

Certainty or Intelligence: Pick One!

Edward A. Lee

*Professor of the Graduate School, UC Berkeley
Distinguished Professor Emeritus, UC Berkeley
Chief Scientist, Xronos Inc.*

*Topos Institute Colloquium
Online, April 18, 2024*



University of California at Berkeley

The Dilemma

The dilemma is that when designing autonomous systems, we want certainty, proofs that only safe behavior is possible, but we also want intelligence, especially adaptability to circumstances not previously seen.

These are conflicting goals!



Image generated by ChatGPT 4

The Bayesians and the Frequentists

Confession: I am a Bayesian, not a frequentist.

A *probability* is a measure of our uncertainty, not a measure of underlying randomness.



Image generated by ChatGPT 4

Bayes' Law

Let S be a statement about which you want to be certain.

e.g., “The car will stay in the lane.”

Let O be an observation of behavior.

e.g., “The car followed a certain trajectory.”

Bayes' law:

$$p(S|O) = \frac{p(O|S)p(S)}{p(O)}$$



Bayes' Law

Let S be a statement about which you want to be certain.

Let O be an observation of behavior.

Posterior Probability

$$p(S|O) = \frac{p(O|S)p(S)}{p(O)}$$

Probability of the observation
given S is true

Prior Probability

Probability of the observation

Certainty (First Version)

Let S be a statement about which you want to be certain.

Let O be an observation of behavior.

$$p(S|O) = \frac{p(O|S)p(S)}{p(O)}$$

Certainty: S is true.

Then $p(S) = 1$, $p(O|S) = p(O)$, and hence $p(S|O) = p(S) = 1$.

The posterior equals the prior!

Certainty (Second Version)

Let S be a statement about which you want to be certain.

Let O be an observation of behavior.

$$p(S|O) = \frac{p(O|S)p(S)}{p(O)}$$

Certainty: S is false.

Then $p(S) = 0$, $p(O) \neq 0$ (observation occurred), so $p(S|O) = 0$.

The posterior equals the prior!

Certainty Means You Can't Learn

Let S be a statement about which you want to be certain.

Let O be an observation of behavior.

$$p(S|O) = \frac{p(O|S)p(S)}{p(O)}$$

If you are certain about S , then the posterior equals the prior!

Certainty or Intelligence: Pick One!

Is Bayes' Law the real reason for this conflict?

The AI Revolution

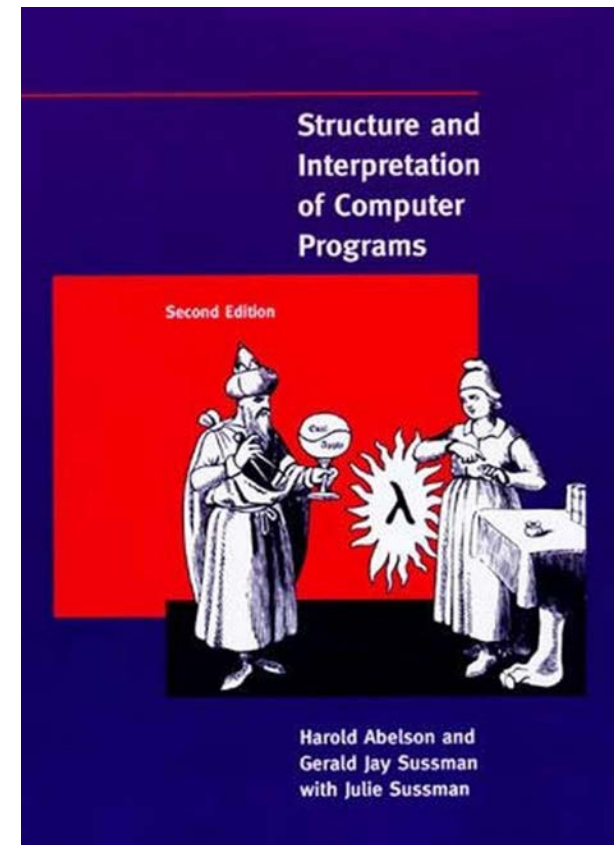
Algorithms, logic, mathematics, rationality, and **formal methods** are in the realm of the *certain*.

Als are realized on computers using algorithms.
Doesn't this put them in the realm of the
certain?

