Limits of Machines, Limits of Humans

Edward A. Lee

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

Digital Humanism Fellow, Institut für die Wissenschaften vom Menschen (IWM), Vienna

Visiting Professor, TU Wien, Vienna

Public Lecture

Technical University of Vienna

May 24, 2022, Vienna, Austria

University of California at Berkeley

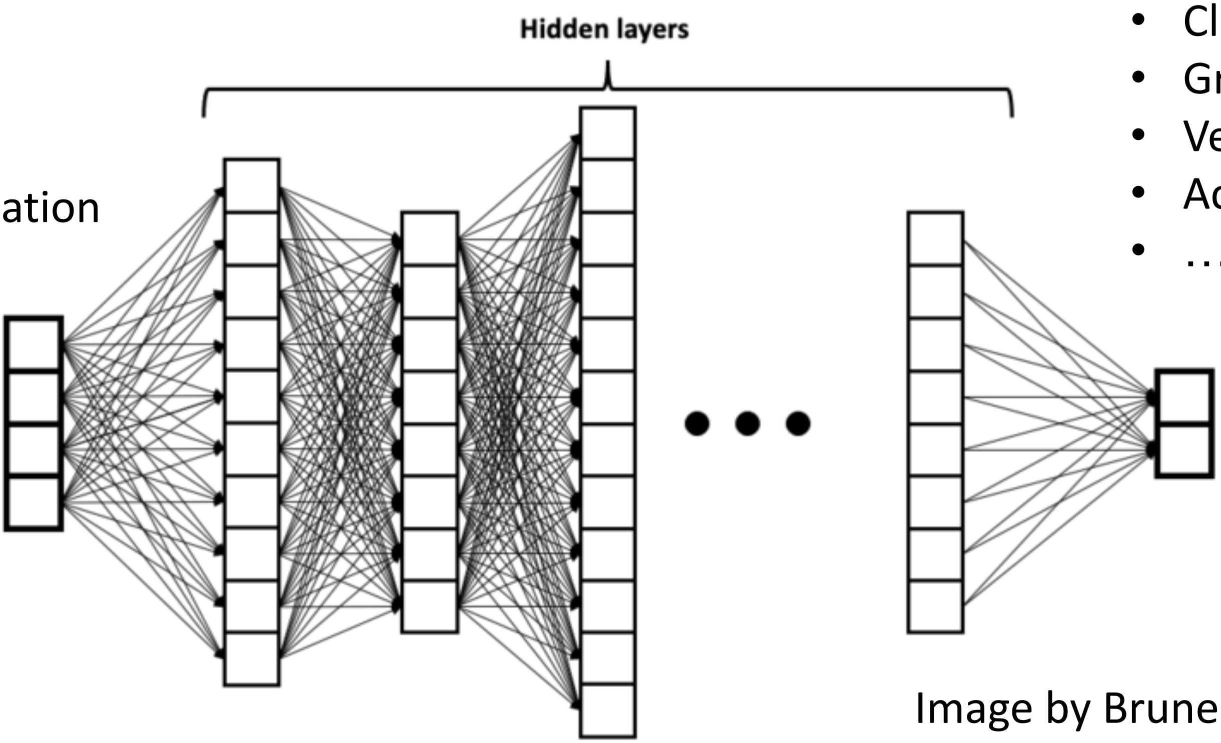
Abstract

"Rationality" in Simon's "bounded rationality" is the principle that humans make decisions based on step-by-step (algorithmic) reasoning using systematic rules of logic to maximize utility. "Bounded rationality" is the observation that the ability of a human brain to handle algorithmic complexity and data is limited. Bounded rationality, in other words, treats a decision maker as a machine carrying out computations with limited resources. In this talk, Edward A. Lee will argue that the recent breakthroughs in AI demonstrate that much of what we consider "intelligence" is not based on algorithmic symbol manipulation, and that what the machines are doing more closely resembles intuitive thinking than rational decision making. Under this model, the goal of "explainable AI" is unachievable in any useful form.

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

The input:

- Image
- Essay
- Case data
- Loan application
- …



The output:

- Classification
- Grade
- Verdict
- Acceptance
- …

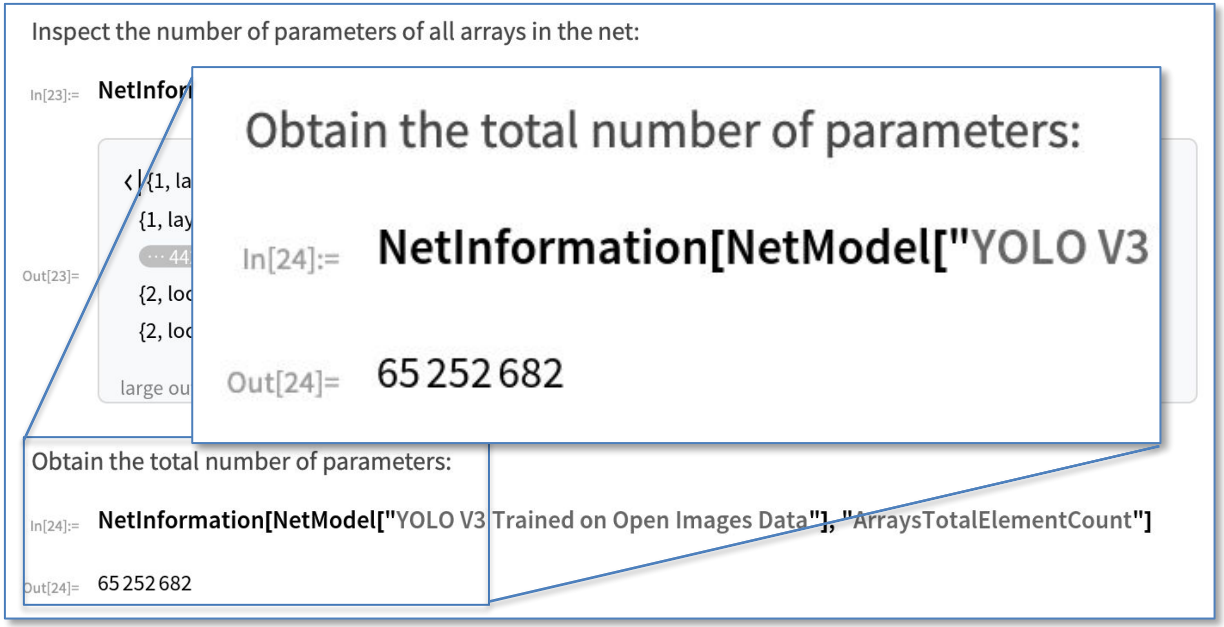
Image by BrunelloN CC-BY-SA 4.0

Does explaining the operations explain the decision?

