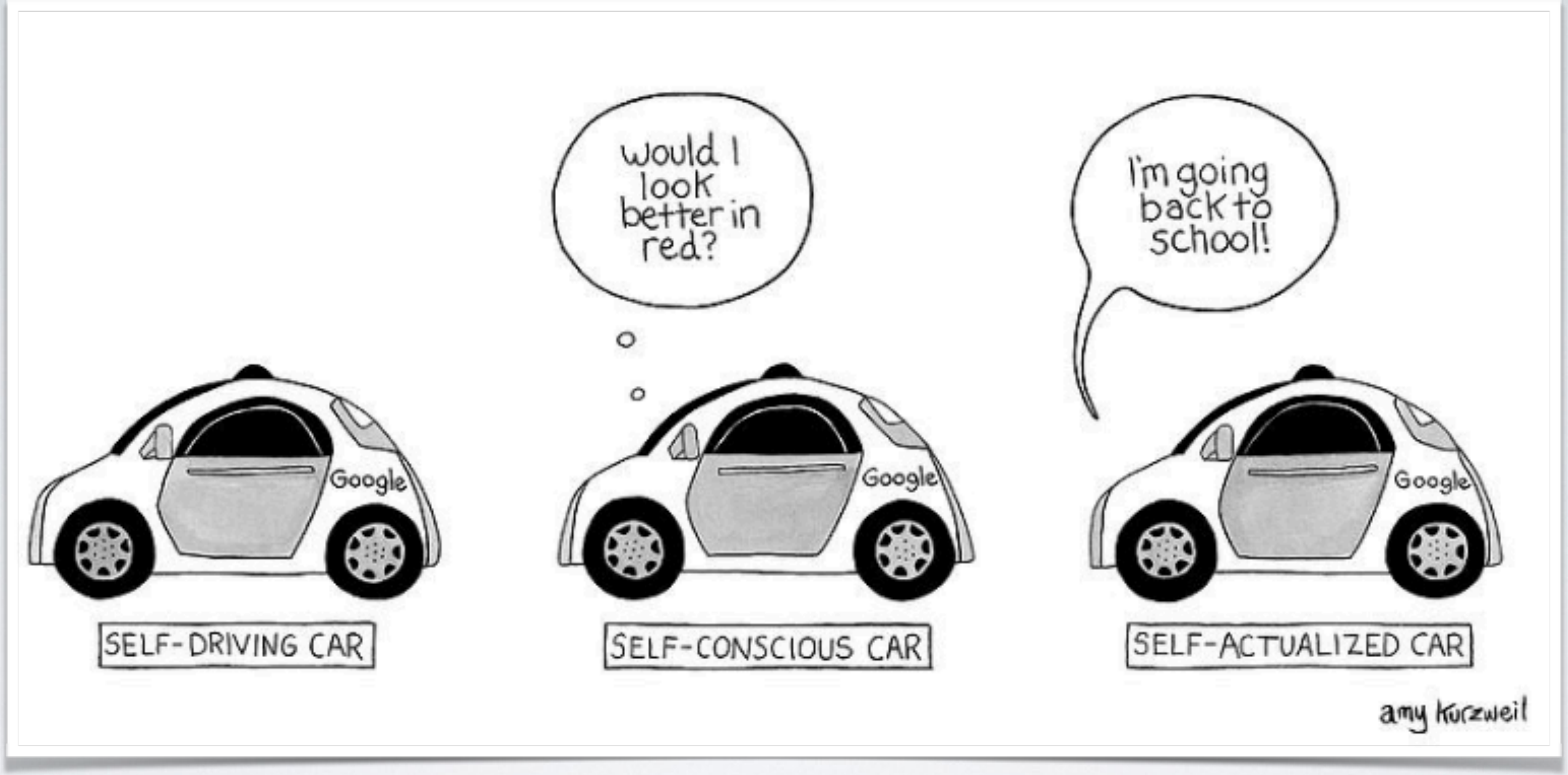


WHAT SHOULD WE FEAR?

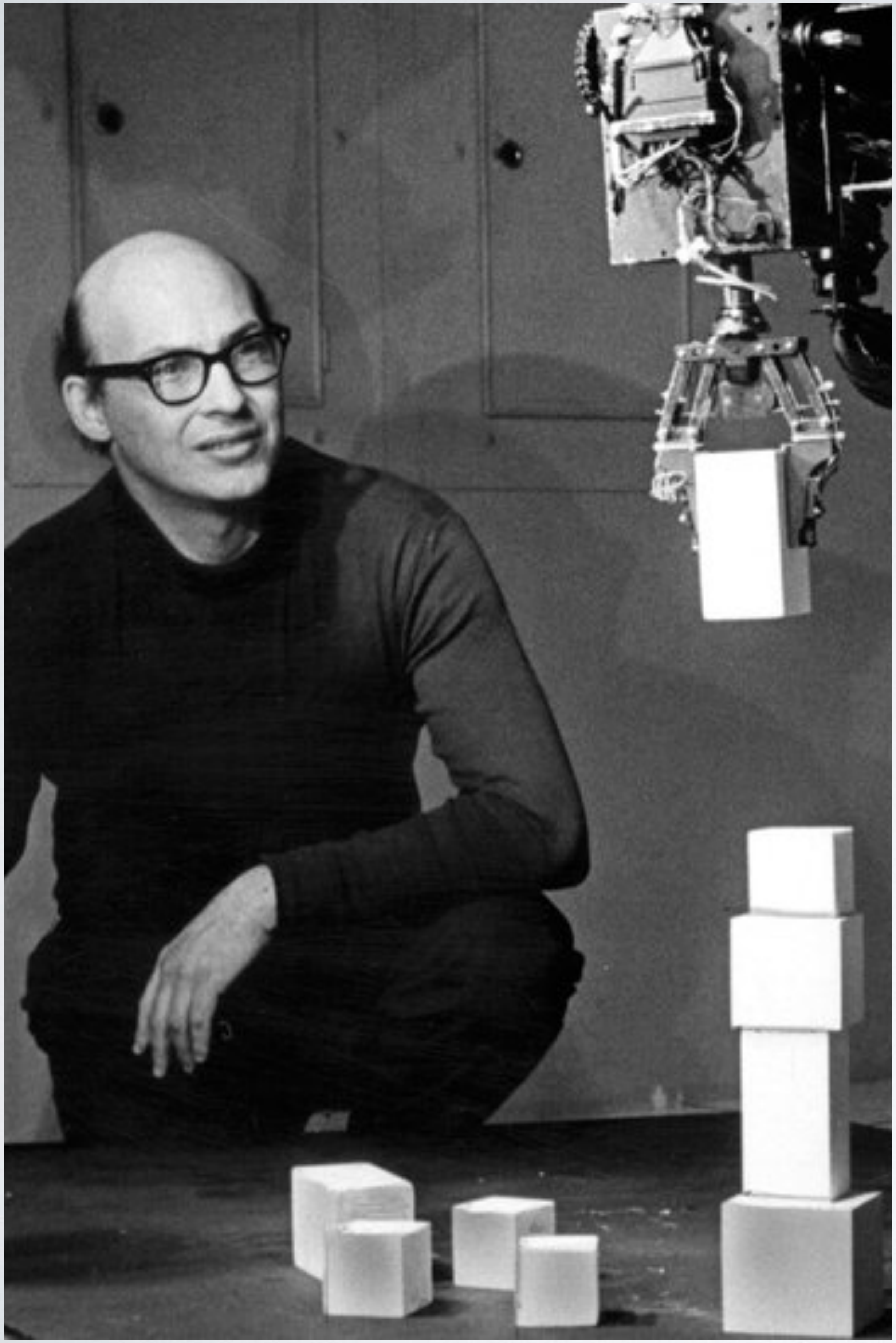
Artificial Intelligence and Global Risks

Leilani H Gilpin (MIT) and Matías Aránguiz (SJTU)

WARNING

This class is interactive

WHO SAID IT?!



“I have exposure to the most cutting edge AI, and I think people should be really concerned by it. AI is a fundamental risk to the existence of human civilization.”

“I have exposure to the most cutting edge AI, and I think people should be really concerned by it. AI is a fundamental risk to the existence of human civilization.”

–Elon Musk

“Every time we improve our AI methods, all of these systems get better. I'm excited about all the progress here and it's potential to make the world better.”

“Every time we improve our AI methods, all of these systems get better. I'm excited about all the progress here and it's potential to make the world better.”

