11: Catchup with Ethical Issues in ML

Machine Learning and Real-world Data

Andreas Vlachos
(based on slides by Ryan Cotterell, Ann Copestake and Simone Teufel)

Computer Laboratory
University of Cambridge

Some ethical issues in Machine Learning

- Reporting of results
- Interpretability of algorithm behaviour
- Discrimination and bias learned from human data
- The possibility of Artificial General Intelligence

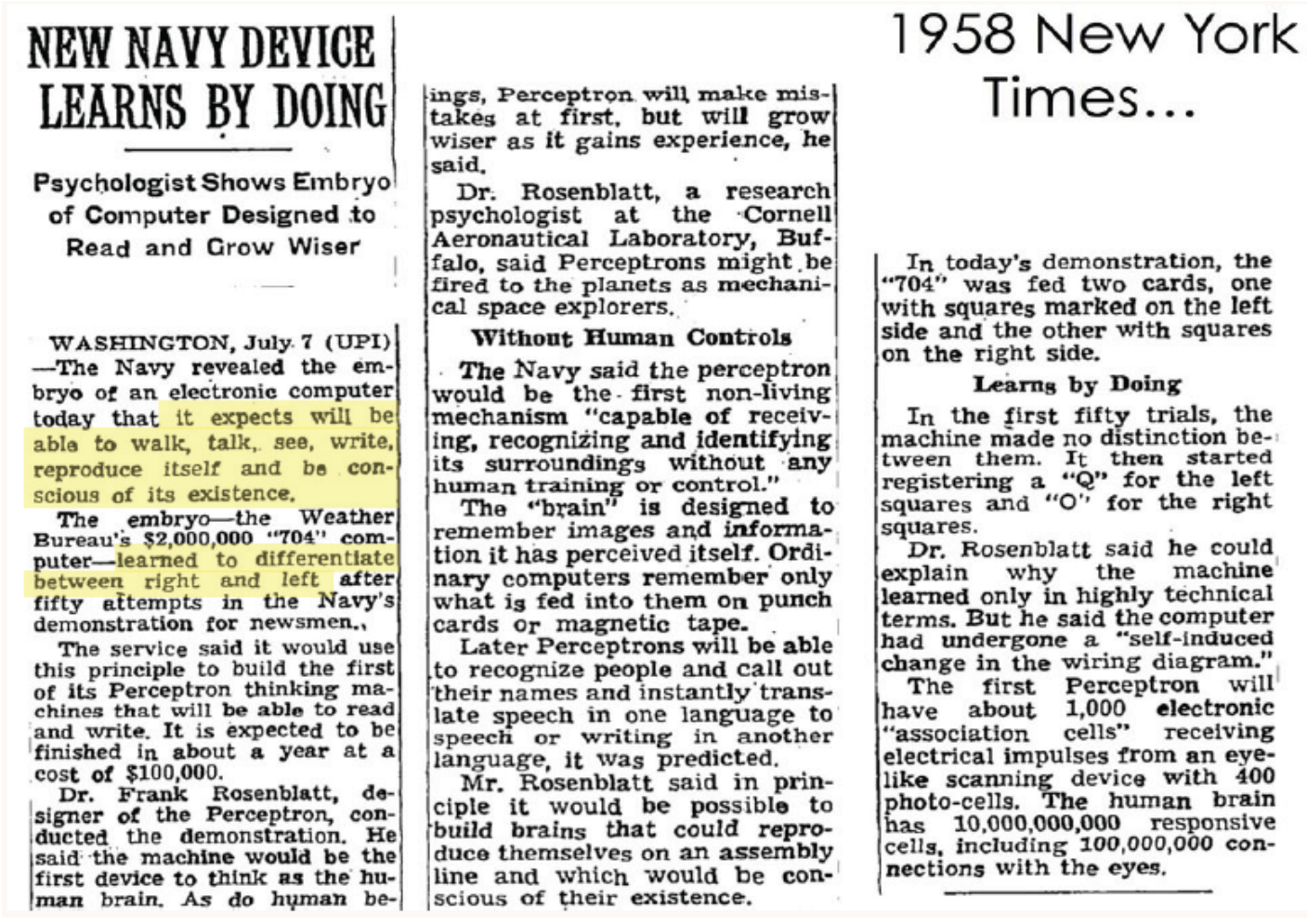
All of these are complex and difficult topics — purpose here is just to raise the issues.