Can the Programmer Explain the Decision of an AI?

- Typically, one decision requires billions of arithmetic operations with millions of parameters.
- Calculating the parameters (“training”) requires many orders of magnitude more.

Describing the operations contributes *nothing* towards anything a human would accept as an “explanation.”



Inspect the number of parameters of all arrays in the net:

```
In[23]:= NetInfor
Out[23]= <|{1, la
{1, lay
··· 44
{2, loc
{2, loc
large ou
```

Obtain the total number of parameters:

```
In[24]:= NetInformation[NetModel["YOLO V3

Out[24]= 65 252 682
```

Obtain the total number of parameters:

```
In[24]:= NetInformation[NetModel["YOLO V3 Trained on Open Images Data"], "ArraysTotalElementCount"]

Out[24]= 65 252 682
```

Source: https://resources.wolframcloud.com/NeuralNetRepository/resources/YOLO-V3-Trained-on-Open-Images-Data

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 a very limited data.

Herb Simon, circa 1981

Silver Bullets?

- 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.”

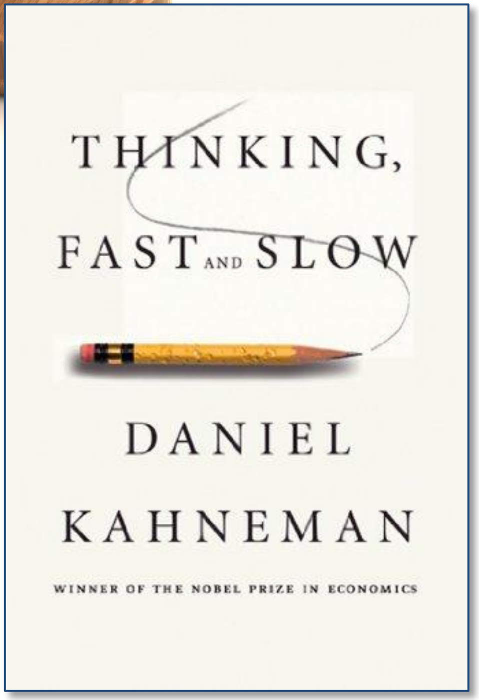
So, how can we find an explanation?

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.

Photo by Chris Potter, CC-BY-AT 2.0



THINKING, FAST AND SLOW

DANIEL KAHNEMAN

WINNER OF THE NOBEL PRIZE IN ECONOMICS

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

A Prediction

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)

Humans are Very Good at Synthesizing Explanations

But even the time since the last food break is probably not a very good explanation for the parole decisions.

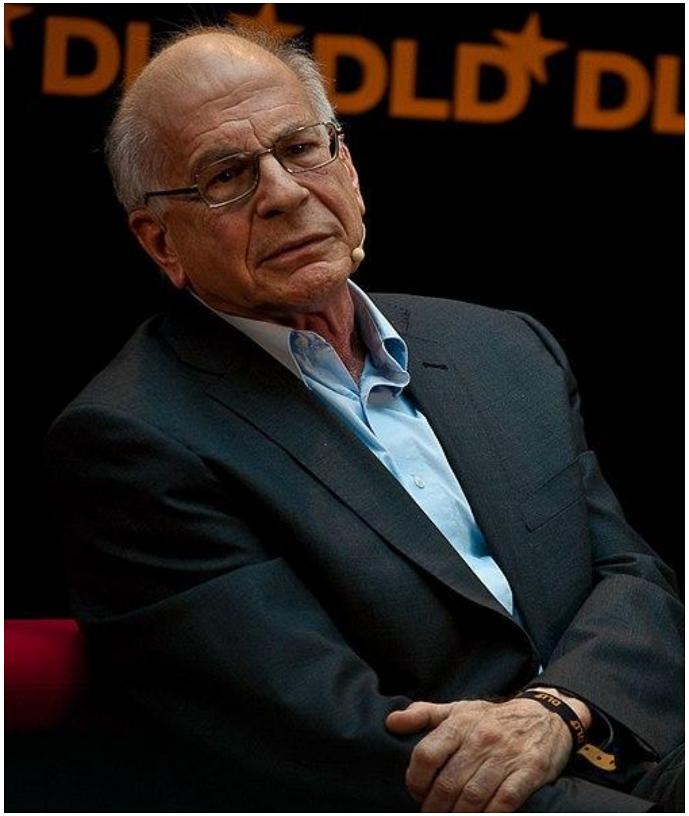
Intuition and experience almost certainly play a huge role.

How Do Humans *Really* 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.



