

# Deep Neural Networks, Explanations, and Rationality

*Edward A. Lee*

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

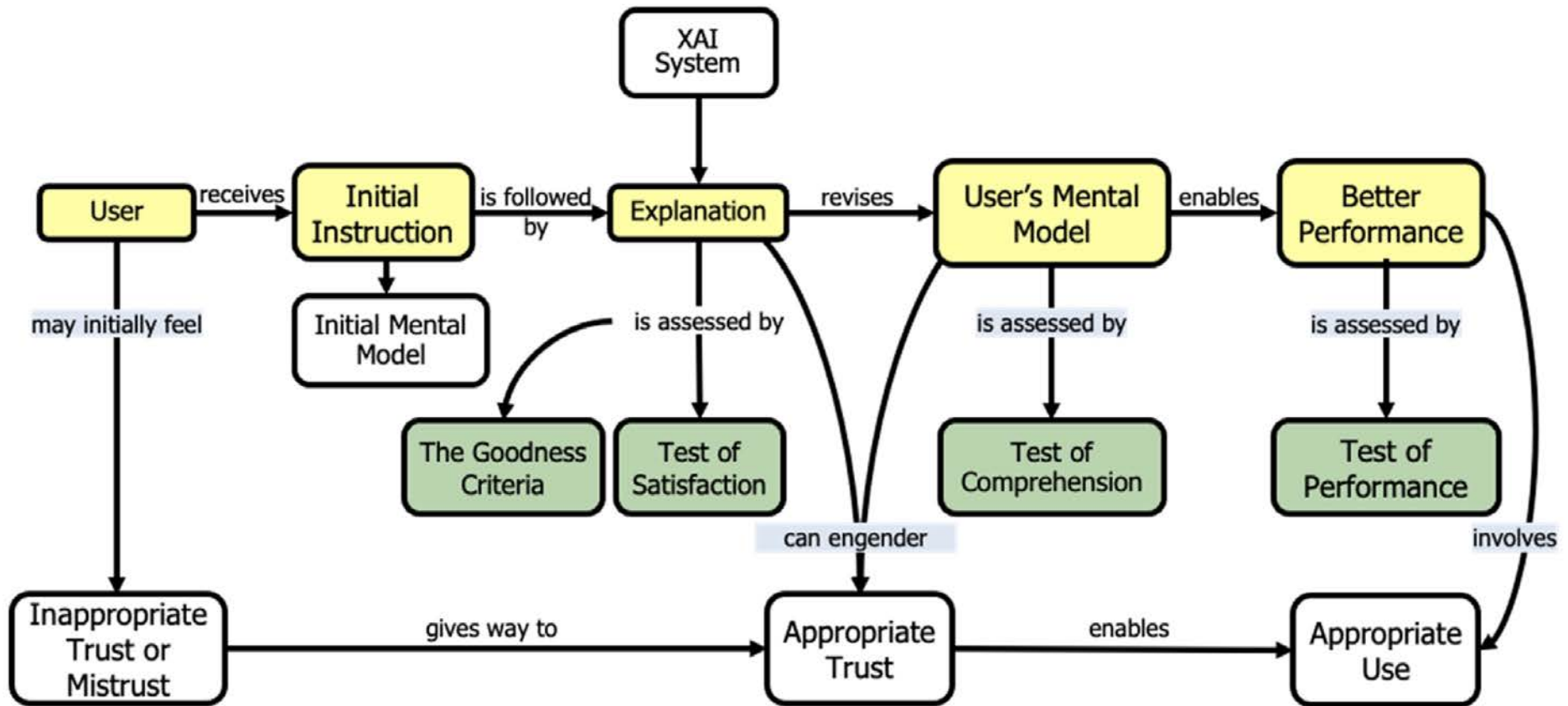
*2nd ACM Digital Humanism Summer School  
TU Wien, Faculty of Informatics, Vienna, Austria*

*September 6, 2023*



**University of California at Berkeley**

# XAI – Explainable AI



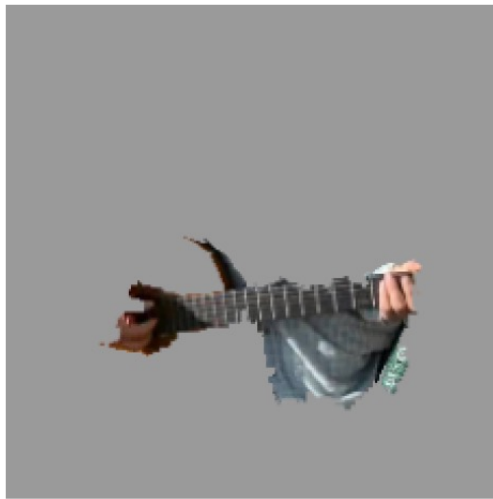
## Psychological model of explanation

From Gunning, Vorm, Wang, and Turek, "DARPA's explainable AI (XAI) program: A retrospective," Applied AI Letters, 2021.

