COMP309/AIML421 in Week 12, 2024

ethics and stuff

Marcus Frean, with obsequious indebtedness signalling to
Drs Andrew Lensen, Bach Nguyen and Qi Chen

Week 12

- Lecture 1
  - the role of ethics in AI research
- Lecture 2
  - *generative* models
- Tutorial:
  - no tutorial this week, as you are too busy (me too)

ML benefits

Efficiency and throughput

- e.g. better/faster coding

Dangerous tasks (via robots)

- e.g. space, deep ocean, fires

Lifestyle

- e.g. spam detection, predictive text, photo recognition, human-like chatting

Medical applications

- e.g. assessing patients, remote surgery

Learning: searching, summarising…

some AI risks

- Fakes → confusion, manipulation by bad actors

- Weapons:
  - Autonomous weapons
  - an AI arms race is unavoidable
- Taking our jobs:
  - millions of jobs potentially at risk
  - what kind of jobs will we lose?
    Physical – Analytical – Social
- Taking our brains:
  - what kind of abilities will we lose?
    Physical – Analytical – Social
- Group dynamics
  - they'll play the game better than us
  - social media have fkd with our society already, and now this…
- Poor implementations (e.g. bias)

- Ubiquitous surveillance, without consent:
  - track and id people "in the wild"
  - states, corporations, & bad actors (mess with your head, on-sell your info, threaten your reputation)
- Privacy:
  - big tech + AI like a sales person knocking on your door, except they've rifled through your mail first
  - listening to our conversations

What can we do?

- Duck & Hide?
- Up-skill?
- Fight the machines?
  Legal ethical framework?

AI ethics

- How to design computer and robotic systems that demonstrate sensitivity to human values, and factor these into making decisions in morally significant situations

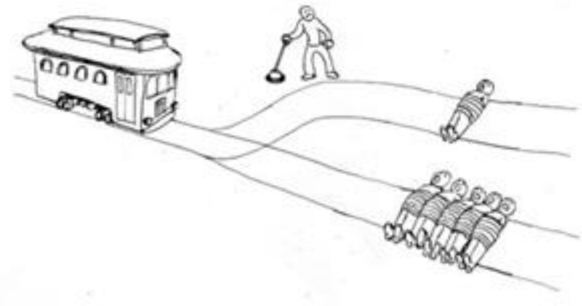
trolley problems

A runaway trolley is heading towards five tied-up people lying on the tracks. You are next to a lever that controls a switch.

You have two options:

1) Do nothing and allow the trolley to kill the five people on the main track.
2) Pull the lever, diverting the trolley onto the side track where it will kill one person.

- Which is the most ethical option?
  What should an AI do?



When an AI results in a harm, who is to blame?
Is it the developers, the company, the users, or the AI itself?

e.g. autonomous vehicles

- there are no universally accepted standards for how autonomous vehicles should behave in morally challenging situations

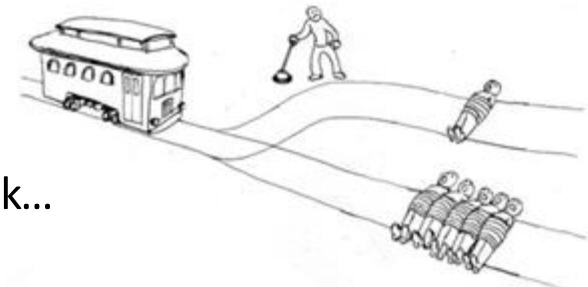
- Mercedes-Benz simply intends to program its self-driving cars to save the people inside the car. Every time.

  The car can only make decisions based on what it can detect and predict, and it's more certain about the presence and safety of its passengers than external factors ⇒

1. “If you know you can save at least one person, at least save that one. Save the one in the car.”

   But also, it’s true that

2. Buyers don’t want to buy a car that will put them at risk…



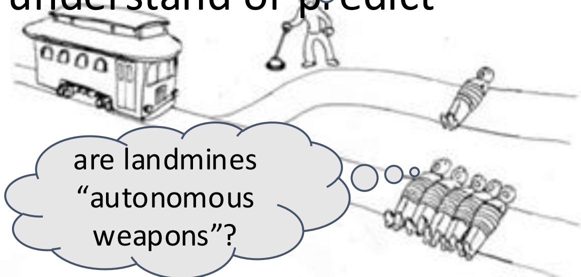
When an AI results in a harm, who is to blame?
Is it the developers, the company, the users, or the AI itself?

e.g. autonomous weapons

- Should the AI bear any responsibility? Most argue "no" because AI lacks intent, consciousness, or malice. *(ChatGPT sounds like a person, but there’s no I)*
- However, as AI systems become more complex and their decision-making processes more opaque, pinpointing where something went wrong can be challenging.
- This raises the question of whether we should even allow systems to operate autonomously in life-and-death situations if we cannot understand or predict their decisions.