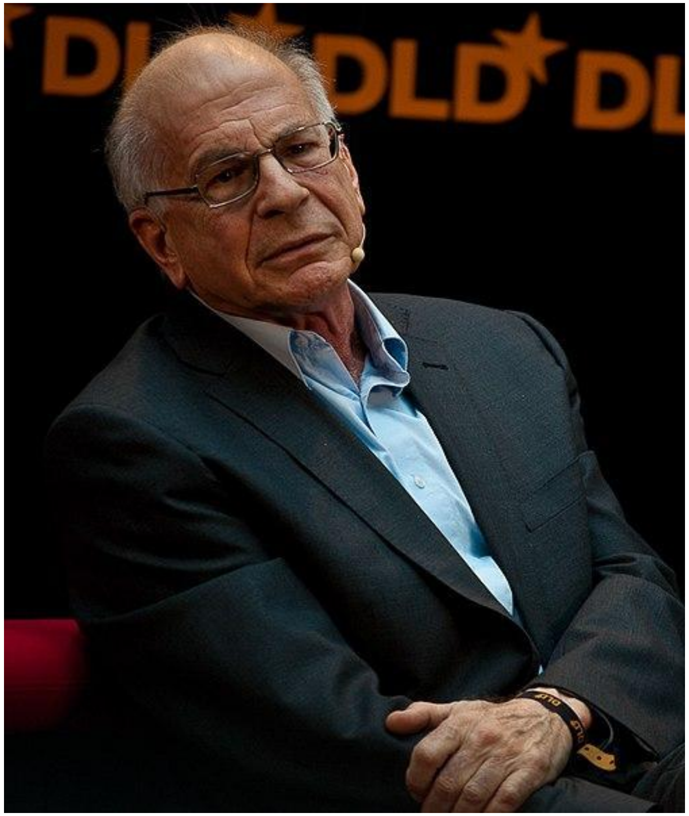
Daniel Kahneman in 2009
Photo by nrkbeta, CC-BY-SA 3.0

How Do Humans *Really* 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.



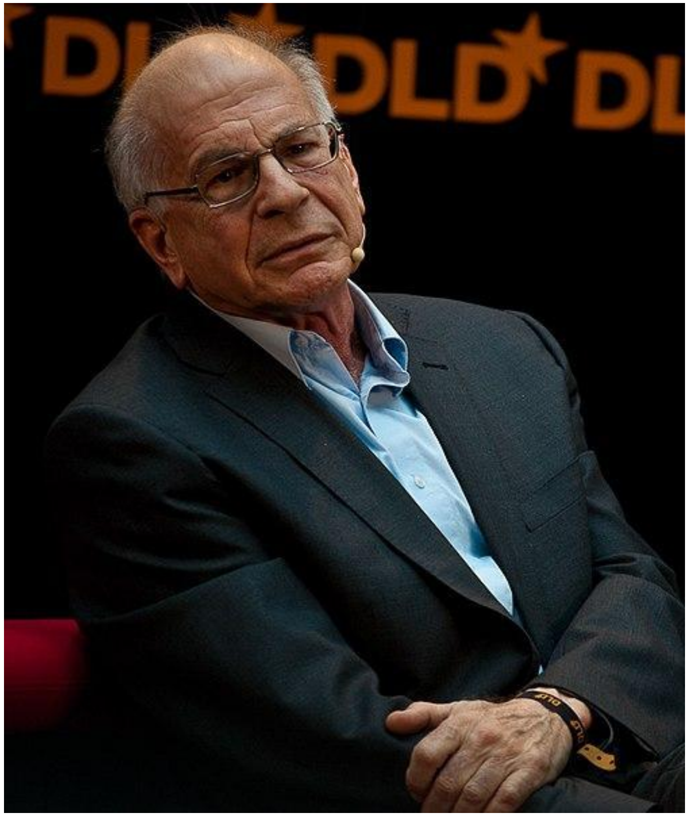
Daniel Kahneman in 2009
Photo by nrkbeta, CC-BY-SA 3.0

How Do Humans *Really* 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.



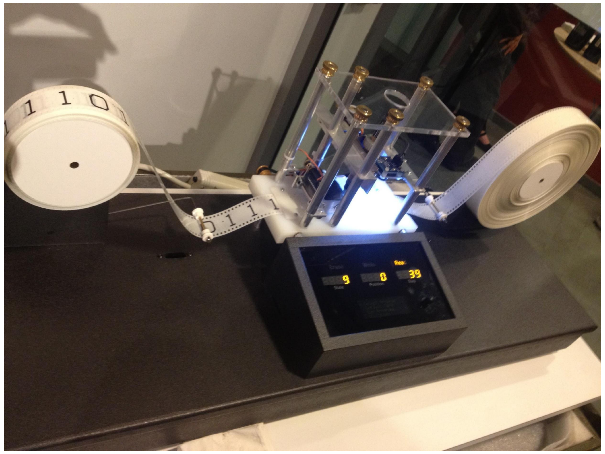
Daniel Kahneman in 2009
Photo by nrkbeta, CC-BY-SA 3.0

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
Photo by GabrielF - Own work, 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.

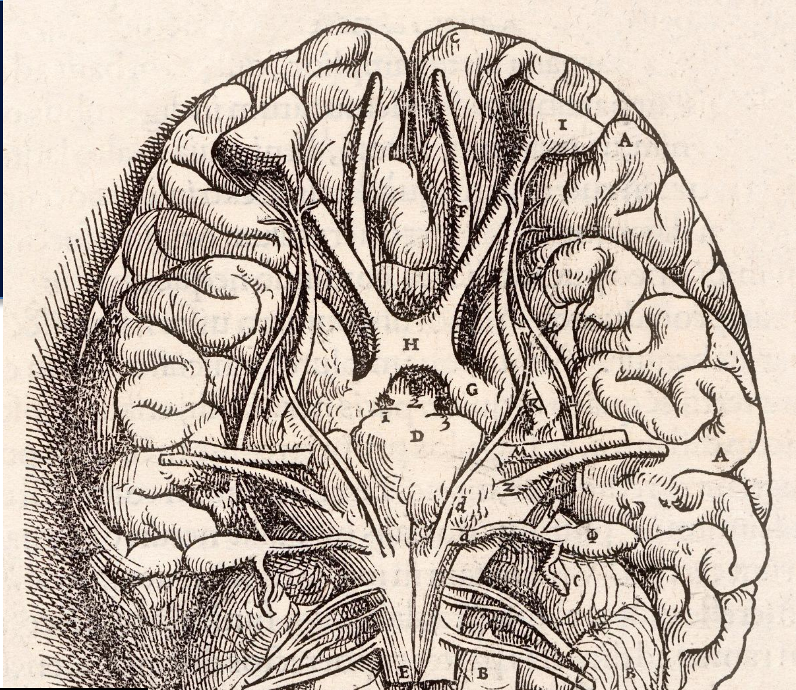
Is System 1 Algorithmic?