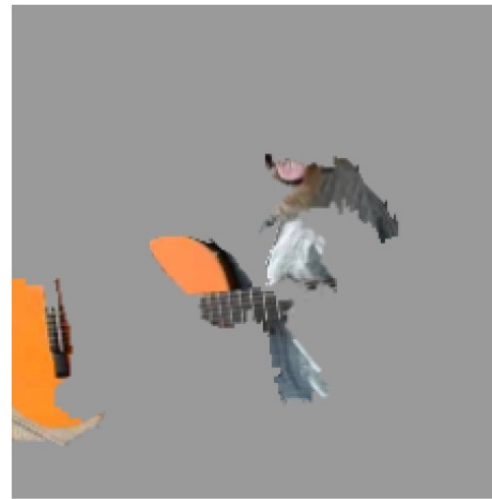
# State of the Art in 2016



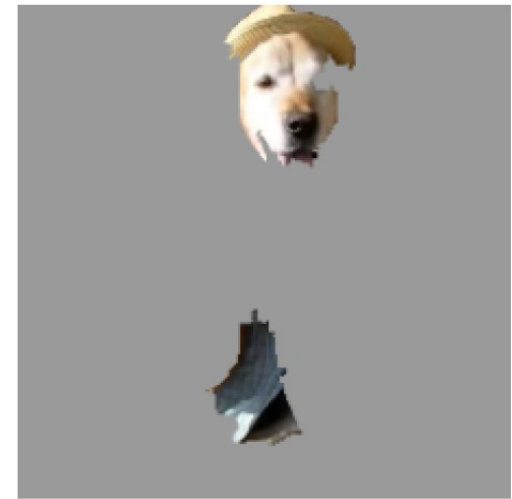
(a) Original Image



(b) Explaining *Electric guitar*



(c) Explaining *Acoustic guitar*

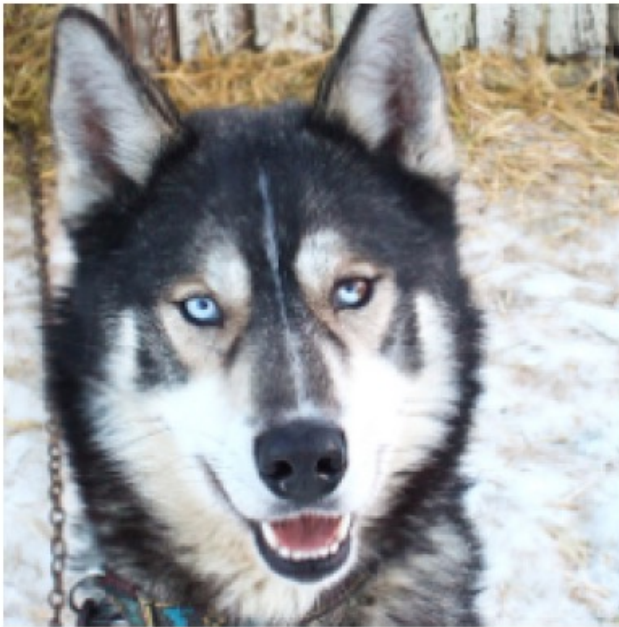


(d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ( $p = 0.32$ ), "Acoustic guitar" ( $p = 0.24$ ) and "Labrador" ( $p = 0.21$ )

Ribeiro, Singh, and Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Knowledge Discovery and Data Mining*, 2016

# Explanations are particularly useful when the AI is incorrect



(a) Husky classified as wolf



(b) Explanation

**Figure 11:** Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Ribeiro, Singh, and Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Knowledge Discovery and Data Mining*, 2016

# Intelligible vs. Accurate

Intelligible models (e.g. rule-based learning), while less accurate than neural networks, can reveal troubling patterns in the training dataset.

Should we sacrifice intelligibility for higher performance?

Caruana, Lou, Koch, Sturm, Gehrke, and Elhadad, “Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission,” *Knowledge Discovery and Data Mining (KDD)*, 2015,

# Fast Forward to LLMs

LLMs:

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



# Dismissals

- “It doesn’t understand like we do.”
- “It makes stuff up.”
- “It makes errors.”

# Dismissals

“It doesn’t reason and use language like humans do.”

“We know from the science of linguistics and the philosophy of knowledge that they differ profoundly from how humans reason and use language.”

The New York Times

<https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>

GUEST ESSAY

## Noam Chomsky: The False Promise of ChatGPT

March 8, 2023

By Noam Chomsky, Ian Roberts and Jeffrey Watumull



# Dismissals

“It is plagiarism, copying from the internet”

EA

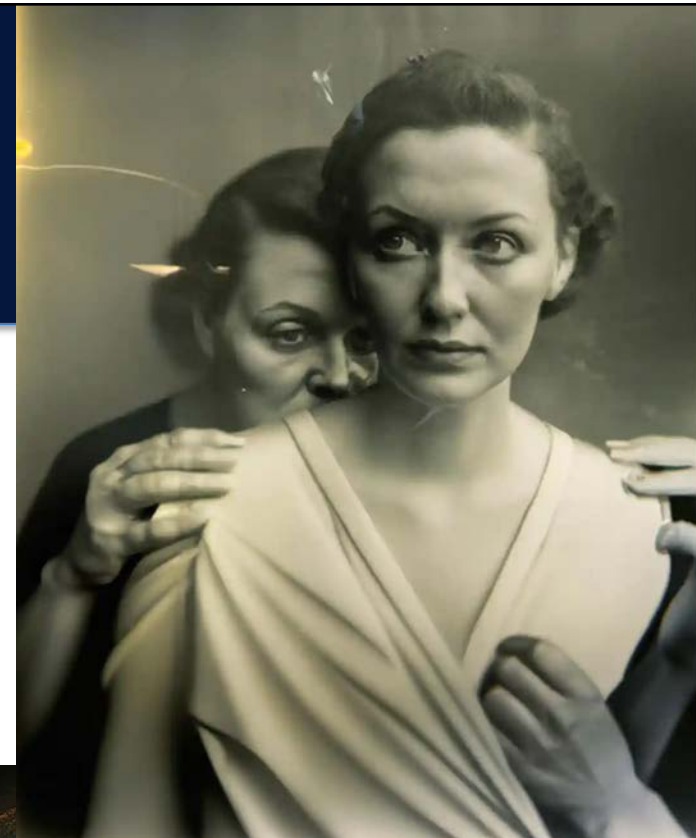
Please express the following sentiment in the style of a pirate: Today's large-language-model AIs are merely repeating content they have harvested from the internet and are therefore plagiarizing material written by humans.



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. Arrrrr!

# The Facts

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



- *Everyone has been surprised.*
- AI experts are less surprised to be surprised
- Emergent properties:
  - Reasoning
  - Arithmetic
  - Mathematics
  - Programming
  - Creativity
  - ...



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

# Can We Address the Fears?

- Algorithmic transparency.

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

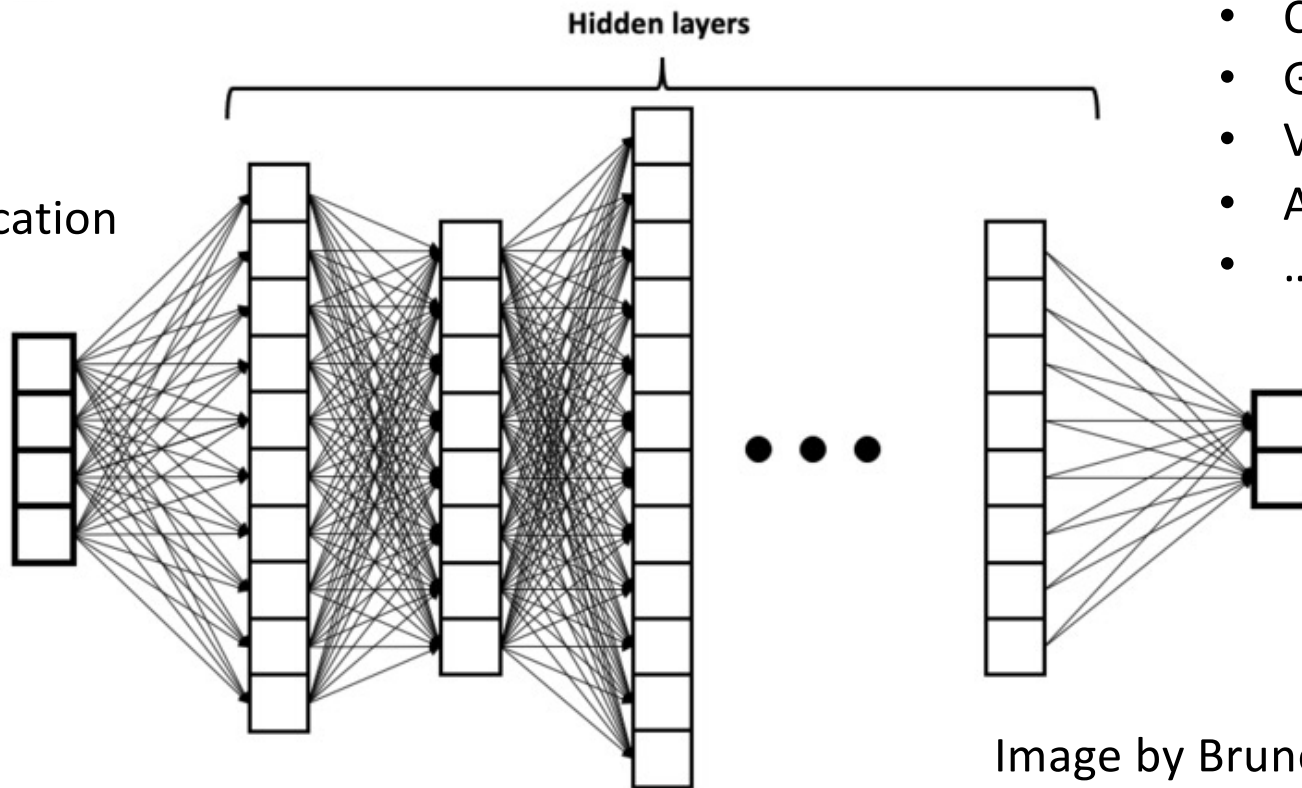
- The right to an explanation.