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

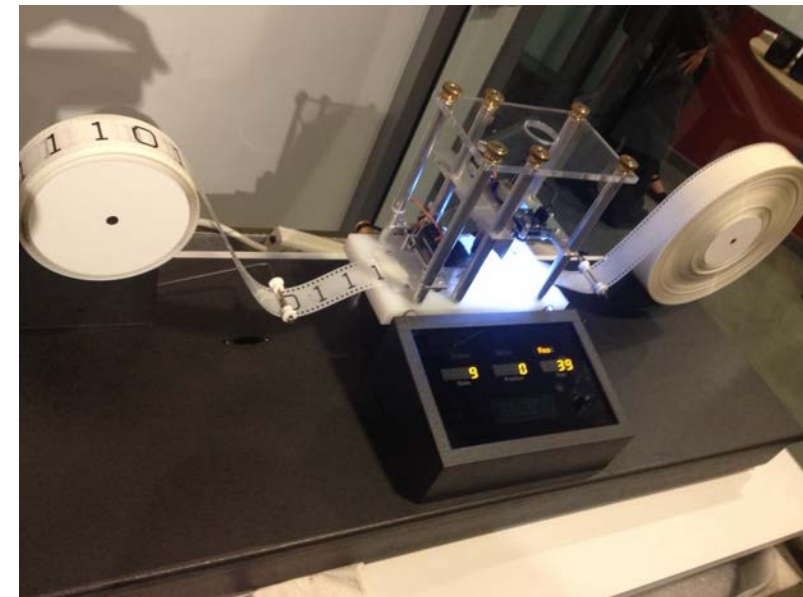


Procedural Epistemology

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

Procedural Epistemology

Proofs:

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

Certainty comes from terminating computations.

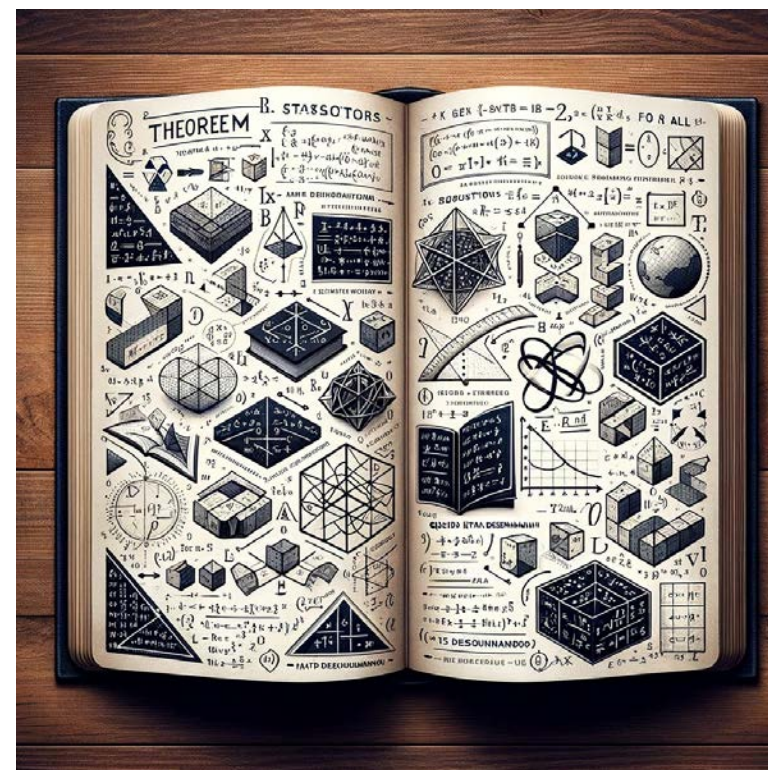


Image generated by ChatGPT 4.0

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

Are these foundation models usefully understood as rational machines?

Text-to-Image:

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



My Essential Claim

*The new machines are not about
procedural epistemology.*

Rationality

Logical \leftrightarrow Intuitive

Reasoned \leftrightarrow Creative

Dispassionate \leftrightarrow Emotional

Predictable \leftrightarrow Adaptive

Certain \leftrightarrow Intelligent?



They make stuff up.

Is it hallucination or creativity?



Edward A. Lee, Berkeley

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.

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



Melting eggs by ChatGPT 4.



Tyler Glaiel @TylerGlaiel · 19h
hope you're all ready for the AI dominated future



can you melt eggs

Images

Videos

How long

Shopping

News

Maps

Books

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

148

3,357

43.1K

1.4M

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: $-12x^2 + 34x + 12 = 0$

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

This is incorrect.

They make mistakes.

GPT-4: Asked to break down the steps:

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

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 ?