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

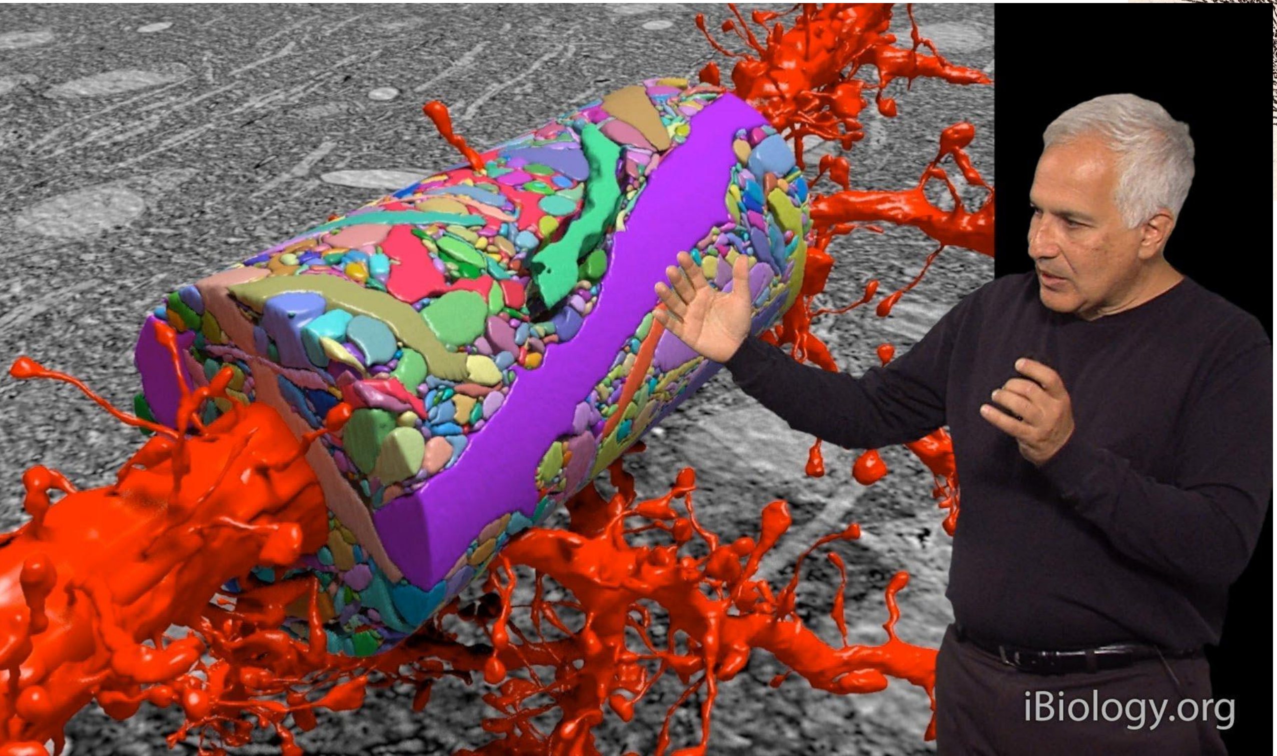
Are system 1 decisions nevertheless algorithmic?

Is the Brain Algorithmic?

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



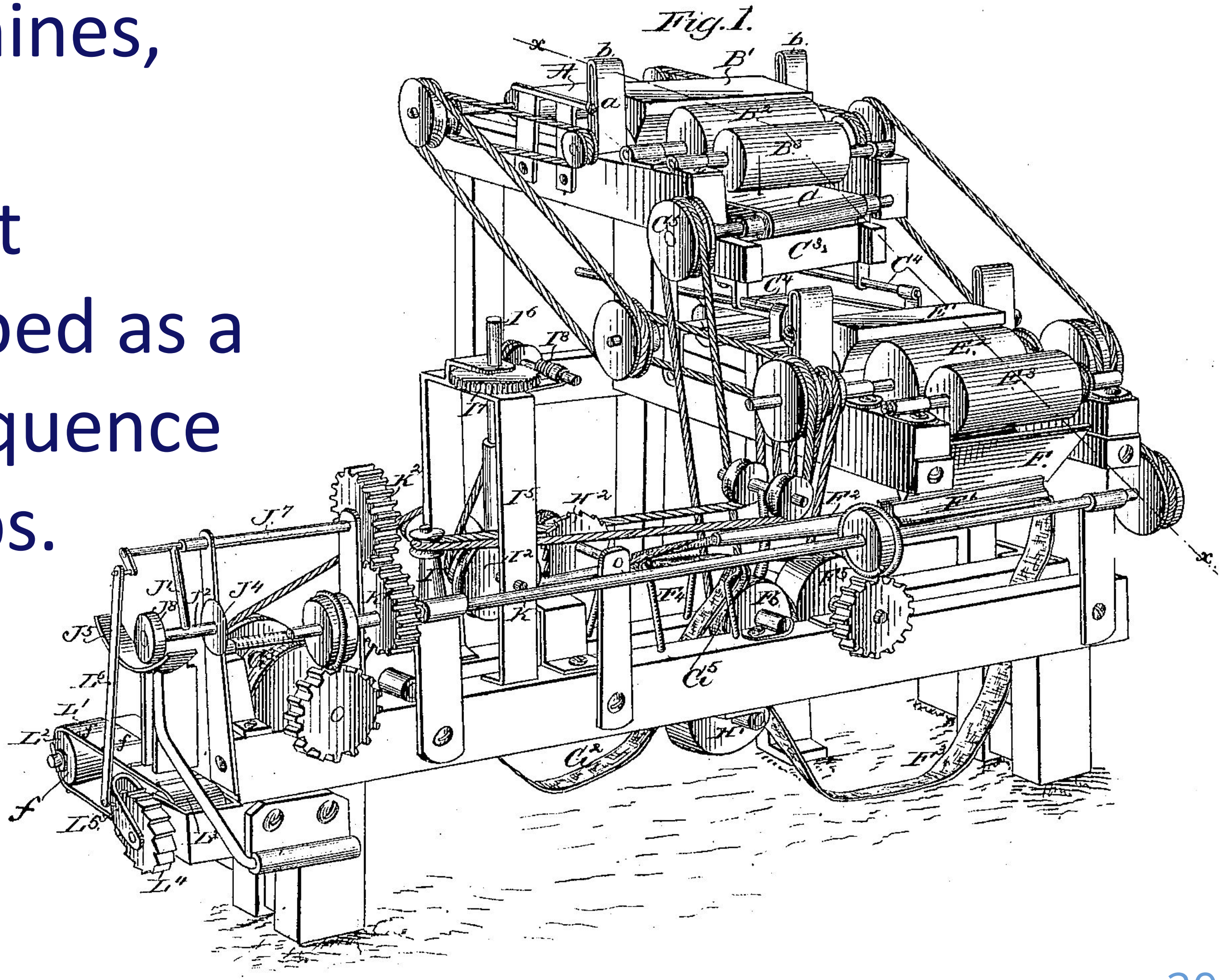
Andreas Vesalius' Fabrica, published in 1543



