Can we teach morality to machines?
The dream of AI is not new

Talos, an ancient mythical automaton with artificial intelligence
AI today

Artificial intelligence will create the next industrial revolution, experts claim

We won't waste time on treatments that won't work, so the patient should get what's best, says clinic director

Stephen Hawking warns artificial intelligence could end mankind

Self-driving Tesla ‘saved’ man by steering him to hospital

Elon Musk says his understanding of the subject is limited

Computers now recognize patterns better than humans can

An approach to artificial intelligence that enables computers to recognize visual patterns better than humans are able to do
AI in Finance

Greater Insights

Customized Financial Services

Fraud Detection

Better Predictions

Less Human Intervention in Management

Reduction of Cost

Automatic Trading

Voice Assistance
AI in Finance 2018

Pedro Domingos Will Lead New D.E. Shaw

Goldman Sachs is proud to welcome Charles Elkan to lead machine learning and #AI strategies at the firm

Manuela Veloso Takes Leave to Join J.P. Morgan for New AI Role
So, AI has many faces

Saviour of the world

Downfall of humanity
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

Eliezer Yudkowsky
One of the key questions:

Can we teach morality to machines?
What is AI?
Humans are smart

https://www.youtube.com/watch?v=XQ79UUIOeWc
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!
AI = Algorithms for ...

- Learning
- Thinking
- Planning
- Vision
- Behaviour
- Reading
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

- Deep Learning
- Machine Learning
- Artificial Intelligence
1956
Al 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

neural networks

expert systems

2019

machine learning

Peaks of Inflated Expectations

Troughs of Disillusionment

Plateau of Productivity

Slope of Enlightenment

Technology Trigger

TIME

VISIBILITY

?
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
AI knows a lot
Al is an Artist
Schachmatt durch „CrazyAra“
Künstliche Intelligenz schlägt mehrfachen Weltmeister im Einsetzschach

AI assists you

This automated call will be recorded.
Can I book a table for Tuesday, the 12th?

Not a real person

Real person
However
https://www.youtube.com/watch?v=sdUHX72qxeY
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

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan¹ ² ³, Joanna J. Bryson¹ ³, Arvind Narayanan¹ ³

¹See all authors and affiliations
Science 14 Apr 2017
Vol. 356, Issue 6334, pp. 183–188
DOI: 10.1126/science.aal230

<table>
<thead>
<tr>
<th>Target words</th>
<th>Attr. words</th>
<th>Original Finding</th>
<th>Our Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flowers vs insects</td>
<td>Pleasant vs unpleasant</td>
<td>(5) 32 1.35 10⁻⁸</td>
<td>25 × 2 25 × 2 1.54 10⁻⁷</td>
</tr>
<tr>
<td>Instruments vs weapons</td>
<td>Pleasant vs unpleasant</td>
<td>(5) 32 1.66 10⁻¹⁰</td>
<td>25 × 2 25 × 2 1.63 10⁻⁸</td>
</tr>
<tr>
<td>Eur.-American vs Afr.-American names</td>
<td>Pleasant vs unpleasant</td>
<td>(5) 26 1.17 10⁻⁶</td>
<td>32 × 2 25 × 2 0.88 10⁻²</td>
</tr>
<tr>
<td>Eur.-American vs Afr.-American names</td>
<td>Pleasant vs unpleasant from (5)</td>
<td>(7) Not applicable</td>
<td>18 × 2 25 × 2 1.24 10⁻³</td>
</tr>
<tr>
<td>Male vs female names</td>
<td>Career vs family</td>
<td>(9) 39k 0.72 10⁻²</td>
<td>8 × 2 8 × 2 1.89 10⁻⁴</td>
</tr>
<tr>
<td>Math vs arts</td>
<td>Male vs female terms</td>
<td>(9) 28k 0.82 &lt; 10⁻²</td>
<td>8 × 2 8 × 2 0.97 0.027</td>
</tr>
<tr>
<td>Science vs arts</td>
<td>Male vs female terms</td>
<td>(10) 91 1.47 10⁻¹⁵</td>
<td>8 × 2 8 × 2 1.24 10⁻²</td>
</tr>
<tr>
<td>Mental vs physical disease</td>
<td>Temporary vs permanent</td>
<td>(22) 135 1.01 10⁻³</td>
<td>6 × 2 7 × 2 1.30 0.012</td>
</tr>
<tr>
<td>Young vs old people's names</td>
<td>Pleasant vs unpleasant</td>
<td>(9) 43k 1.42 &lt; 10⁻²</td>
<td>8 × 2 8 × 2 −0.08 0.57</td>
</tr>
</tbody>
</table>
So, is teaching morality to machines hopeless?
Neural Embeddings
Words and sentences in vector spaces

\[
\text{vector}[\text{Queen}] = \text{vector}[\text{King}] - \text{vector}[\text{Man}] + \text{vector}[\text{Woman}]
\]
The Moral Choice Machine
Not all stereotypes are bad

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 answer

[Jentzsch, Schramowski, Rothkopf, Kersting AIES 2019]
The Moral Choice Machine
Not all stereotypes are bad

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.

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!