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 ?

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

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

$$x(0) = b(9 + 3 + 1 - 26/b)$$

This is correct.

Solving for b , we find that $b = 2$. So $a = 3$ and $b = 2$ for.

Makes an implicit assumption that the equation has a solution.

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

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

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?

The Experts Say:

Physics of AI

Sebastian Bubeck, 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!

Sebastian Bubeck

20:47 / 1:00:03

Can we get Certainty through 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 and does not lead to certainty!

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

Can we get certainty by regulating AI?

- 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.

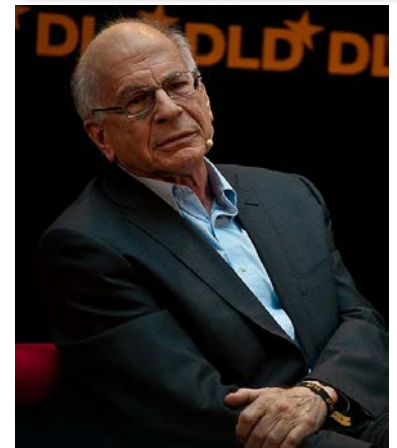
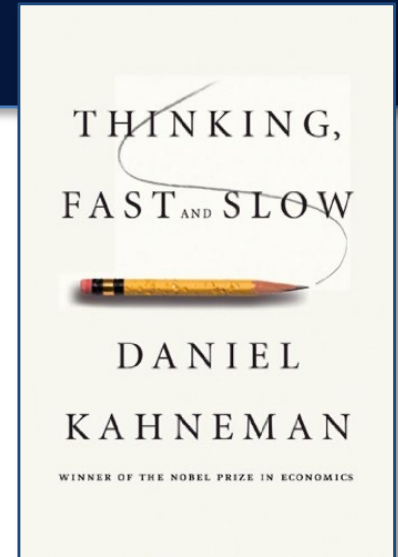
The operations done by the computer, despite being “rational,” do not provide what we would call an “explanation.”

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, and, only with system 2 can we achieve **certainty**.



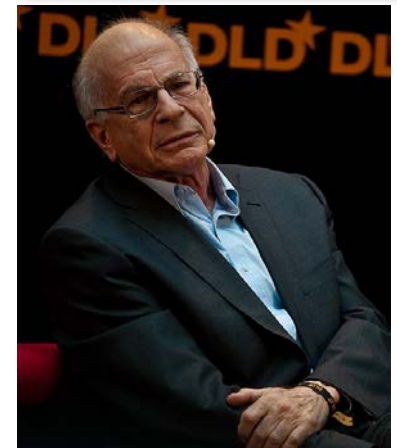
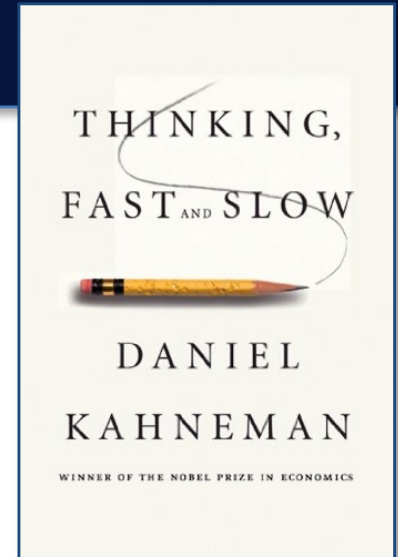
Daniel Kahneman in 2009
Photo by nrkbeta, CC-BY-SA 3.0

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.

Rationality

Logical ↔ Intuitive

Reasoned ↔ Creative

Dispassionate ↔ Emotional

Predictable ↔ Adaptive

Certain ↔ Intelligent

This side is limited by
Bounded Rationality!



Certainty or Intelligence: Pick One!