Jeff Lichtman,
Harvard, 2015

Many Machines Are Not Usefully Modeled by Algorithms.

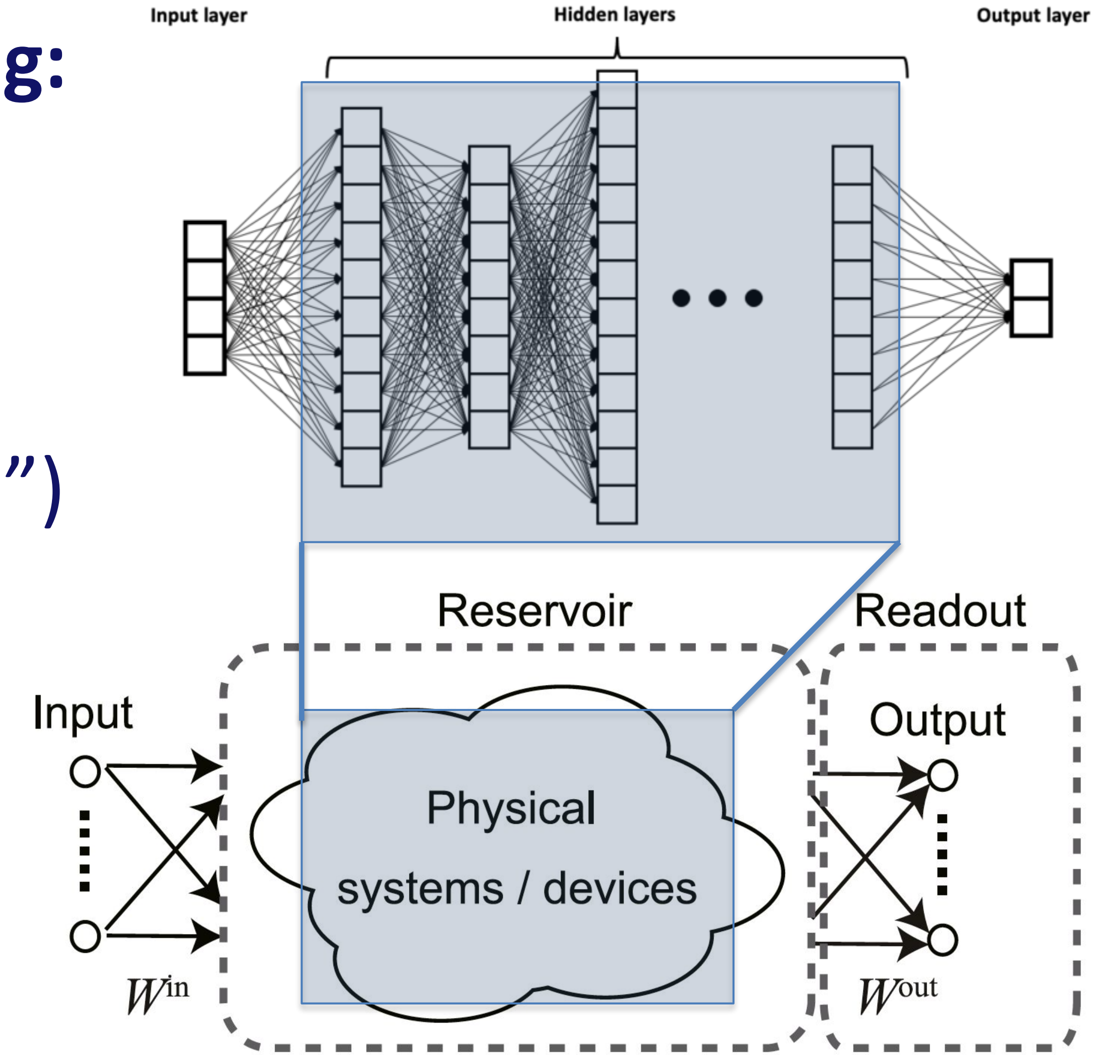
For many machines, their essential operation is not usefully described as a terminating sequence of discrete steps.



Even DNNs may not be fundamentally algorithmic.

Reservoir Computing:

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

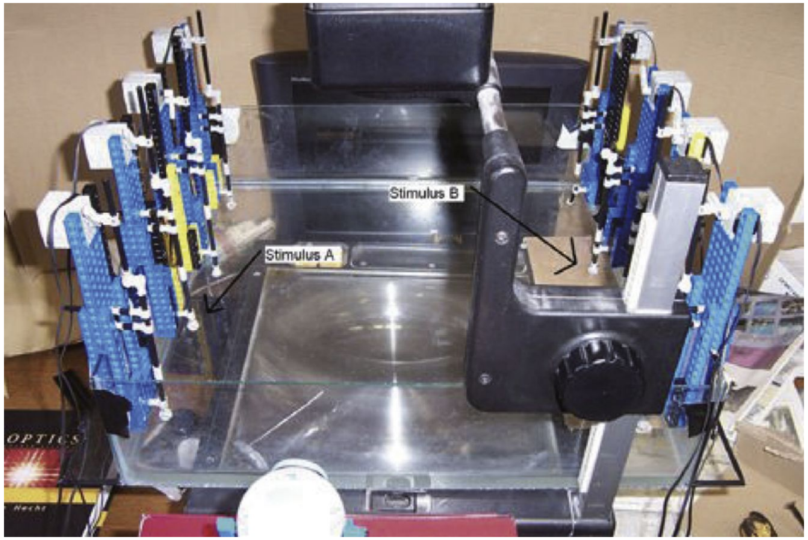


Tanaka, et al., "Recent advances in physical reservoir computing: A review," *Neural Networks, 2019*

