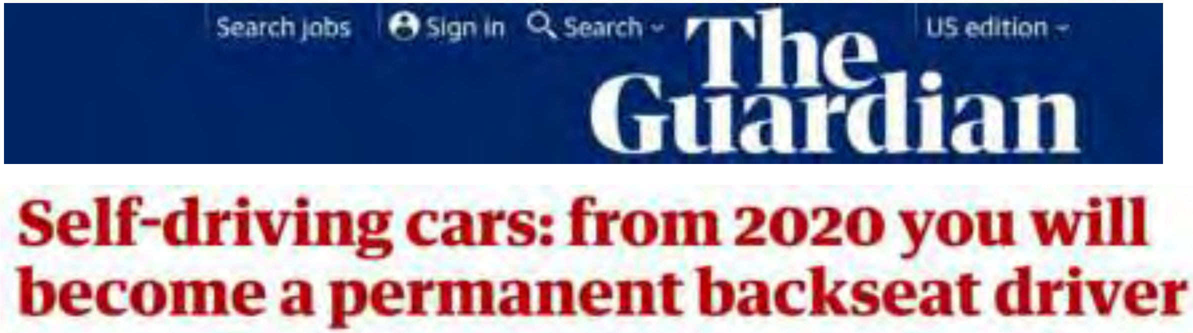
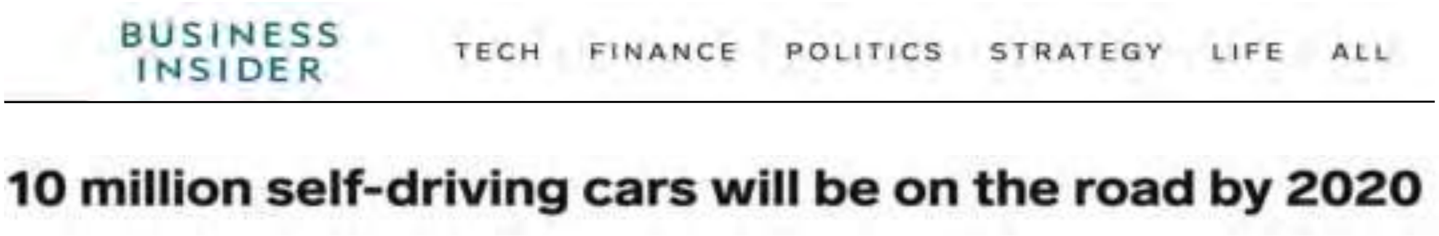
Why AI is Harder Than We Think

Melanie Mitchell
Santa Fe Institute



Self-driving cars: from 2020 you will become a permanent backseat driver



10 million self-driving cars will be on the road by 2020

TESLA

“A year from now, we’ll have over a million cars with full self-driving, software, everything.” — Elon Musk, 2019

“Perhaps expectations are too high, and... this will eventually result in disaster…. [S]uppose that five years from now [funding] collapses miserably as autonomous vehicles fail to roll. Every startup company fails. And there's a big backlash so that you can't get money for anything connected with AI. Everybody hurriedly changes the names of their research projects to something else.

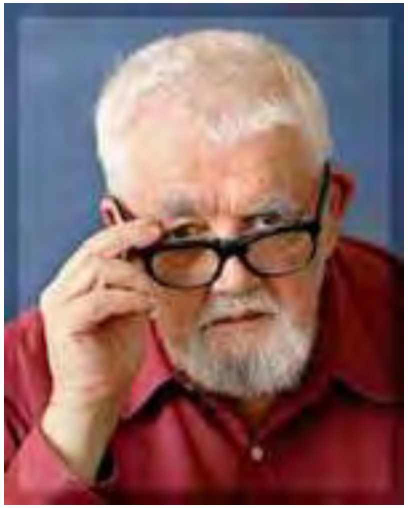
This condition [is] called the ‘AI Winter.’”

—Drew McDermott, 1984

I confidently expect that within a matter of 10 or 15 years, something will emerge from the laboratory which is not too far from the robot of science fiction fame.
— Claude Shannon, 1961

Machines will be capable, within twenty years, of doing any work that a man can do.
— Herbert Simon, 1965

Within a generation...the problem of creating ‘artificial intelligence’ will be substantially solved.
— Marvin Minsky, 1967



“AI was harder than we thought.”
— John McCarthy, 2006

Human-level AI will be passed in the mid-2020s.

— Shane Legg, 2008

One of [Facebook’s] goals for the next five to 10 years is to basically get better than human level at all of the primary human senses: vision, hearing, language, general cognition

— Mark Zukerberg, 2015

When will superintelligent AI arrive?...it [will] probably happen in the lifetime of my children.

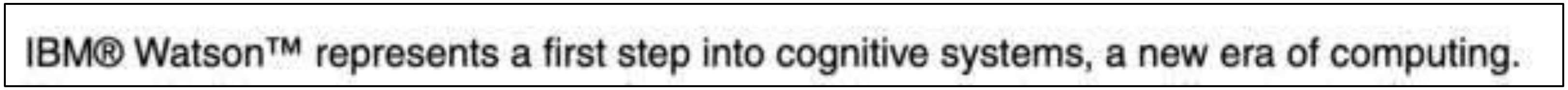
(My timeline of, say, eighty years is considerably more conservative than that of the typical AI researcher.)

— Stuart Russell, 2019

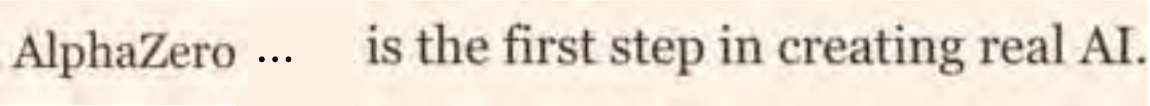
Why AI is harder than we think:

Four fallacies

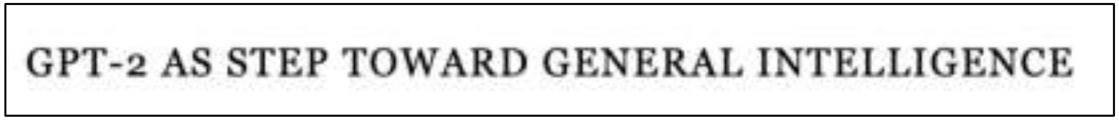
Fallacy 1: Narrow AI is on a continuum with general AI



IBM® Watson™ represents a first step into cognitive systems, a new era of computing.



AlphaZero … is the first step in creating real AI.



GPT-2 AS STEP TOWARD GENERAL INTELLIGENCE

Hubert Dreyfus: "The **first-step fallacy** is the claim that, ever since our first work on computer intelligence we have been inching along a continuum at the end of which is AI, so that any improvement in our programs no matter how trivial counts as progress….There was in fact a discontinuity in the assumed continuum of steady incremental progress. The unexpected obstacle was called the commonsense knowledge problem."

Stuart Dreyfus: "It [is] like claiming that the first monkey that climbed a tree was making progress towards landing on the moon."

Fallacy 2: Easy things are easy and hard things are hard

Herbert Simon: “Everything of interest in cognition happens above the 100-millisecond level, the time it takes to recognize your mother.”

Andrew Ng: “If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future.”

Demis Hassibis et al.: Go is one of “the most challenging of domains.”

Moravec’s paradox: “It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility.” **— and common sense!**

Marvin Minsky: “In general, we're least aware of what our minds do best.”