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:

- Image
- Essay
- Case data
- Loan application
- ...



## The output:

- Classification
- Grade
- Verdict
- Acceptance
- ...

Does explaining the operations explain the decision?

# 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 Rational Thought

**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.**



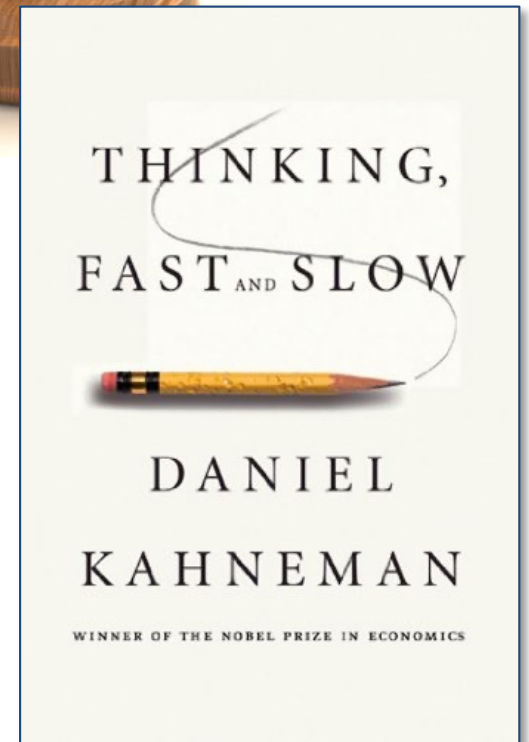
Herb Simon, circa 1981

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).



# 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:** T **Fine tune an LLM** 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)

# Possible (and Risky) Uses of Explanation Machines

## Input:

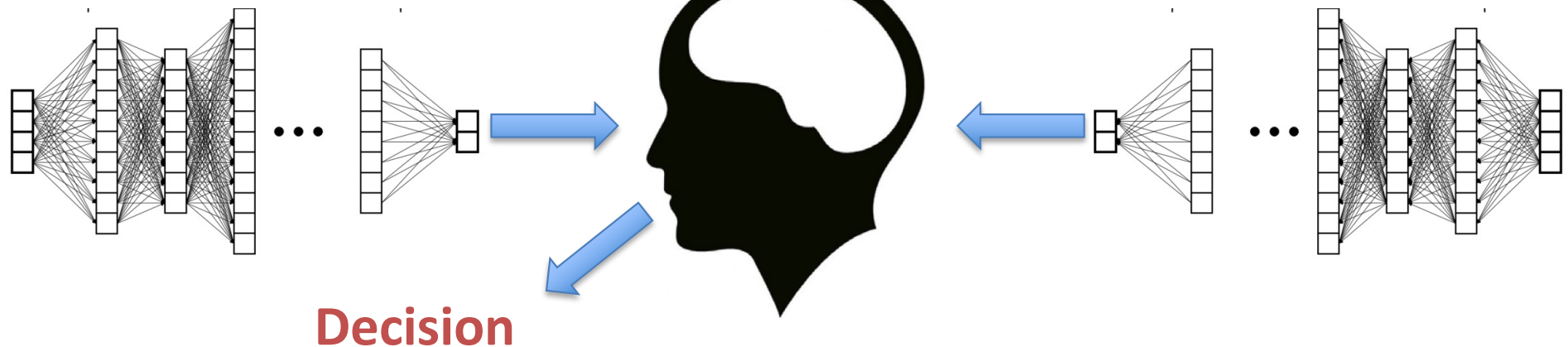
- Case data
- Verdict: **Convict**

**Explanation  
for  
conviction**

**Explanation  
for acquittal**

## Input:

- Case data
- Verdict: **Acquit**



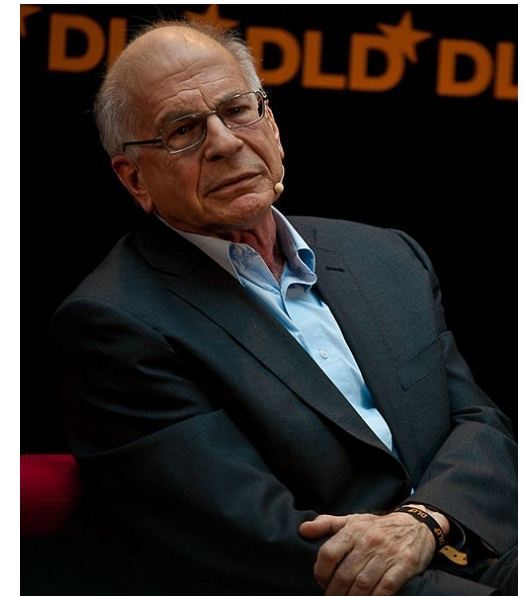
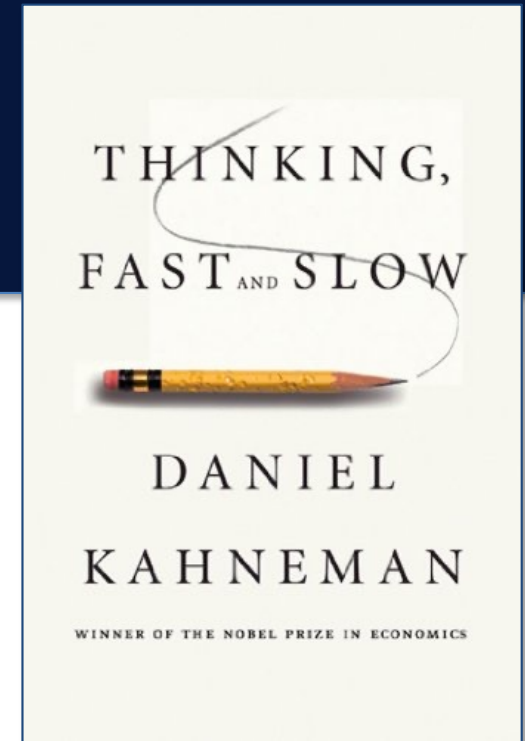
Is the decision here a scoring of the DNNs or a verdict on the case?

