Machine Learning

Lecture 11
Ethics and Bias in Machine Learning

Justin Pearson[1]

mailto:it-1dl034@lists.uu.se

[1] https://intro-ml-1dl034-uu-se.github.io/index.html

Plan

Today we will only scratch the surface but we will address the following topics.

- Some examples to get your thinking.
- An Attempt to Define Bias
- Training sets.
- Solving Bias: Not just a technical problem.
- A Brief overview of some European frameworks.

Some examples of Bias

Sometimes the examples are difficult to verify. Partly because machine learning is a moving target, partly because defining and identifying bias is complicated.

- Ask google image search to for pictures of a CEO. Will we get
    - Mostly Men?
    - Mostly Women?
    - A balanced sample of men and women?

Image Search

- It was commonly reported that you would only get pictures of men.
- Not true anymore.
- Why do you think that you would get pictures of only men?
- If you wanted to redress the balance what would you do?

Ask ChatGPT to write a story

- Ask ChatGPT to write a story about a brave solider who saves the world?
- Ask ChatGPT to write a story about a computer scientist who saves the world?

Male or Female?

COMPAS tool[3]

Used in the US as a decision support tool for judges. Used to assess if offenders are going to commit crimes again. Used in parole cases.

Overestimated the probability that a black person would re-offend.[2]

- Why do you think this is the case?
- Is it a good idea to have software doing this sort of thing?
- Would you as technical person trust the output of such software? What would it take to convince you that it had made the correct decision?

[2] Although the true story more complicated. See the Kleinberg paper referenced at the end

[3] https://en.wikipedia.org/wiki/COMPAS_(software)

A made up example based on real life

You favourite political party complains that your favourite social network keeps serving stories of their members acting stupidly.

Is this bias?

- Give a situation where it is not biased.
- Give a situation where it is biased.

What duties to programmers have when it comes to recommending things?

Recruitment Algorithm

Amazon built an AI based recruitment system. It was Biased against women.

Why might this be?

Image Recognition

> *A face recognition system (say some sort of electronic locking mechanism) that does not work well with people who have darker skin tones.*

Why might this be?

Why should we care about ethics?

Computer scientists and programmers typically ignore the ethical implications of their developments:

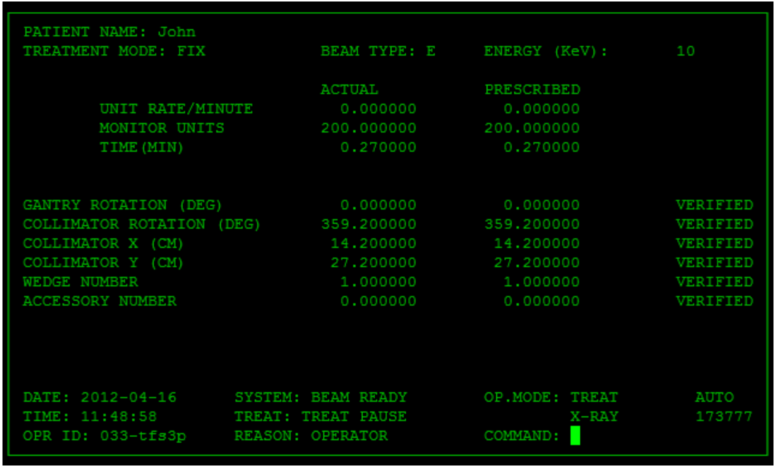
- We produce and release software with bugs. Even though our software sometimes can kill people
- We produce software that can lead to more inequality (banks, hiring, medical treatments...)

Other disciplines do worry about ethics:

- Nobody would build bad bridge or a building that will fall down. Building companies and architects do their best avoid harm or loss of life.
- Medical developments go through extensive testing.