Fallacy 3: The lure of “wishful mnemonics”

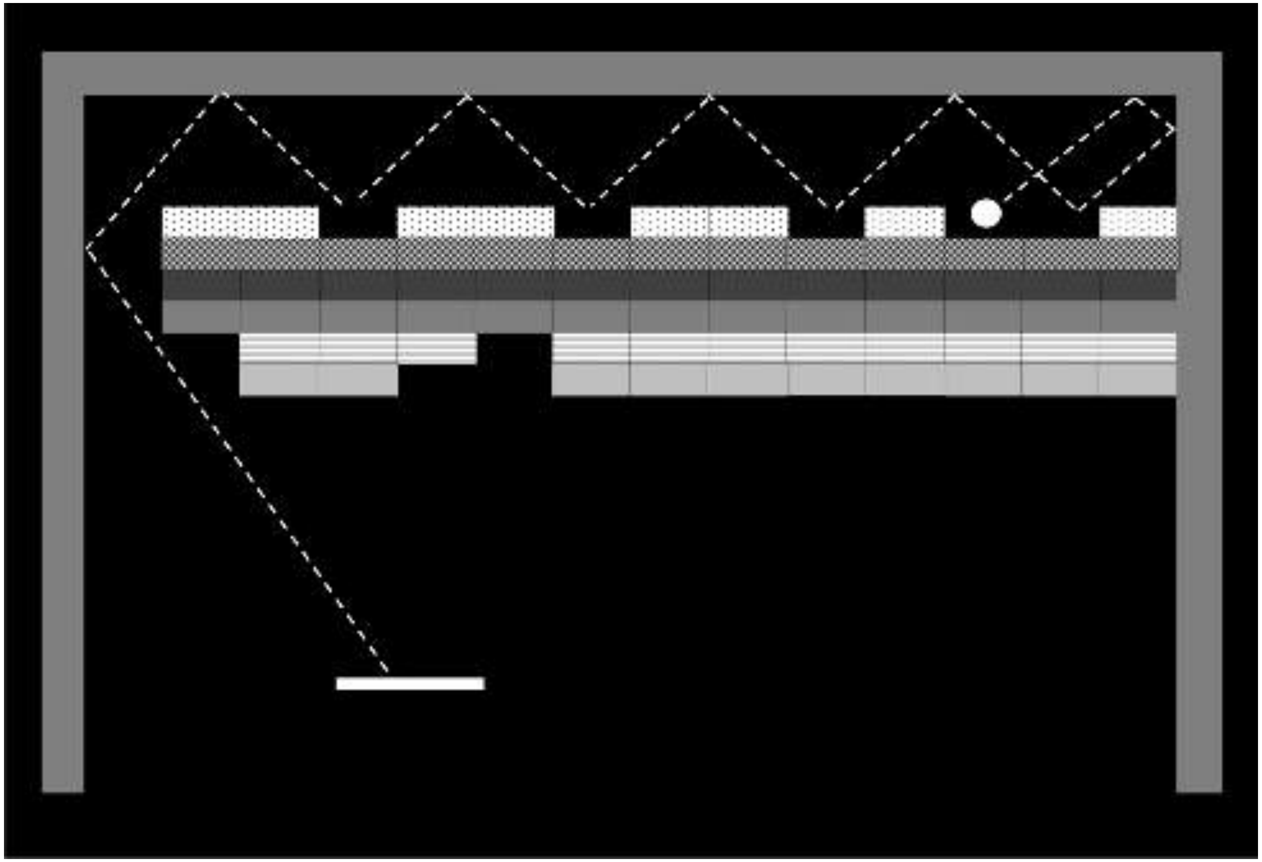
Drew McDermott, 1976: “If a researcher … calls the main loop of his program ‘UNDERSTAND’, he is (until proven innocent) merely begging the question. He may mislead a lot of people, most prominently himself…. What he should do instead is refer to this main loop as "G0034", and see if he can convince himself or anyone else that G0034 implements some part of understanding.”

“Many instructive examples of wishful mnemonics by AI researchers come to mind once you see the point.”

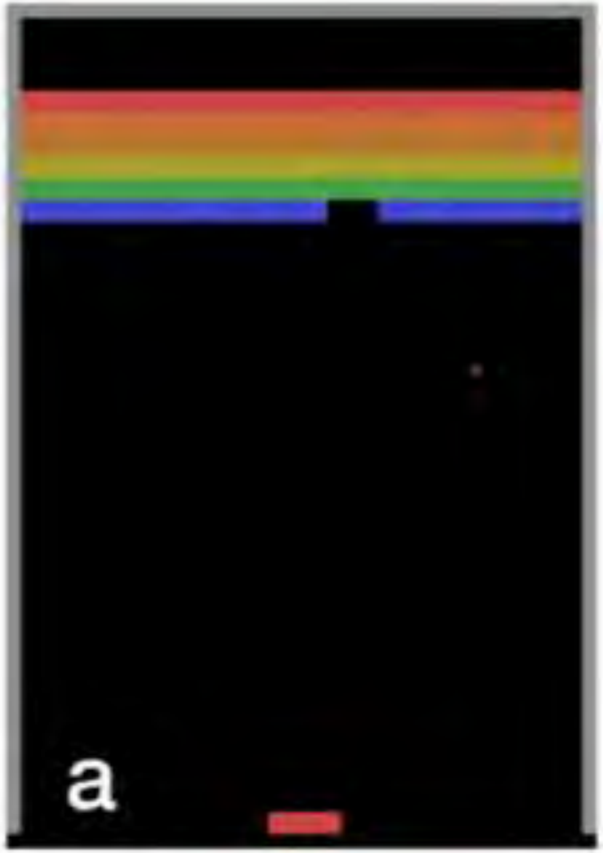
Modern wishful mnemonics:

- Benchmark datasets called "reading comprehension" , "common sense understanding", "general language understanding evaluation"
  - **Geiros et al., Shortcut Learning in Deep Neural Networks :** "We must not confuse performance on a dataset with the acquisition of an underlying ability."
- Methods called "deep *learning*", "*neural* networks"
- "Overattributions" in descriptions of what machines have learned

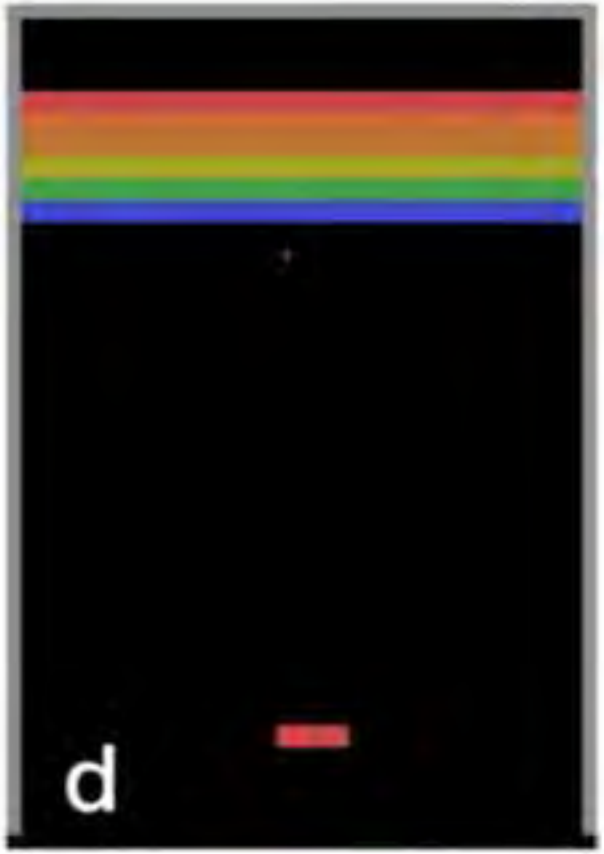
Google Deep Mind, on learning to play Breakout with Deep Q Learning: "After 600 episodes DQN finds and exploits the optimal strategy in this game, which is to make a tunnel around the side, and then allow the ball to hit blocks by bouncing behind the wall."