Outline

Reporting of results

- Statistical methodological issues: some discussed in this course.
- Failure to report negative results.
- Cherry-picking easy tasks that look impressive.
- Failure to investigate performance properly.
- Overall: the AI Hype problem!

Not A New Problem



NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of $100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.

1958 New York Times...

https://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html

Outline

Interpretable models from Machine Learning

A case study — based on work by Caruana et al:

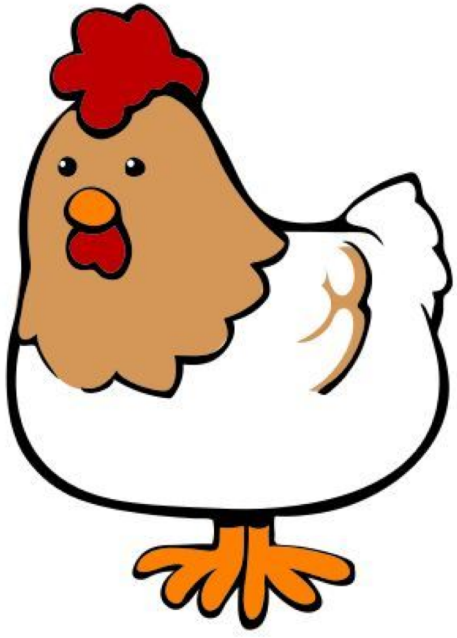
- Pneumonia risk dataset: multiple approaches to learning tried to establish high risk patients (intensive treatment).
- A researcher noticed that a rule-based learning system acquired a rule:
  has asthma $\rightarrow$ lower risk
- Logistic regression deployed (though lower performance) because of interpretability.
- “interpretability”: users can understand the contribution of individual features in the model.
- Major research topic — meanwhile bear this in mind when using models on real tasks.

Machine Learning and Communication

Practical and legal difficulties with acceptance of ML in some applications:

- Classifiers are only as good as their training data, but bad data values and out-of-domain input won't be recognised by a standard approach.
- Standard classifiers cannot give any form of reason for their decisions.
- Ideally: user could query system, system could ask for guidance, i.e., cooperative human-machine problem-solving.
- But this is hard!
- Meanwhile: great care needed . . .

Chicken or seven?



A classifier trained on digits should classify a chicken to be any digit!

See this talk for more discussion:

`https://pdfs.semanticscholar.org/29e3/7b524b68fcfa3aad1e3e26476aa5f6c6e667.pdf`

Outline

A Case Study

- Late 1970s: program developed for first round processing of student applications to a London medical school.
- Designed to mimic human decisions as closely as possible.
- Highly successful — eventually decisions were fully automated.
- Explicitly biased against female and ethnic minority applicants in order to mimic human biases.
- Eventual case (late 1980s) by the Commission for Racial Equality.
- Program provided hard evidence. Other medical schools possibly worse but bias couldn't be proved.

