11: Catch-up: Ethical Issues in ML

Eric Chamoun (based on earlier slides by Andreas Vlachos)

Some ethical issues in Machine Learning

- Reporting of results
- Intended uses of ML systems
- 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.

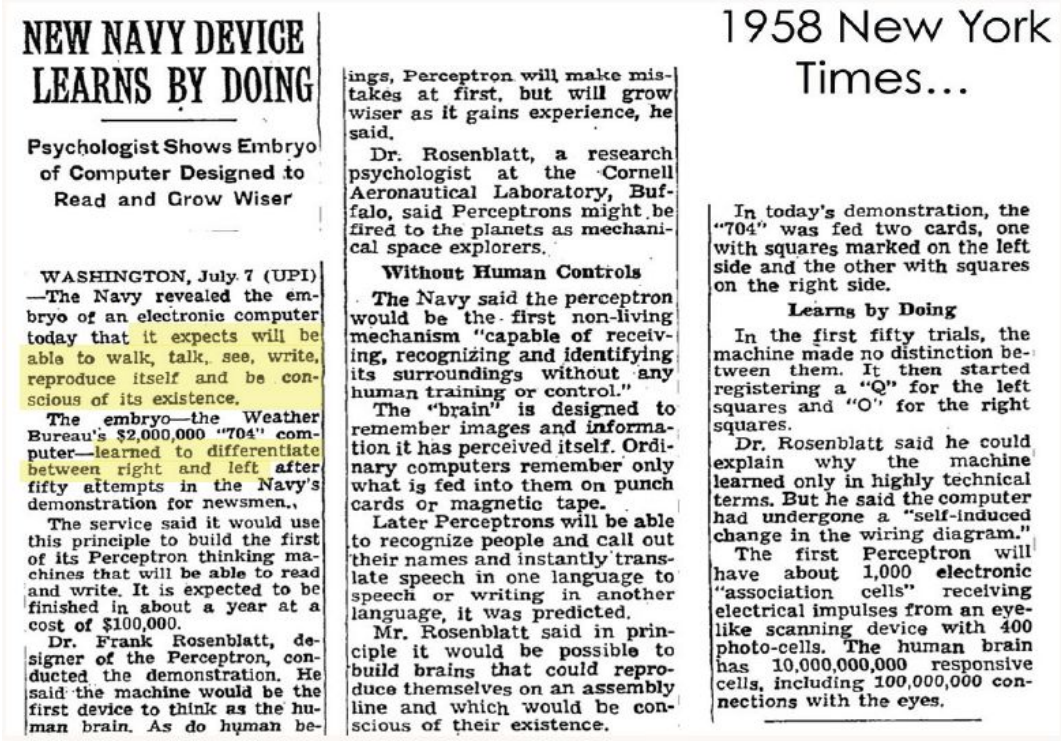
Outline

1. **Reporting of results**
2. Intended uses of ML systems
3. Interpretability of algorithm behaviour
4. Discrimination and bias learned from human data
5. The possibility of Artificial General Intelligence
6. Intended uses of ML systems

Reporting of results

- Statistical methodological issues: some discussed in this module already.
- Failure to report negative results
- Cherry-picking easy tasks to make system look impressive
- Failure to investigate performance properly
- 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.

1958 New York Times...

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.

Outline

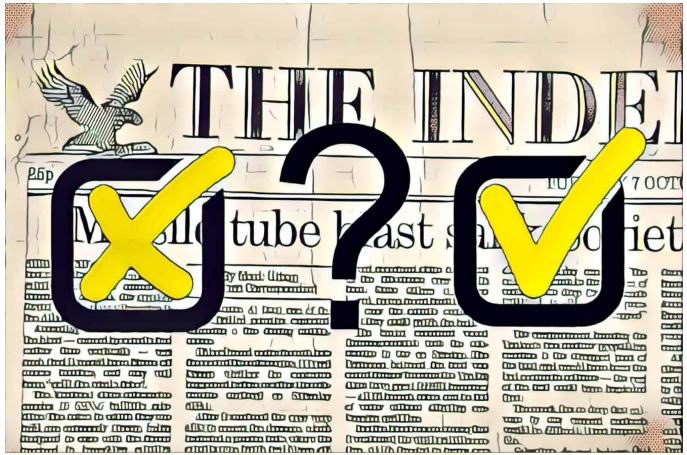
1. Reporting of results
2. **Intended uses of ML systems**
3. Interpretability of algorithm behaviour
4. Discrimination and bias learned from human data
5. The possibility of Artificial General Intelligence

Who?

- When building ML systems for applied tasks, important to be specific about:
  - **Who** is it intended for? E.g. Tool providing scientific feedback for *writers* (vs e.g. *reviewers* that could use them to automate their jobs)
  - **Who** would be the model owners responsible for decisions surrounding its deployment? E.g. In fact-checking, social media companies that could potentially use them to censor people? Or journalists to assist their jobs?