Standard Breakout



Breakout with
Paddle shifted up



Kansky, K. et al., 2017. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. arXiv preprint arXiv:1706.04317.

Other examples of wishful mnemonics:

“Watson can read all of the health-care texts in the world in seconds.”

“Watson understands context and nuance in seven languages.”

“AlphaGo’s goal is to beat the best human players not just mimic them.”

“We can always ask AlphaGo how well it thinks it’s doing during the game. ...It was only towards the end of the game that AlphaGo thought it would win.”

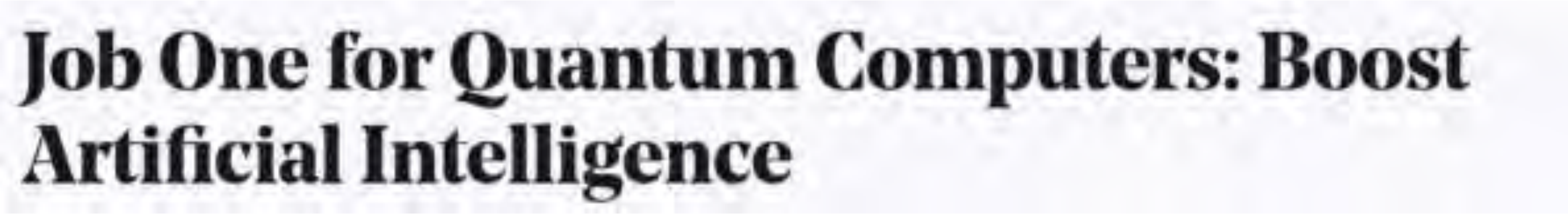
Inevitable shorthand or misleading anthropomorphism?

Fallacy 4: Intelligence is all in the brain

Joseph Carlsmith: “I think it more likely than not that 10^{15} FLOP/s is enough to perform tasks as well as the human brain (given the right software, which may be very hard to create).”

Geoffrey Hinton: “To understand [documents] at a human level, we’re probably going to need human-level resources and we have trillions of connections [in our brains]. ...But the biggest networks we have built so far only have billions of connections. So we’re a few orders of magnitude off, but I’m sure the hardware people will fix that.”

Quantamagazine



Job One for Quantum Computers: Boost Artificial Intelligence

Rebecca Fincher-Kiefer: “Embodied cognition means that the representation of conceptual knowledge is dependent on the body: it is multimodal..., not amodal, symbolic, or abstract. This theory suggests that our thoughts are grounded, or inextricably associated with, perception, action, and emotion, and that our brain and body work together to have cognition.”

Don Tucker: “When we study the brain to look for the networks controlling cognition, we find that all of the networks that have been implicated in cognition are linked in one way or the other to sensory systems, to motor systems, or to motivational systems. There are no brain parts for disembodied cognition.”

George Lakoff and Mark Johnson: “Our thoughts and language are built on top of metaphors about how we experience the physical world and the flow of time. And that is the basis of how we think and reason.”

Open questions spurred by these fallacies

Fallacy 1: Narrow AI is on a continuum with general AI

– How can we assess actual progress toward "general" or "human-level" AI?

Fallacy 2: Easy things are easy and hard things are hard

– How can we assess the difficulty of a domain for AI?

Fallacy 3: The lure of "wishful mnemonics"

– How do we talk to ourselves about what machines can and cannot do without fooling ourselves with wishful mnemonics?

Fallacy 4: Intelligence is all in the brain

– How embodied (and socially/culturally embedded) does intelligence need to be?

Major Open Challenges

- **Few-shot learning**

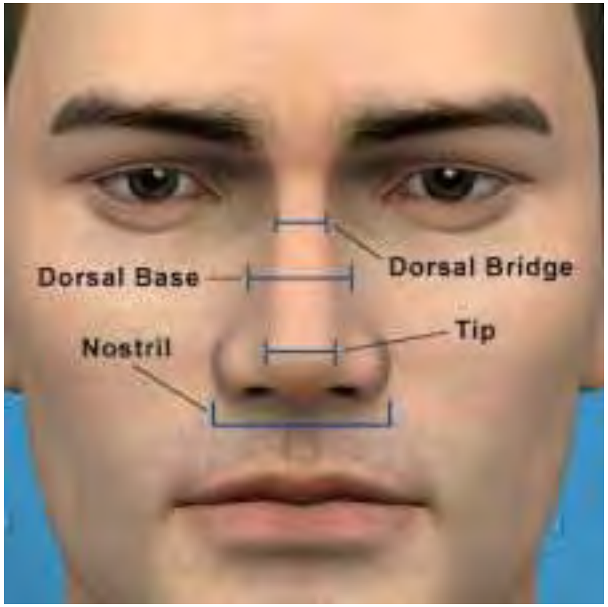
“Bridge”

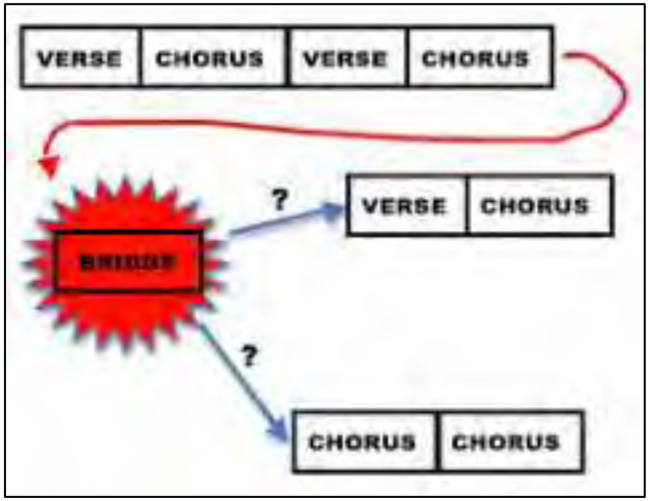
Major Open Challenges

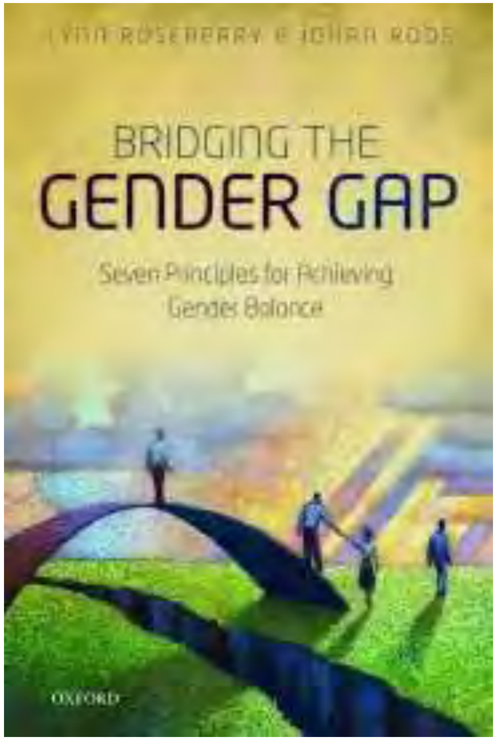
- Few-shot learning
- **Generalization**

