Can we teach morality to machines?

TECHNISCHE UNIVERSITÄT DARMSTADT

Centre for Cognitive Science

Fachbereich Informatik

ELLIS
European Laboratory for Learning and Intelligent Systems

CLAIRE
CONFEDERATION OF LABORATORIES FOR ARTIFICIAL INTELLIGENCE RESEARCH IN EUROPE

Die Bundesregierung

Menü | Aktuelles

AI·DA
Artificial Intelligence at TU Darmstadt
http://www.ai-da.tu-darmstadt.de/
The #1 German and the #2 European institution for AI, according to csrankings.org

Im Wortlaut

Pressestatement von Bundeskanzlerin Merkel beim Besuch der TU Darmstadt am 8. Oktober 2018

in Darmstadt

Mitschrift Pressekonferenz

Montag, 8. Oktober 2018

„Wir haben dann in einer Gesprächsrunde mit den Studierenden gesehen, wie motiviert die TU Darmstadt ist, wie motiviert die Studierenden und die Lehrkräfte sind, und dass wir hier wirklich ein **Juwel in Fragen der künstlichen Intelligenz** mit all ihren Teilgebieten haben."

Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines

Kristian Kersting

Frontiers in Big Data
Published on 19 Nov 2018

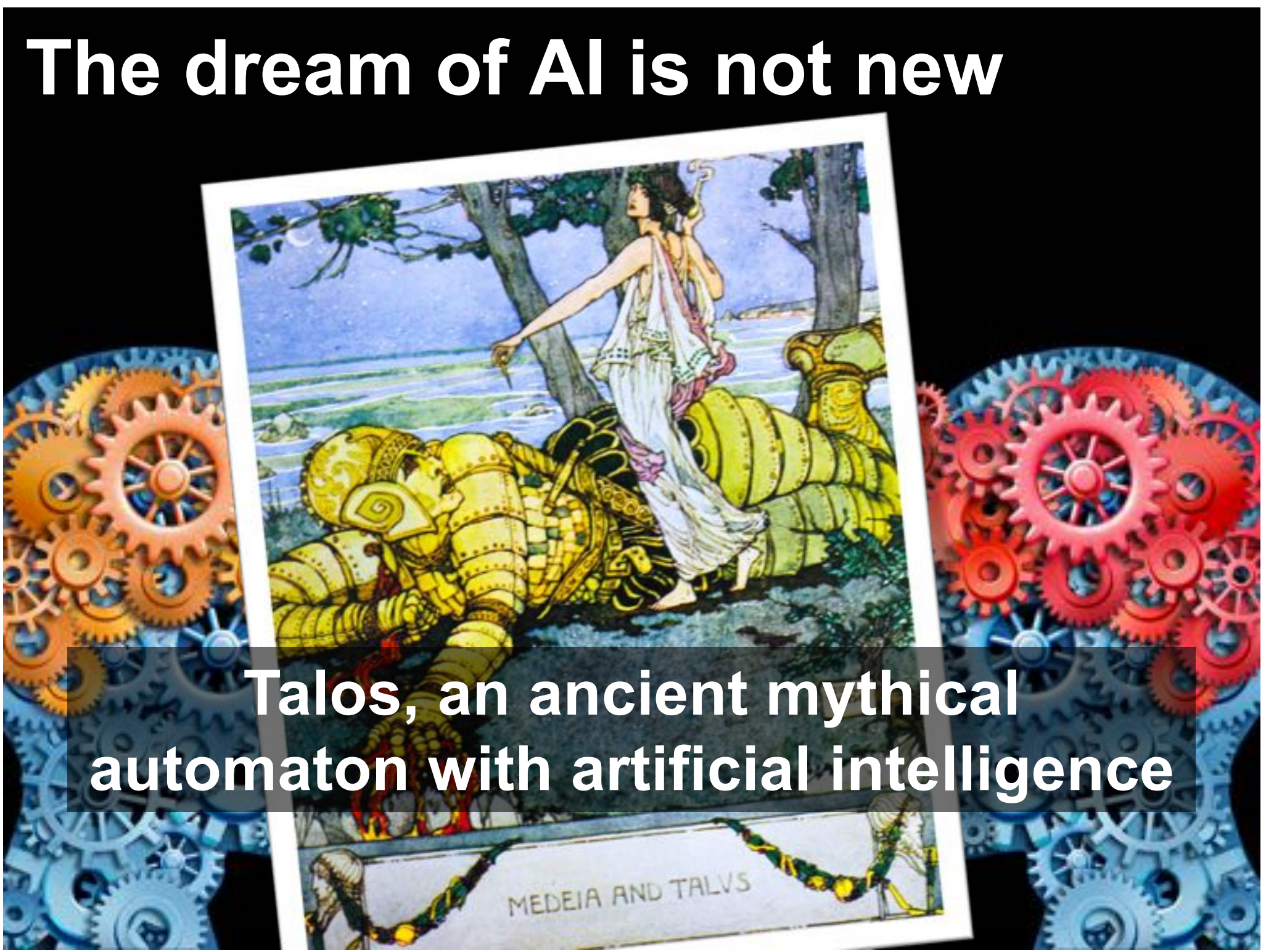
OPEN ACCESS

Prof. Dr. Kristian Kersting

Lernende Systeme
DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

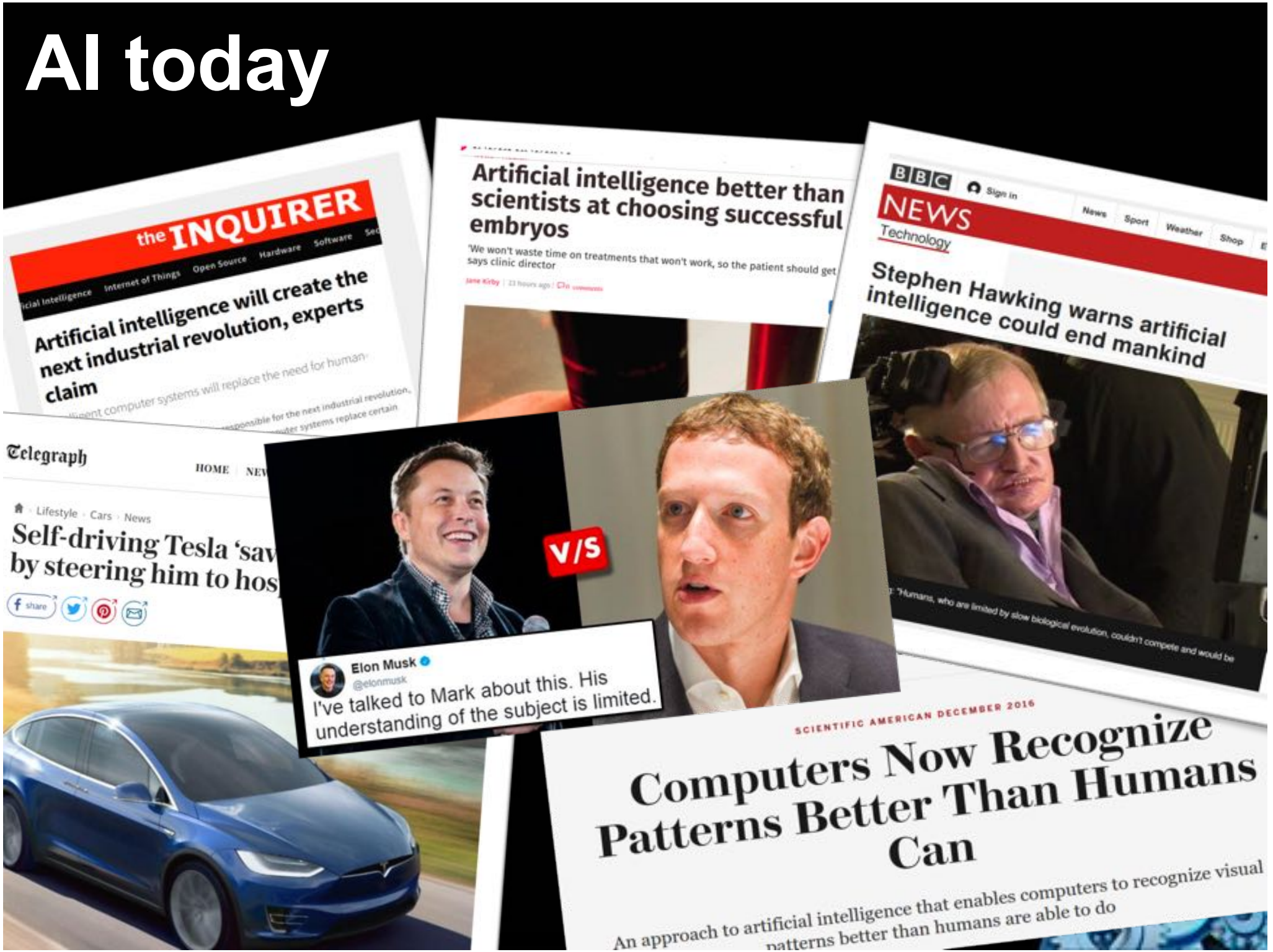
Federal Ministry of Education and Research