# 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.



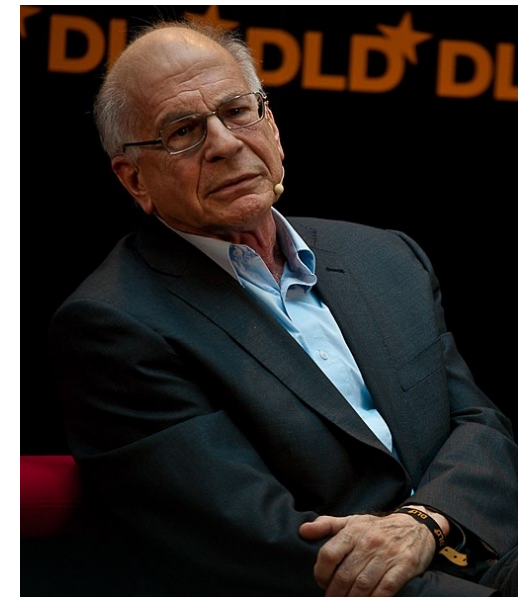
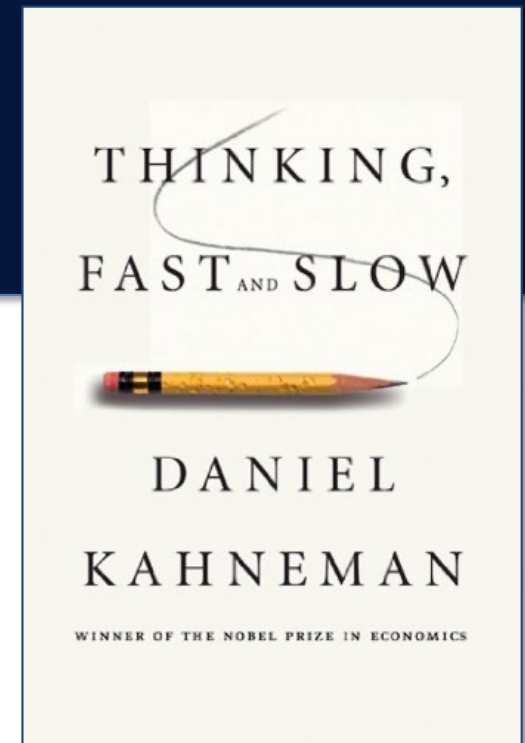
Daniel Kahneman in 2009

# 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.



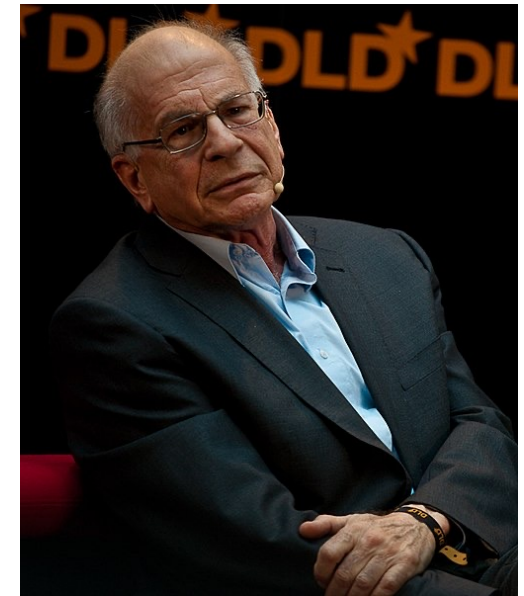
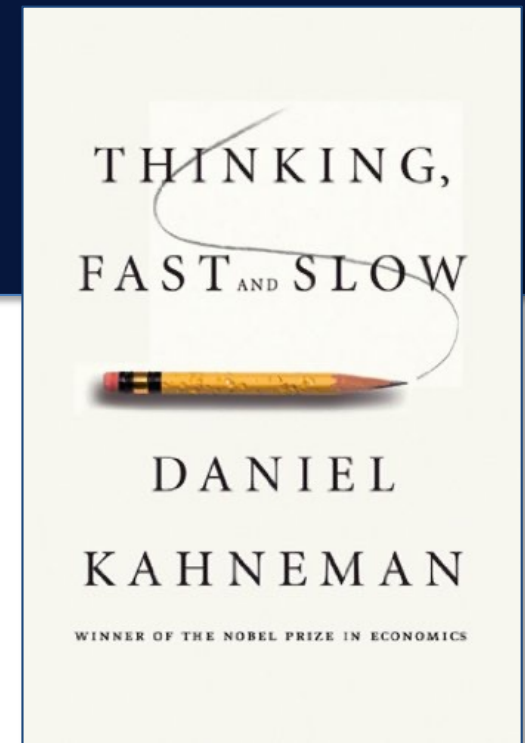
Daniel Kahneman in 2009  
Photo by nrkbeta, CC-BY-SA 3.0 22

# 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

# Rational Decision Making (System 2) Is Algorithmic

## What is an algorithm?

- 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.



Machine designed by Mike Davey



# 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.

# Is System 1 Algorithmic?

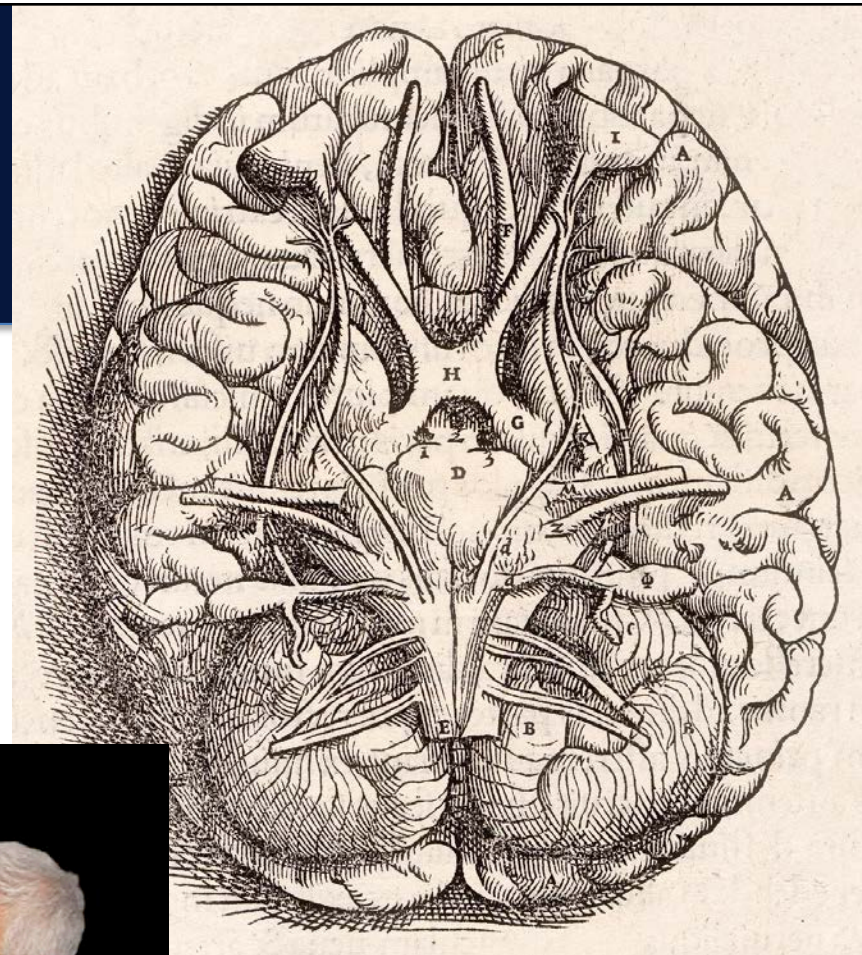
According to Simon and Kahneman, system 1 decisions do not arise from **rational processes**, step-by-step reasoning using clearly explicable rules of logic.