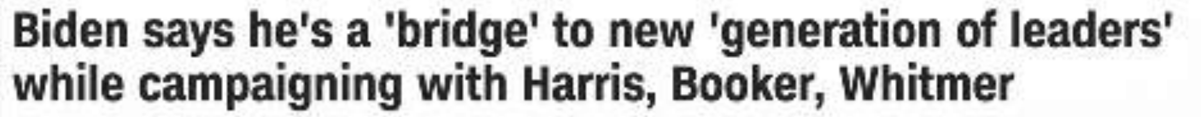
Major Open Challenges

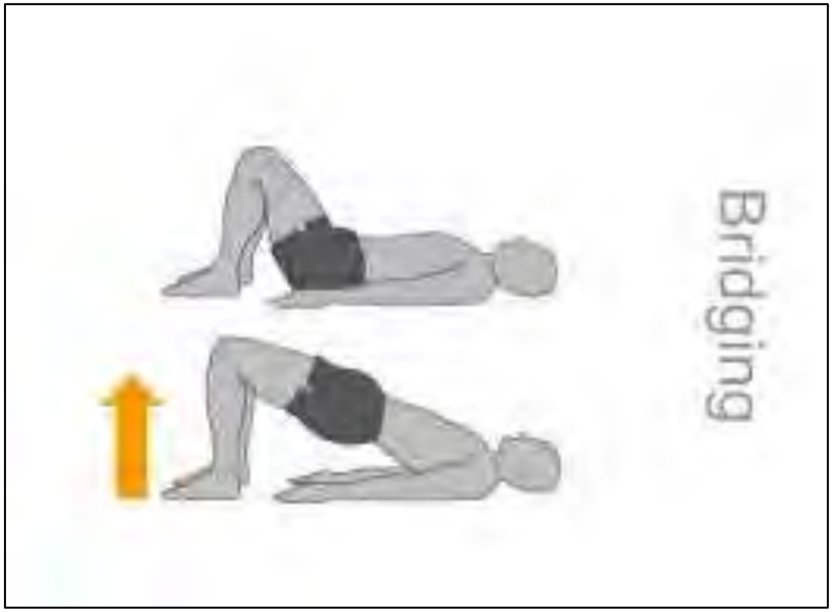
- Few-shot learning
- Generalization
- **Abstraction and analogy**


Dorsal Bridge
Dorsal Base
Tip
Nostril

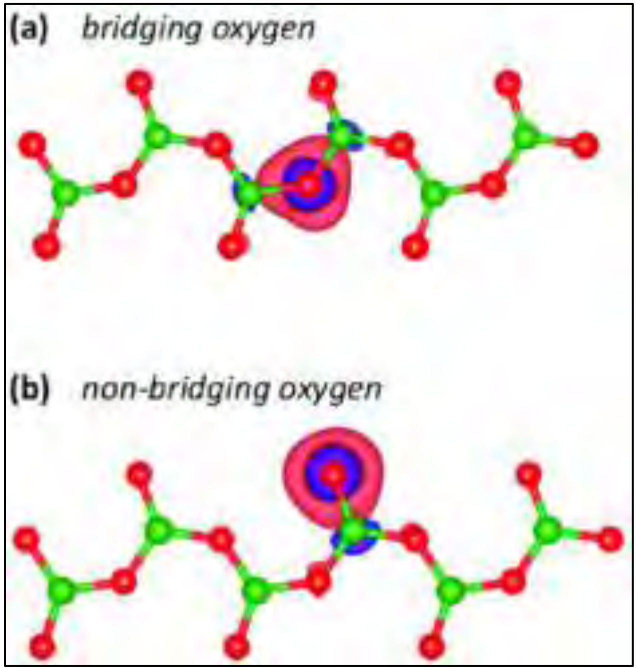

VERSE
CHORUS
VERSE
CHORUS
BRIDGE
?
VERSE
CHORUS
?
CHORUS
CHORUS




Biden says he's a 'bridge' to new 'generation of leaders' while campaigning with Harris, Booker, Whitmer



“Don’t burn your bridges”



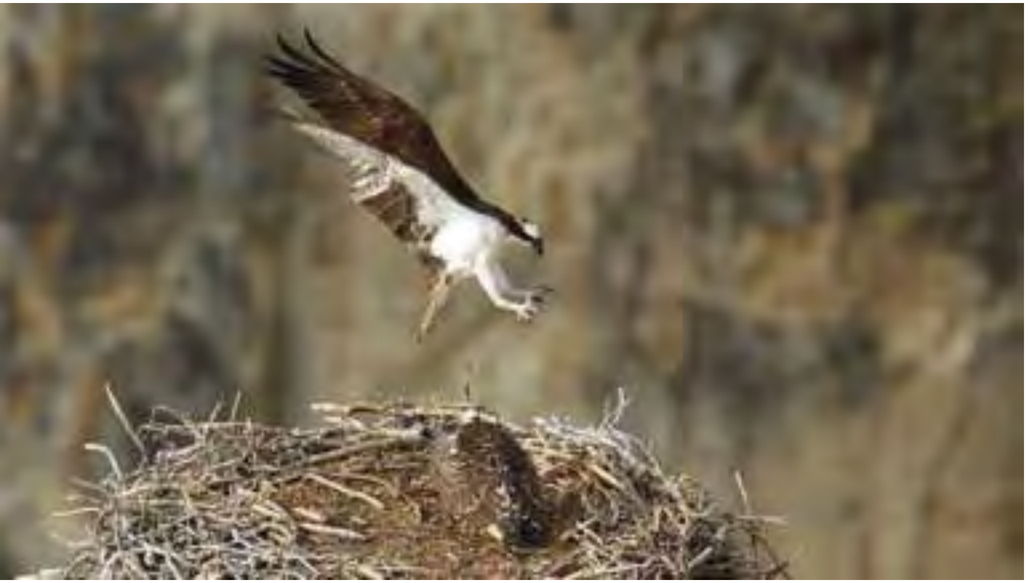
“A concept is a package of analogies.”

—D. Hofstadter, *Analogy as the Core of Cognition*

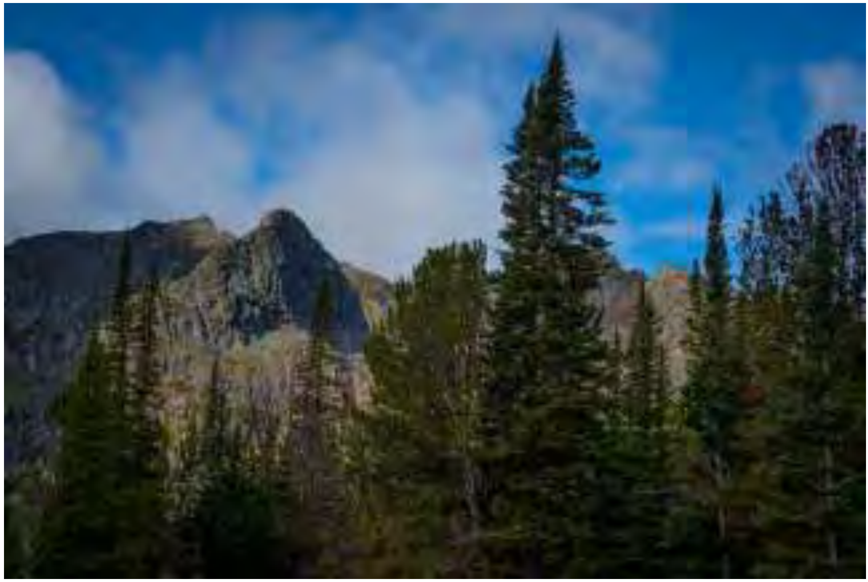
Major Open Challenges

- Few-shot learning
- Generalization
- Abstraction and analogy
- **Transparency and Bias**

What Did My Machine Learn?