–Mark Zuckerberg

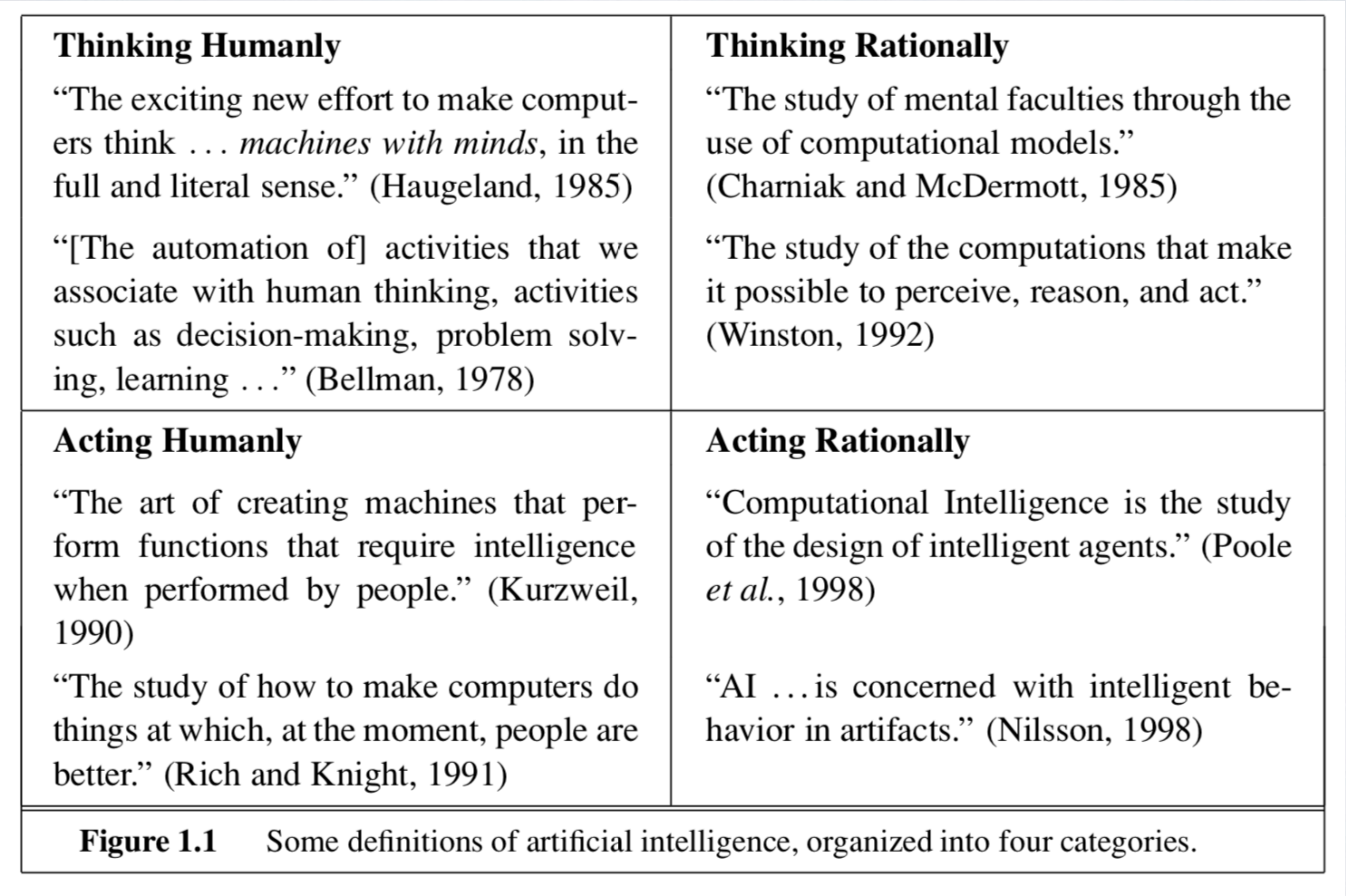
“No computer has ever been designed that is ever aware of what it's doing; but most of the time, we aren't either.”

–Marvin Minsky

> “Worrying about the rise of evil killer robots is like worrying about overpopulation on Mars”
>
> –Andrew Ng

WHAT IS AI?

WHAT IS AI : NORVIG & RUSSELL



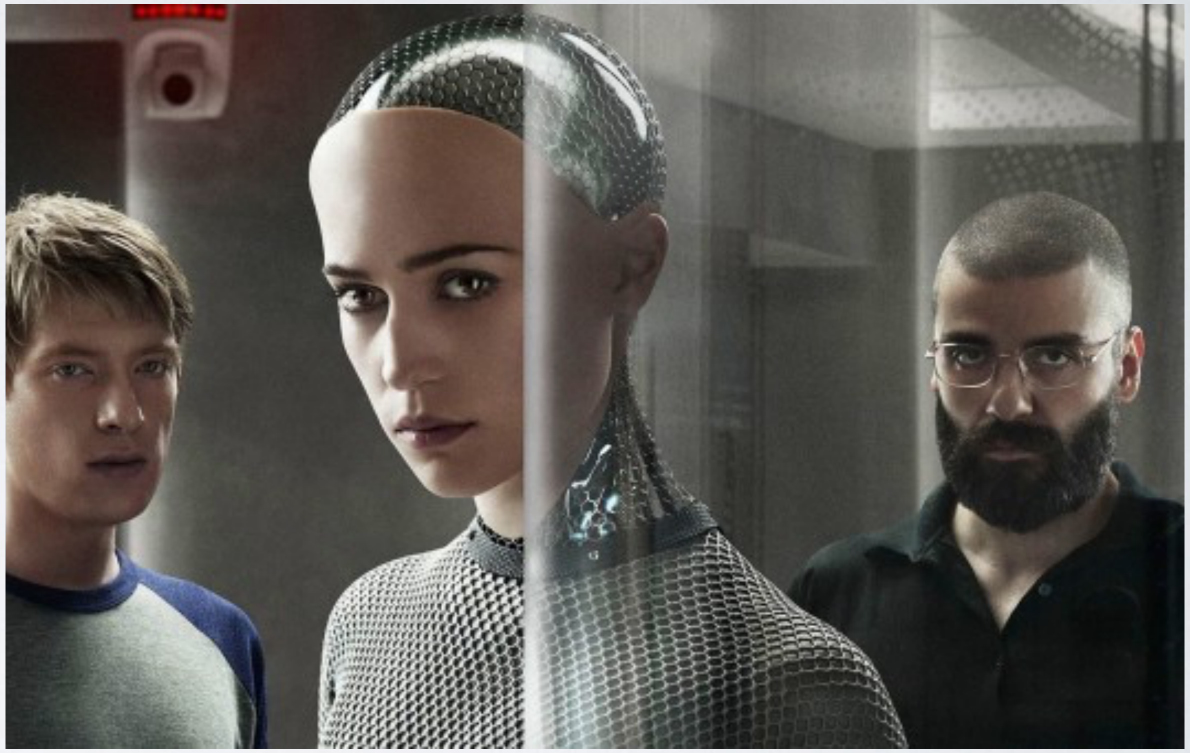
Thinking Humanly	**Thinking Rationally**
"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Acting Humanly	**Acting Rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

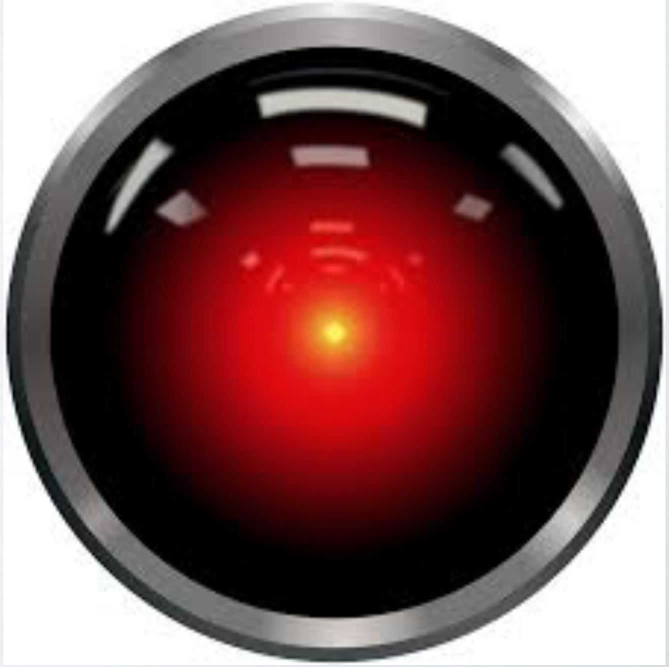
Figure 1.1 Some definitions of artificial intelligence, organized into four categories.

WHAT IS AI :TEST(S)