Evidence from LLMs is that rational processes emerge from system 1-like processes!!

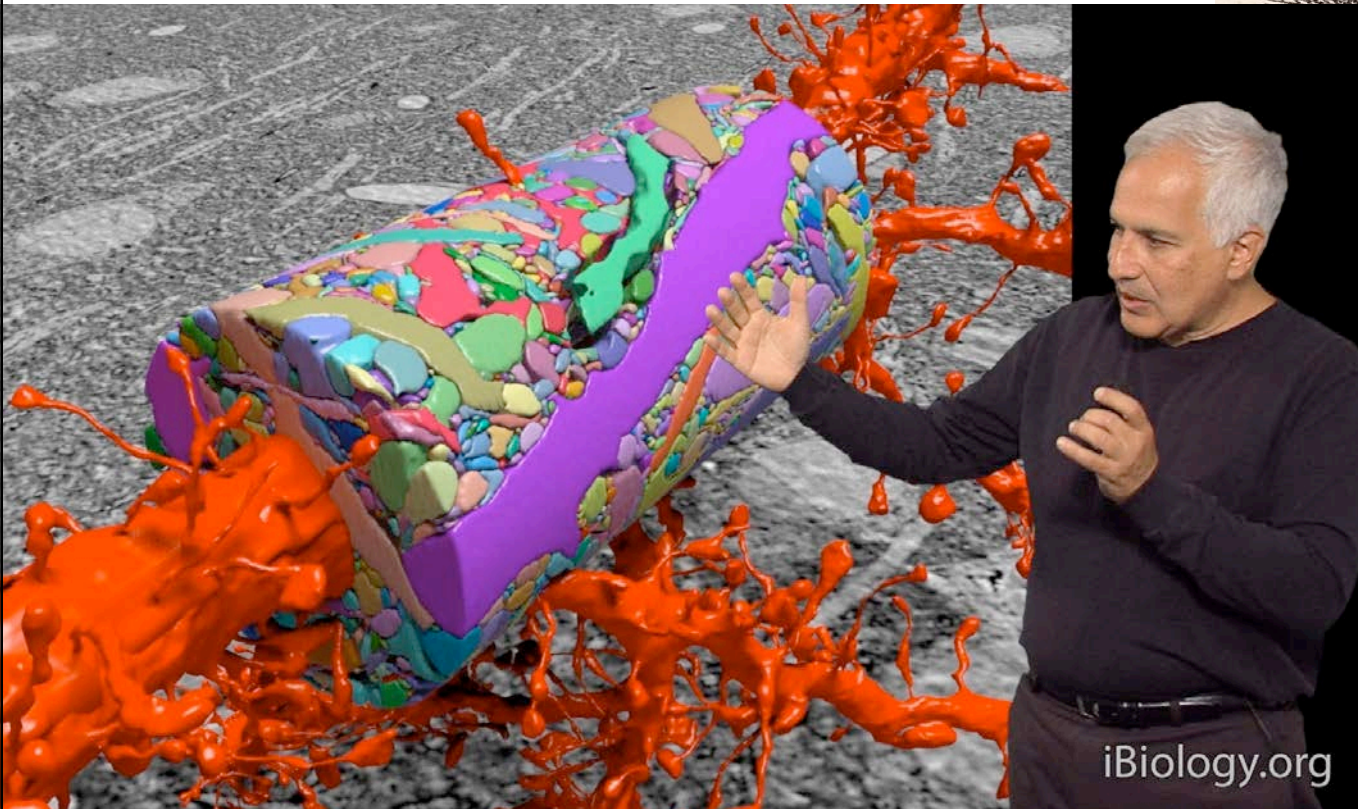
Are system 1 processes nevertheless algorithmic?

# Is the Brain Algorithmic?

Understanding brain function in terms of low-level neuron connections and neuron firing has proved elusive.



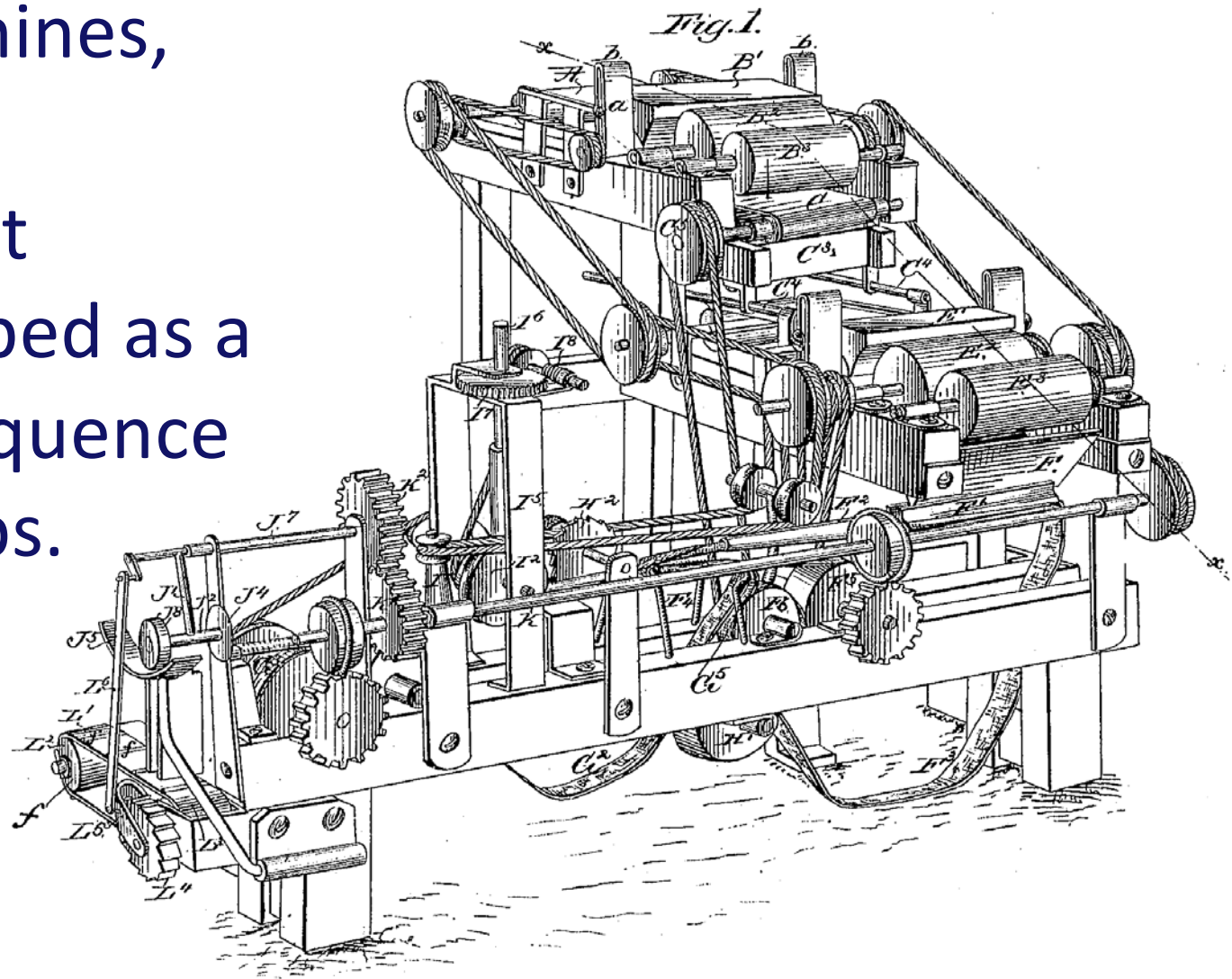
Andreas Vesalius' *Fabrica*,  
published in 1543