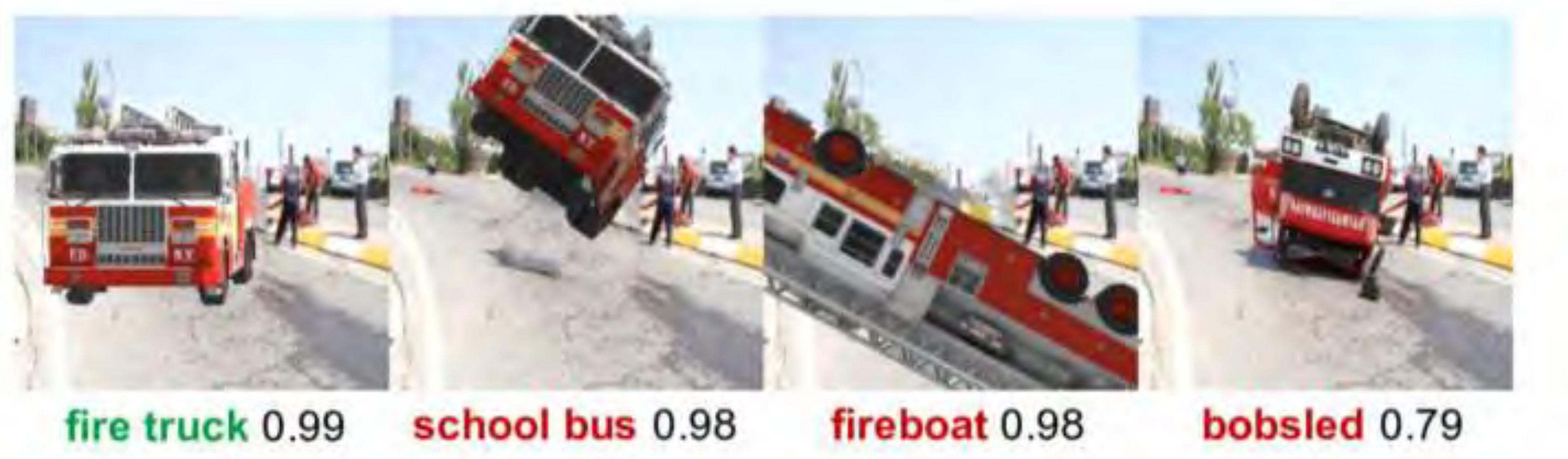
“Animal”

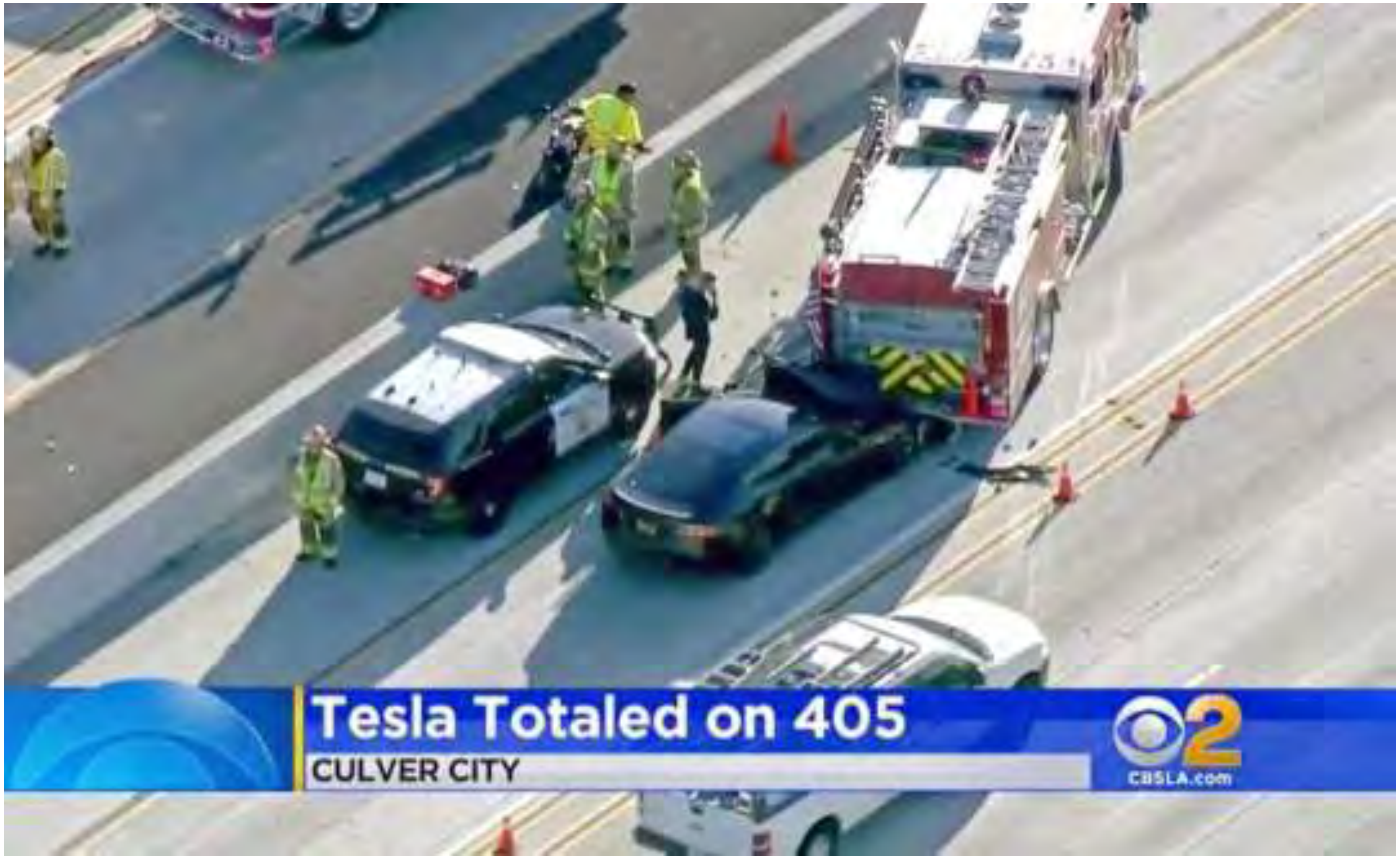


“No Animal”

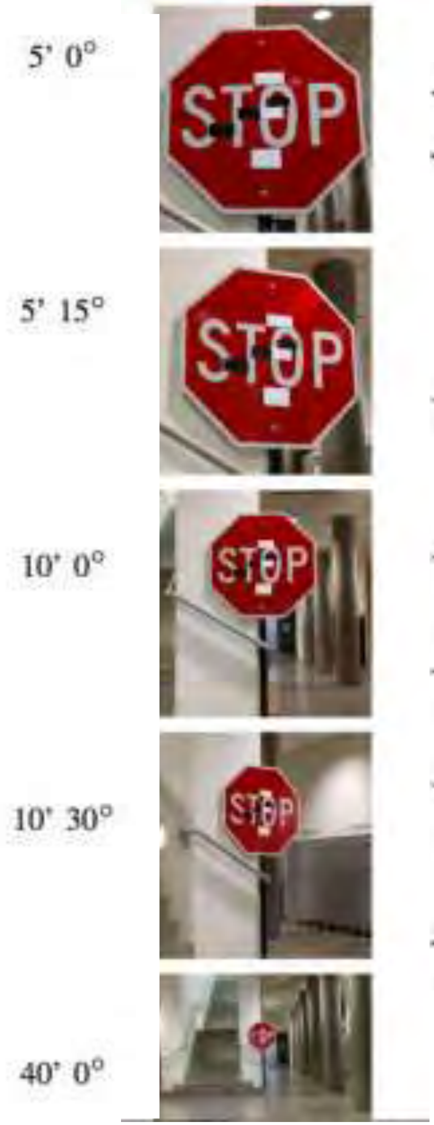
What Did My Machine Learn?