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
- …

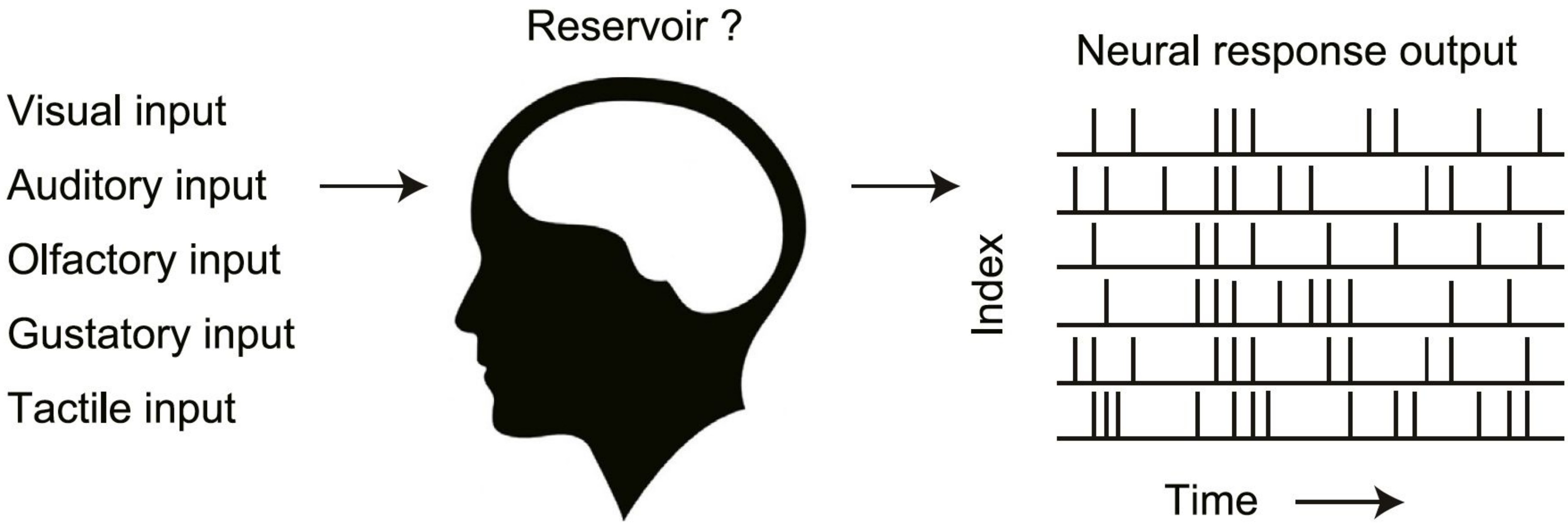


Fernando and Sojakka, Pattern recognition in a bucket," In European conference on artificial life, 2003.

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

A Provocative Conjecture

Is it possible that some brain regions work as reservoirs?



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

More Silver Bullets?

- 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?

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).

More Silver Bullets?

- Give the machines human-like intelligence.

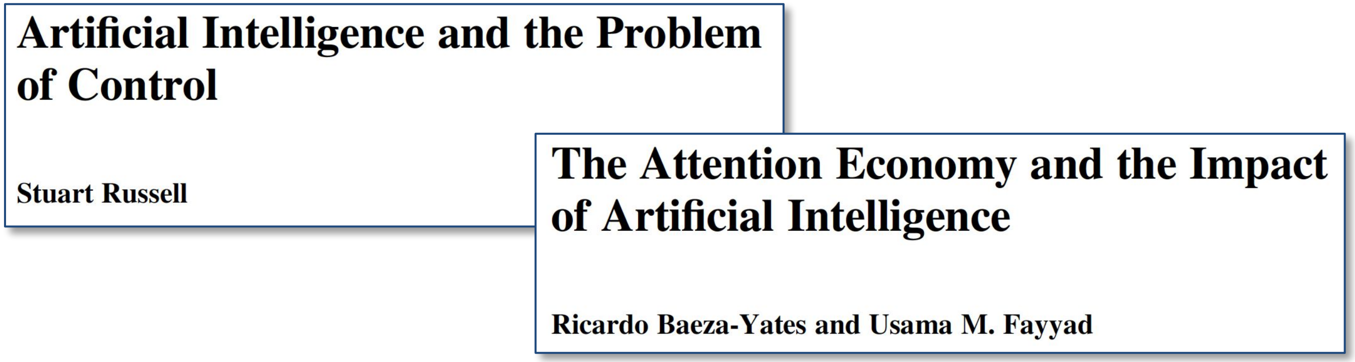
  Artificial General Intelligence (AGI) is about achieving human-like intelligence. Is that a good idea?