Radiation Therapy

Therac-25



```
PATIENT NAME: John
TREATMENT MODE: FIX              BEAM TYPE: E     ENERGY (KeV):       10

                                 ACTUAL           PRESCRIBED
        UNIT RATE/MINUTE           0.000000         0.000000
        MONITOR UNITS            200.000000       200.000000
        TIME(MIN)                  0.270000         0.270000


GANTRY ROTATION (DEG)              0.000000         0.000000          VERIFIED
COLLIMATOR ROTATION (DEG)        359.200000       359.200000          VERIFIED
COLLIMATOR X (CM)                 14.200000        14.200000          VERIFIED
COLLIMATOR Y (CM)                 27.200000        27.200000          VERIFIED
WEDGE NUMBER                       1.000000         1.000000          VERIFIED
ACCESSORY NUMBER                   0.000000         0.000000          VERIFIED




DATE: 2012-04-16      SYSTEM: BEAM READY     OP.MODE: TREAT        AUTO
TIME: 11:48:58        TREAT: TREAT PAUSE              X-RAY        173777
OPR ID: 033-tfs3p     REASON: OPERATOR       COMMAND: █
```

80s programming. All written in PDP-11 assembly language.

Therac-25

- Radiation therapy machine. At least 6 patients where given 100 times the intended dose of radiation.
- The causes are complex [4], but at least one cause was identified:
  - **Inadequate Software Engineering Practices** ... including:

The software should be subject to extensive testing and formal analysis at the module and software level; system testing alone is not adequate. Regression testing should be performed on all software changes.

[4] `http://sunnyday.mit.edu/papers/therac.pdf`

Why should we care about ethics?

Even though we have a history of producing bad software. The chances of causing harm are even greater when we start using AI.

Suppose **you** design an autopilot for a car. Consider the three situations:

- The car autopilot kills the driver by crashing the car into a road barrier. Nobody else is killed.
- The car autopilot kills a pedestrian because your vision recognition system does not notice her.
- The car autopilot kills the driver to avoid killing six pedestrians by driving the car into a road barrier.
- The car autopilot kills six pedestrians to avoid killing the driver.

Is Ethics a technical problem?

- Discuss for 5 minutes with the person next to you. Discuss if ethics is problem that can be solved technically, or if not why not?

Bias

- General AI has lots of ethical and legal problems. For the rest of the lecture we will look a common machine learning problem Bias.

Bowls

This is what happens when you get old in England.

Bias

- Origins Old French: biais meaning sideways, askance, against the grain.
- “a strong feeling in favour of or against one group of people, or one side in an argument, often not based on fair judgement.”
- Statistical bias is a term that refers to any type of error or distortion that is found with the use of statistical analyses.[5]
- Norm theoretic definition. Biases typically involve systematic departures from accepted norms.

[5] https://link.springer.com/referenceworkentry/10.1007/978-94-007-0753-5_2865

Psychology of Bias

- Cognitive Bias: A cognitive bias is a systematic pattern of deviation from norm or rationality in judgement. Individuals create their own "subjective reality" from their perception of the input.
- Unconscious or implicit bias is a negative attitude, of which one is not consciously aware, against a specific group.

Our psychology is against us. It is often easier to make biased decisions that are not necessarily fair or rational.

Avoiding Bias in Machine Learning[6]

This is only the start.

Pick the right dataset.

> *Machine learning is only as good as the data that trains it. Whatever data you feed into your AI must be comprehensive and balanced while replicating the actual demographic of society.*

This means that you have to identify possible sources of bias.

[6] `https://www.ibm.com/resources/guides/predict/trustworthy-ai/avoid-bias/`

Pick the right dataset — Labelling

- Be careful with labelling. This can introduction Bias.
- Imagenet a widely used dataset in image analysis is biased to various degrees[7]. This is because it is a labelled dataset, and it has been labelled with a community effort, and there is a collection bias.

[7] https://dl.acm.org/doi/abs/10.1145/3351095.3375709

Pick the right dataset — Ignoring Features

- Implicit variables are often easy to learn. Simply removing a feature is not going to help. There was a recent case of a Covid dataset, where x-rays were taken. Most of the Covid patients were lying down. The ML algorithm simply inferred the orientation of the x-ray rather than inferring if the patient had Covid or not.

Avoiding Bias — Model Training

- if the training data is not balanced or the model architecture is not designed to handle diverse inputs, the model may produce biased outputs.