How?

- Also important to specify **how** the designed system is intended to be used.
- E.g. automated fact-checking system that identifies potential information, **how** will misinformation be dealt once identified?



- Crucial to specify **how** to assess potential impact of the designed system.

Why?

- Also important to be specific about **why** the system has been designed
- E.g. a system achieving good performance in generating legal cases should be specific about the end goal of this system:
    - Intended to replace lawyers? Or assistants?
    - What can the system do well and what can it not? E.g. Weaknesses of LLMs when it comes to referencing cases.
- Otherwise, your system will be all over the news :)

Outline

1. Reporting of results
2. Intended uses of ML systems
3. **Interpretability of algorithm behaviour**
4. Discrimination and bias learned from human data
5. The possibility of Artificial General Intelligence

Interpretability of results

- Important to understand reasons behind a model's performance - these may not always be good ones.
- A case study - based on work by Caruana et al.:
  - Pneumonia risk dataset: multiple approaches to learning tried to establish high-risk patients.
  - A researcher noticed that a rule-based system had acquired the following rule

    has asthma $\longrightarrow$ lower risk

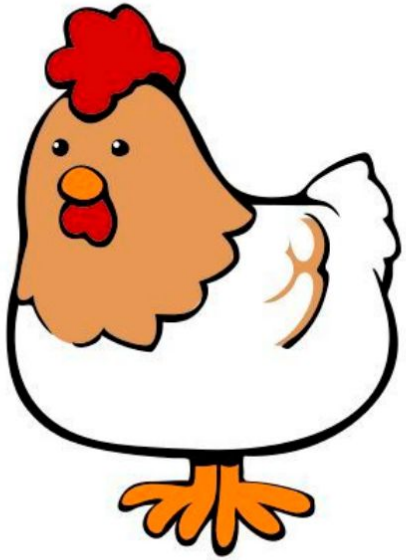
  - Logistic regression deployed (though lower performance) due to interpretability

Major research topic - meanwhile bear this in mind when using models on real tasks

Machine Learning and Communication

- Practical and legal difficulties with application of ML systems in real-world settings:
  - Classifiers only as good as training data! E.g.: bad data values/out-of-domain inputs won't be recognized by standard approaches.
  - 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 this chicken to be any digit!

Outline

1. Reporting of results
2. Intended uses of ML systems
3. Interpretability of algorithm behaviour
4. **Discrimination and bias learned from human data**
5. The possibility of Artificial General Intelligence

A Case Study

- Late 1970s: program developed for processing the first round of student applications to 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 (1980s) by the Commission for Racial Equality.
- Program provided hard evidence. Other schools possibly worse but bias couldn’t be proved.

Machine learning from real world data

- Medical school admissions program did not use machine learning.
- Techniques such as word embeddings (distributional semantics) implicitly pick up human biases (even if trained on Wikipedia).
- Problem comes with how it is used.
- “We’re just reflecting what’s in the data” isn’t a reasonable response: e.g. bias in many contexts would violate 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.

Outline

1. Reporting of results
2. Intended uses of ML systems
3. Interpretability of algorithm behaviour
4. Discrimination and bias learned from human data
5. **The possibility of Artificial General Intelligence**

Artificial Intelligence as an existential threat?

- Currently extremely rapid progress in deep learning and probabilistic programming - AI hype is stronger than ever since release of ChatGPT.
- Leading AI researchers and others are thinking seriously about what might happen if Artificial General Intelligence is achieved (‘superintelligence’)
- Centre for 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 AGI and megalomaniac AIs.

Exploration of ethical issues

- Various attempts to define appropriate ethical codes for AI/ML/Robotics
- Asimov's Three laws of Robotics:
    - A robot may not injure a human being or, through inaction, allow a human being to come to harm.
    - A robot must obey orders given to it by human beings except where such orders would conflict with the First Law.
    - A robot must protect its own existence as long as such protection does not conflict with the First and Second Laws.
- Added later:
    - Zeroth law: a robot may not harm humanity, or, by inaction, allow humanity to come to harm.

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

References for further reading

NY Times 1958 article (AI Hype):

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

Talk with the chicken and seven example:

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

Intended uses of ML systems

https://aclanthology.org/2023.findings-emnlp.577.pdf

Case study: Bias in the application processing algorithm for the London Medical School

https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias

Ethics in Computer Science: “The Ethical Algorithm: The Science of Socially Aware Algorithm Design” Book

https://www.amazon.co.uk/Ethical-Algorithm-Science-Socially-Design/dp/0190948205

Asimov’s Three laws of Robotics:

https://www.amazon.co.uk/I-Robot-Isaac-Asimov/dp/0008279551/ref=tmm_pap_swatch_0?_encoding=UTF8&qid=&sr=

Thank you for your attention!