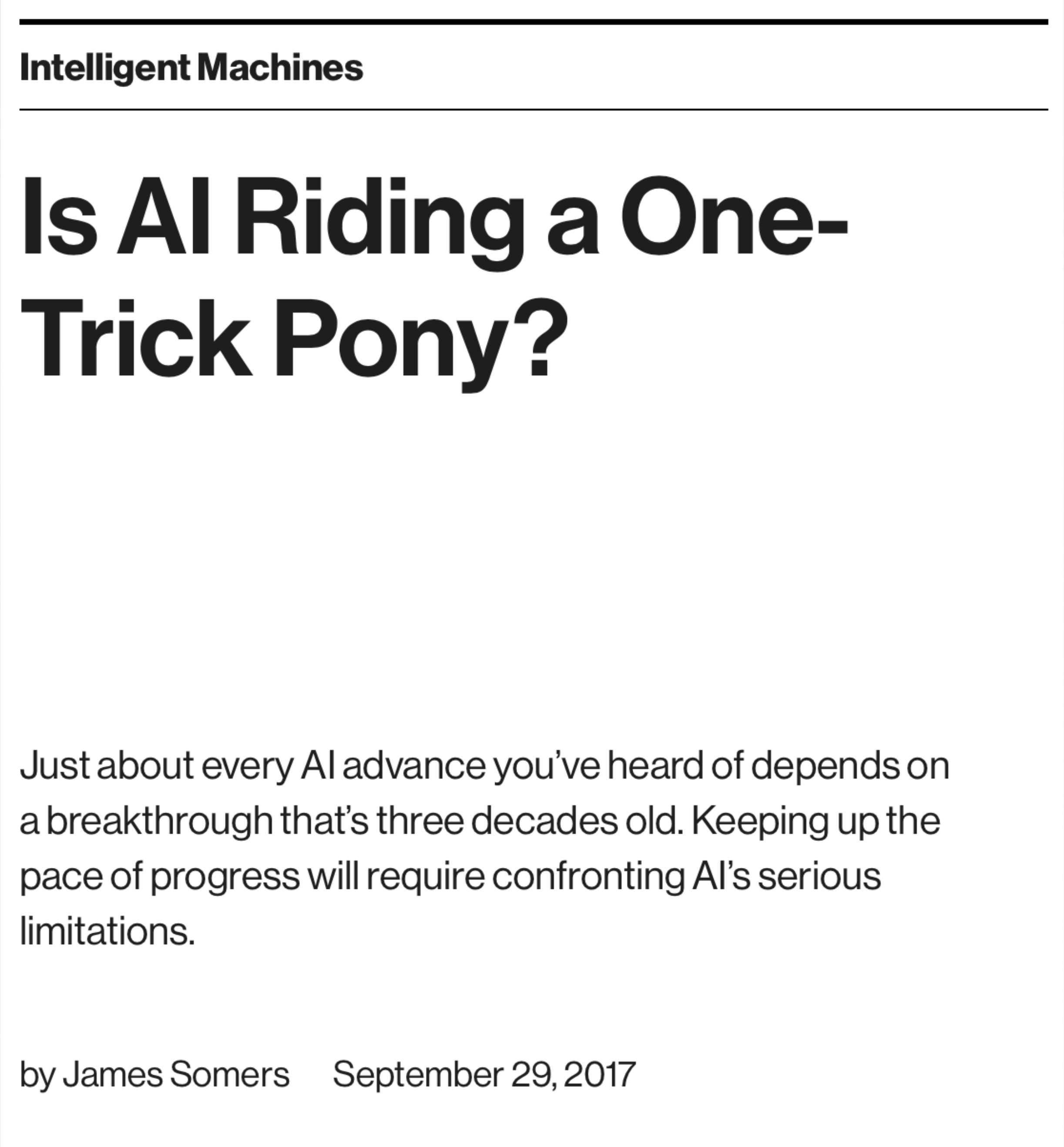
- Proposed by Alan Turing in 1950
  - A computer **passes** if (2 out of 3) human interrogators cannot tell whether the responses to questions come from a computer or human.
  - Some* have claimed to pass the test…
- Winograd schema
  - Proposed "in the spirit" of the Turing test
  - Example : "Jim comforted Kevin because he was so upset. Who was upset?"
  - Multiple choice - binary decision problem

WHAT IS AI : POPULAR CULTURE





IN CURRENT MEDIA



Intelligent Machines

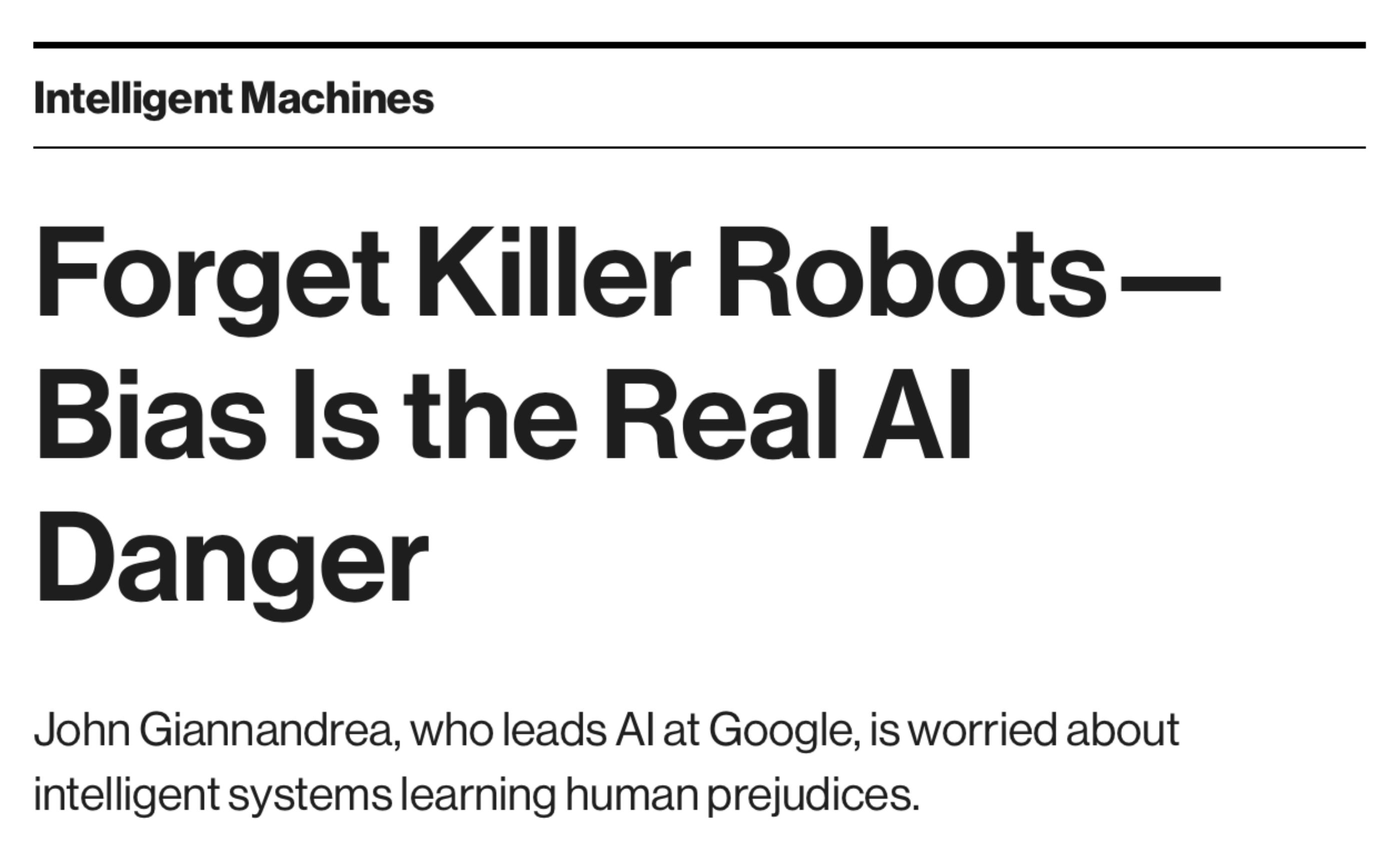
Is AI Riding a One-Trick Pony?

Just about every AI advance you've heard of depends on a breakthrough that's three decades old. Keeping up the pace of progress will require confronting AI's serious limitations.

by James Somers September 29, 2017

Bernard Parker, left, was rated high risk; Dylan Fugett was rate

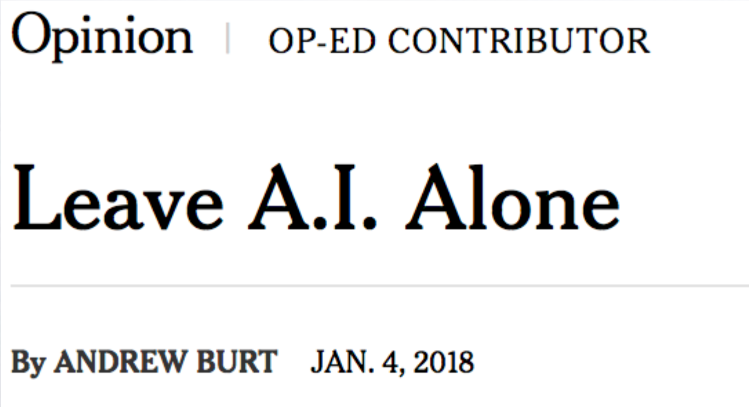
Machine Bias



Intelligent Machines

Forget Killer Robots — Bias Is the Real AI Danger

John Giannandrea, who leads AI at Google, is worried about intelligent systems learning human prejudices.



Opinion | OP-ED CONTRIBUTOR

Leave A.I. Alone

By ANDREW BURT JAN. 4, 2018

NO RIGHT ANSWER

“Suitcase words” - PHW

GOALS OF THE COURSE

1. Learn about classic AI breakthroughs and debunk common myths about artificial intelligence

2. Pose (and try to answer) relevant and pressing questions about the risks of artificial intelligence in many disciplines (medicine, finance, etc.)

3. What is the global perspective on this issue? How can we help? (E.g. "Space race" for AI?)

CREDIT

You can get credit in two ways

1. Write one extended abstract per week

    - 1 page
    - Must reflect that week's information or opinions
    - Goal - document could be sent to a conference and / or turned into a paper

CREDIT

2. Write one report for the whole class

    - Around ~5 pages
    - Describe a case of AI not mentioned in class that produced or can produce harm

OPTIONAL

There will be programming exercises on the website that you can explore (if you want). But they are completely optional.

WHAT IS AI?