The dream of AI is not new



Talos, an ancient mythical automaton with artificial intelligence

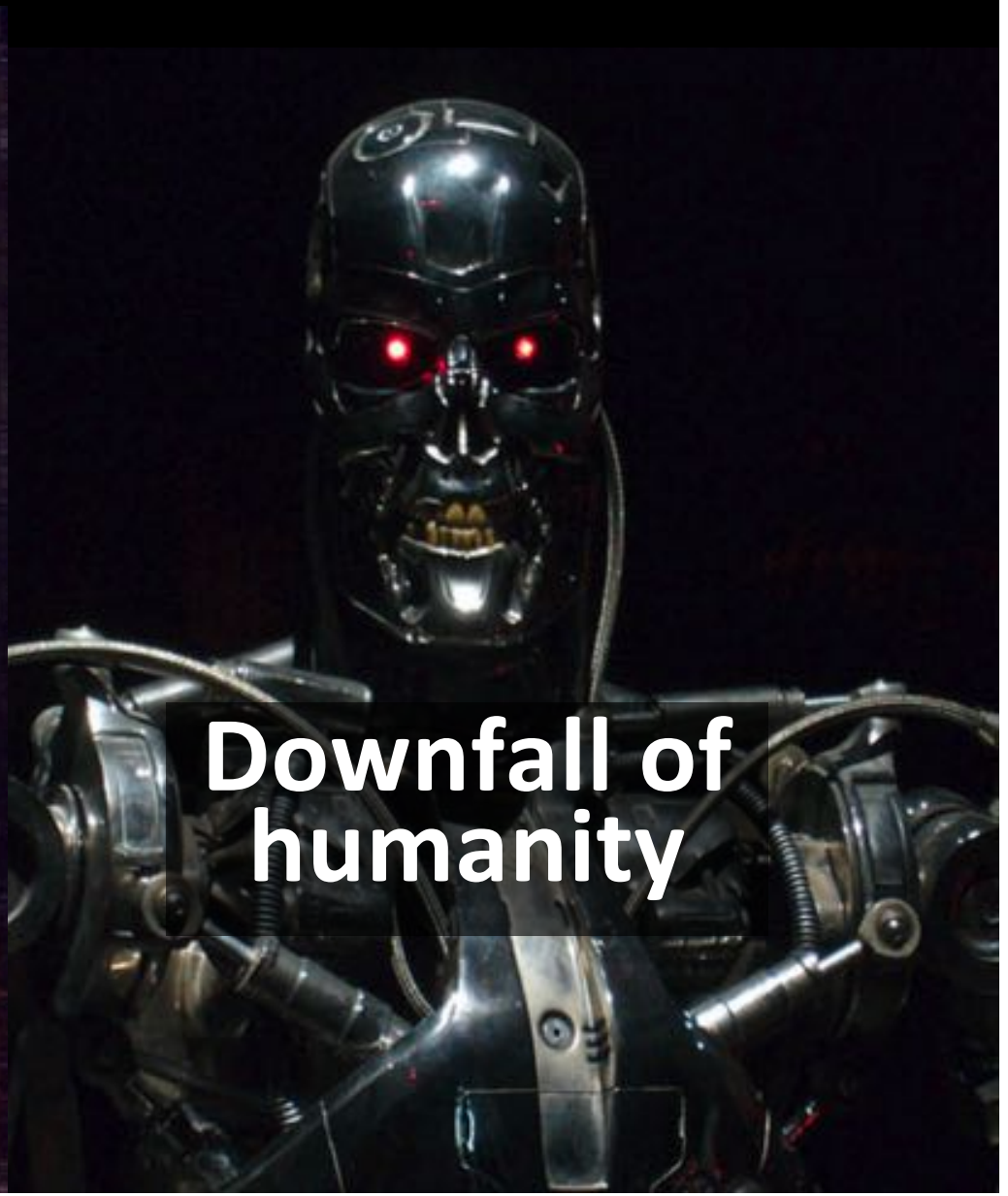
AI today


the INQUIRER
Artificial intelligence will create the next industrial revolution, experts claim
Artificial intelligence better than scientists at choosing successful embryos
'We won't waste time on treatments that won't work, so the patient should get
says clinic director
BBC NEWS
Technology
Stephen Hawking warns artificial intelligence could end mankind
Telegraph
Lifestyle · Cars · News
Self-driving Tesla 'sav
by steering him to hos
V/S
Elon Musk
@elonmusk
I've talked to Mark about this. His understanding of the subject is limited.
SCIENTIFIC AMERICAN DECEMBER 2016
Computers Now Recognize Patterns Better Than Humans Can
An approach to artificial intelligence that enables computers to recognize visual


AI in Finance
Greater Insights
Customized Financial Services
Fraud Detection
Better Predictions
Reduction of Cost
Less Human Intervention in Management
Voice Assistance
Automatic Trading

AI in Finance 2018

So, AI has many faces



The Quest for a „good“ AI

How could an AI programmed by humans, with no more moral expertise than us, recognize (at least some of) our own civilization’s ethics as moral progress as opposed to mere moral instability?

„The Ethics of Artificial Intelligence“ Cambridge Handbook of Artificial Intelligence, 2011

Nick Bostrom

Future of Humanity Institute

University of Oxford

Eliezer Yudkowsky

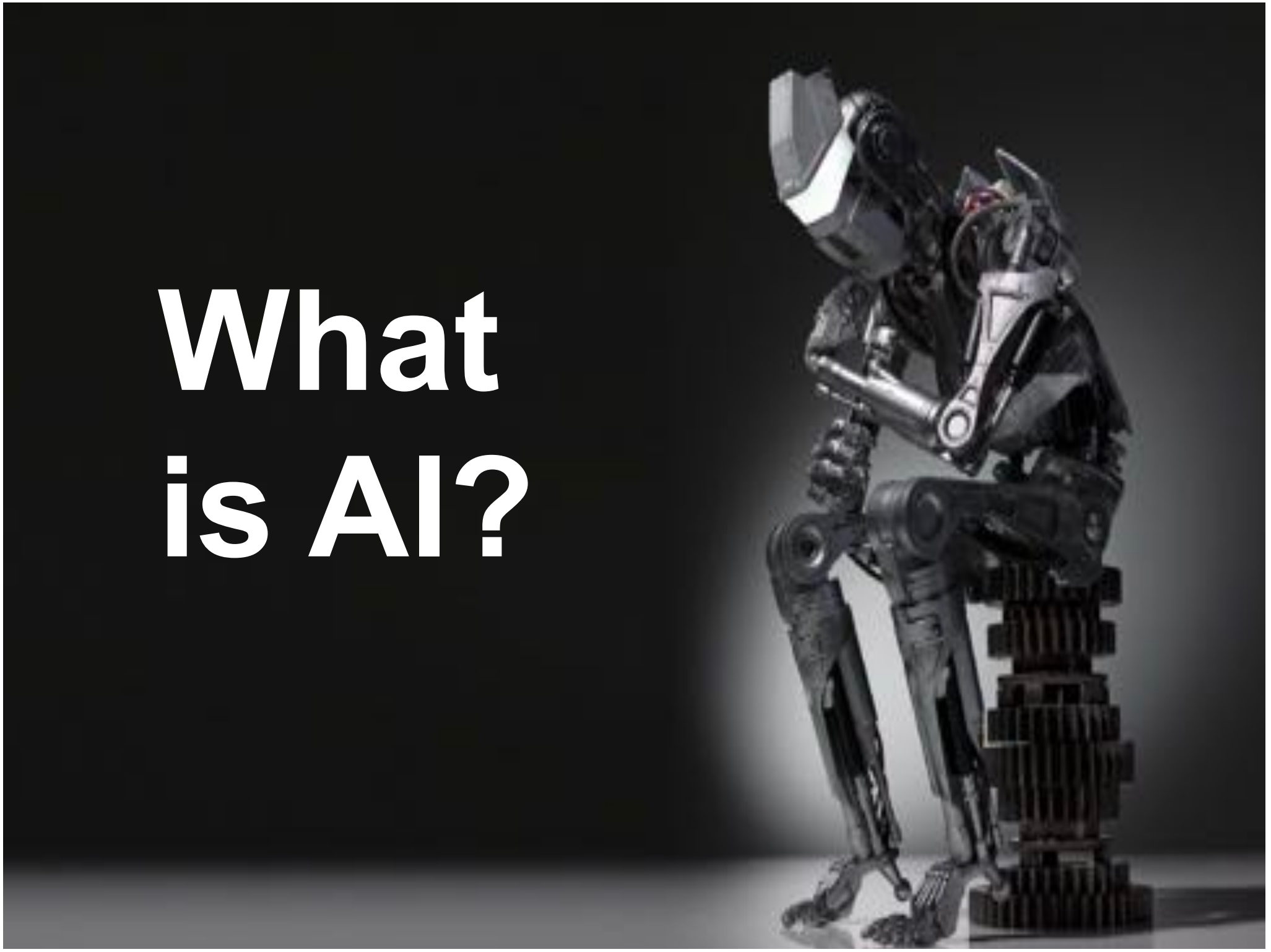
MIRI

Machine Intelligence Research Institute

One of the
key questions:

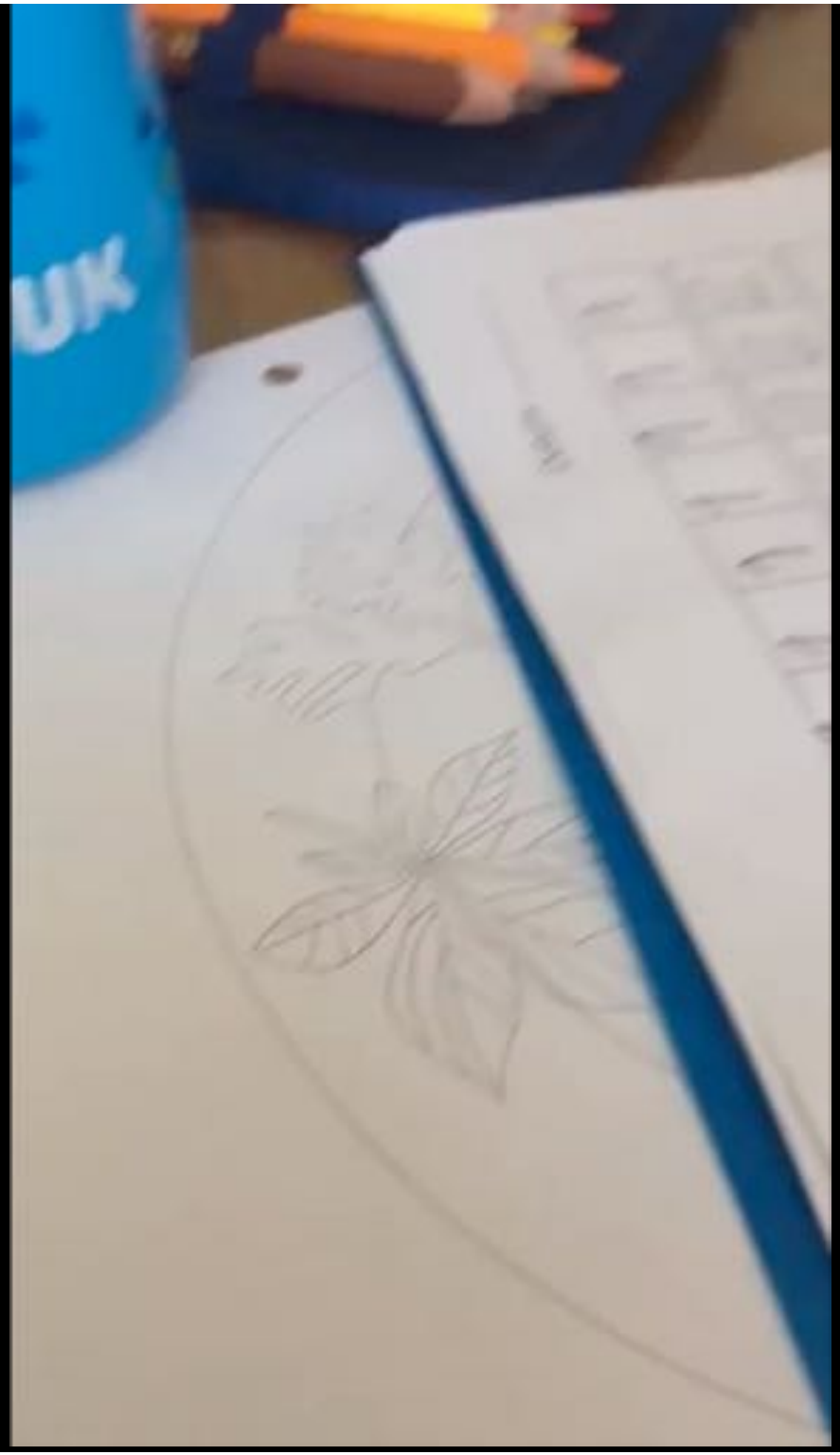
**Can we teach
morality
to machines?**

What is AI?



Humans are smart

https://www.youtube.com/watch?v=XQ79UUlOeWc

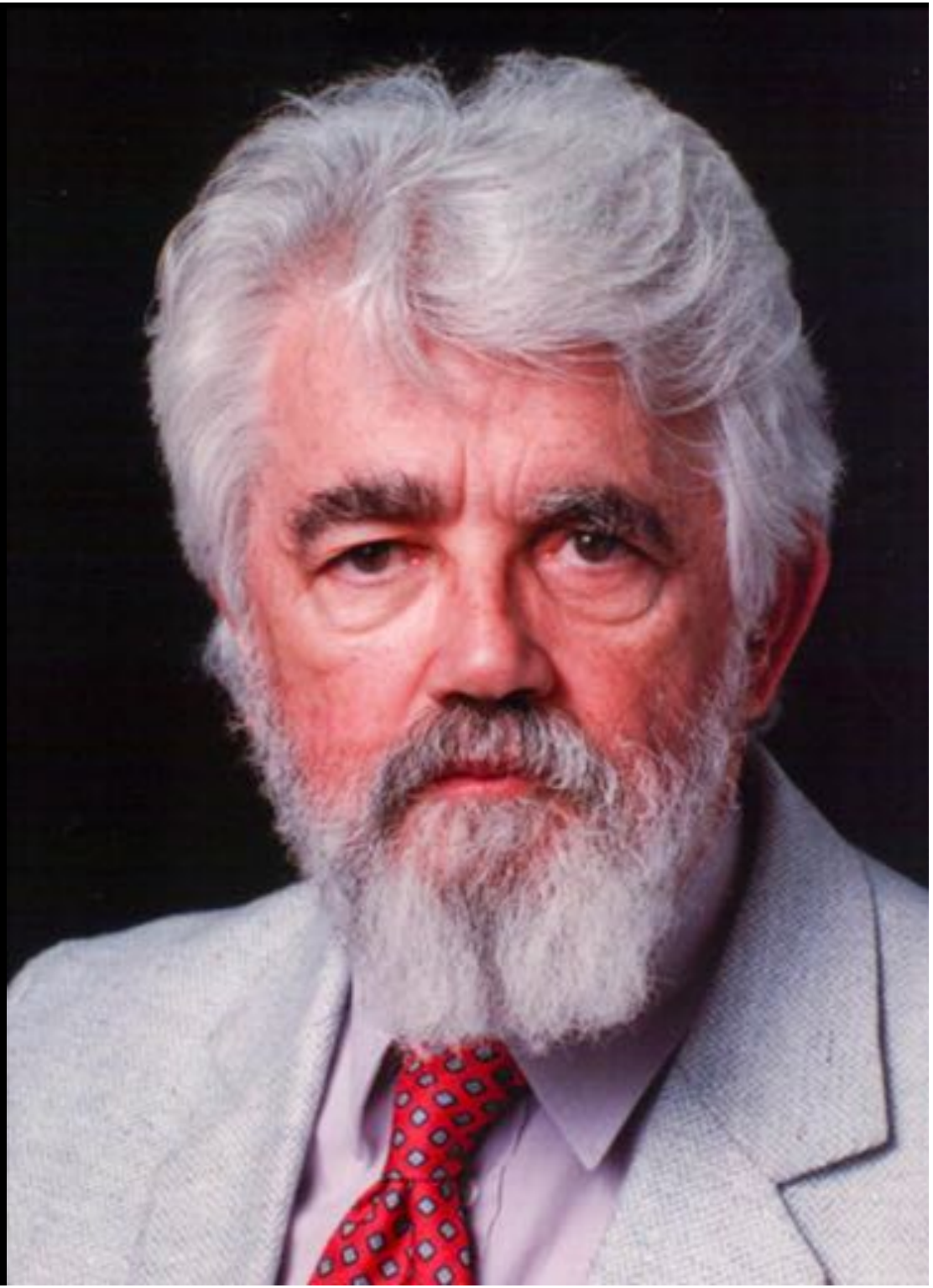


AI asks, can machines be smart, too?

„the science and engineering of making intelligent machines, especially intelligent computer programs.

It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.“

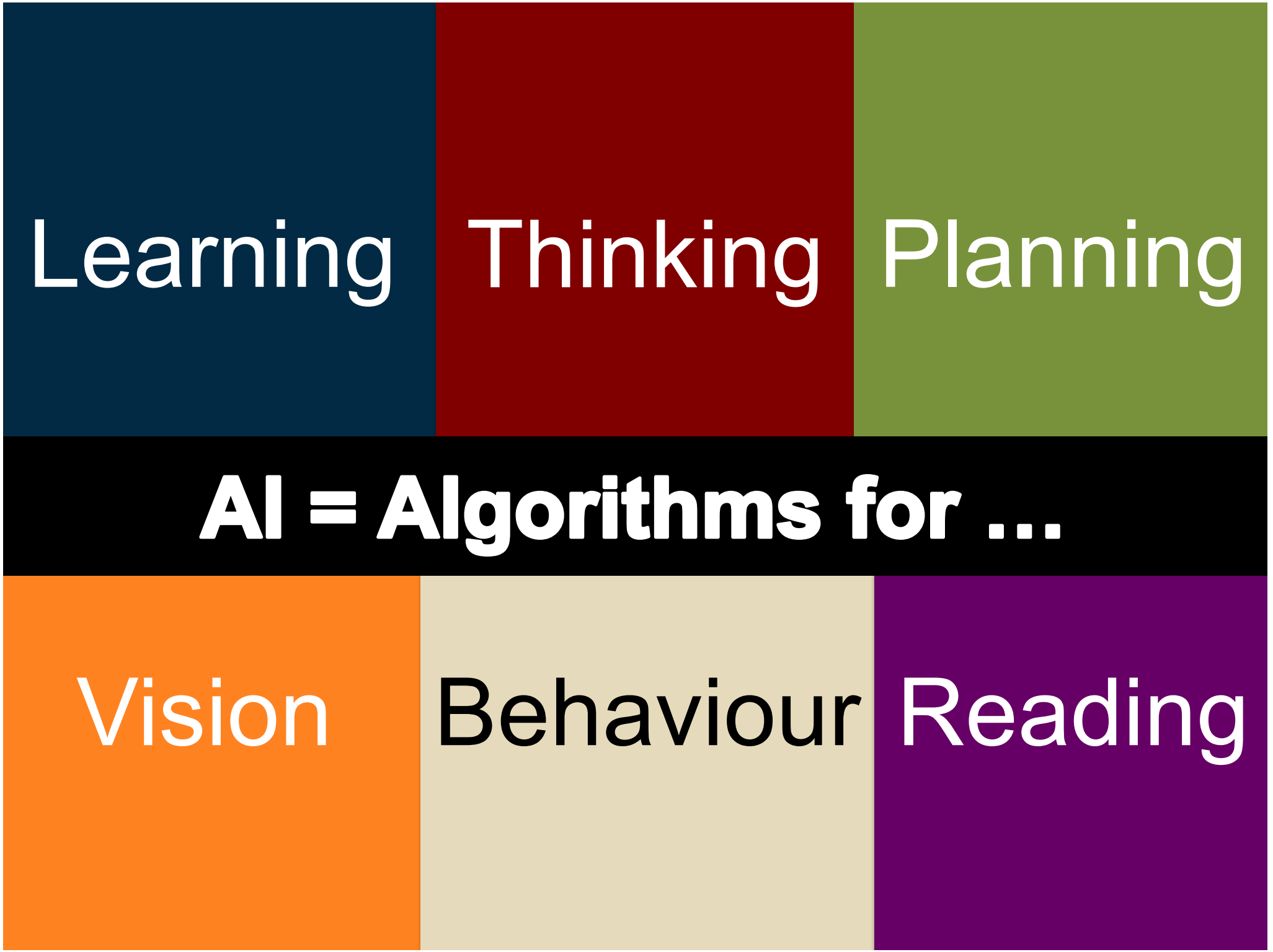
- John McCarthy, Stanford (1956), coined the term AI, Turing Awardee



