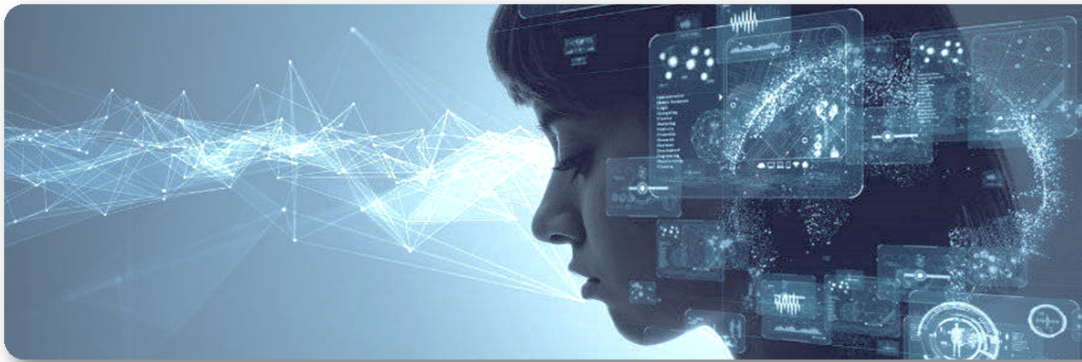


COMP 2211 Exploring Artificial Intelligence

Ethics of Artificial Intelligence

Prof. Song Guo, Dr. Desmond Tsoi & Dr. Huiru Xiao

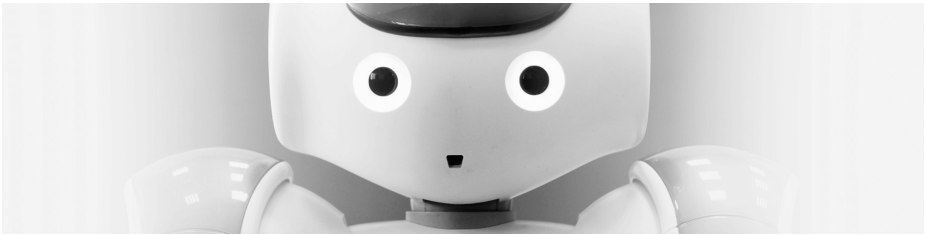
Department of Computer Science & Engineering
The Hong Kong University of Science and Technology, Hong Kong SAR, China

What is AI Ethics?

- AI ethics is a term that refers to the ethical issues surrounding the use of Artificial Intelligence (AI) technology.
- As the use of AI systems becomes more prevalent across the globe, governments, industry groups and AI-focused executives are grappling with how to ensure the technology is used ethically.

Key Questions In the Field of AI Ethics

- What are applications for AI systems ethical for a given organization?
- How can companies ensure AI systems are built to operate in a fair and unbiased way?
- What processes do companies need to ensure AI systems continue to function ethically over time?



Definition of AI Ethics

Definition

The UK's Alan Turing Institute defines AI ethics as a set of values, principles and techniques that employ widely accepted standards of 'right' and 'wrong' to guide the development and use of AI technologies.

- In practice, this means ensuring that organizations using AI have the right AI ethics policy and governance practices to ensure the technology is used for good and does not unintentionally harm people.

Potential 'Harms' That AI Systems May Cause

- Invading people's right to privacy by processing data without consent or handling it in a way that reveals personal information without an individual's consent.
- Making biased or unfair decisions or recommendations about certain populations or demographics.
- Make decisions in a way that can't be explained in plain language, so it is unclear if their conclusions are fair and unbiased.
- Making unreliable decisions or delivering poor quality outcomes due to model implementation issues.
- Denying people their right to accountability for the decisions AI systems make about them.

Is AI Ethical?

- AI technology itself is neither ethical nor unethical.
- Instead, enterprises must establish principles or frameworks to ensure that they use AI systems ethically and responsibly and guard against AI misuse.
- Problems:
  - Non-consensus about what ethical responsibilities enterprises have for different applications for AI technology.
  - Different AI-focused executives can look at the same use case for AI and draw different conclusions about their moral responsibilities.