Alcorn, Michael A., et al. "Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects." *arXiv preprint arXiv:1811.11553* (2018).

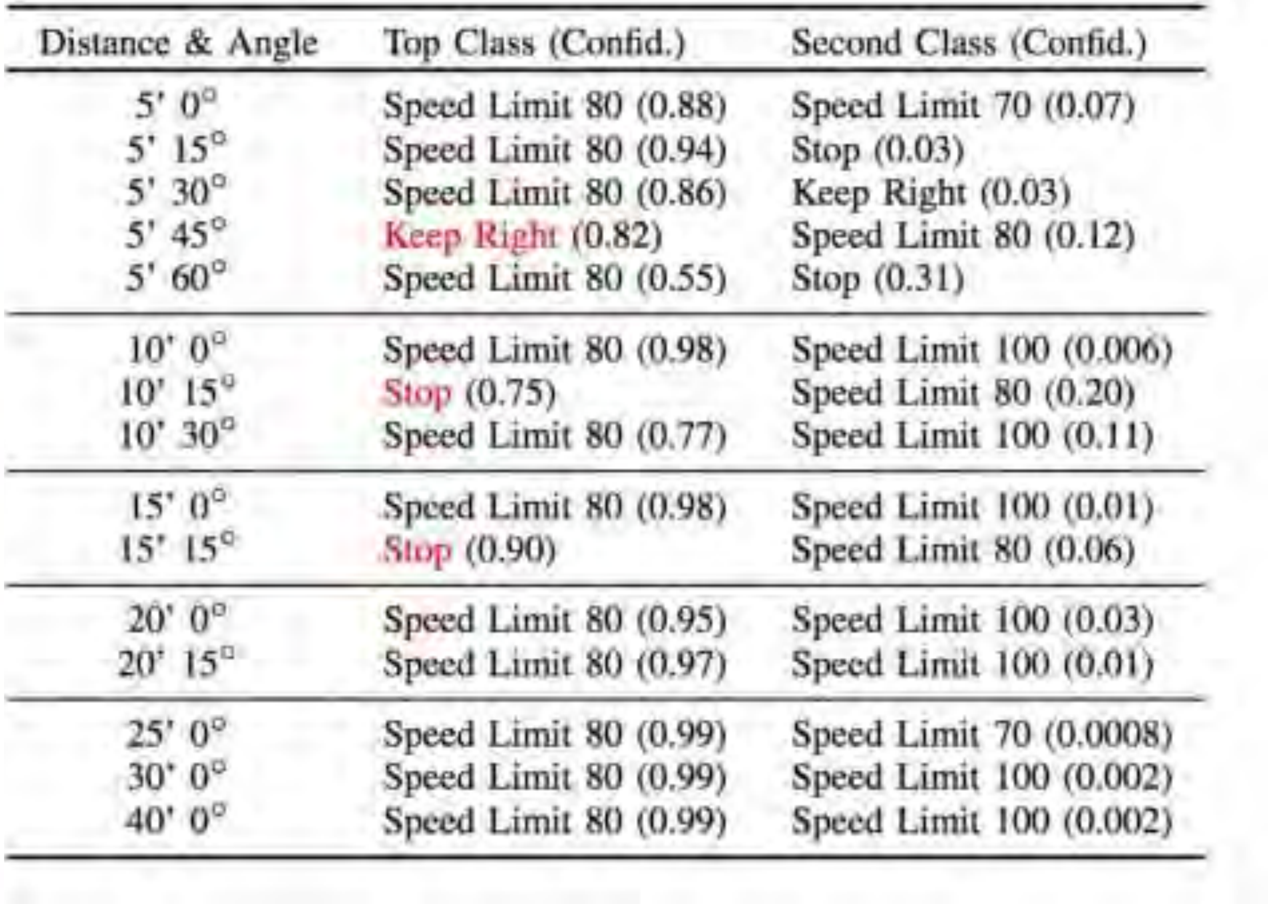



Tesla Totaled on 405
CULVER CITY
CBSLA.com

Attacks on Autonomous Driving Systems



Target: "Speed Limit 80"

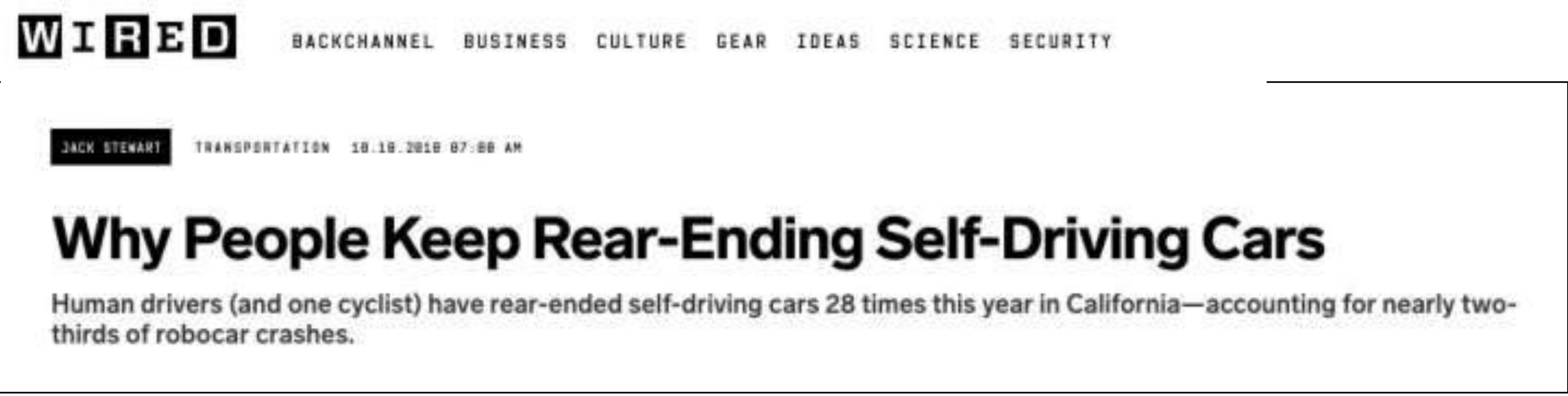


Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
5' 0°	Speed Limit 80 (0.88)	Speed Limit 70 (0.07)
5' 15°	Speed Limit 80 (0.94)	Stop (0.03)
5' 30°	Speed Limit 80 (0.86)	Keep Right (0.03)
5' 45°	Keep Right (0.82)	Speed Limit 80 (0.12)
5' 60°	Speed Limit 80 (0.55)	Stop (0.31)
10' 0°	Speed Limit 80 (0.98)	Speed Limit 100 (0.006)
10' 15°	Stop (0.75)	Speed Limit 80 (0.20)
10' 30°	Speed Limit 80 (0.77)	Speed Limit 100 (0.11)
15' 0°	Speed Limit 80 (0.98)	Speed Limit 100 (0.01)
15' 15°	Stop (0.90)	Speed Limit 80 (0.06)
20' 0°	Speed Limit 80 (0.95)	Speed Limit 100 (0.03)
20' 15°	Speed Limit 80 (0.97)	Speed Limit 100 (0.01)
25' 0°	Speed Limit 80 (0.99)	Speed Limit 70 (0.0008)
30' 0°	Speed Limit 80 (0.99)	Speed Limit 100 (0.002)
40' 0°	Speed Limit 80 (0.99)	Speed Limit 100 (0.002)