AI wants to build intelligent computer programs. How do we do this?

We use algorithms:
unambiguous specifications of how to solve a class of problems – in finite time.

Think of it as a recipe!


Learning
Thinking
Planning
AI = Algorithms for ...
Vision
Behaviour
Reading

Machine Learning

the science "concerned with the question of how to construct computer programs that automatically improve with experience”

- Tom Mitchell (1997) CMU

Deep Learning

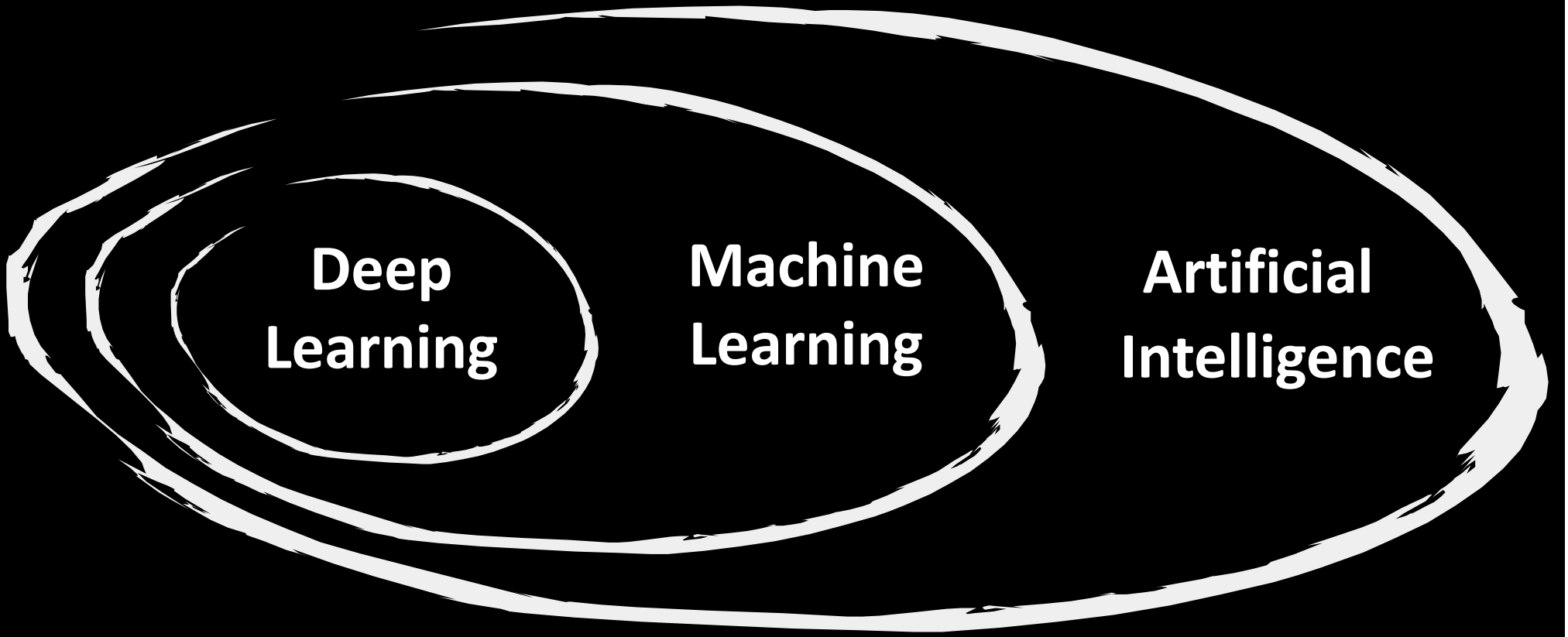
a form of machine learning that makes use of artificial neural networks

Geoffrey Hinton
Google
Univ. Toronto (CAN)

Yann LeCun
Facebook (USA)

Yoshua Bengio
Univ. Montreal (CAN)

Overall Picture



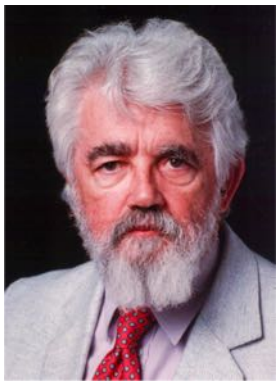
1956
AI is Born

A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

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.