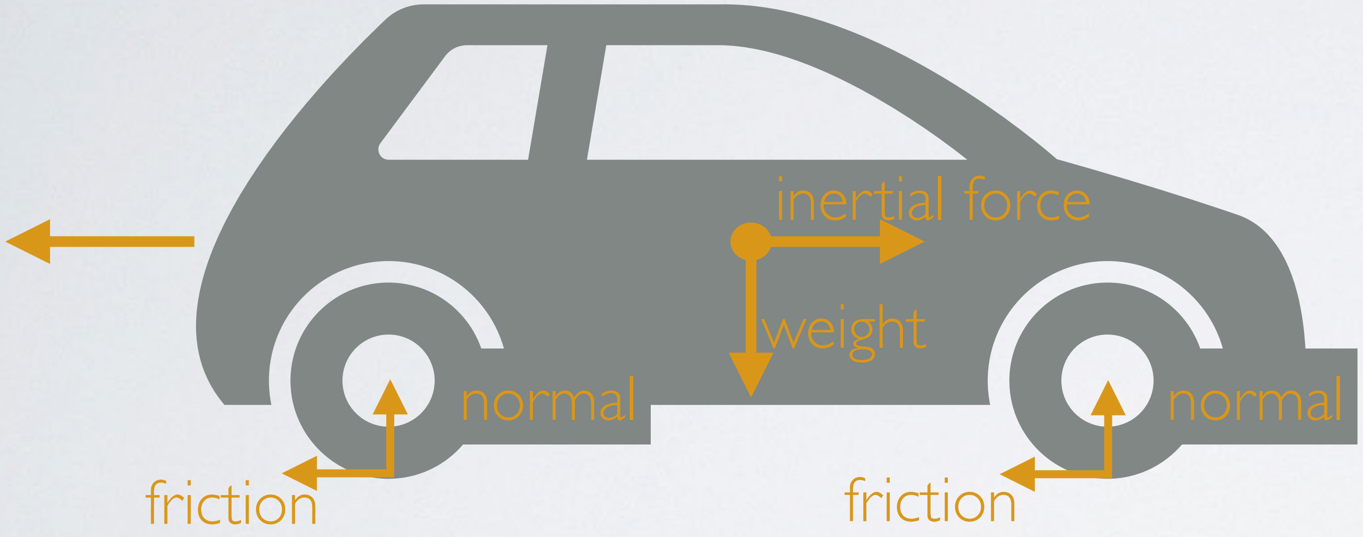
or…back to the class

“OLD” PROBLEM SOLVING

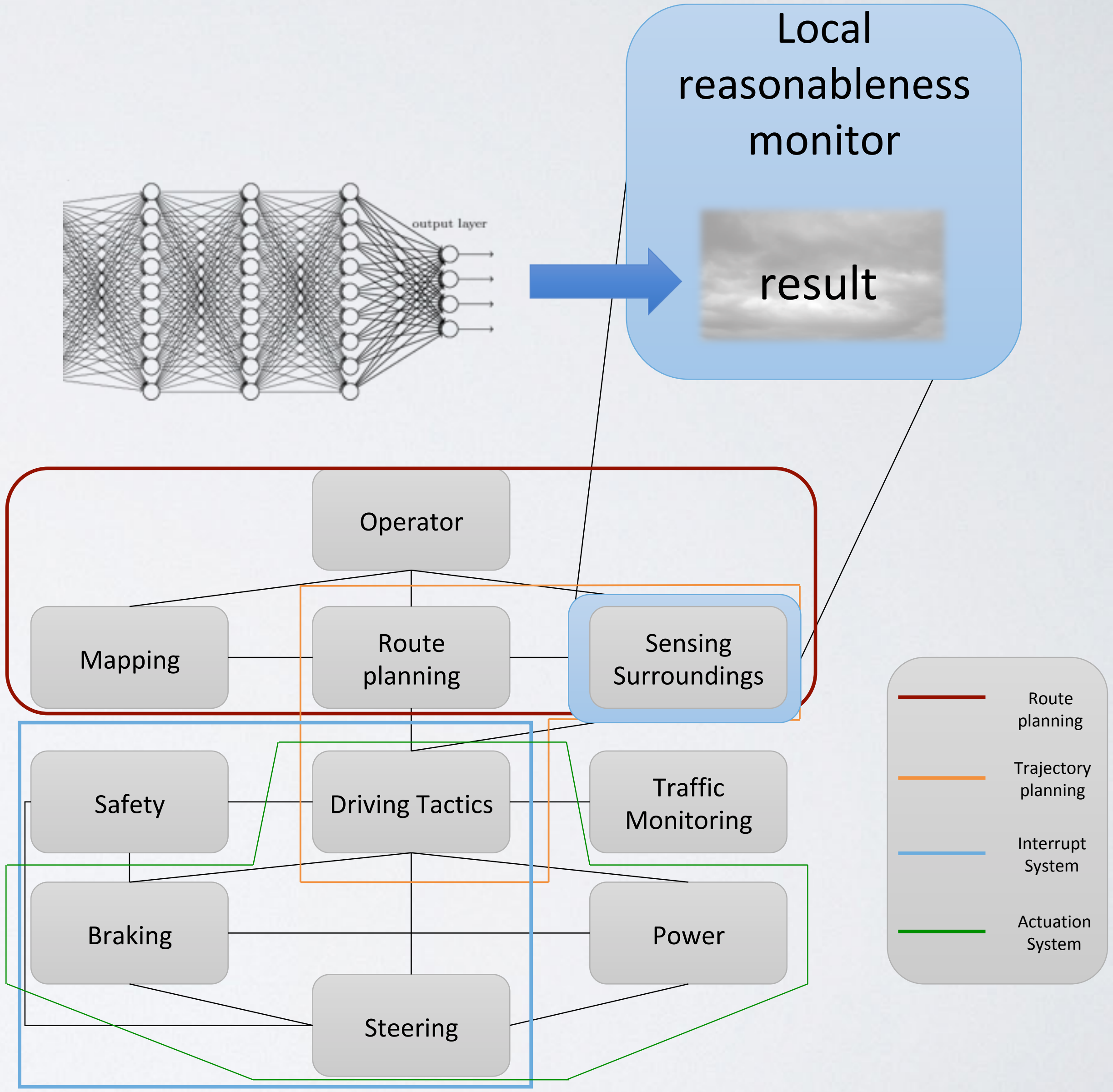
Cognitive Systems / expert systems / GOFAI

- **Logic**al systems / Knowledge-based
  - Scheduling
- Model-based reasoning (**casual** rules)
  - Diagnostic rules

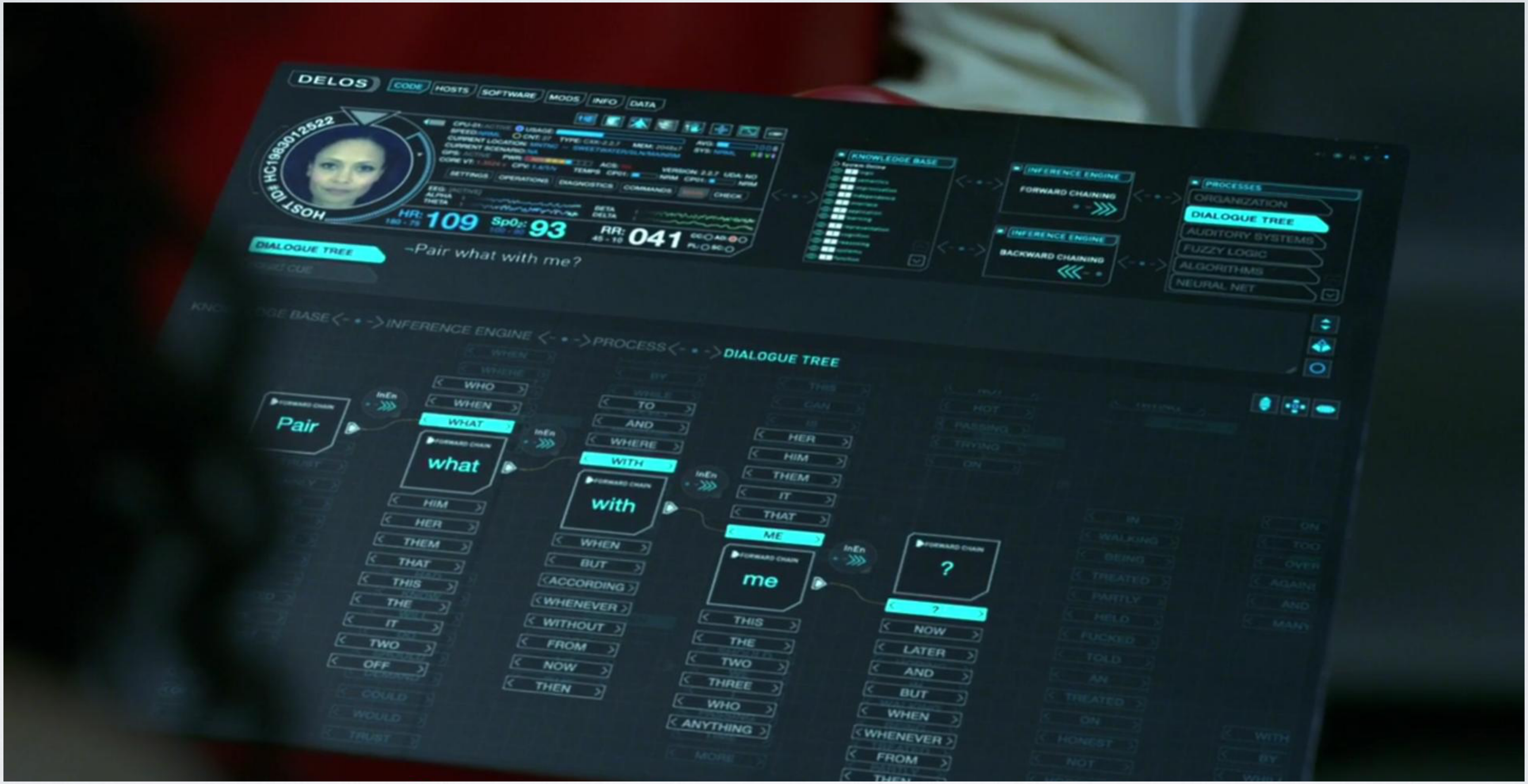
"OLD" RESEARCH TODAY



```
==> (explain normal-forces)
REASON: rear-wheels-force decreased AND
    its magnitude exceeds the traction
    threshold.
Since the rear wheels lost traction
    the friction of the contact patches
        MUST HAVE decreased;
    so, the normal forces MUST HAVE
    decreased.
Consistent with the accelerometers.
```



“OLD” RESEARCH TODAY



-Pair what with me?