- 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.*

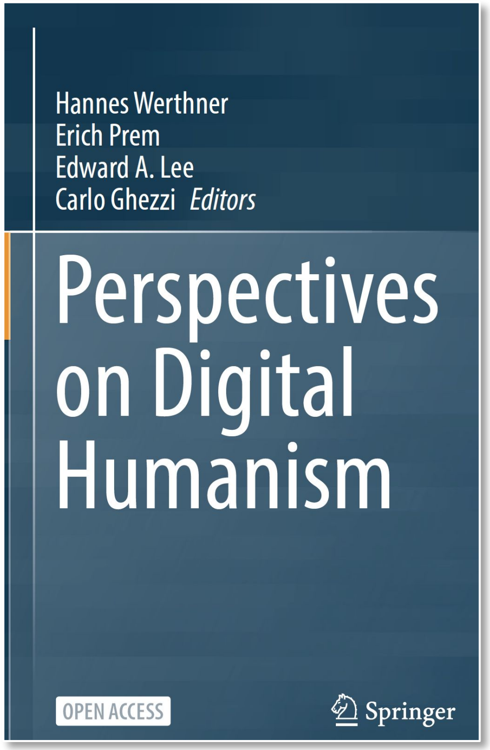


Artificial Intelligence and the Problem of Control

Stuart Russell

The Attention Economy and the Impact of Artificial Intelligence

Ricardo Baeza-Yates and Usama M. Fayyad



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

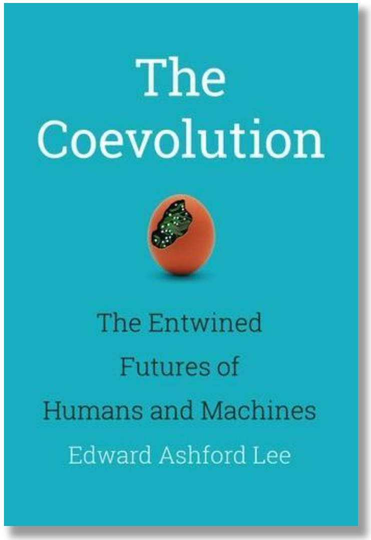
Perspectives on Digital Humanism

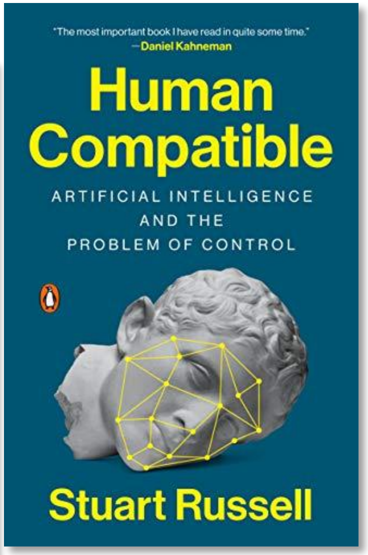
OPEN ACCESS

Springer

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**.

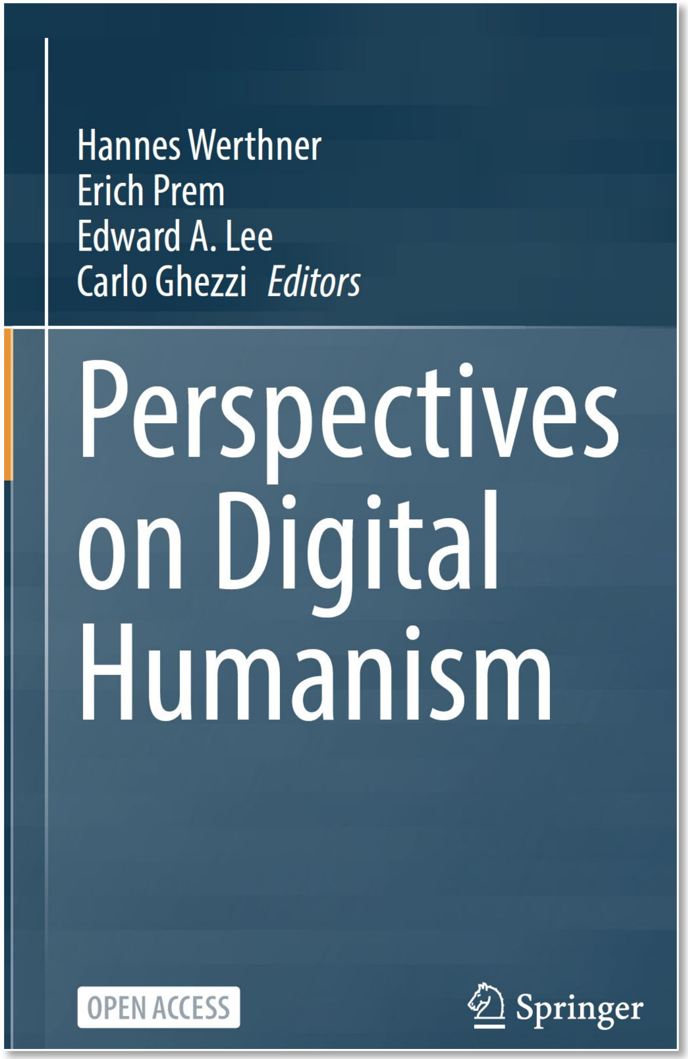



"The most important book I have read in quite some time." —Daniel Kahneman
Human Compatible
ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL
Stuart Russell

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?
- … multidisciplinary engagement …

Next time I come to Vienna,
I hope you will explain it to me.



