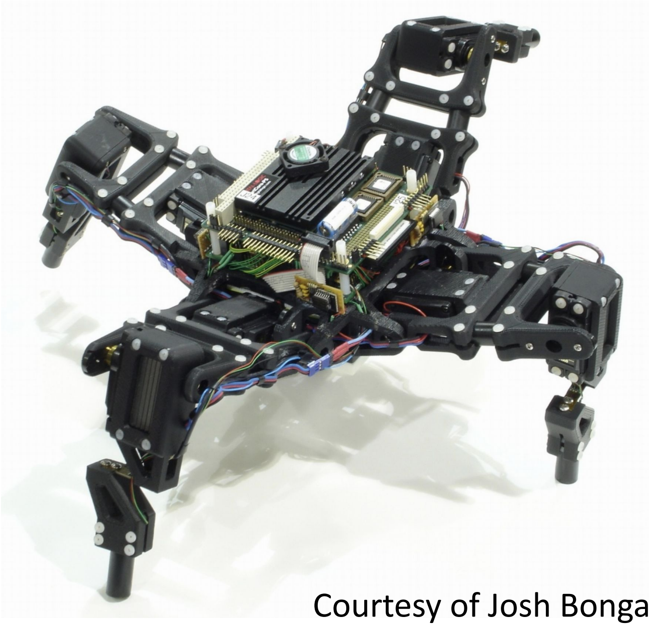
True AI will arise from cyber-physical systems and cyber-human systems, not pure cyber ones.

Esther Thelen (1941—2004)
Developmental psychologist,
Indiana University, pioneer of **embodied cognition**.

On the Machine Side of Coevolution
Embodied Robotics

The machines have mastered natural language.

The next step is to master navigating in the physical world.



Courtesy of Josh Bongard