15: Ethics in Machine Learning, plus Artificial General Intelligence and some old Science Fiction

Machine Learning and Real-world Data

Ryan Cotterell, after slides by Ann Copestake and Simone Teufel

Computer Laboratory
University of Cambridge

Lent 2020

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.

1 Reporting results

2 Interpretability of Results

3 Discrimination and bias

4 Artificial General Intelligence / Superintelligence

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!

Outline.

1. Reporting results
2. Interpretability of Results
3. Discrimination and bias
4. Artificial General Intelligence / Superintelligence

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 → 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 …

Outline.

1 Reporting results

2 Interpretability of Results

3 Discrimination and bias

4 Artificial General Intelligence / Superintelligence

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.
- Interpretability etc

Outline.

1 Reporting results

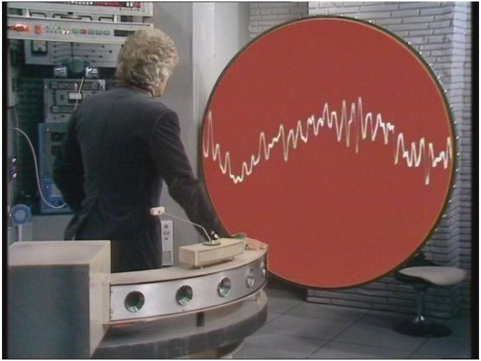
2 Interpretability of Results

3 Discrimination and bias

4 Artificial General Intelligence / Superintelligence

Dr Who, The Green Death, episode 5 (1973)

BOSS (Bimorphic Organisational Systems Supervisor), a megalomaniac supercomputer.



The Doctor: *"If I were to tell you that the next thing I say would be true, but the last thing I said was a lie, would you believe me?"*

Some assumptions in that episode

- Real AI was close: believed by many people in 1970s.
- An AI might be malevolent towards humans.
- An AI would be able to control some people and subvert other computers.
- The AI would be able to communicate in fluent natural language.
- The AI would be logic-based: to the extent it could be confused (briefly!) by a paradox.

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.

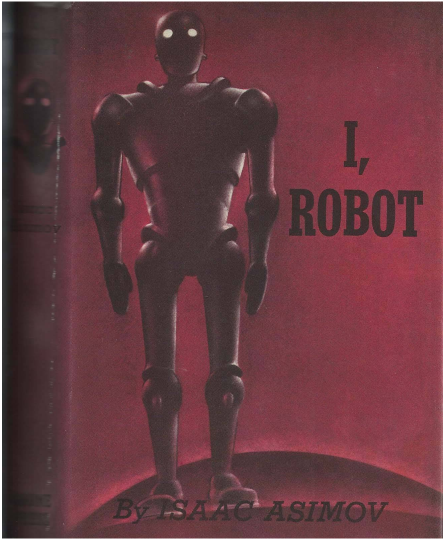
Exploration of ethical issues

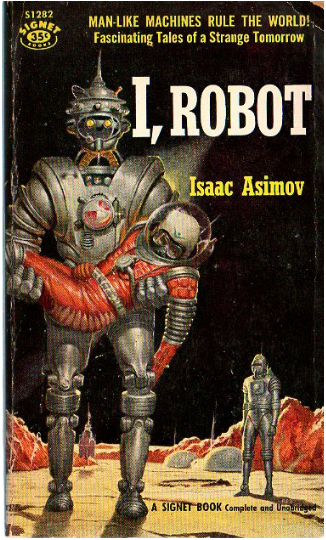
- Various attempts are being made to define appropriate ethical codes for AI/Machine Learning/Robotics.
- Asimov's 'Three laws of Robotics' are discussed seriously:
  1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  2. A robot must obey the orders given 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:

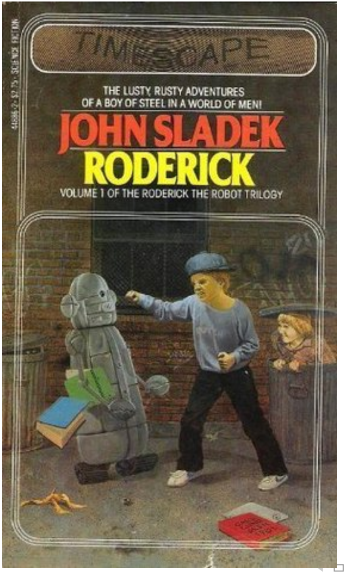
  - Zeroth law: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

I, Robot (Asimov, 1940–1950)

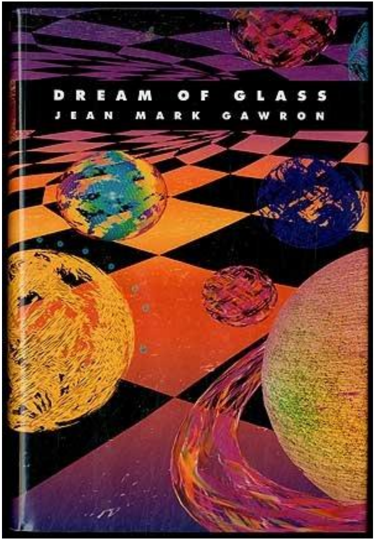



S1282
SIGNET
35¢
BOOKS
MAN-LIKE MACHINES RULE THE WORLD!
Fascinating Tales of a Strange Tomorrow
I, ROBOT
Isaac Asimov
A SIGNET BOOK Complete and Unabridged

Roderick or The Education of a Young Machine (Sladek 1980), Roderick at Random (Sladek 1983)


TIMESCAPE
THE LUSTY, RUSTY ADVENTURES
OF A BOY OF STEEL IN A WORLD OF MEN!
JOHN SLADEK
RODERICK
VOLUME 1 OF THE RODERICK THE ROBOT TRILOGY

Dream of Glass (Gawron, 1993)



Schedule

- Today: last lecture (not examinable!) and ticking.
- Monday March 9: demonstrators available from 14:05 for final ticks, no lecture.
- Question sets: will release qset4 to supervisors soon.
- All qsets (but no answers) available to students after supervisions finished.