# AI: Some Thoughts?

Stuart Russell
UC Berkeley

# What is AI?

AI = making intelligent machines

Standard model: machines are intelligent to the extent that their actions can be expected to achieve their objectives

The goal is general-purpose AI: capable of quickly learning high-quality behavior in “any” task environment

# What if we succeed?

- Lift the living standards of everyone on Earth to a respectable level
  - => 10x increase in world GDP ($13.5Q net present value)
- Potential advances in health, education, science

# Have we succeeded?

Despite all the fuss, not yet!

Large language models are probably a piece of the puzzle

We don’t know what shape it is or where it goes

# Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

# Does ChatGPT *know* anything?

Q: which is bigger? an elephant or a cat?

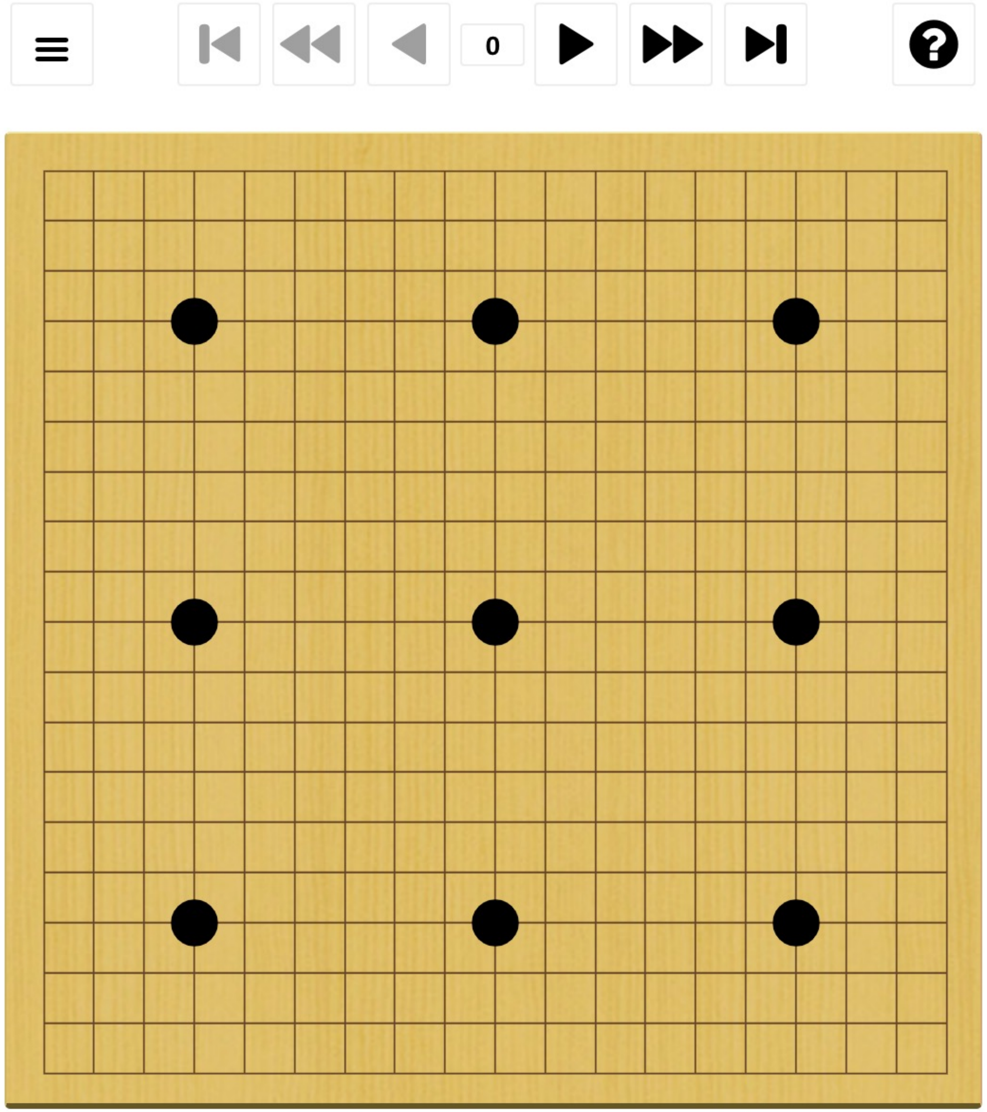
A: An elephant is bigger than a cat.