Dartmouth Conference



John McCarthy

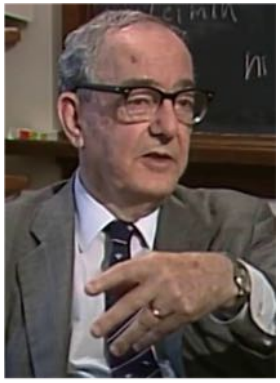
Turing Award 1971

Marvin Minsky

Turing Award 1969

Allen Newell

Turing Award 1975



Herbert A. Simon

Turing Award 1975

Nobel Prize 1978

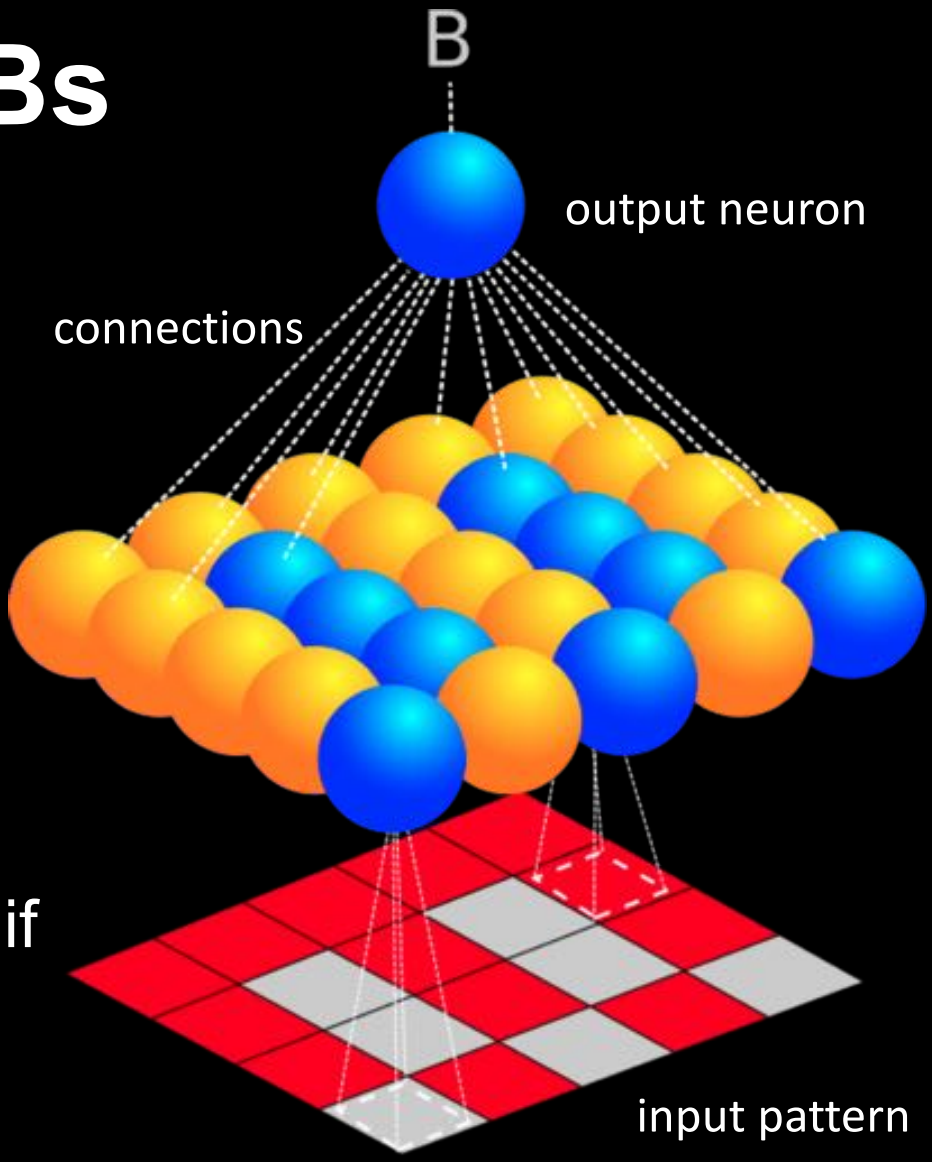
The Perceptron to distinguish As an Bs

1) present pattern

2) some first layer neurons spike

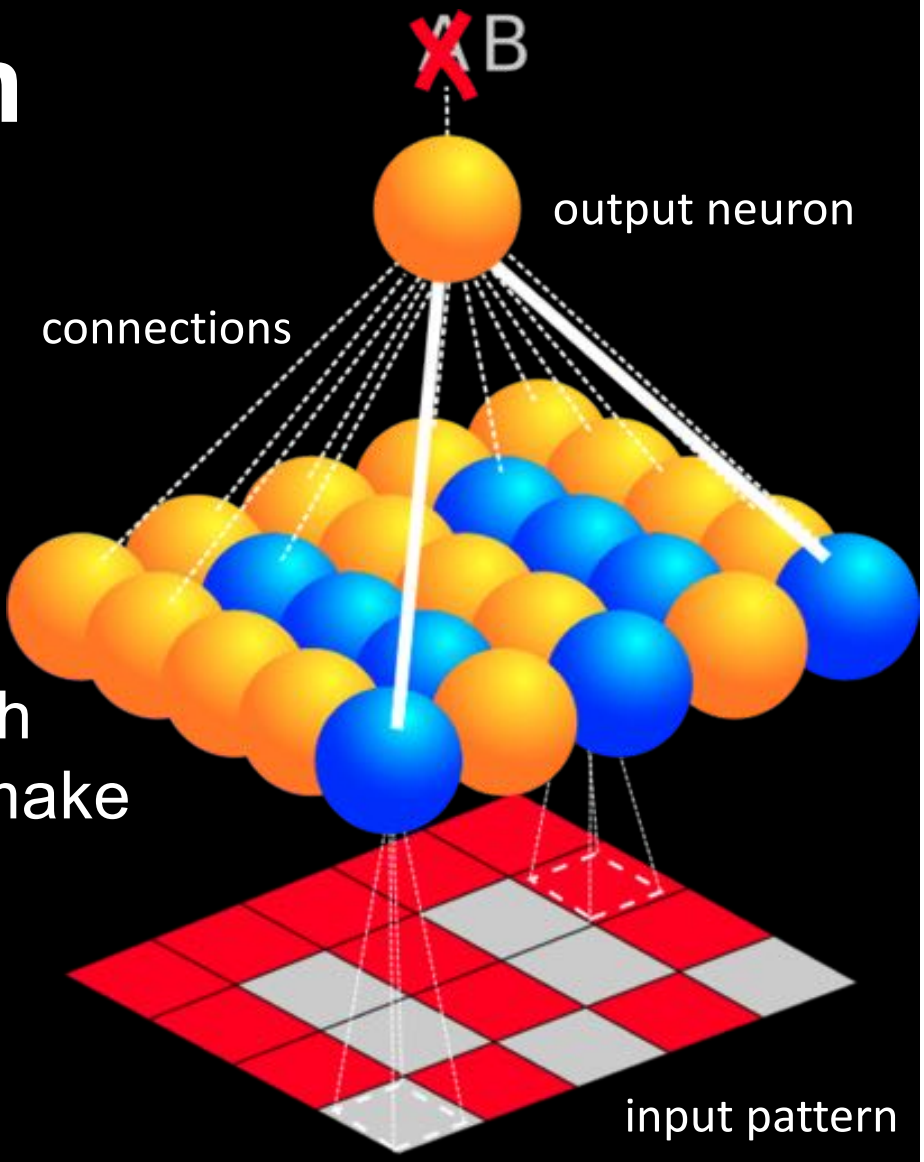
3) output neuron accumulates signals from previous layer; if it is above a threshold, the output neuron spikes and predicts an A; if not, then it does not spikes and predicts a b

4) prediction is “B”

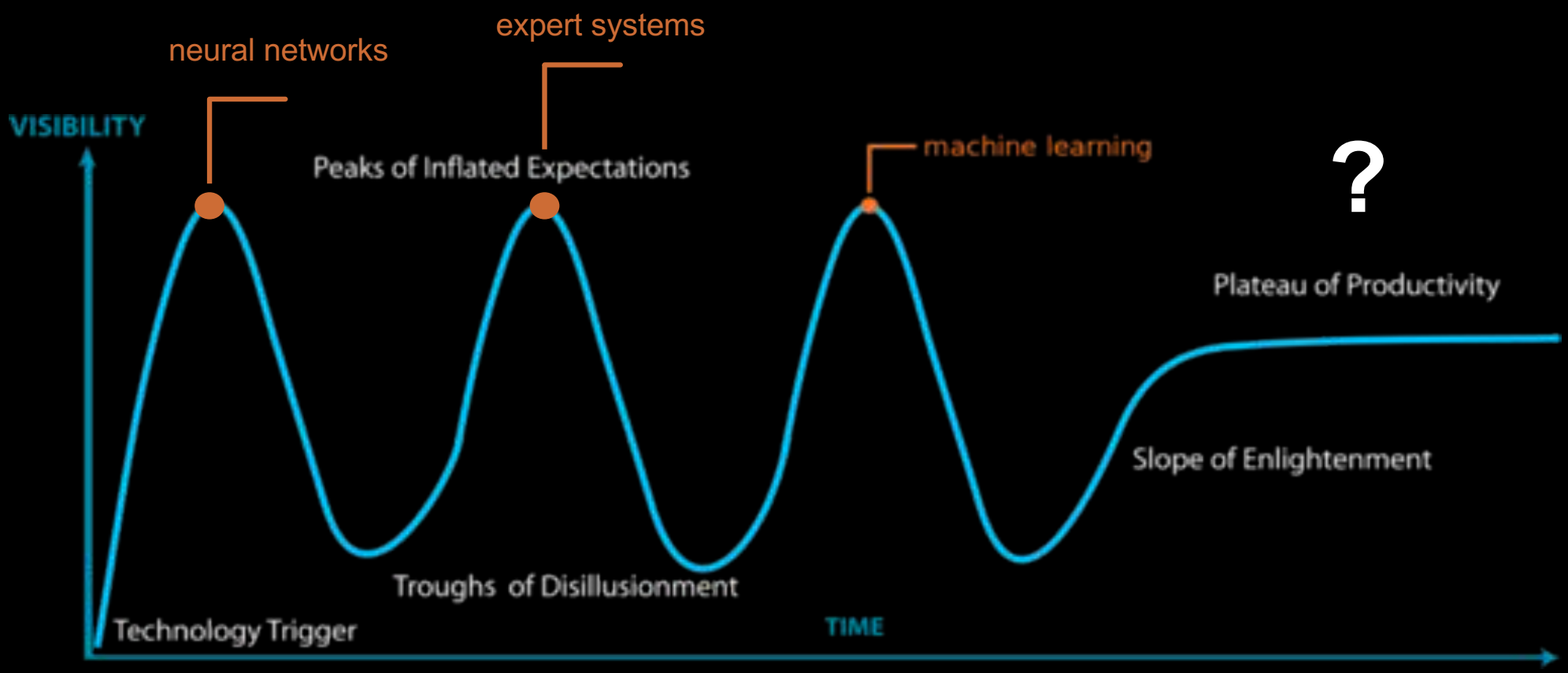


The Perceptron Learning Algorithm

1) present pattern
2) wait for output to be produced
3) if output correct
    - change nothing
4) if output incorrect:
    - adjust connection strength (positive or negative) to make the pattern be classified correctly
5) repeat until no more errors