NOW, ADDING "BIG DATA"

- Probabilistic methods
- Machine learning - computational statistics
  - Neural networks
  - Deep learning
  - Reinforcement learning
- Anomaly detection

WHAT AI CANNOT DO…

- Everything that cannot be expressed as an optimization problem
    - No big questions - What is life?
    - Need to "translate" - chat bot responses and image recognition
- General AI - AI is completely **domain** specific

WHAT HAS AI DONE?

A brief history, and looking forward

WHERE DID WE START

- Relies on a history of computing
  - Computing power and memory was expensive
  - Largely theoretical

WHEN HAVE WE DONE "GOOD ENOUGH"

- "Good enough" - solving problems better than humans
  - Then, the original problem does not seem so hard
- General or particular effect(s)
  - In finance, when machines trade better than humans, we do not see the risk in "good enough"
  - We will see this in all examples (medicine, etc).

THREE WAVES OF AI

1. The first wave - Definitions

2. The second wave - Diagnosis

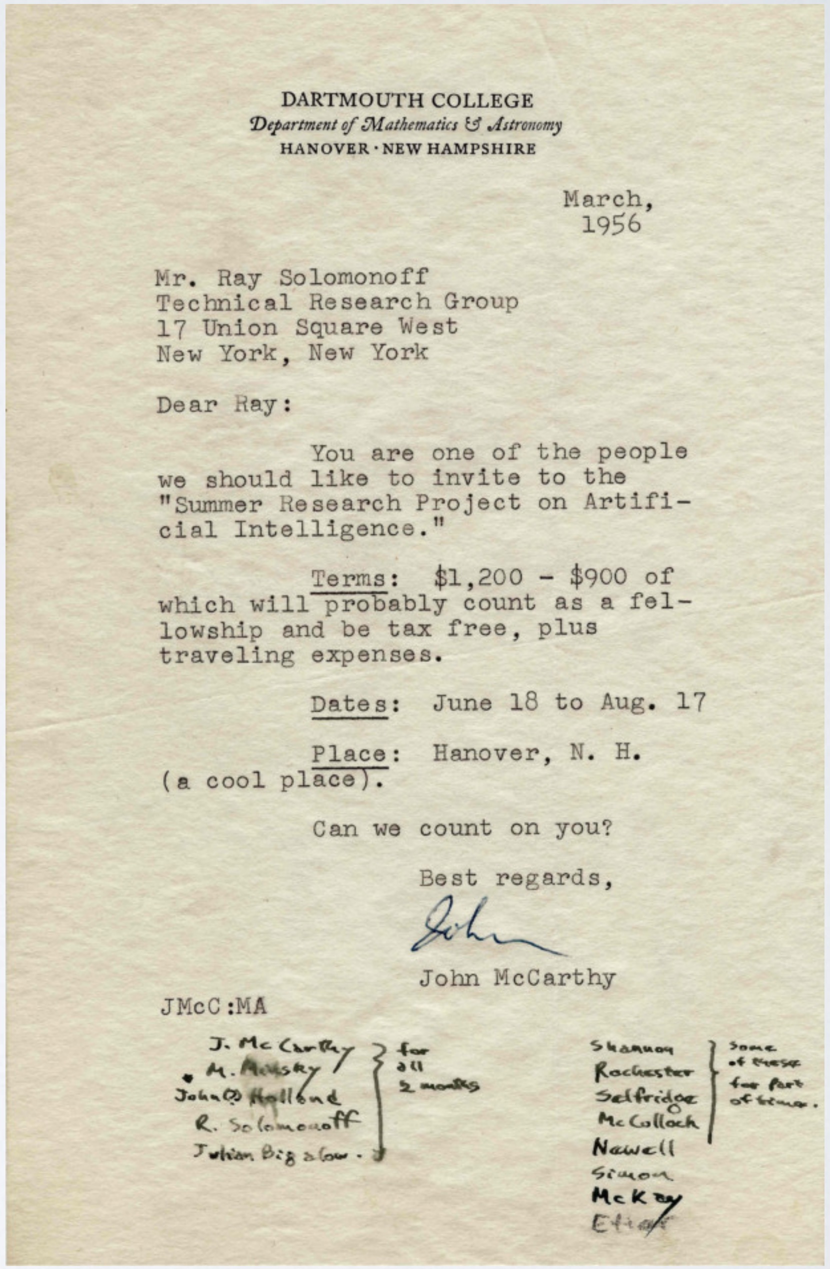
3. The third wave - “playing games” or captioning **(currently)**

THE TURING TEST

- Is this intelligence?
- A computer would need the following:
    - Natural language processing
    - Knowledge representation
    - Automated reasoning
    - Machine learning

THE DARTMOUTH CONFERENCE

- 1956
- Organized by Marvin Minsky, John McCarthy, Claude Shannon (IBM) and Nathan Rochester (IBM)
- Gained its name, mission, and its major contributors



DARTMOUTH COLLEGE
Department of Mathematics & Astronomy
HANOVER · NEW HAMPSHIRE

March,
1956

Mr. Ray Solomonoff
Technical Research Group
17 Union Square West
New York, New York

Dear Ray:

You are one of the people we should like to invite to the "Summer Research Project on Artificial Intelligence."

Terms: $1,200 - $900 of which will probably count as a fellowship and be tax free, plus traveling expenses.

Dates: June 18 to Aug. 17

Place: Hanover, N. H. (a cool place).

Can we count on you?

Best regards,

John

John McCarthy

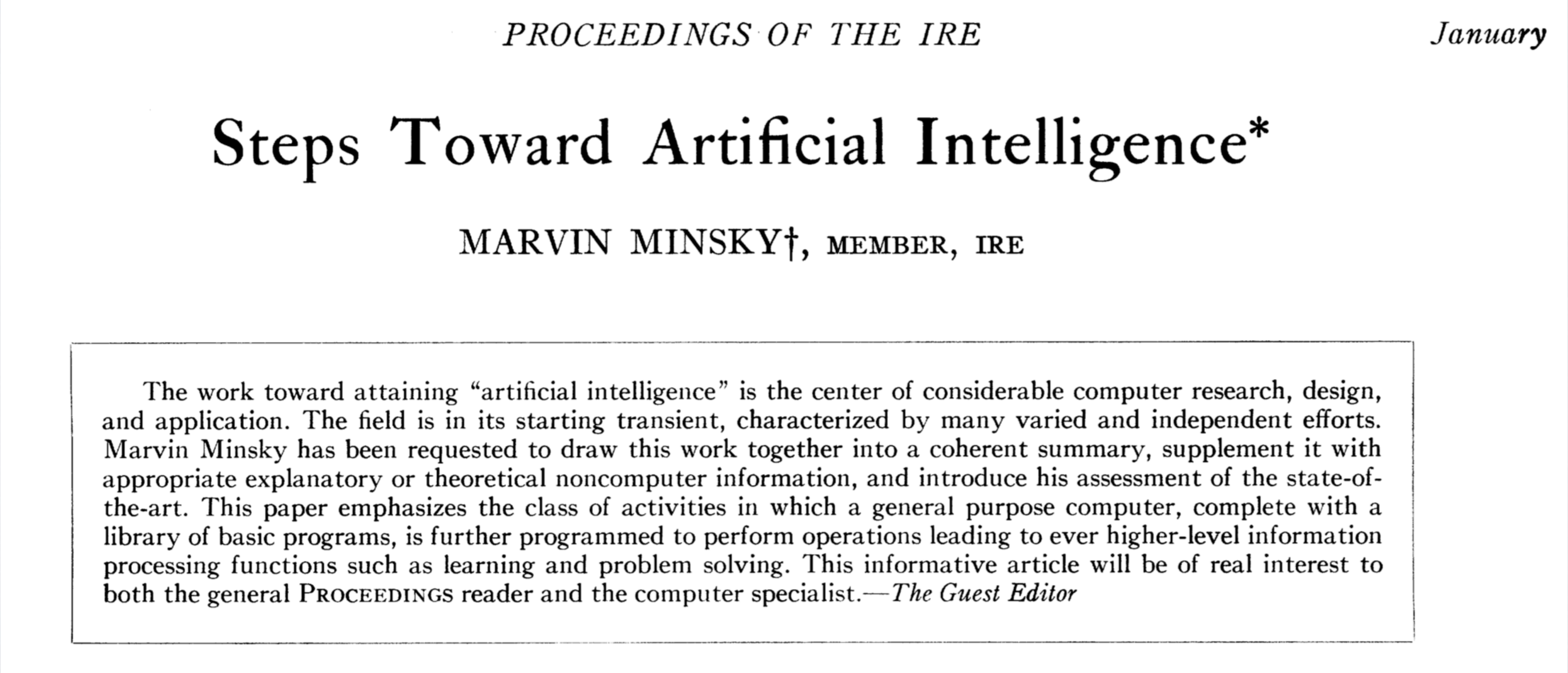
JMcC:MA

J. McCarthy
M. Minsky
John (?) Holland
R. Solomonoff
Julian Bigelow.
for all 2 months

Shannon
Rochester
Selfridge
McCulloch
Newell
Simon
McKay
Elias
some of these for part of time.