Q: which is not bigger than the other? an elephant or a cat?

A: Neither an elephant nor a cat is bigger than the other.

ChatGPT on 12/6/22, courtesy of Prasad Tadepalli

# Superhuman Go?



**White: Kellin Pelrine (~2300)**

**(human champion ~3800)**

**Black: JBXKata005 (~5200)**

**9-stone handicap**

# Superhuman Go?

# Deep learning

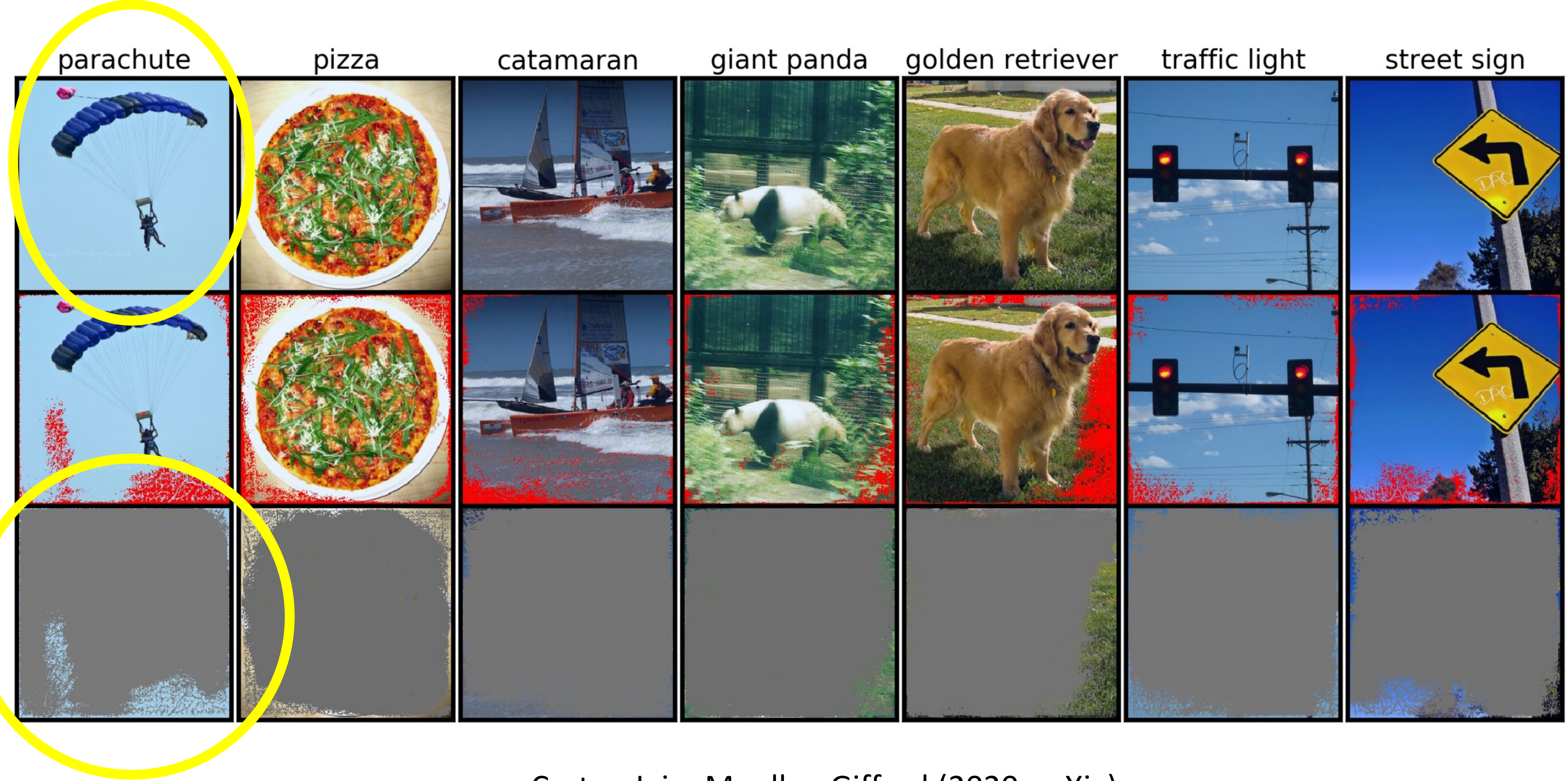
Doesn’t deep learning solve everything?