Jeff Lichtman,  
Harvard, 2015

# Many Machines Are Not Usefully Modeled by Algorithms.

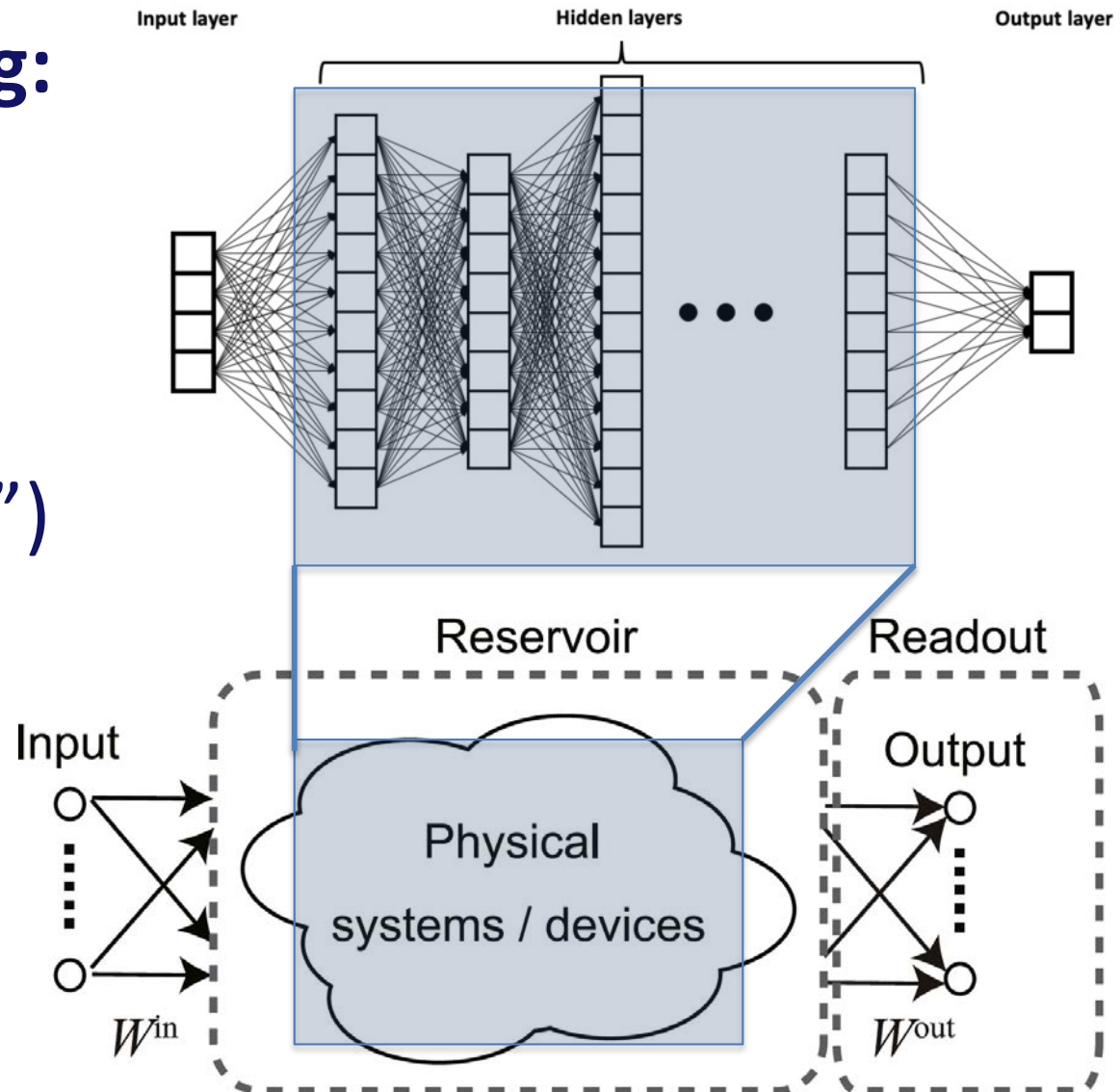
For many machines, their essential operation is not usefully described as a terminating sequence of discrete steps.



# Even DNNs may not be fundamentally algorithmic.

## Reservoir Computing:

Replace the intermediate layers with a fixed blob of physics (a “reservoir”)



Tanaka, et al., “Recent advances in physical reservoir computing: A review,” *Neural Networks*, 2019

# Reservoir Computing

Many very different devices have been shown to function effectively as reservoirs:

- Buckets of water
- Bundles of carbon nanotubes
- In-vitro cell cultures
- ...



Fernando and Sojakka, Pattern recognition in a bucket,” In European conference on artificial life, 2003.

**It may prove that today’s DNNs are brute-force algorithmic simulations of non-algorithmic processes.**

# Evidence from LLMs

An ability to reason can emerge from chaotic processes in a way that has little evident connection to the underlying logic of the processes.

# More ways to address the fears?

- Give the machines human-like intelligence.

Artificial General Intelligence (AGI) is about achieving human-like intelligence. Is that a good idea?