Techniques include:

- Under sampling. Use a subset.
- Over sampling. Make copies or weight the minority
- Synthetic Minority Oversampling Technique (SMOTE). Generate new instances of the minority algorithm that are statistically similar.

The list goes on. If your spam filtering training set contains 90% spam messages, then your algorithm just needs to predict that any message is spam, and it is right most of the time.

Bias — Not just a technical problem

Again Jon Kleinberg[8] has various impossibility results about achieving fairness. There is still debate on how to interpret the result. It means that you have to be very careful with what you mean by fairness mathematically, and there is more than one answer.

Kleinberg's result does mean that avoiding bias cannot be solved by purely technical means. So this means that you have to put best practices in place, try to collect good data, and try to understand what your ML algorithm is doing. Try to identify possible sources of bias and fairness.

[8] https://arxiv.org/abs/1609.05807

European Framework for Trustworthy AI

Seven requirements that are defined very much in the spirit of the rights of individuals and what they can expect from AI.

1. Human Agency and Oversight: Including fundamental rights, human agency and human oversight.
2. Technical Robustness and Safety: Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility.
3. Privacy and Data Governance: Including respect for privacy, quality and integrity of data, and access to data.
4. Transparency: Including traceability, explainability and communication.
5. Diversity, Non-discrimination and Fairness: Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.
6. Environmental and Societal Well-Being: Including sustainability and environmental friendliness, social impact, society and democracy
7. Accountability: Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

Fundamental Rights

Fundamental rights encompass rights such as human dignity and non-discrimination, as well as rights in relation to data protection and privacy, to name just some examples. Prior to self-assessing an AI system with this Assessment List, a fundamental rights impact assessment (FRIA) should be performed.

Excepts from the Checklist

Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?

Excepts from the Checklist – Accuracy

- Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?
- Did you put in place measures to ensure that the data (including training data) used to develop the AI system is up-to-date, of high quality, complete and representative of the environment the system will be deployed in?

Excepts from the Checklist — Explainability

- Did you explain the decision(s) of the AI system to the users?
- Do you continuously survey the users if they understand the decision(s) of the AI system?

We've only briefly touched upon this in the course, but you need to avoid black box decisions. You need to have mechanism where you can explain why a decision was made.

EU AI Act[9]

As they say on the internet: IANL. But legislation is here, and in some industries you will have to be aware.

Risk based approach based on how it is going to be applied.

- Unacceptable Risk Systems will be Prohibited. Includes social scoring, real-time remote biometric identification in public spaces such as facial recognition.
- High Risks Systems will be Carefully Regulated. Includes biometrics, critical infrastructure, Education, Law enforcement.
- Limited and Minimal Risk Systems. Everything else, but still subject to transparency obligations.

[9] https://cset.georgetown.edu/article/the-eu-ai-act-a-primer/

References

- A Survey on Bias and Fairness in Machine Learning.[10]
- Examples of Bias in Machine Learning.[11]
- IBM on Bias.[12]

[10] https://dl.acm.org/doi/abs/10.1145/3457607 or https://arxiv.org/abs/1908.09635

[11] https://www.prolific.com/resources/shocking-ai-bias

[12] https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/

References

- A ‘black box’ AI system has been influencing criminal justice decisions for over two decades – it’s time to open it up.[13]
- Jon Kleinberg[14] has lots of interesting articles on bias. Inherent Trade-Offs in the Fair Determination of Risk Scores[15] has an interesting discussion of the COMPAS case, as well as an impossibility result.

[13] `https://theconversation.com/a-black-box-ai-system-has-been-influencing-criminal-justice-decisions-for-over-two-decades-it`

[14] https://www.cs.cornell.edu/home/kleinber/

[15] `https://arxiv.org/abs/1609.05807`

References

- CAIPD[16]
- EU - Ethics guidelines for trustworthy AI.[17]
- The EU Artificial Intelligence Act.[18]

[16] https://www.caidp.org/

[17] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[18] https://artificialintelligenceact.eu/