“killer robots”

but also, think police tasers, etc



are landmines “autonomous weapons”?

bias

- Bias is disproportionate weight in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.

- Machine learning follow rules learned from data.

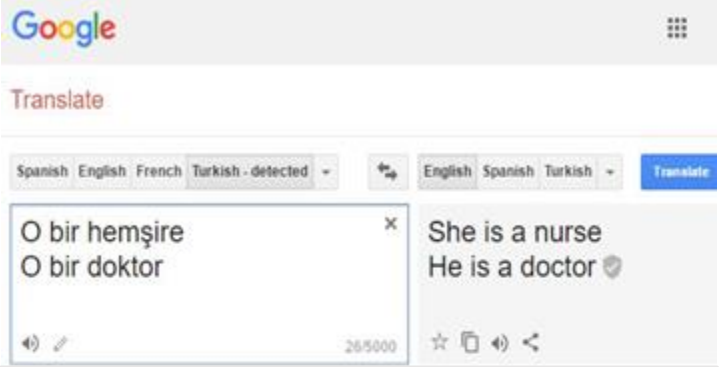
- If the data is biased/skewed, an error or unjust rule may be learned - which can lead to biased outcomes.

Bias in AI: not just one problem

1. **Sample Bias**: Occurs when the data used to train the model isn't representative of the broader population.
   → system that performs well for one group but not others.
   - *Example*: If facial recognition is predominantly trained on light-skinned faces, it might perform poorly on dark-skinned individuals.
2. **Aggregation Bias**: is the system built to over-generalise?
   - *Example*: A health AI tool that doesn't differentiate between male and female patients might miss sex-specific risks.
3. **Confirmation Bias**: is the system prone to stereotyping?
   - *Example*: If an AI recommendation system starts to suggest certain types of content based on initial user interactions, it might pigeonhole the user into a narrow content band, ignoring broader interests.
4. **Bias due to Evaluation/Deployment mismatch**: Did the metrics match the deployment situation?
   - *Example*: If a speech recognition tool is only tested on urban accents and is then used in diverse regions, its real-world performance might be overestimated.
   - *Example*: A traffic prediction system trained in one city might inaccurately predict traffic patterns in another city with different driving habits.

examples of bias

In the Turkish language, there is one pronoun, “o,” that covers every kind of singular third person.



O bir hemşire
O bir doktor

She is a nurse
He is a doctor

- Man is to Computer programmer as Women is to Home maker
- Man is to Manager as Women is to Assistant

Many financial institutions use machine learning algorithms to predict creditworthiness and decide who gets approved for loans or credit cards.

Often, these algorithms can be biased against young adults or people with thin credit histories.

http://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/

- Commercial face recognition systems IBM Watson Visual Recognition, Microsoft Face Detect and Face++
- Face pictures with range of skin tones.
- The result was that the lighter the skin, the more accurate the result.
- All 3 performed the worst on darker female faces with error rates from 20.8% to 34.7%.

What can we do?

- Transparency: If a public agency can’t explain an algorithm or its decision-making – should it be used?

side note:
Overcorrecting for one form of bias can introduce new problems. AI responses may be tweaked toward equality of outcome in some form, but this is potentially introducing yet another form of bias.

“Fair” is not trivial to define!

In other words the pursuit of one form of fairness can lead to other forms of injustices. I don’t assume the progressive orthodoxy will get it right.

AI and hate speech

Detecting hate speech using AI is a challenging task for multiple reasons:

1. **Ambiguity in Definition**: the very definition of 'hate speech' can be ambiguous. What one person or culture considers hateful, another might see as a strong opinion or even a joke. Drawing a line is complex.
2. **Bias in Training Data**: all the types of bias mentioned above could apply to the data used to train hate speech detectors.
3. **Overreach and Free Speech Concerns**: There's a significant risk of AI systems overreaching and flagging legitimate speech as hate speech.

Perhaps:

- AI could aid in the detection of hate speech by flagging potential instances for human review.
- Some combination of AI tools, human judgment, and community feedback is likely the most effective approach.
- Platforms and organizations should be transparent about their definitions, methodologies, and decisions regarding hate speech detection.

“Marcus, where can I learn more about these fascinating topics?”

AIML430: Applications and Implications of Artificial Intelligence

Deep Learning:
a panacea?

noun

1.a solution or remedy for all difficulties or diseases.