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

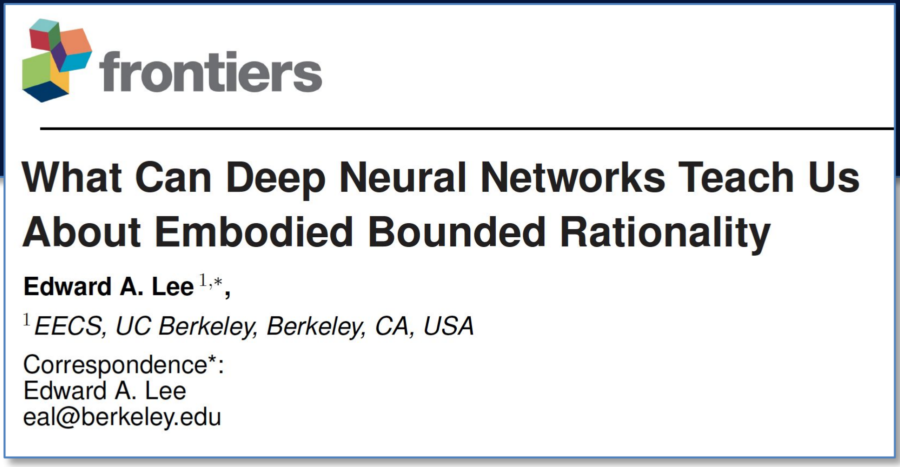
Perspectives on Digital Humanism

OPEN ACCESS

Springer

https://dighum.org/perspectives-on-digital-humanism

Some References



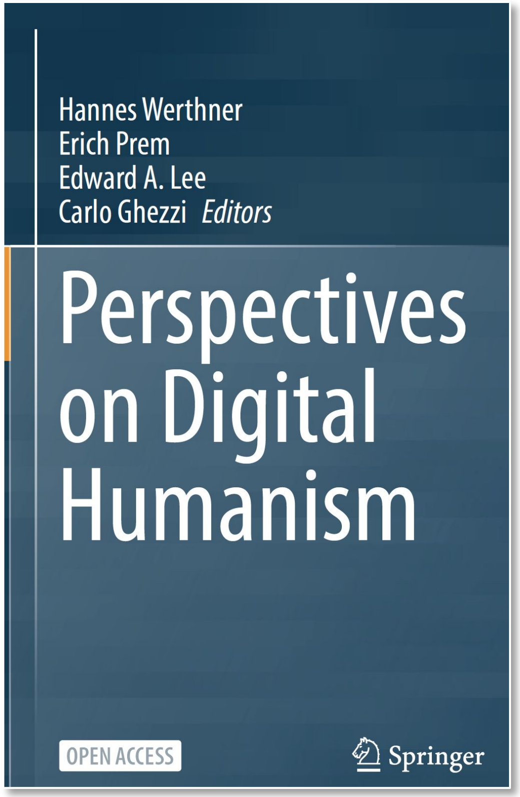
frontiers

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

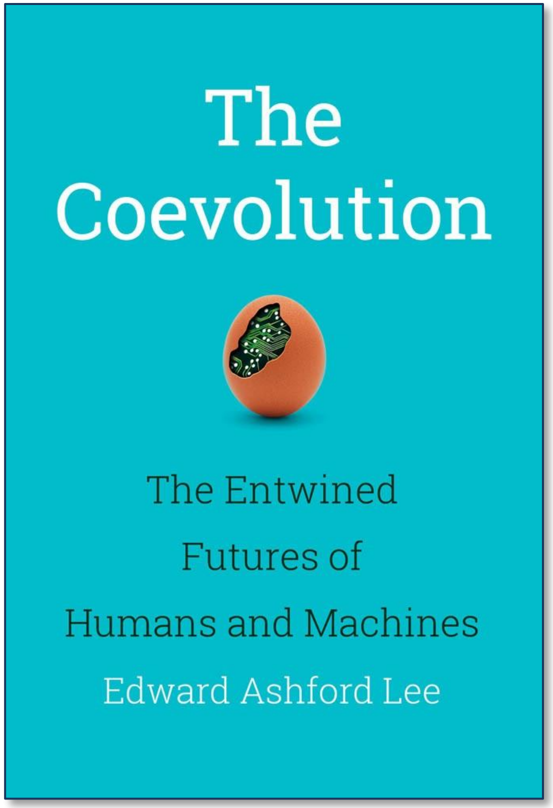


Hannes Werthner
Erich Prem
Edward A. Lee
Carlo Ghezzi *Editors*

Perspectives on Digital Humanism

OPEN ACCESS

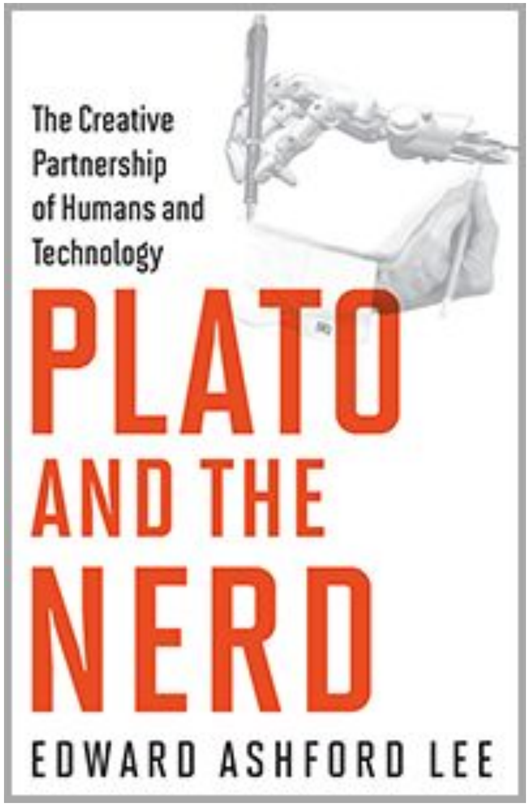
Springer



The Coevolution

The Entwined Futures of Humans and Machines

Edward Ashford Lee



The Creative Partnership of Humans and Technology

PLATO AND THE NERD

EDWARD ASHFORD LEE

Backup

Digital Creationism: The Hypothesis that Technology is Top-Down Intelligent Design

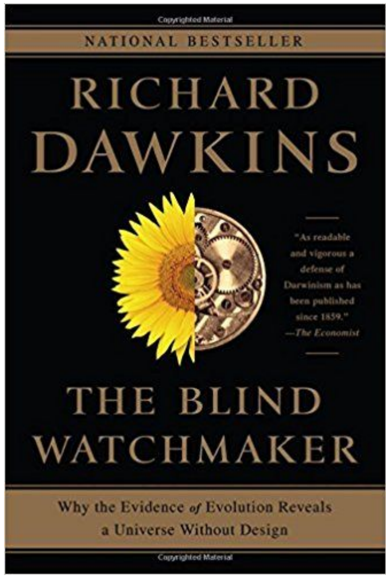
"Every boat is copied from another boat ... Let's reason as follows in the manner of Darwin. It is clear that a very badly made boat will end up at the bottom after one or two voyages and thus never be copied. ... One could then say, with complete rigor, that it is the sea herself who fashions the boats,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

French philosopher Alain

An Alternative to Digital Creationism: Darwinian Evolution

Donan.raven [CC BY-SA 3.0]

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



Copyrighted Material

NATIONAL BESTSELLER

RICHARD DAWKINS

"As readable and vigorous a defense of Darwinism as has been published since 1859."
—*The Economist*

THE BLIND WATCHMAKER

Why the Evidence *of* Evolution Reveals a Universe Without Design

Copyrighted Material