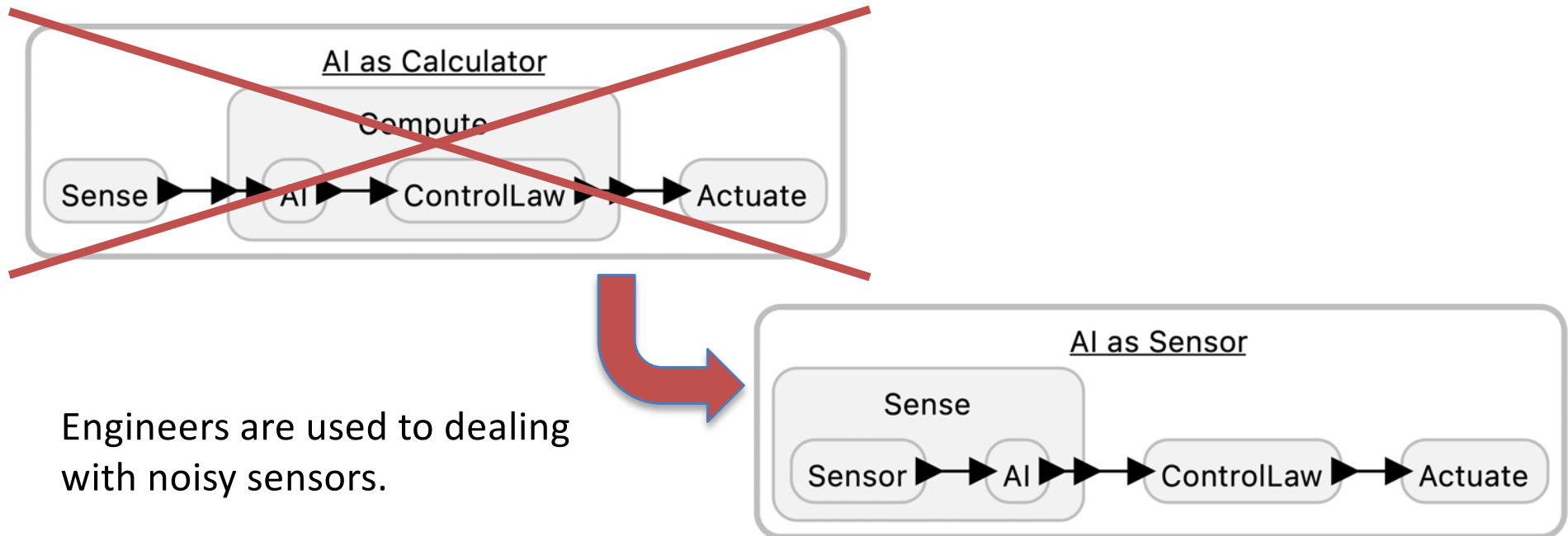
These goals are contradictory!

(I'm quite certain about that.)

How to Deal With Contradictory Goals?

Approach 1: Refactor

Treat AI as a sensor system (works for perception):



How to Deal With Contradictory Goals?

Approach 2: Envelope Protection

Limit the actions an AI can take.

Analogous to flight-envelope protection, which limits the actions a *human* can take.

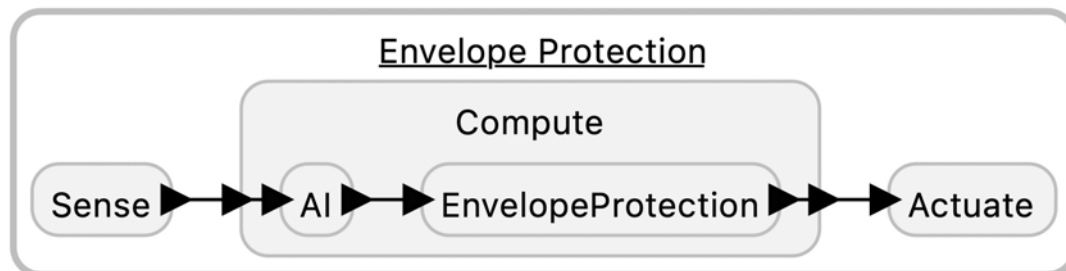


Image by ChatGPT 4

How to Deal With Contradictory Goals?