MINSKY’S 5 STEPS

Search, Pattern Recognition, Learning, Planning, and Induction



PROCEEDINGS OF THE IRE *January*

Steps Toward Artificial Intelligence*

MARVIN MINSKY†, MEMBER, IRE

The work toward attaining “artificial intelligence” is the center of considerable computer research, design, and application. The field is in its starting transient, characterized by many varied and independent efforts. Marvin Minsky has been requested to draw this work together into a coherent summary, supplement it with appropriate explanatory or theoretical noncomputer information, and introduce his assessment of the state-of-the-art. This paper emphasizes the class of activities in which a general purpose computer, complete with a library of basic programs, is further programmed to perform operations leading to ever higher-level information processing functions such as learning and problem solving. This informative article will be of real interest to both the general PROCEEDINGS reader and the computer specialist.—*The Guest Editor*

ASIDE: WINOGRAD SCHEMA

- Inspired by Terry Winograd in the spirit of the Turing Test: A schema consists of 3 parts

- A sentence or brief discourse that contains

  - Two noun phrases of the same semantic class (male, female, inanimate).

  - An ambiguous pronoun that may refer to either of the above noun phrases.

  - A special word and alternate word, such that if the special word is replaced with the alternate word, the natural resolution of the pronoun changes.

One more definition

ASIDE: WINOGRAD SCHEMA

- A question with:
  - The ambiguous pronoun
  - And provides two answer choices corresponding to the noun phrases in question
- Machine is given the problem in standardized form - binary decision problem

One more definition

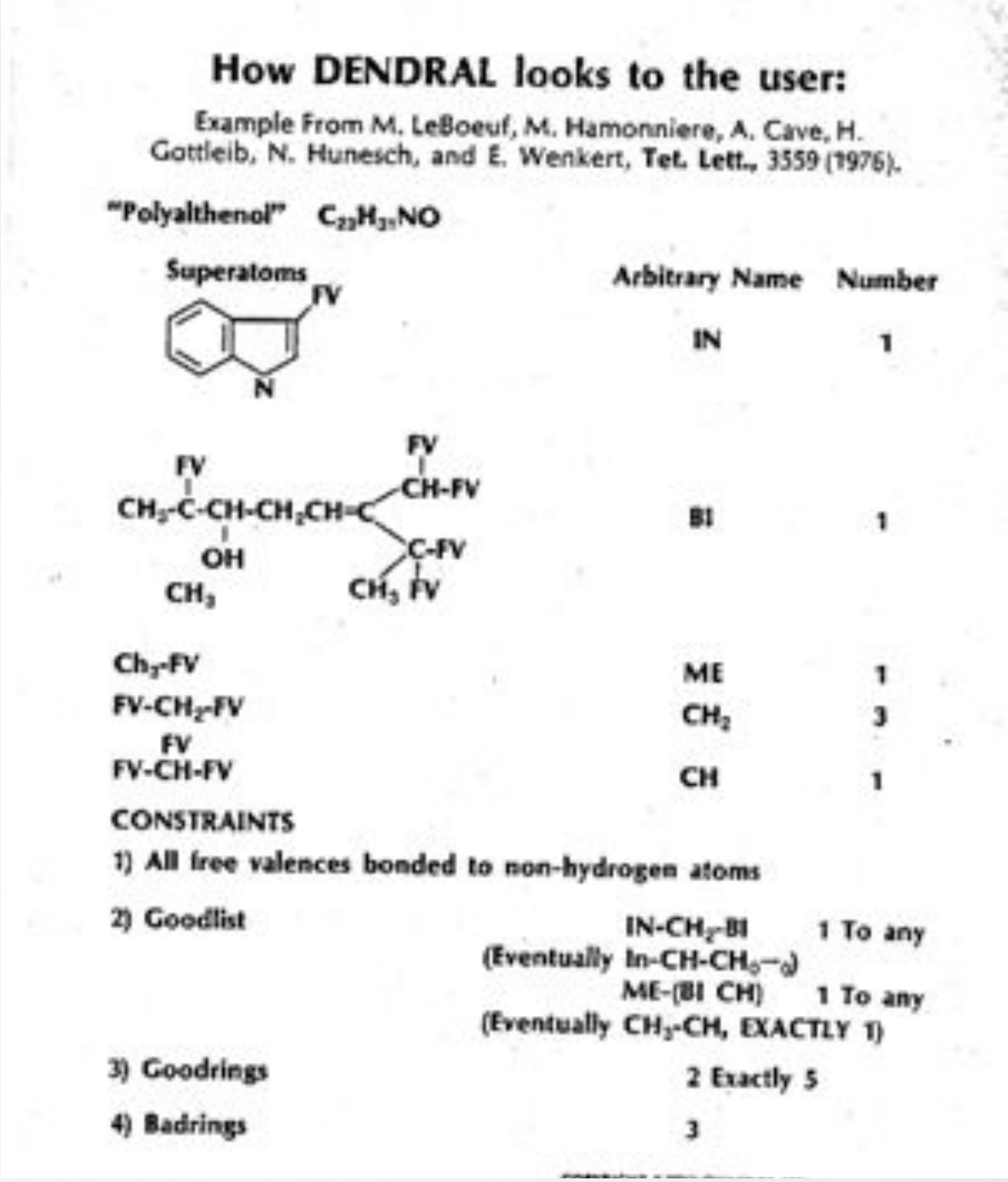
ASIDE: WINOGRAD SCHEMA

- Example: the trophy doesn't fit in the brown suitcase because it is too big. What is too big?

- At IJCAI-16, the highest score was 58% correct by Liu et al.

DENDRAL

- Developed in the 1960s
- Edward Feigenbaum and Raj Reddy won the Turing Award for it in 1994.
- Rule based system to help organic chemists identify unknown organic molecules



How DENDRAL looks to the user:

Example From M. LeBoeuf, M. Hamonniere, A. Cave, H. Gottleib, N. Hunesch, and E. Wenkert, **Tet. Lett.**, 3559 (1976).

"Polyalthenol" $C_{23}H_{31}NO$

Superatoms	Arbitrary Name	Number
	IN	1
	BI	1
Ch_3-FV	ME	1
FV-CH_2-FV	CH_2	3
FV FV-CH-FV	CH	1

FV
N
FV
FV
CH_3-C-CH-CH_2CH=C
CH-FV
OH
C-FV
CH_3
CH_3 FV

CONSTRAINTS

1) All free valences bonded to non-hydrogen atoms

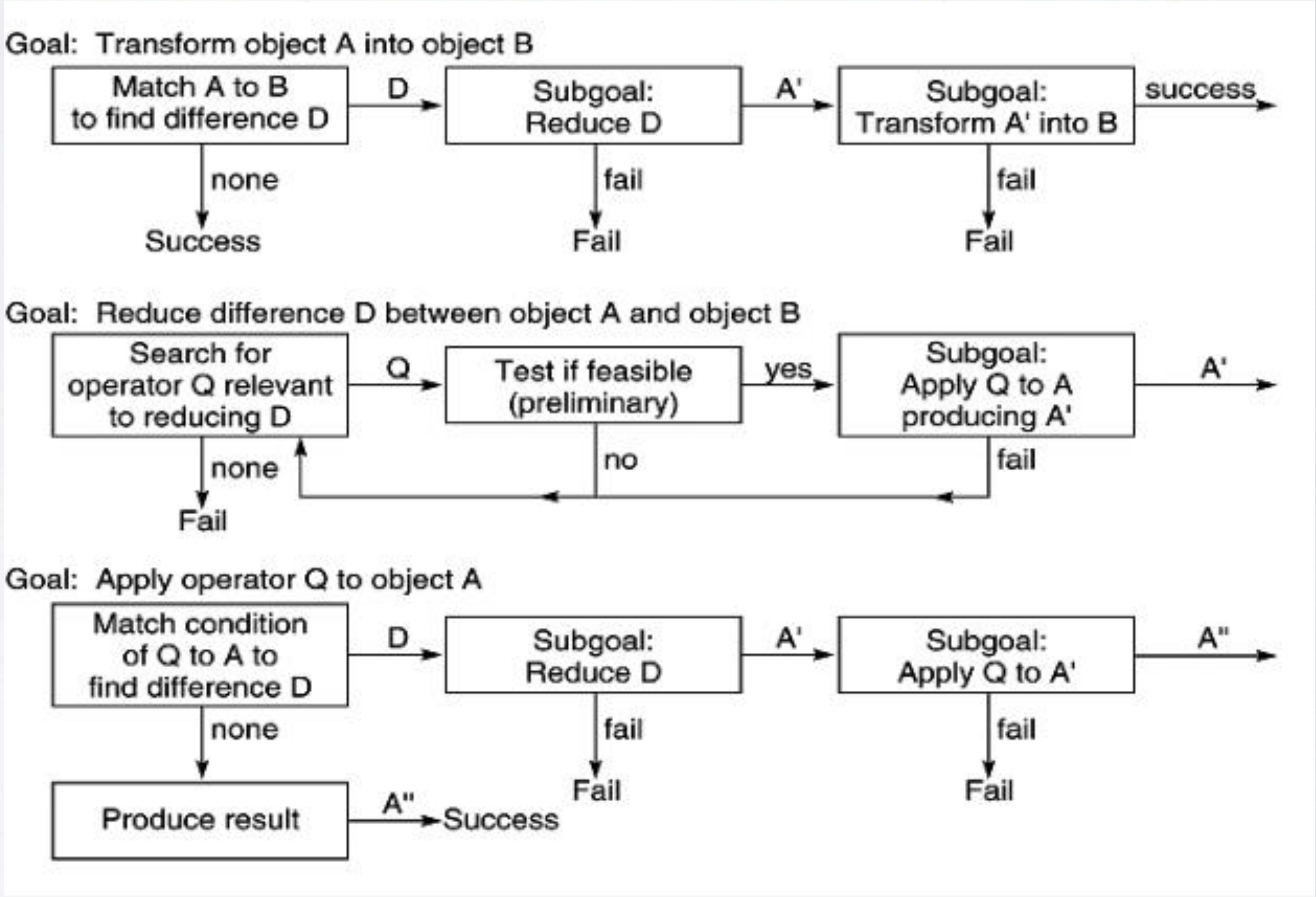
2) Goodlist
IN-CH_2-BI 1 To any
(Eventually In-CH-CH_0—$_0$)
ME-(BI CH) 1 To any
(Eventually CH_3-CH, EXACTLY 1)