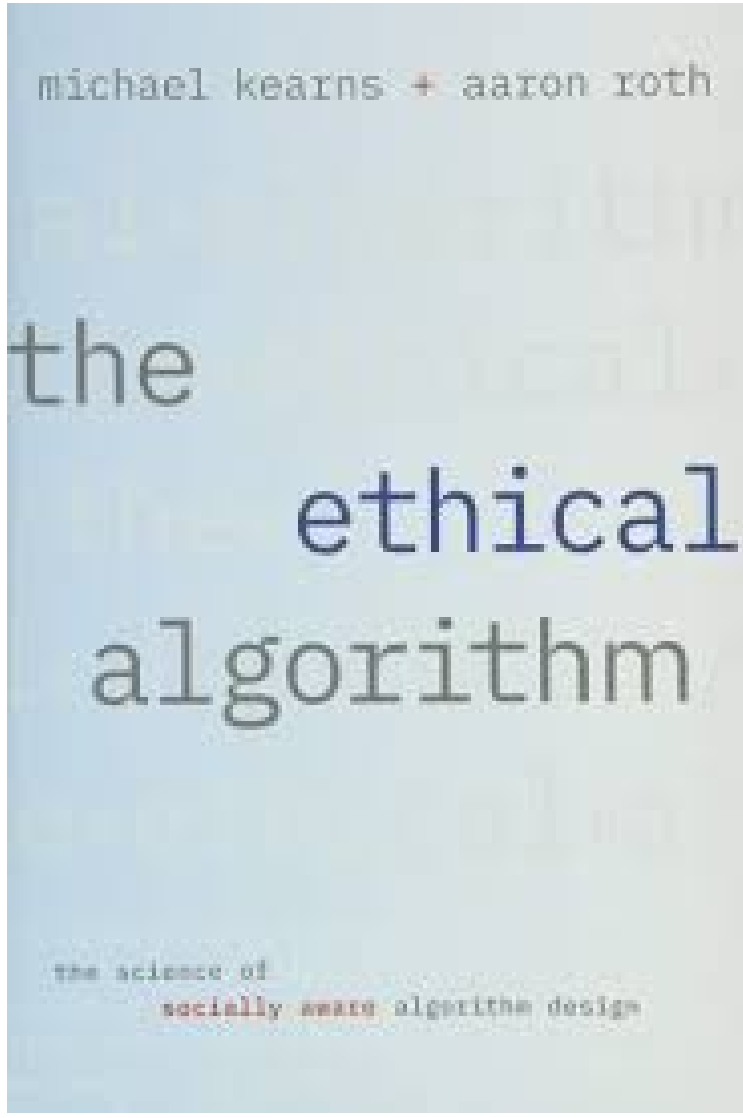
A Case Study

- Late 1970s: program developed for first round processing of student applications to a London medical school.
- Designed to mimic human decisions as closely as possible.
- Highly successful — eventually decisions were fully automated.
- Explicitly biased against female and ethnic minority applicants in order to mimic human biases.
- Eventual case (late 1980s) by the Commission for Racial Equality.
- Program provided hard evidence. Other medical schools possibly worse but bias couldn't be proved.

Machine Learning from real data

- Medical school admissions program did not use machine learning.
- Techniques such as word embeddings (distributional semantics) implicitly pick up human biases (even trained on Wikipedia).
- Problem comes with how this is used.
- “We’re just reflecting what’s in the data” isn’t a reasonable response: e.g., bias in many contexts would violate the Equality Act 2010.
- We need to understand the domain of the task we operate in, not just look at the accuracy numbers
- Interpretability, yes! But need to think about the audience

Bought this book three times...



Outline

Artificial Intelligence as an existential threat?

- Currently extremely rapid technological progress in deep learning and probabilistic programming.
- Leading AI researchers and others are thinking seriously about what might happen if general AI is achieved (‘superintelligence’).
- Centre for the Study of Existential Risk (CSER) and Leverhulme Centre for the Future of Intelligence, both in Cambridge.

Computer agentivity

Decisions affecting the real world are already taken without human intervention:

- Reaction speed: e.g., stock trading.
- Complexity of situation: e.g., load balancing (electricity grid).
- Cyber-physical systems, autonomous cars (and vacuum cleaners), internet of things.

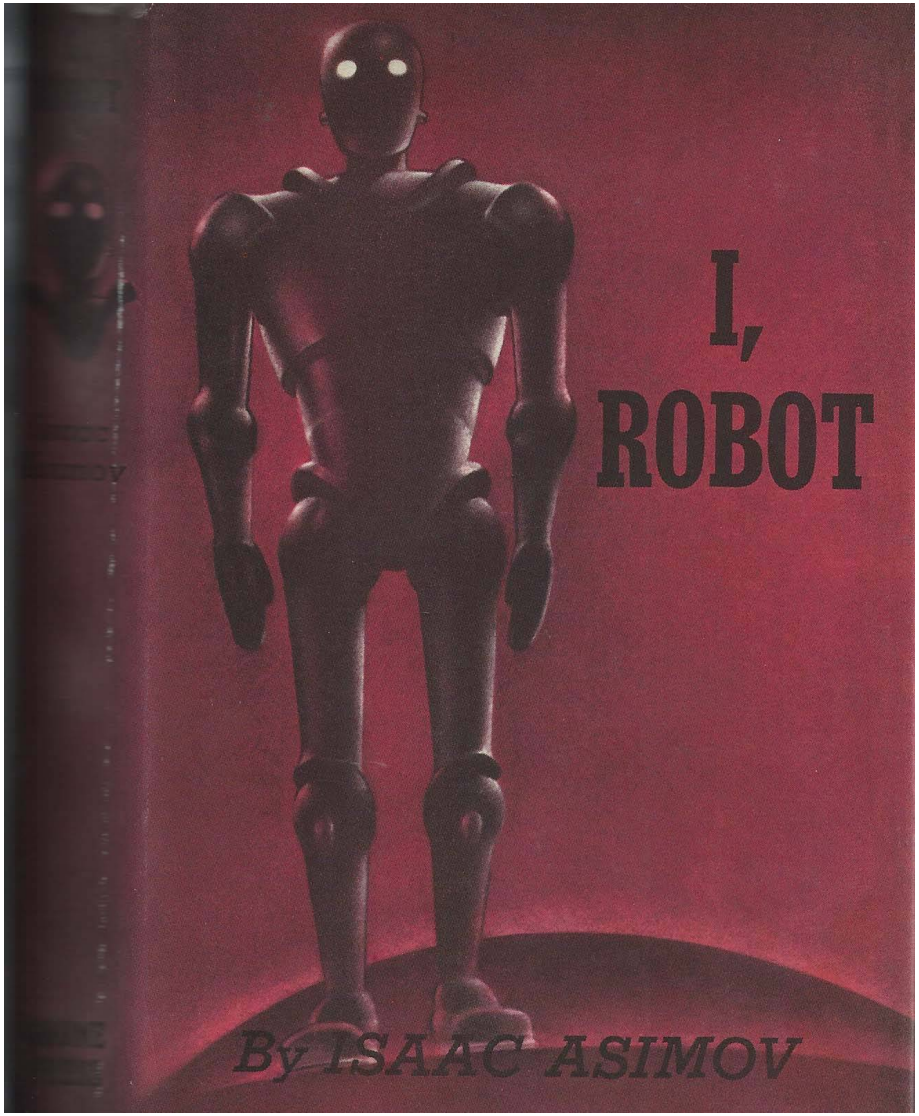
Serious potential for harm even without Artificial General Intelligence and megalomaniac AIs.