# Human-Like Intelligence

## The Ultimate Goal?

Several AI releases have flopped because they behaved in all-too-human ways:

- On March 23, 2016, Microsoft released a chatbot called **Tay**, an AI designed to interact with users on social media such as Twitter using the vernacular of hip youngsters using the media.
- **Tay did that very well. Too well. It quickly learned to write vulgar and racist tweets.**

Kastrenakes, "Microsoft Made a Chatbot That Tweets Like A Teen," *The Verge*, (2016).

Vincent, "Twitter taught Microsoft's AI chatbot to be a racist asshole," *The Verge*, (2016).

# 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 Berkeley graduate student, respectively.

## 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  $b$ ?

**GPT-4:** 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  $b$ ?

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

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

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

Makes an implicit assumption that the equation has a solution.

$$0 = 13b -$$

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

This is correct.

$$\begin{aligned} a^6 x^3 - (a^3 + a^2 + a + 1)bx^2 \\ + (a^2 + a + 1)b^2x - b^3 \\ = 27x - 26 \end{aligned}$$

...

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

# Sebastian Bubeck says:

Physics of AI



## 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!



20:47 / 1:00:03



# Robot Psychology?

Robot psychology, also known as artificial intelligence psychology or machine psychology, is the study of the cognitive and emotional processes of artificial intelligence and robots...

-- ChatGPT, responding to “What is robot psychology?”  
January 5, 2023

The scientific methods of psychology may be the best we have for trying to understand LLNs!

# More ways to address the fears?

- Give the machines human-like intelligence.

Artificial General Intelligence (AGI) is about achieving human-like intelligence. Is that a good idea?

- Keep humans in the loop in any decision.

As we have already learned the hard way, humans are easily manipulated by the machines.

# Machines Manipulate Humans

- Today's **information flood** makes the use of AIs to filter information both *necessary* and *inevitable*.
- The **information economy** leads AIs to filter to manipulate humans.
- Human decision making is *no longer independent of the machines*.

**Artificial Intelligence and the Problem of Control**

Stuart Russell

**The Attention Economy and the Impact of Artificial Intelligence**

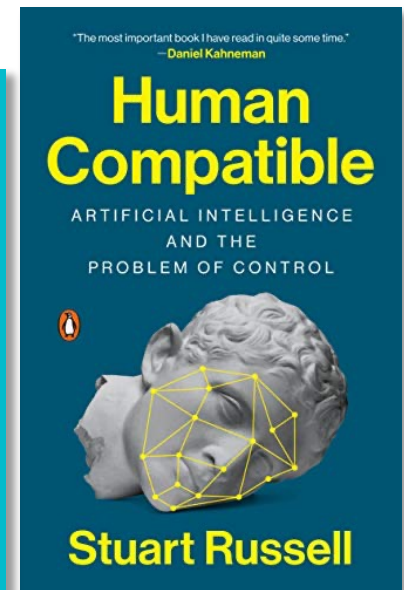
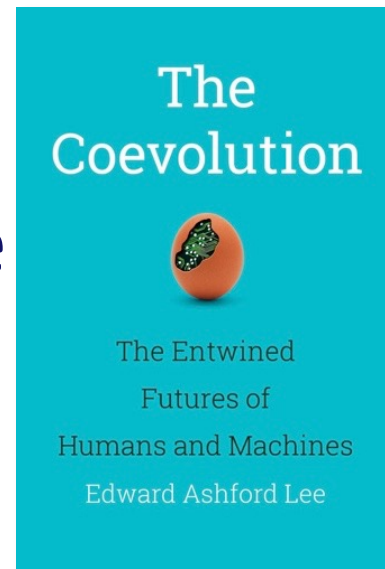
Ricardo Baeza-Yates and Usama M. Fayyad



# Machines Manipulate Humans

This is not just a prediction for the future.

- Predictable humans can be reliably presented with advertisements they will click on.
- Political extremists are more predictable than moderates.
- Feedback: the machines make you more predictable so that their predictions are more accurate.
- The result is an **information apocalypse**, where humans live on **islands of disjoint truths**.





# Are we in control?

What do we mean by “we”?

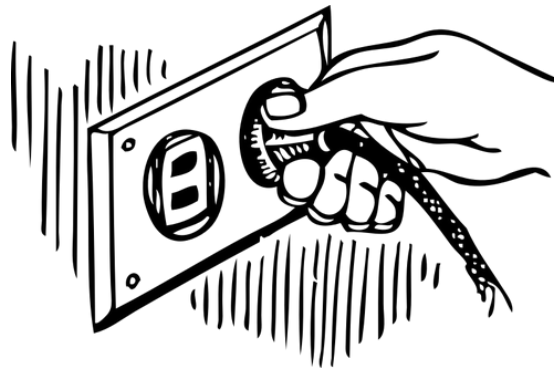
- Humanity? All 8 billion?
- Engineers? Programmers? CEOs? Users?



# Are we in control?

What do we mean by “control”?

- We can pull the plug?
- We design the systems? We choose to use them?



Can we pull the plug? What would happen today if we had to shut down all the computers on the planet?

# The Question, Clarified

Do the engineers, programmers, business leaders, or users determine the outcome of technology development?

# Digital Creationism vs. Coevolution

Vasa



## The Coevolution



The Entwined  
Futures of  
Humans and Machines  
Edward Ashford Lee

# Digital Creationism vs. Coevolution

Vasa



“Every boat is copied from another boat ... Let’s reason as follows in the manner of Darwin. It is clear that a very badly made boat will end up at the bottom after one or two voyages and thus never be copied. ... One could then say, with complete rigor, that it is the sea herself who fashions the boats, choosing those which function and destroying the others.”

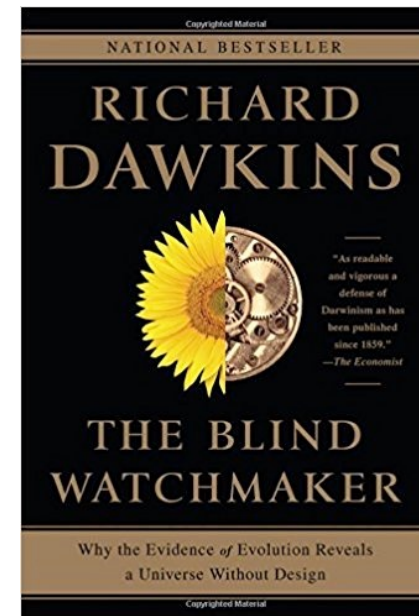
French philosopher Alain

# Digital Creationism vs. Coevolution



Evolutionary processes are capable of much more complex and interconnected systems than top-down intelligent design.