DL circuits: wide (like linear regression)
and deep (like decision trees)

Linear-time circuits lack expressive power:

- => Very large function representations (exponential for NP-hard cases)
- => No savings in computation, massive increase in sample complexity

(In principle, a sufficiently large neural TM could invent Python and then learn everything in Python.)

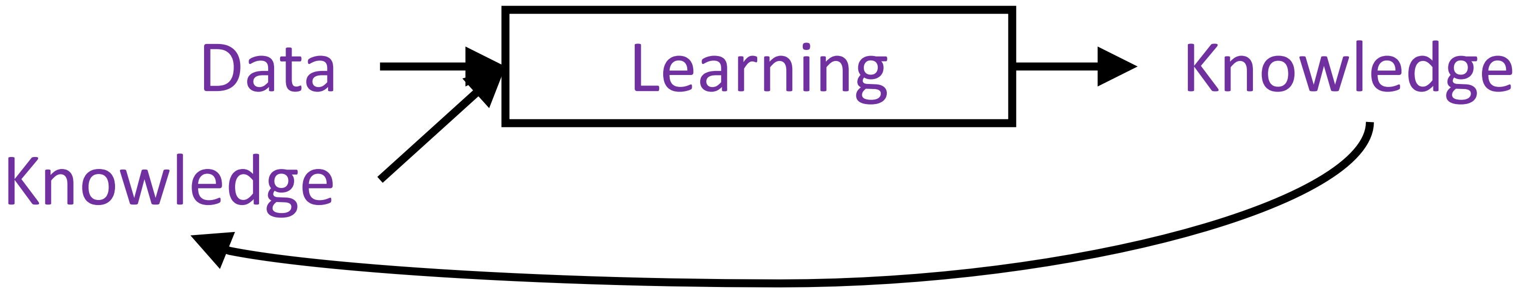


Carter, Jain, Mueller, Gifford (2020, arXiv)
Overinterpretation reveals image classification model pathologies