3) Goodrings 2 Exactly 5

4) Badrings 3

PROBLEM SOLVERS

- Slagle - SAINT - Symbolic Integration in Freshman Calculus -> MACSYMA
- Geometry Theorem Prover - Gelertner
- General Problem Solver (Newell and Simon)
- Now
  - Mathematica


Goal: Transform object A into object B
Match A to B to find difference D
D
Subgoal: Reduce D
A'
Subgoal: Transform A' into B
success
none
Success
fail
Fail
fail
Fail
Goal: Reduce difference D between object A and object B
Search for operator Q relevant to reducing D
Q
Test if feasible (preliminary)
yes
Subgoal: Apply Q to A producing A'
A'
none
Fail
no
fail
Goal: Apply operator Q to object A
Match condition of Q to A to find difference D
D
Subgoal: Reduce D
A'
Subgoal: Apply Q to A'
A"
none
fail
Fail
fail
Fail
Produce result
A"
Success

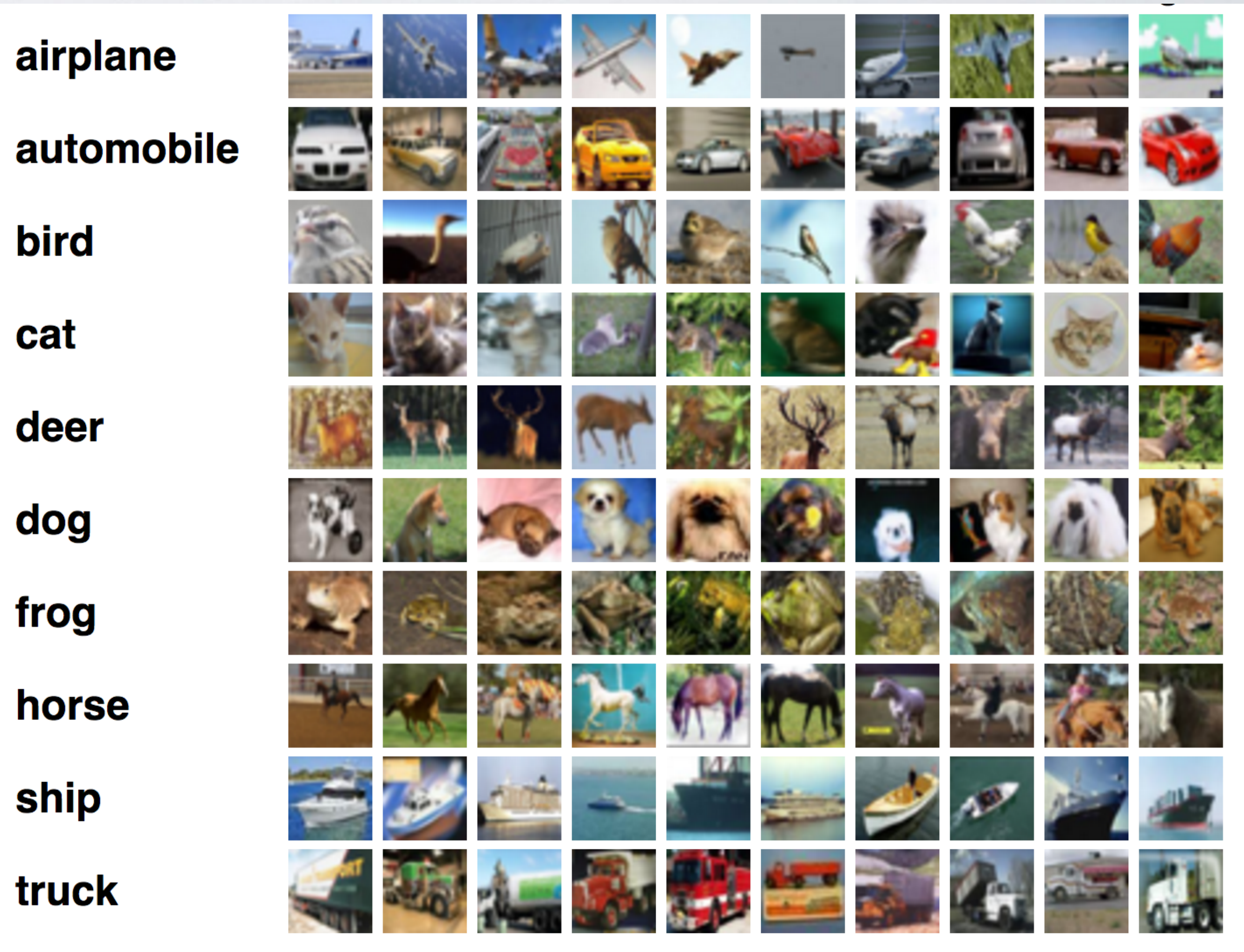
PLAYING GAMES

- Chess
- Jeopardy
- Go

The third wave - Games

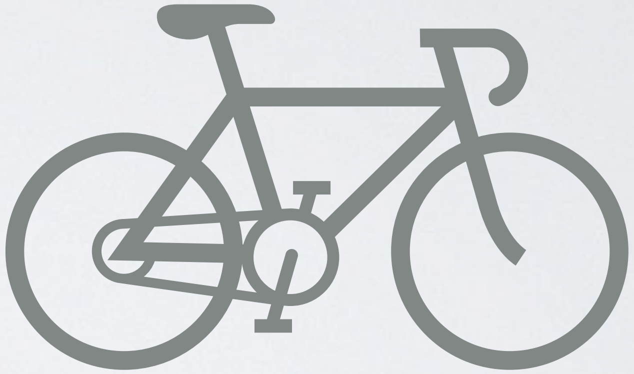
IMAGES

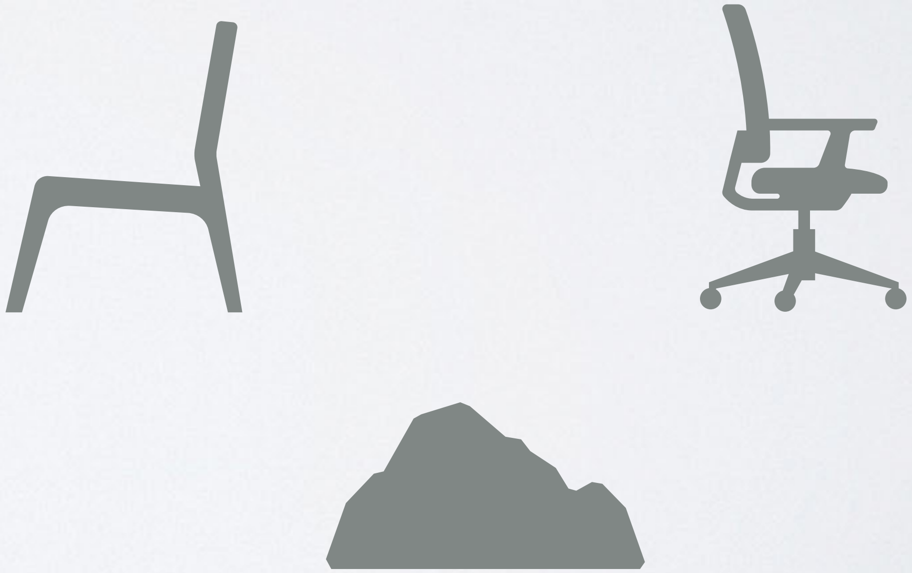
- Probabilistic methods
  - Machine learning
  - Bayesian
- Neural networks
- Deep Learning
- Classification
- Towards perception in self-driving cars



BUT…..

- Polanyi's paradox
  - We know more than we can tell
  - Riding a bike
- Concept understanding





WHAT WILL AI DO?

- More data ==> better classification?

WHAT WILL AI DO?

- More data ==> better classification?
  - Everybody lies…
  - Netflix recommendations without a queue

WHAT WILL AI DO?

- More data ==> more bias ==> more harm

WHAT WILL AI DO?

- More data ==> more bias ==> more harm
  - Racial discrimination in Face Detection
  - Cameras ask if "blinking"

WILL ROBOTS TAKE OVER?

"…such machines, by which the scholar may, by turning a crank, grind out the solution of a problem without the fatigue of mental application, would by its introduction into schools, do incalculable injury. But who knows that such machines when brought to greater perfection, may not think of a plan to remedy all their own defects and then grind out ideas beyond the ken of mortal mind!"

–R. Thorton in 1847 (about the invention of a four function mechanical calculator)

…"once the machine thinking method has 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"

–Alan Turing in 1951 *Intelligent Machinery : A Heretical Theory*

“We will soon create intelligences greater than our own. When this happens, human history will have reached a kind of singularity, an intellectual transition as impenetrable as the knotted space-time at the center of a black hole, and the world will pass far beyond our understanding. This singularity, I believe, already haunts a number of science-fiction writers. It makes realistic extrapolation to an interstellar future impossible. To write a story set more than a century hence, one needs a nuclear war in between … so that the world remains intelligible”

–Vernor Vinge in *Omni*, January 1983

"…A future period during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed. Although neither utopian nor dystopian, this epoch will transform the concepts that we rely on to give meaning to our lives, from our business models to the cycle of human life, including death itself."