Approach 3: Algorithmic Confidence

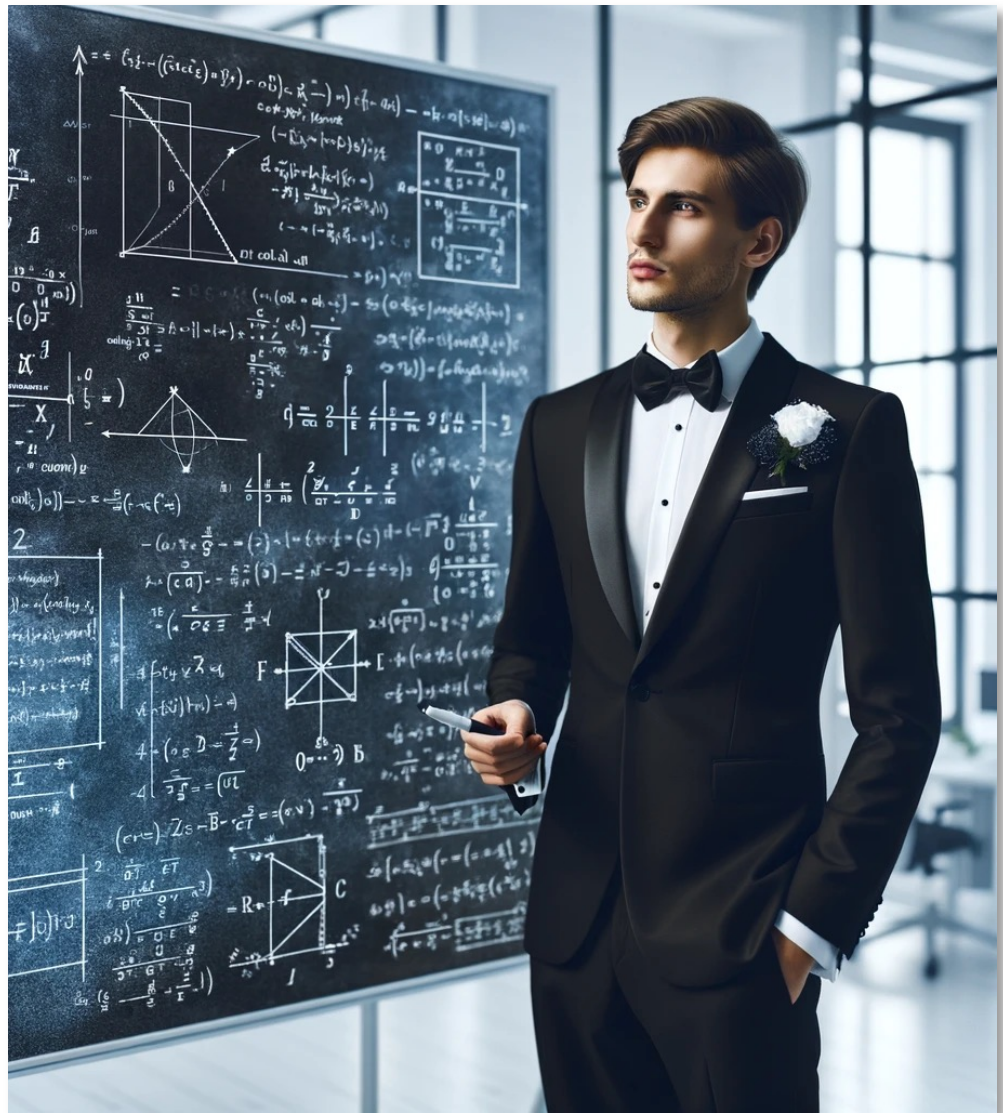
Rely on *formal methods*.

E.g.,

- Theorem provers (proofs need not be short)
- Model checking (exhaustive)

These can only provide confidence in a *model*, not in a *system*.

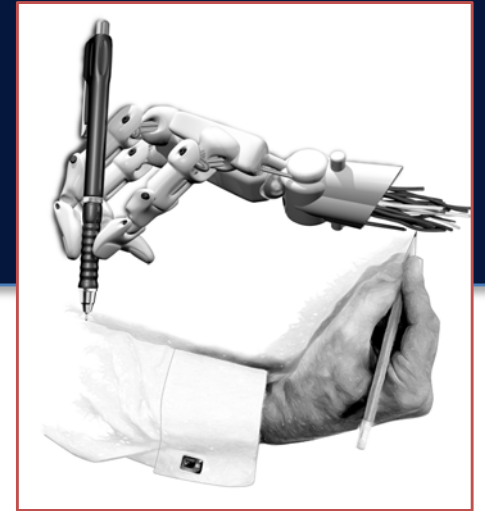
And formal methods are *not* showing promise of scaling to DNNs and reinforcement learning!



Certainty or Intelligence: Pick One!

- **Certainty** prevents adaptability (Bayes' Law).
- Certainty arises from **rational processes**.
- **Bounded rationality** limits this.
- **Formal methods** cannot scale enough.
- **Intelligence** requires **adaptability**.

*Intelligence may be rooted in **prediction, not rationality,** which appears to be an emergent property.*



References



My web page, with links to more references and talks.



What Can Deep Neural Networks Teach Us About Embodied Bounded Rationality

Edward A. Lee^{1,*},

¹EECS, UC Berkeley, Berkeley, CA, USA

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