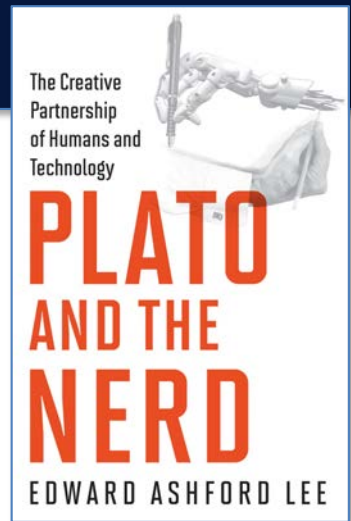
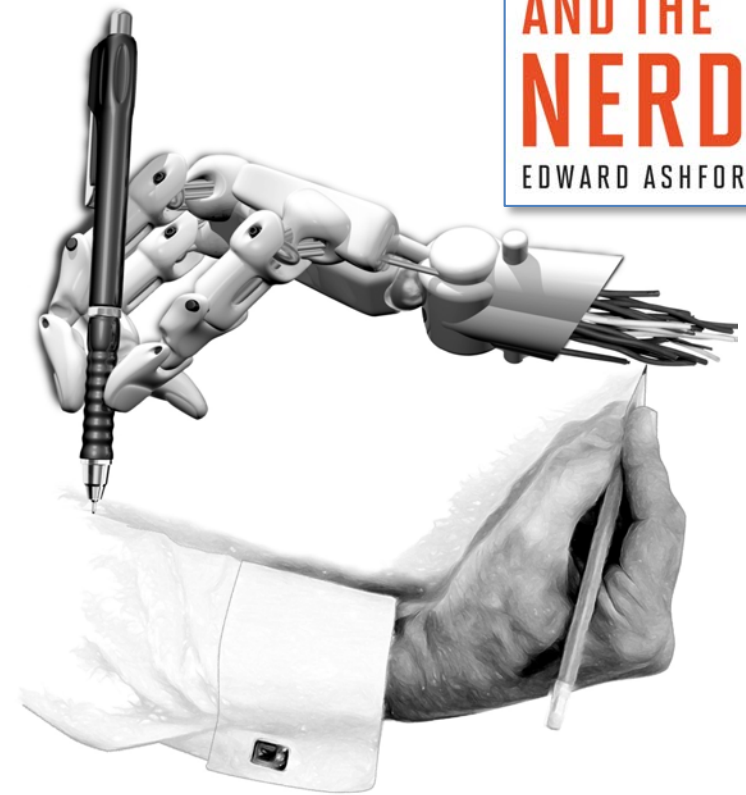
Donan.raven [CC BY-SA 3.0]



# Coevolution

“Are we playing God, creating a new life form in our own image, or are we being played by a Darwinian evolution of a symbiotic new species?”

“Are humans the purveyors of the ‘noisy channel’ of mutation, facilitating sex between software beings by recombining and mutating programs into new ones?”



# On the Human Side of Coevolution: Intellectual Prostheses





# 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]

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

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

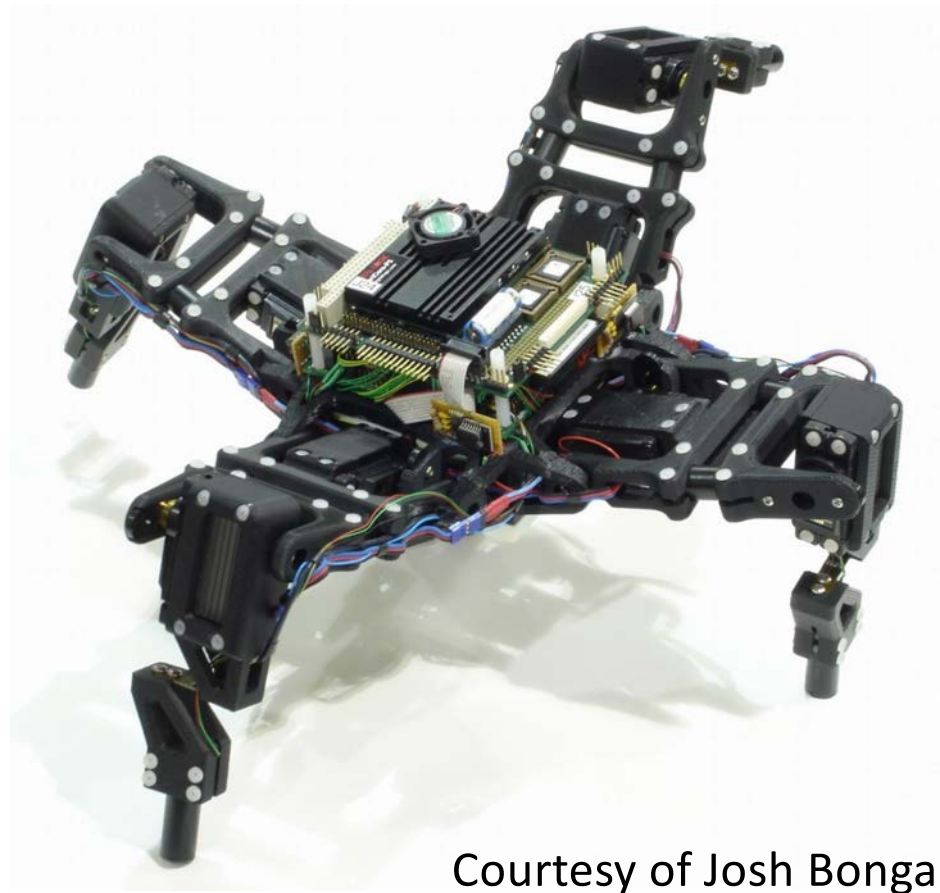


# On the Machine Side of Coevolution

## Embodied Robotics

The machines have mastered natural language.

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

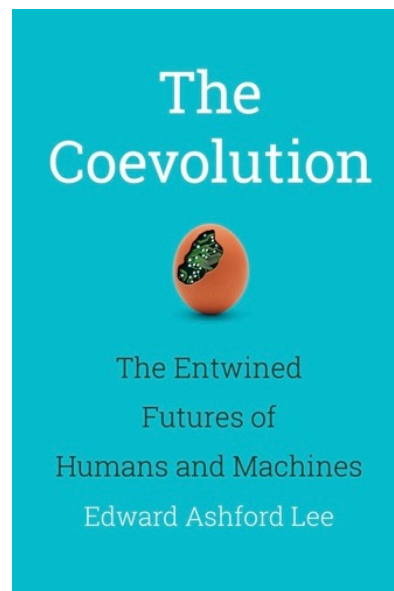


Courtesy of Josh Bongard

# Interaction, Embodiment, Feedback, and Subjectivity

- Interaction is more powerful than observation.
- Interaction requires a *first-person self*.
- A *first-person self* requires interaction.

Feedback, interaction, causation, free will, and accountability are all tied together.



Silvio Micali



Robin Milner



Shafi Goldwasser Judea Pearl

# The Curse of Recursion

What happens when most of the data used to train LLMs was created by LLMs?

Shumailov, Shumaylov, Zhao, Gal, Papernot, and Anderson, “The Curse of Recursion: Training on Generated Data Makes Models Forget,” *arXiv:2305.17493v2 [cs.LG]*, May 31 2023.

# Opportunities

- Licensing or certification of AIs?
- Use AIs to expose human biases?
- Better recommender systems?
- Use explanation machines to explain *all* possible decisions?
- Use AIs to expose deliberate abuse of information filtering?
- Use AIs to teach?
- ... multidisciplinary engagement ...