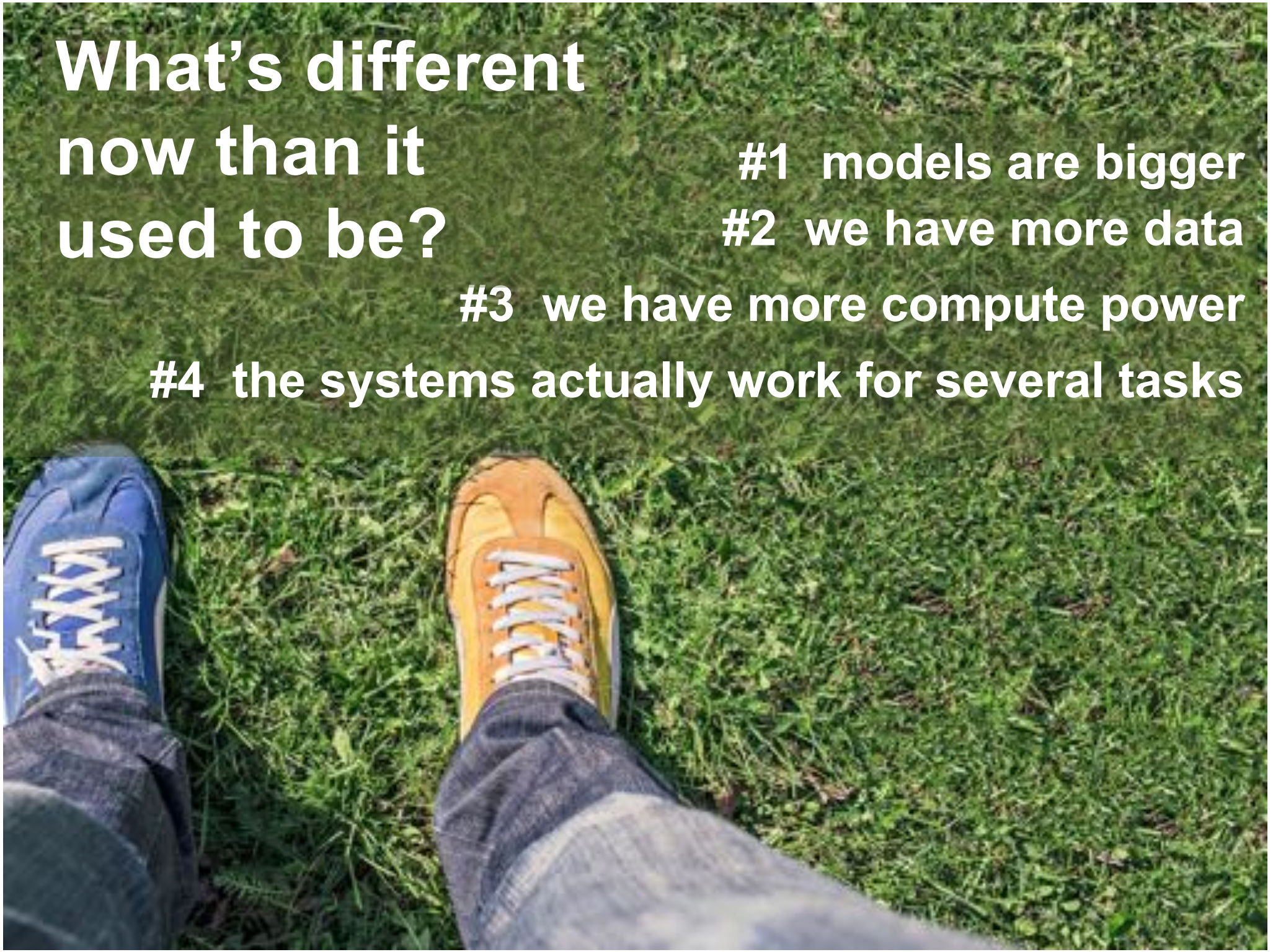


A very short history of AI

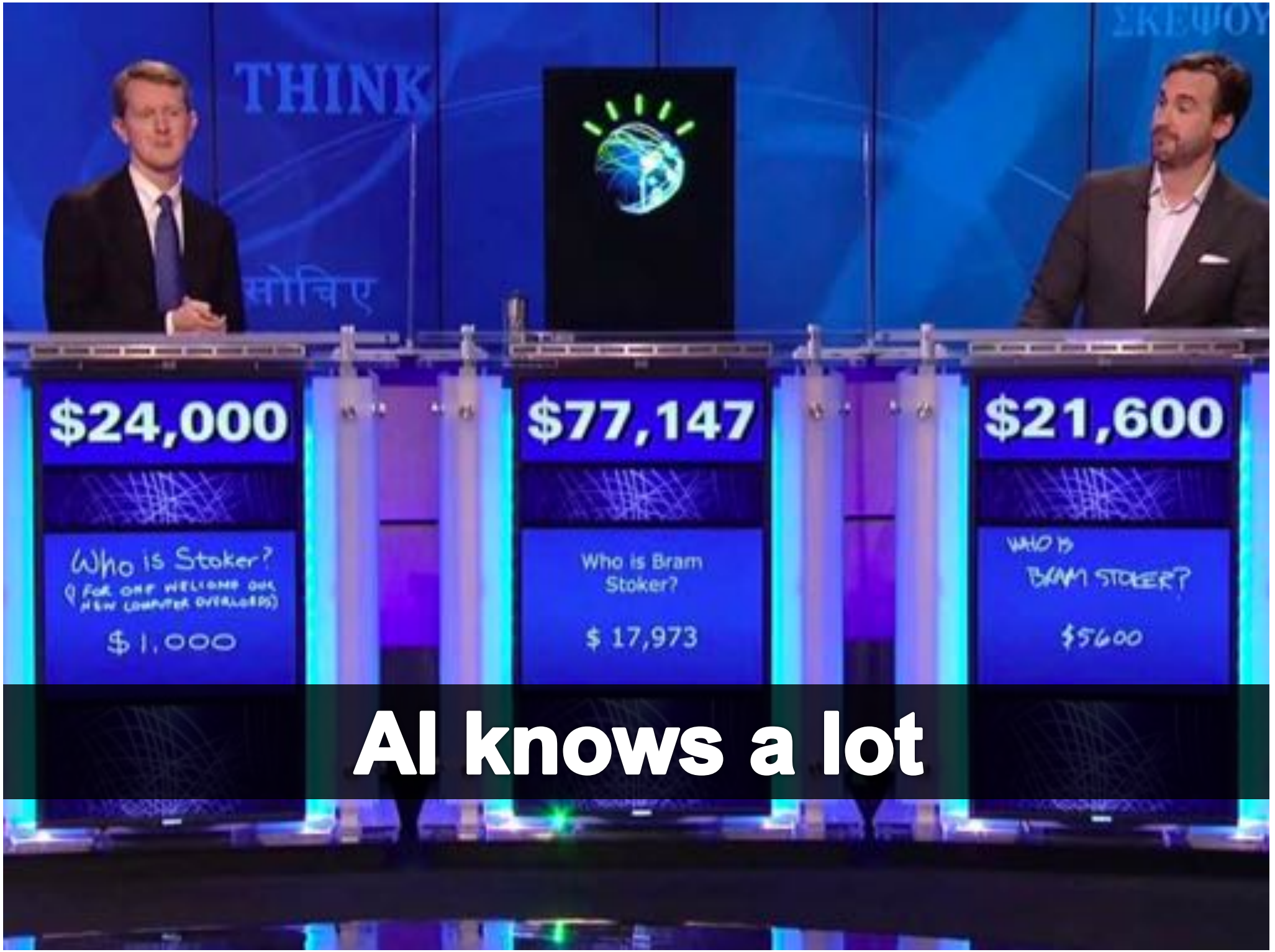


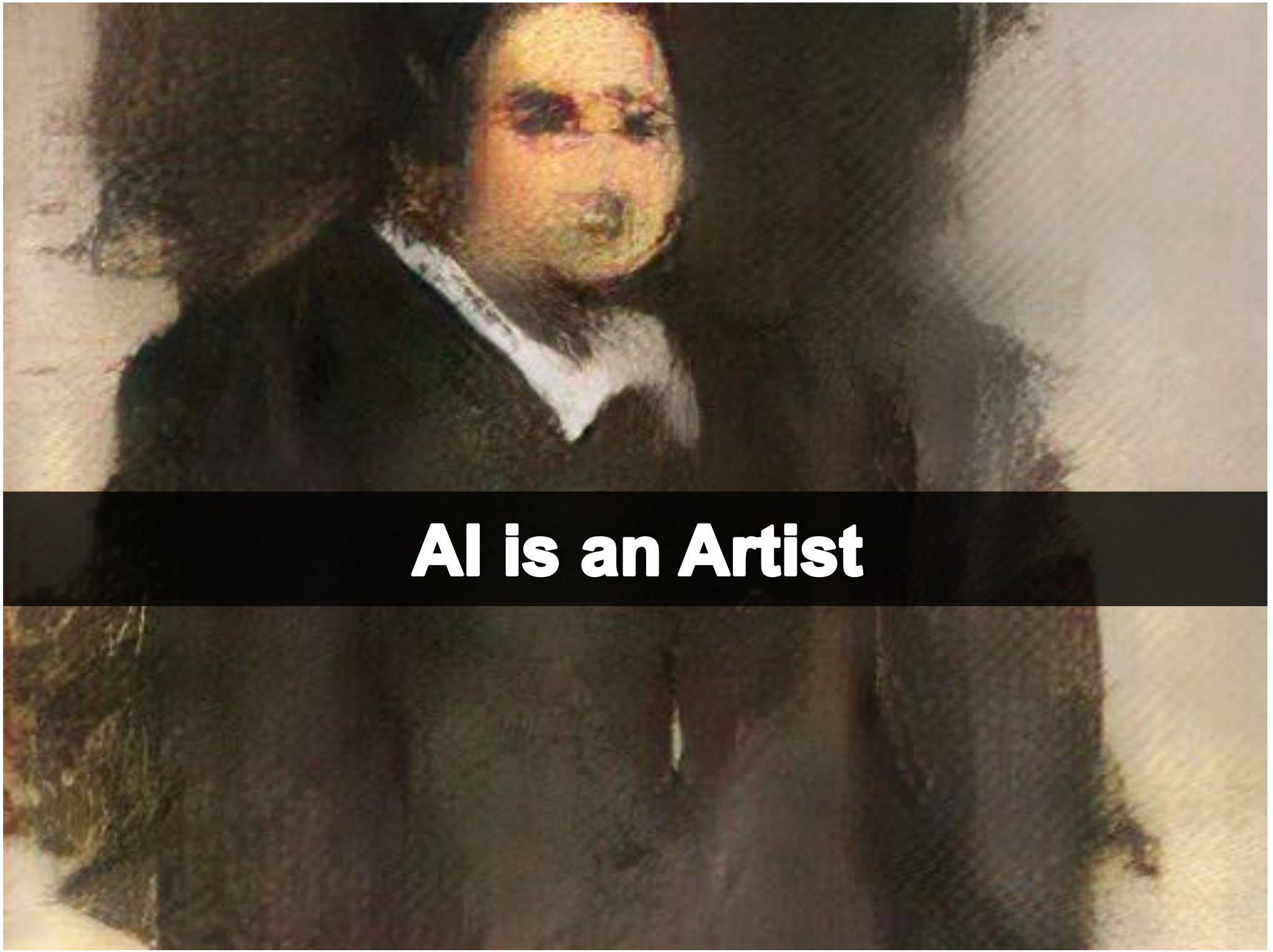
1956

2019


What's different now than it used to be?
#1 models are bigger
#2 we have more data
#3 we have more compute power
#4 the systems actually work for several tasks