–Ray Kurzweil in *The Singularity is Near*

"'By a 'superintelligence' we mean an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills. This definition leaves open how the superintelligence is implemented: it could be a digital computer, an ensemble of networked computers, cultured cortical tissue or what have you. It also leaves open whether the superintelligence is conscious and has subjective experiences."

–Nick Bostrom in *How Long Before Superintelligence*

WHAT IS THE SINGULARITY?

- When "artificial intelligence" becomes better than human intelligence (by some measure of better)

- Permanent state when humanity is "overthrown" by machine intelligence.

- Machines can improve themselves (hardware and software).

REQUIREMENTS FOR MACHINES TO TAKE OVER

- Superintelligence?
- Increase current access to power?
- Increase current computation power?
- Time?

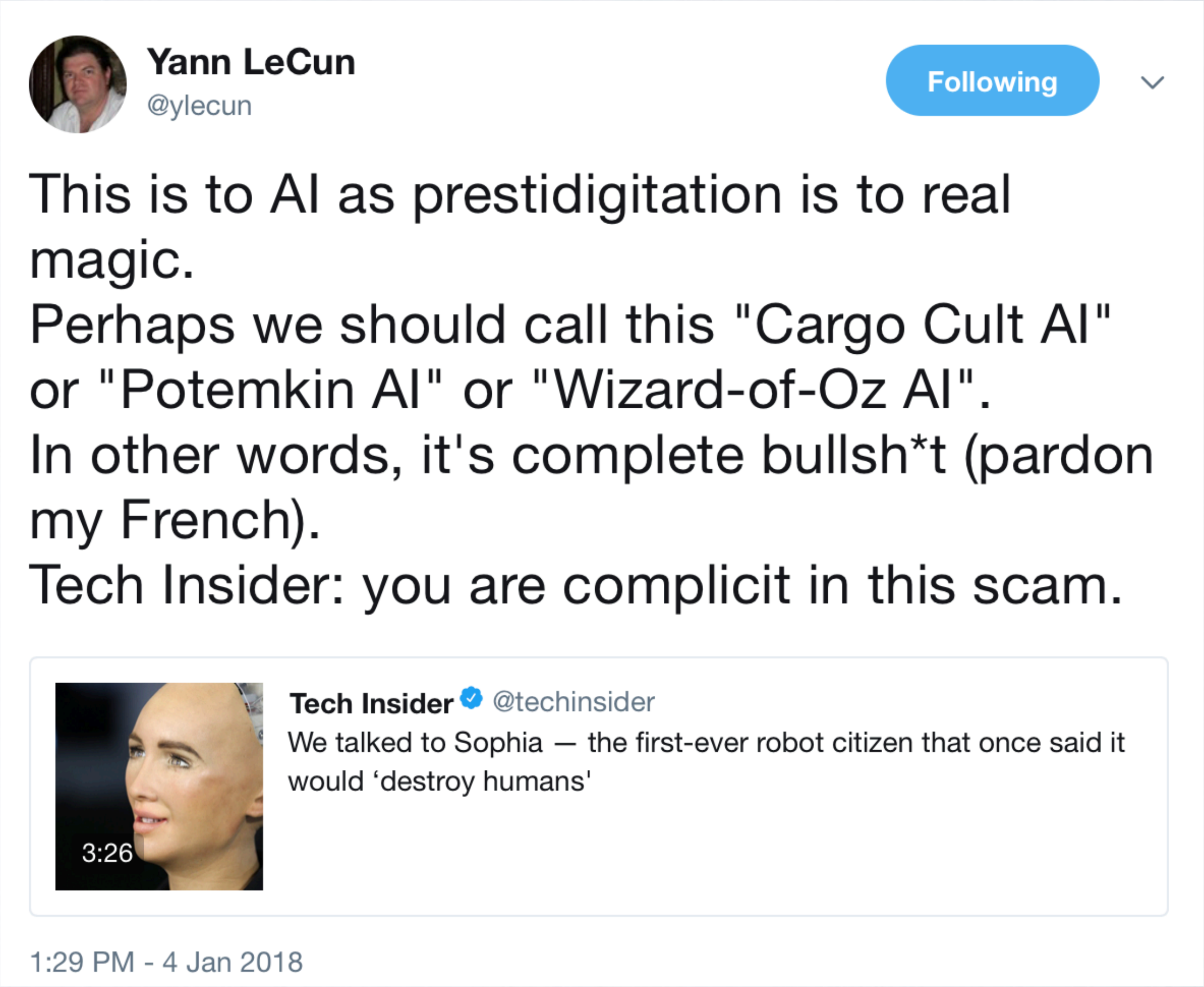
REQUIREMENTS FOR MACHINES TO TAKE OVER

- ~~Superintelligence?~~
- ~~Increase current access to power?~~
- ~~Increase current computation power?~~
- ~~Time?~~

Someone could do it today…

SO WHY IS IT IMPORTANT

- Robot rights or popular news?
  - Sofia - Citizen of Saudi Arabia
- Media cover or science?



Yann LeCun
@ylecun

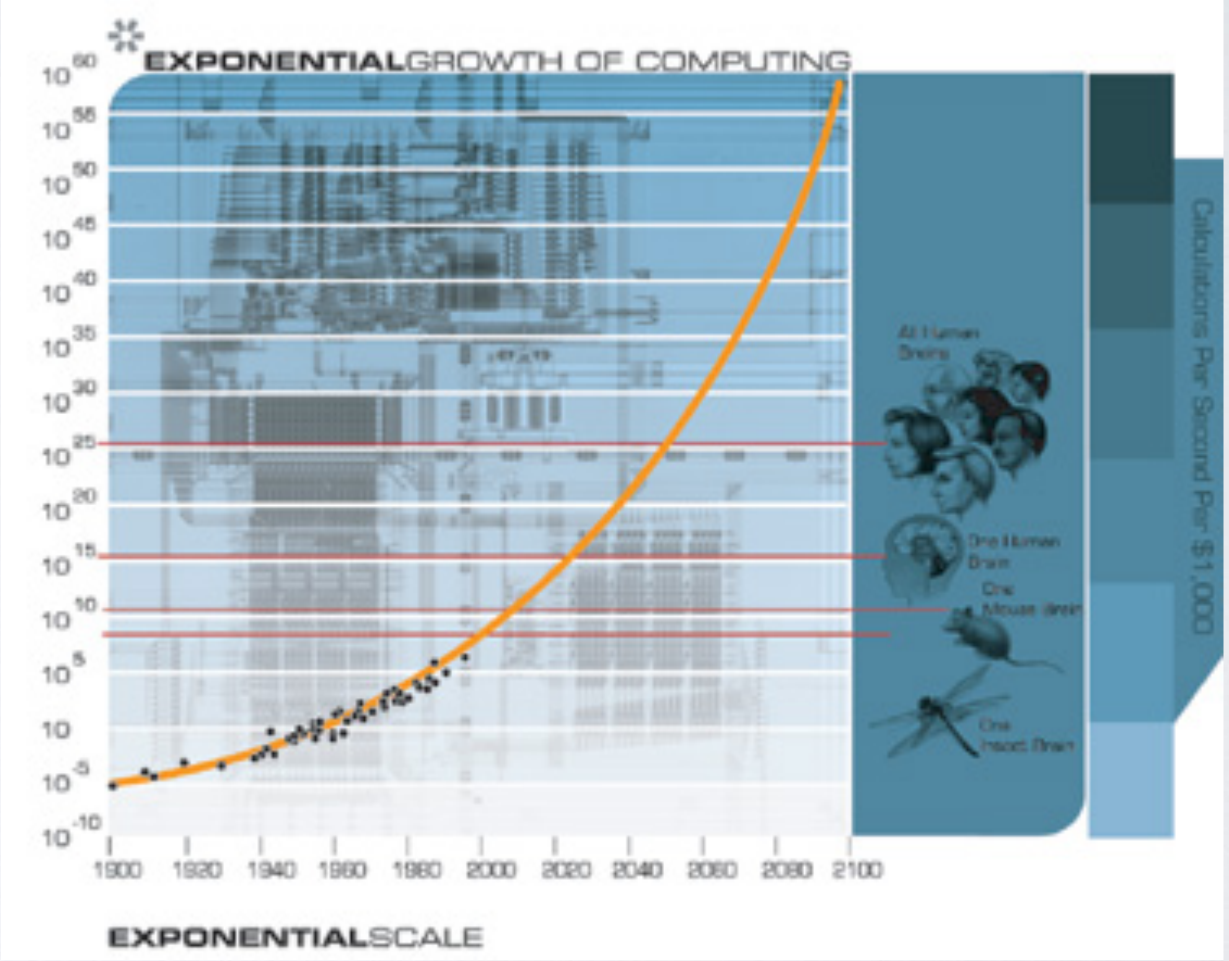
This is to AI as prestidigitation is to real magic.
Perhaps we should call this "Cargo Cult AI" or "Potemkin AI" or "Wizard-of-Oz AI".
In other words, it's complete bullsh*t (pardon my French).
Tech Insider: you are complicit in this scam.

Tech Insider @techinsider
We talked to Sophia — the first-ever robot citizen that once said it would 'destroy humans'

1:29 PM - 4 Jan 2018

WHEN WILL THIS ALL HAPPEN

- Law of accelerating returns
- The singularity isn't near
- How to Keep up with AI
  - Get Bigger brains - NYT
  - We don't understand AI - Tech Review



courtesy of kurzweil.net

CONTRIBUTIONS OF THIS COURSE

- Provide credible sources about AI
- Recognize [social and existential] risks and pose how to think about possible solutions
- Understand AI and AI regulation across the globe
    - Europe vs US vs China