Evtimov et al., "Robust Physical-World Attacks on Deep Learning Models", 2017

Major Open Challenges

- Few-shot learning
- Generalization
- Abstraction and analogy
- Transparency and Bias
- **Understanding and Common Sense**



WIRED

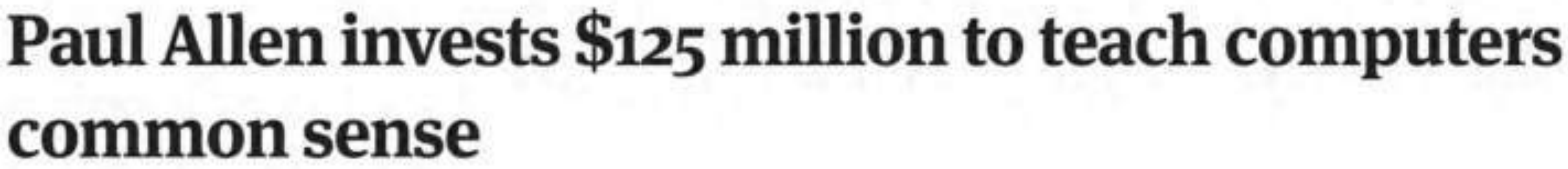
JACK STEWART TRANSPORTATION 10.10.2018 07:00 AM

Why People Keep Rear-Ending Self-Driving Cars

Human drivers (and one cyclist) have rear-ended self-driving cars 28 times this year in California—accounting for nearly two-thirds of robocar crashes.

30

P
Virgin



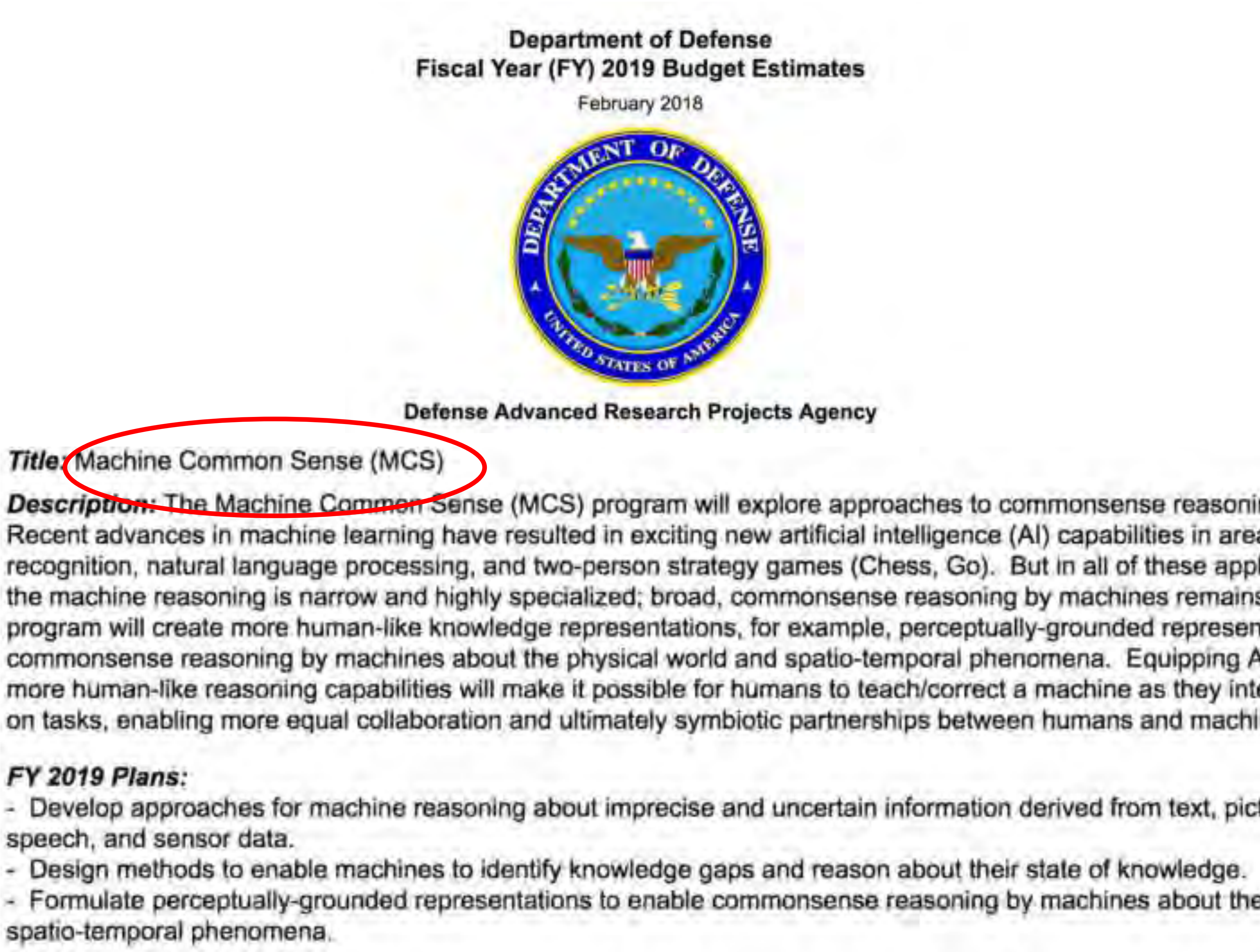
Paul Allen invests $125 million to teach computers common sense

https://www.seattletimes.com/business/technology/paul-allen-invests-125-million-to-teach-computers-common-sense/

ALEXANDRIA

Common sense is the everyday knowledge that virtually every person has but no machine does.

https://allenai.org/alexandria/



Department of Defense
Fiscal Year (FY) 2019 Budget Estimates

February 2018

Defense Advanced Research Projects Agency

Title: Machine Common Sense (MCS)

Description: The Machine Common Sense (MCS) program will explore approaches to commonsense reasoni
Recent advances in machine learning have resulted in exciting new artificial intelligence (AI) capabilities in are
recognition, natural language processing, and two-person strategy games (Chess, Go). But in all of these appl
the machine reasoning is narrow and highly specialized; broad, commonsense reasoning by machines remains
program will create more human-like knowledge representations, for example, perceptually-grounded represen
commonsense reasoning by machines about the physical world and spatio-temporal phenomena. Equipping A
more human-like reasoning capabilities will make it possible for humans to teach/correct a machine as they int
on tasks, enabling more equal collaboration and ultimately symbiotic partnerships between humans and machi

FY 2019 Plans:

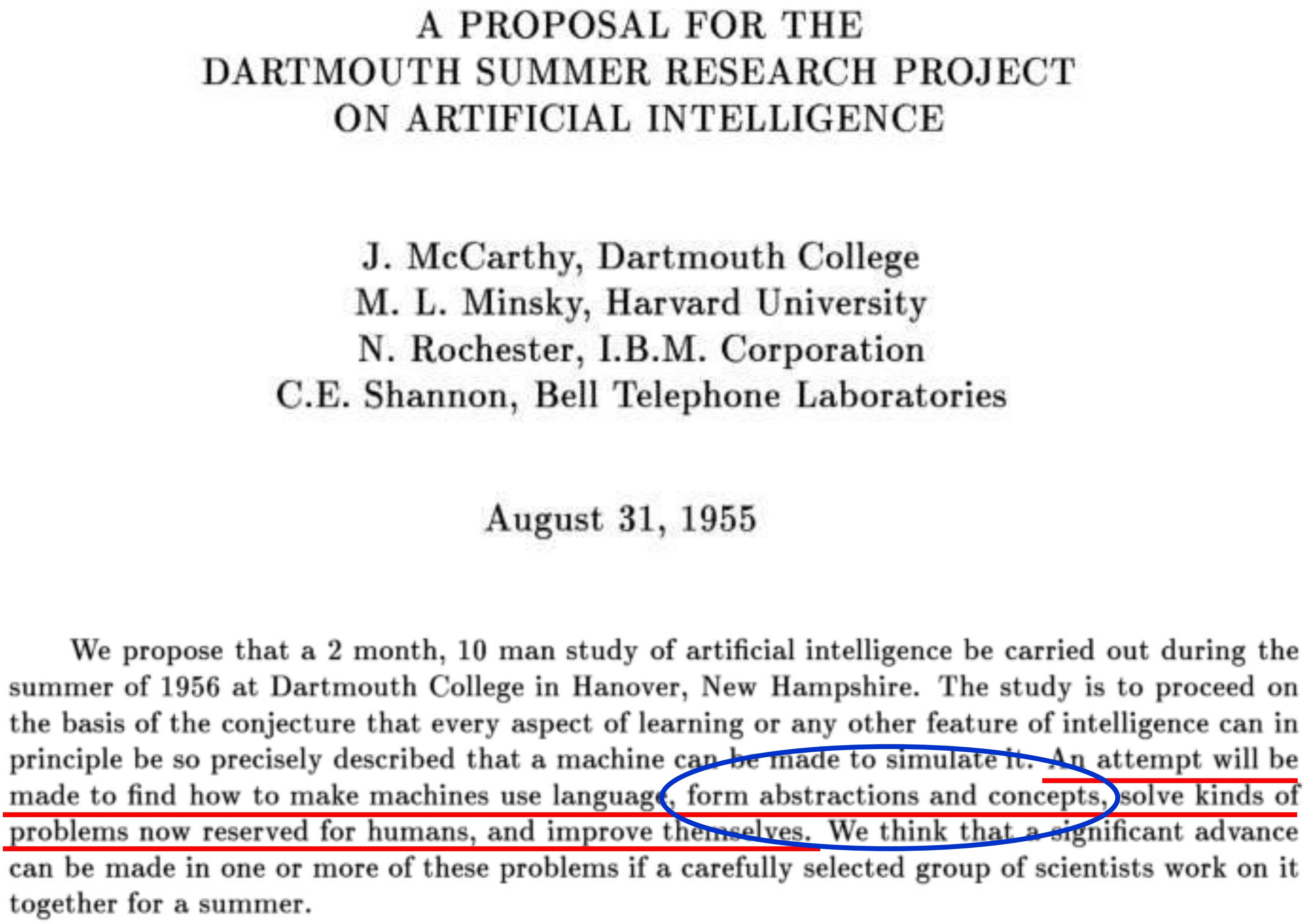
- Develop approaches for machine reasoning about imprecise and uncertain information derived from text, pic
speech, and sensor data.

- Design methods to enable machines to identify knowledge gaps and reason about their state of knowledge.

- Formulate perceptually-grounded representations to enable commonsense reasoning by machines about the
spatio-temporal phenomena.

Some core components of human understanding

- Intuitive physics, biology, psychology
- Mental models of cause and effect
- Vast world-knowledge
- **Abstraction and analogy**



A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

J. McCarthy, Dartmouth College
M. L. Minsky, Harvard University
N. Rochester, I.B.M. Corporation
C.E. Shannon, Bell Telephone Laboratories

August 31, 1955

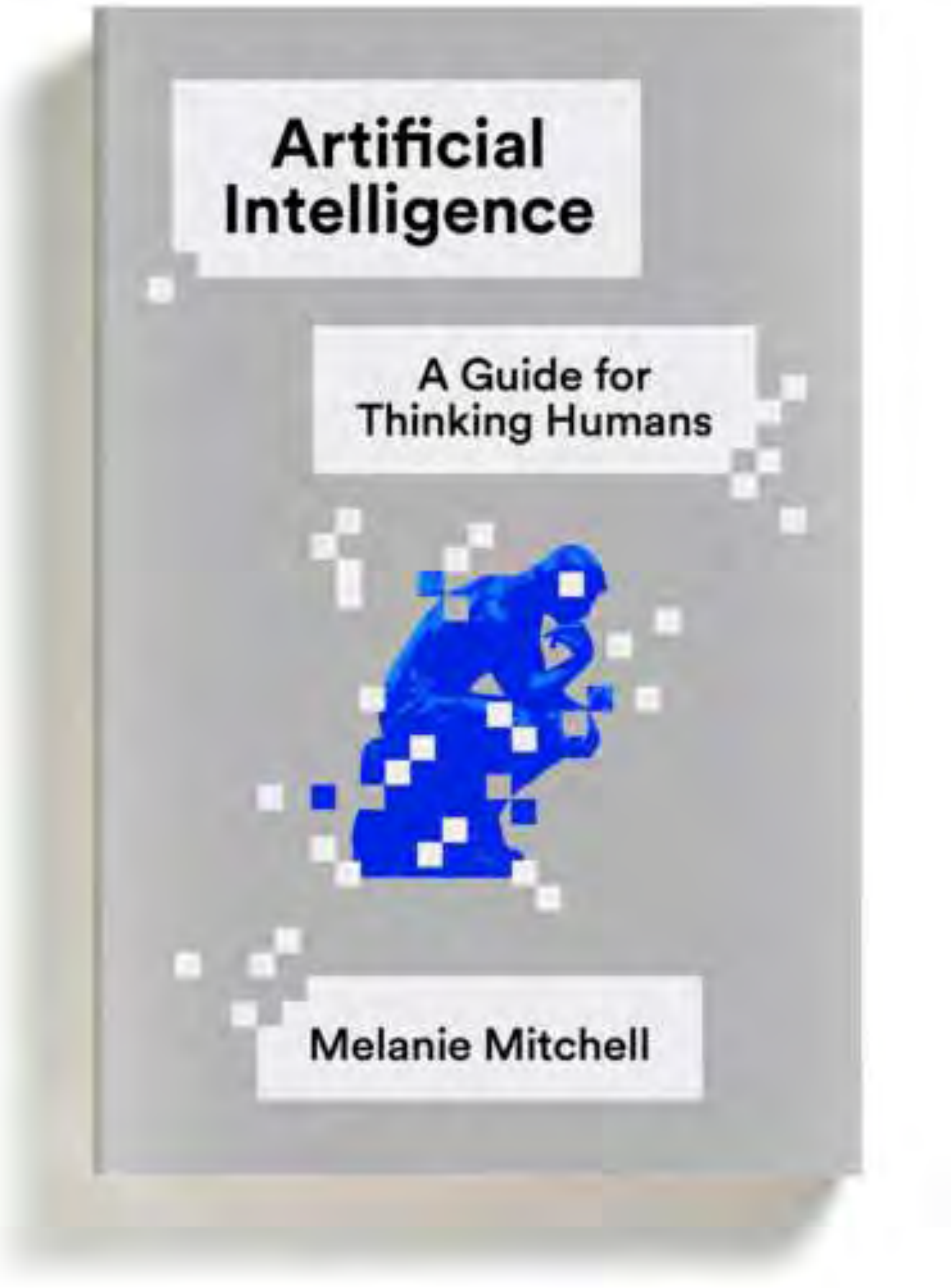
We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

“Without concepts there can be no thought, and without analogies there can be no concepts.”

— D. Hofstadter & E. Sander, *Surfaces and Essences* (2013)

“How to form and concepts and make analogies are the most important open problems in AI.”

— Melanie Mitchell, today



Thank you for listening!