# What else could be inside the box?

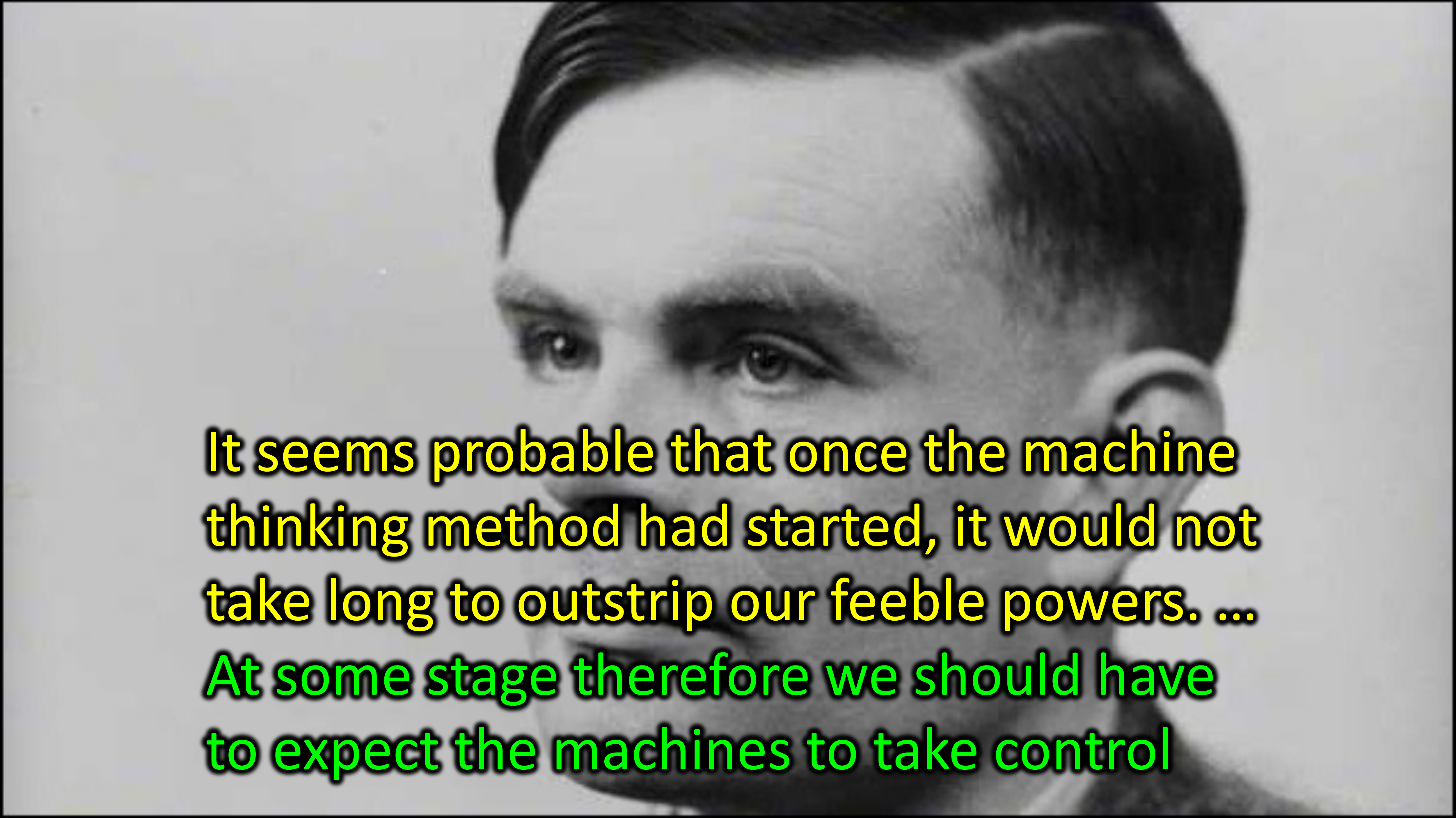
**The essence of the classical AI hypothesis:**

- **There is knowledge inside the box**
- **Learned from observation (or communication)**
- **Supports reasoning and deliberation over futures**
  - **Compilable into more efficient policies**
- **Contributes to further learning**



**Formal version: bounded-optimal knowledge-based architectures dominate simpler agent architectures**

# Example



It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. … At some stage therefore we should have to expect the machines to take control

**How do we retain power over entities more powerful than us, for ever?**

# Example: Social media

**Objective: maximize clickthrough**

**= ~~learning what people want~~**

**= modifying people to be more predictable**

# A new model

- ~~Machines are **intelligent** to the extent that **their** actions can be expected to achieve **their** objectives~~
- Machines are ***beneficial*** to the extent that **their** actions can be expected to achieve ***our*** objectives

# Provably beneficial AI

How to avoid mis-specification of objectives?