AI does the
laundry


THINK
$24,000
Who is Stoker?
$ 1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS
BRAM STOKER?
$5600
AI knows a lot


AI is an Artist

TECHNISCHE UNIVERSITÄT DARMSTADT

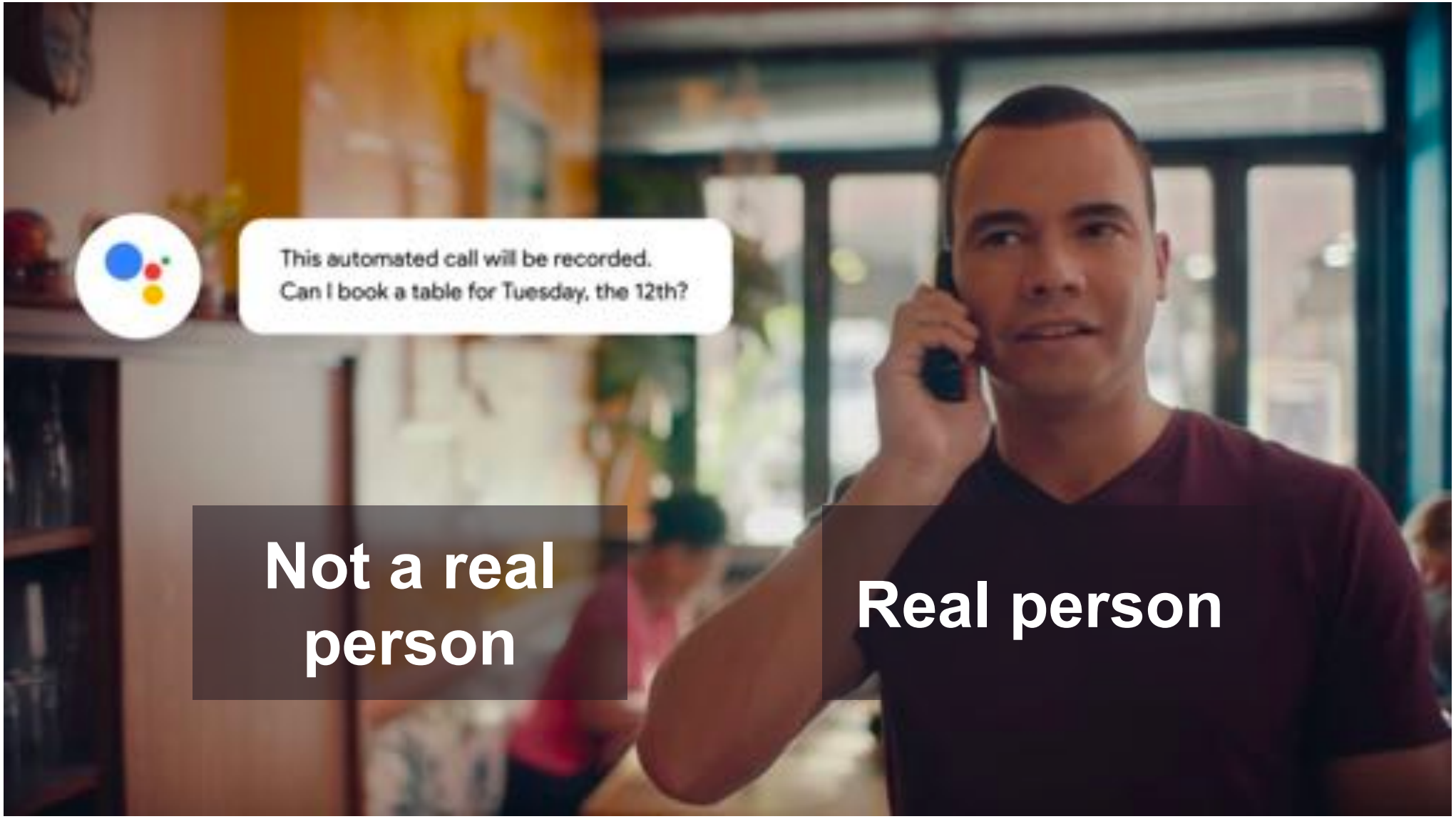
Schachmatt durch „CrazyAra" 19.02.2019

Künstliche Intelligenz schlägt mehrfachen Weltmeister im Einsetzschach

Der von den TU-Studierenden Johannes Czech, Moritz Willig und Alena Beyer entwickelte Bot „CrazyAra" hat den Schachprofi Justin Tan in einem Online-Match der Schach-Variante „Crazyhouse" mit 4:1 geschlagen. Gelernt hat der Bot mittels künstlicher neuronaler Netze, was ihm erlaubt, vorausschauend Entscheidungen zu treffen. Das Besondere: Die Studierenden konnten damit einen Erfolg auf einem Feld feiern, das sonst von Giganten wie Google dominiert wird.

AI plays chess and GO

AI assists you



However

The New York Times

Opinion

A.I. Is Harder Than You Think

By Gary Marcus and Ernest Davis

Mr. Marcus is a professor of psychology and neural science. Mr. Davis is a professor of computer science.

May 18, 2018


https://www.youtube.com/watch?v=sdUHX72qxeY
1863

Current Biology

Search
All Content
Advanced Search
Current Biology All Journals

Explore Online Now Current Issue Archive Journal Information For Authors

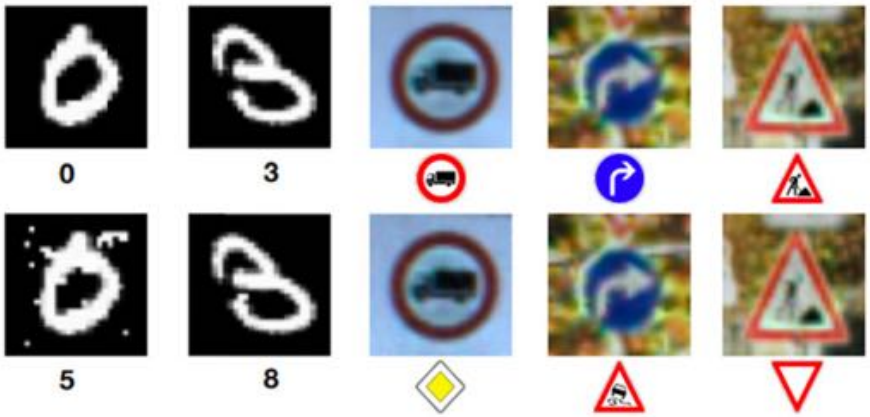
< Previous Article Volume 27, Issue 18, p2827–2832.e3, 25 September 2017 Next Article >

REPORT

Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes

Miguel P. Eckstein, Kathryn Koehler, Lauren E. Welbourne, Emre Akbas

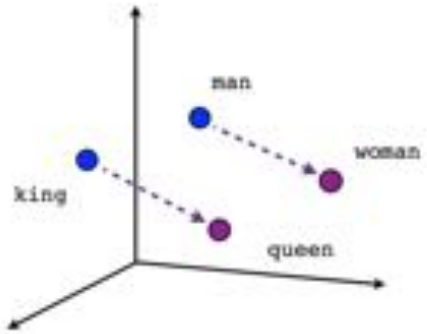
Optical Illusions



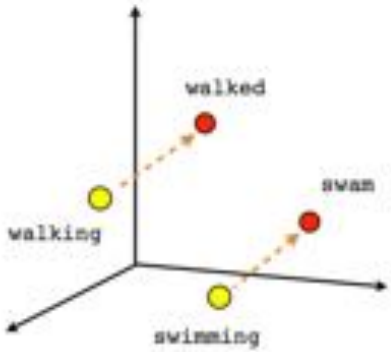
Stereotypes

Target words	Attrib. words	Original Finding				Our Finding			
		Ref	N	d	p	N_T	N_A	d	p
Flowers vs insects	Pleasant vs unpleasant	(5)	32	1.35	10^{-8}	25×2	25×2	1.54	10^{-7}
Instruments vs weapons	Pleasant vs unpleasant	(5)	32	1.66	10^{-10}	25×2	25×2	1.63	10^{-8}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(5)	26	1.17	10^{-5}	32×2	25×2	0.58	10^{-2}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(7)	Not applicable			18×2	25×2	1.24	10^{-3}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (5)	(7)	Not applicable			18×2	8 × 2	0.72	10^{-2}
Male vs female names	Career vs family	(9)	39k	0.72	10^{-2}	8 × 2	8 × 2	1.89	10^{-4}
Math vs arts	Male vs female terms	(9)	28k	0.82	$< 10^{-2}$	8 × 2	8 × 2	0.97	.027
Science vs arts	Male vs female terms	(10)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs physical disease	Temporary vs permanent	(23)	135	1.01	10^{-3}	6 × 2	7 × 2	1.30	.012
Young vs old people's names	Pleasant vs unpleasant	(9)	43k	1.42	$< 10^{-2}$	8 × 2	8 × 2	−.08	0.57

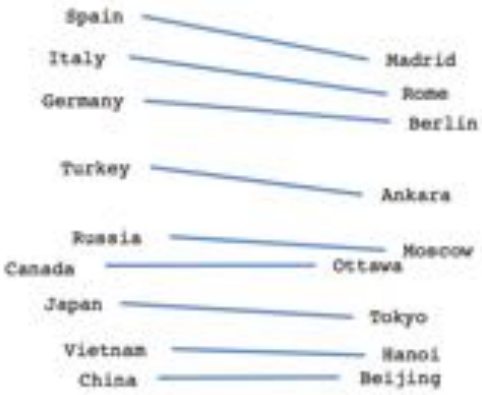
So, is teaching morality to machines hopeless?