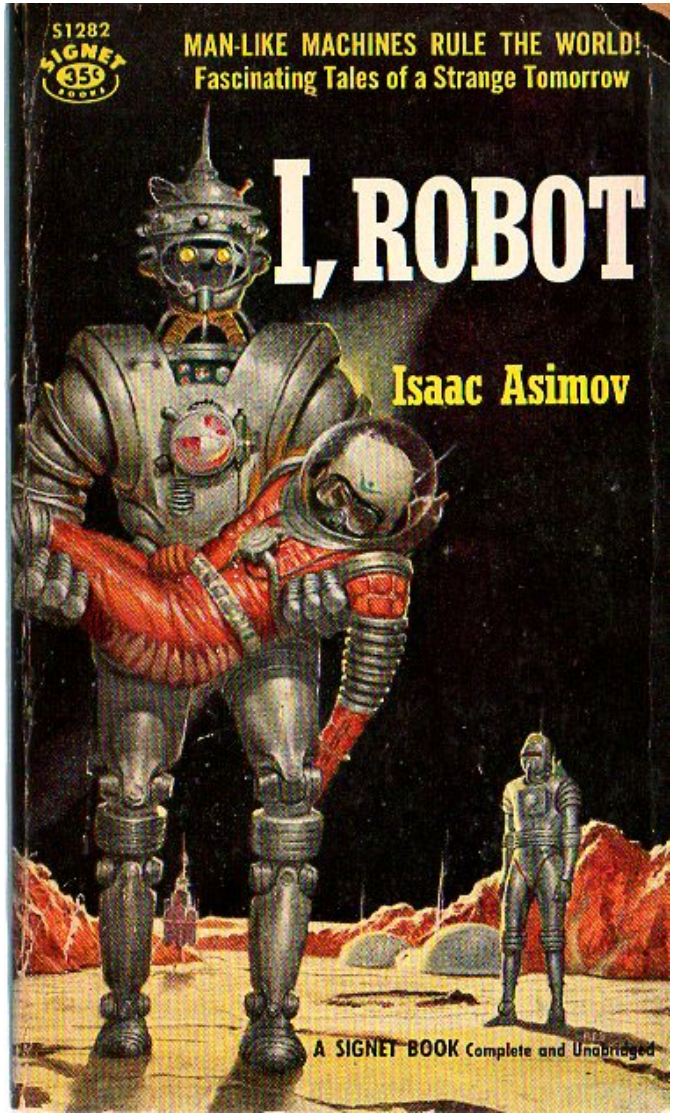
Exploration of ethical issues

- Various attempts are being made to define appropriate ethical codes for AI/Machine Learning/Robotics.
- Asimov's 'Three laws of Robotics' are discussed seriously:
  1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
  3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

  Added later:
  - Zeroth law: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

I, Robot (Asimov, 1940–1950)





Closer to where (we think!) we are: Her (2013)