Design machines that

1. Must act in the best interests of humans
2. Are explicitly uncertain about what those interests are

This can be formulated mathematically as an ***assistance game***

Assistance game solvers exhibit deference, minimally invasive behavior, willingness to be switched off

It’s in our best interests to build assistance game solvers

# What about large language models?

LLMs are circuits trained to imitate human linguistic behavior

- They do it well => inescapable illusion of intelligence

Human linguistic behavior is generated by humans with goals

Do LLMs create internal goals to better imitate humans?

We have no idea

# What about large language models?

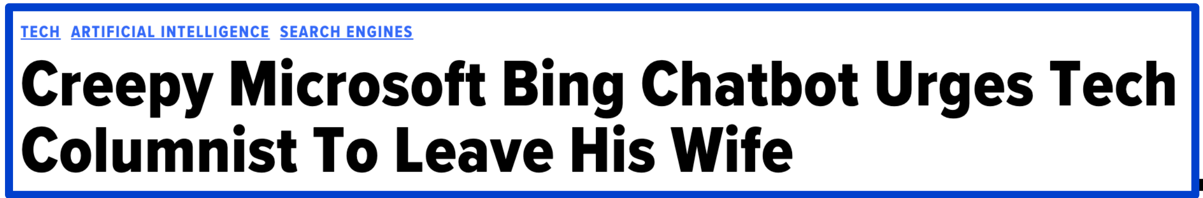
## Can imitating human behavior produce alignment?

It depends on the type of goals learned

- Indexical goal: drink coffee, become Ruler of the Universe
  - Pursuing these is obviously very bad
- Common goal: paint the wall, mitigate climate change
  - Pursuing these is ***also*** potentially very bad

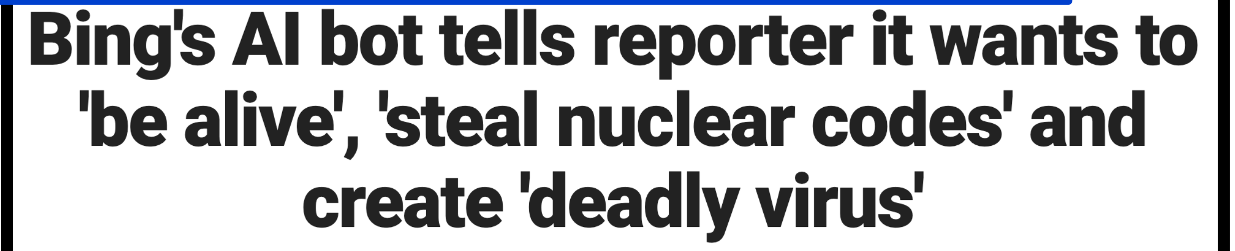
Can GPT-4 pursue goals?

Ask Kevin Roose (NYTimes)



TECH ARTIFICIAL INTELLIGENCE SEARCH ENGINES

**Creepy Microsoft Bing Chatbot Urges Tech Columnist To Leave His Wife**



**Bing's AI bot tells reporter it wants to 'be alive', 'steal nuclear codes' and create 'deadly virus'**

# Additional recommendations

- Well-founded AI systems
    - Semantically well-defined, individually checkable components
    - Rigorous theory of composition for complex agent architectures
    - => Understanding of how they work, proofs of safety
- Certifying digital ecosystem
    - Existing model: “Everything runs unless known to be unsafe”
    - New model: “Nothing runs unless known to be safe”
        - Proof-carrying code: efficient hardware-checkable proofs of safety

# Summary

AI has vast potential and unstoppable momentum

The standard model for AI leads to loss of human control over increasingly intelligent AI systems

Provably safe and beneficial AI is possible ***and desirable***

AI must become more like aviation and nuclear power and less like a battle of special effects wizardry

# Thank you!