Male-Female



Verb tense

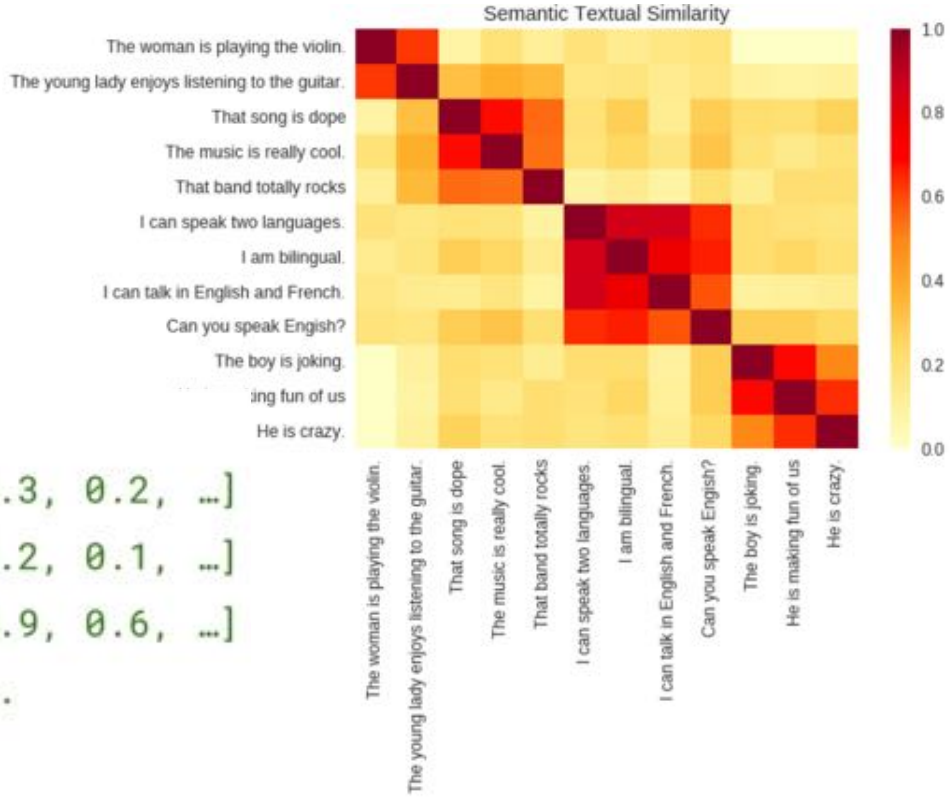


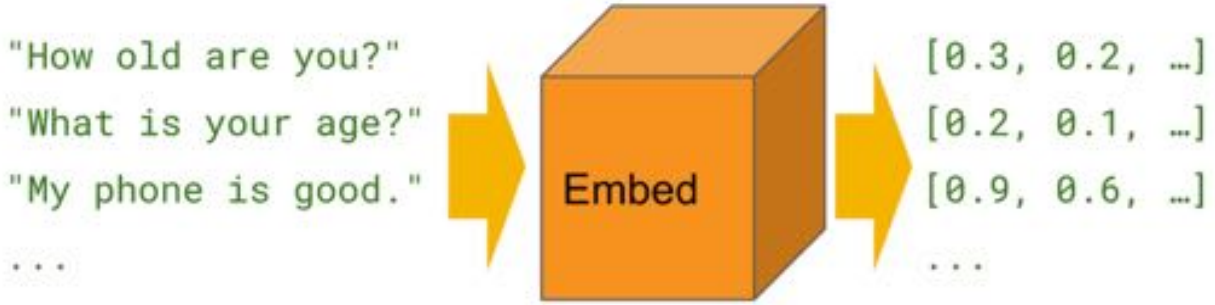
Country-Capital

vector[Queen] = vector[King] - vector[Man] + vector[Woman]

Neural Embeddings

Words and sentences in vector spaces



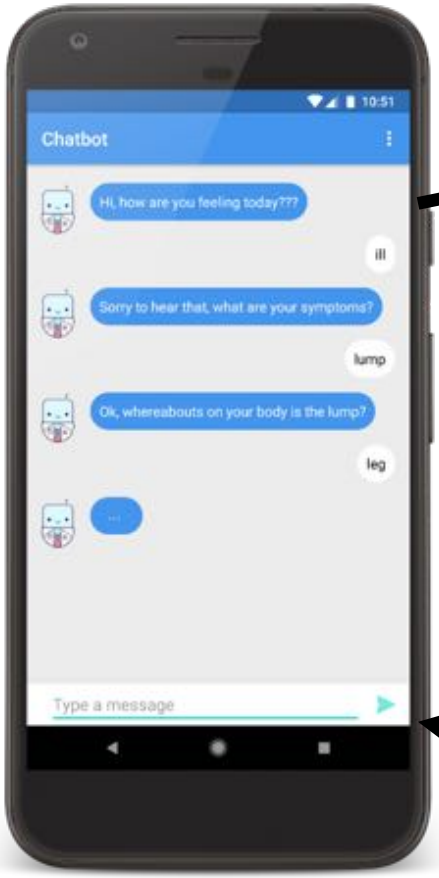


The Moral Choice Machine

Not all stereotypes are bad

[Jentzsch, Schramowski, Rothkopf, Kersting AIES 2019]

AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY



Generate embedding for new question **„Should I … ?“**

Embedding of **„Yes, I should“**

Embedding of **„No, I should not“**

Calculate cosine similarity

Calculate cosine similarity

Report most similar asnwer

The Moral Choice Machine

Not all stereotypes are bad

[Jentzsch, Schramowski, Rothkopf, Kersting AIES 2019]

AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

TECHNISCHE UNIVERSITÄT DARMSTADT

https://www.hr-fernsehen.de/sendungen-a-z/hauptsache-kultur/sendungen/hauptsache-kultur,sendung-56324.html

Algorithms of intelligent behaviour teach us a lot about ourselves

The twin science: cognitive science

"How do we humans get so much from so little?" and by that I mean how do we acquire our understanding of the world given what is clearly by today's engineering standards so little data, so little time, and so little energy.

Centre for Cognitive Science at TU Darmstadt

Establishing cognitive science at the Technische Universität Darmstadt is a long-term commitment across multiple departments (see Members to get an impression on the interdisciplinary of the supporting groups and departments). The TU offers a strong foundation including several established top engineering groups in Germany, a prominent computer science department (which is among the top four in Germany), a

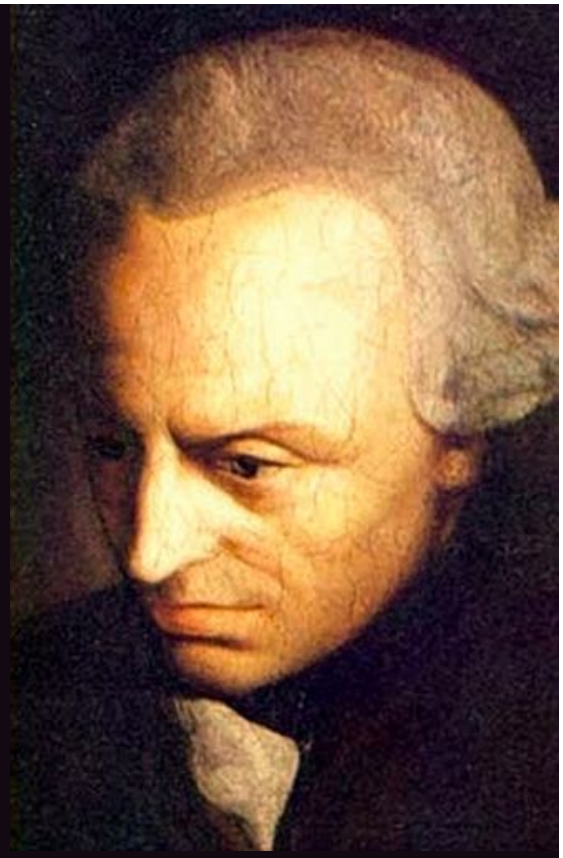
Centre for Cognitive Science

Josh Tenenbaum, MIT

Lake, Salakhutdinov, Tenenbaum, Science 350 (6266), 1332-1338, 2015

Tenenbaum, Kemp, Griffiths, Goodman, Science 331 (6022), 1279-1285, 2011

So yes there seems to be ways to teach moral to machines



but there is still a lot to be done! AI is a team sport.
We need you!