AI Ethics Issues and Considerations

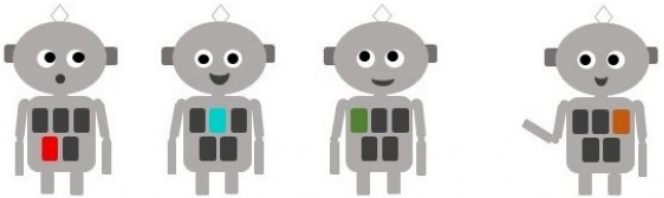
- There are many ethical dilemmas associated with AI use.
- These range from
  - deciding whose lives autonomous vehicles should prioritize saving in a multi-person crash situation to
  - ensuring that credit scoring AIs do not discriminate against people unfairly based on factors such as gender.

Can Artificial Intelligence be Ethical?

- The general public and business community are becoming increasingly aware of the ethical issues surrounding AI use.
- Three main areas to ensure the AI models in production in their organizations function ethically:
  1. Data Ethics
  2. AI Model Bias
  3. AI Model Monitoring and Maintenance

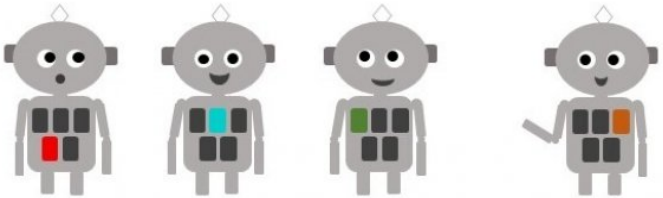
1. Data Ethics

- Ensuring AI models function in a way that is fair, unbiased and in customers' best interests starts with ensuring the data that feeds into them is collected, governed and used ethically.
  - This means ensuring companies secure the proper consent from customers before using their data and handle it in a secure way that respects their privacy.
  - This means taking proactive steps to address data bias in those datasets and ensure the populations being analyzed are fairly represented in the data.



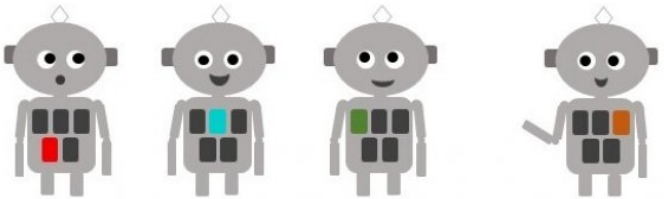
2. AI Model Bias

- Ensuring the data that feeds AI models is relatively unbiased.
- AI and ML ethics requires enterprises to ensure that those models are built in a way that ensures they make their decisions or recommendation fairly.



3. AI Model Monitoring and Maintenance

- Some machine learning models adjust themselves to improve their accuracy.
- Others may see their accuracy change over time due to changes in the data flowing into them, in a phenomenon known as 'AI model drift'.
- That means enterprises must also ensure AI models are effectively monitored and maintained to continue functioning as intended over time.



The Emergence of AI Principles

In the last few years, a number of institutions have published AI principles:

- Principles for Algorithmic Transparency and Accountability (ACM 2017)
- IEEEs General Principles of Ethical Autonomous and Intelligent Systems (IEEE 2017)
- Five principles for cross-sector AI code (UK House of Lords, 2018)
- Ethics Guidelines for Trustworthy AI (European Commission, 2019)
- AI Ethics Principles (Google, 2019, 2020, 2021)
  `https://ai.google/principles/`

The Seven European Union Principles

1. Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions ...
2. Technical robustness and safety: AI systems need to be resilient and secure. They need to be safe, ensuring a fallback plan if something goes wrong ...
3. Privacy and data governance: Besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured ...
4. Transparency: The data, system and AI business models should be transparent ...
5. Diversity, non-discrimination and fairness: Unfair bias must be avoided ...
6. Societal and environmental well-being: AI systems should benefit all human beings ...
7. Accountability: Mechanisms should be implemented to ensure responsibility and accountability for AI systems ...

Common Grounds

There are many different lists of principles, but it seems that they can be synthesized into five fundamental principles:

1. Autonomy: People should be able to make their own decisions, e.g. human-in-the-loop, privacy protection
2. Beneficence: Society at large should benefit
3. Non-maleficence: Harmful consequences should be avoided, e.g. systems should be robust
4. Justice: Diversity, non-discrimination and fairness
5. Explicability: Transparency and explainability

The Problem with Principles

It is good to state principles. However, they also create problems since they are very high-level.

- They can be interpreted in different ways
  For example, autonomous killer drones can be considered beneficial for the soldiers, or morally impermissible, because machines decide about life and death.
- They can conflict with each other in concrete cases
  For example, privacy and data collection for health science can conflict.
- They can come into conflict in practice
  For example, an excellent diagnosis might still be preferable even if its reasoning cannot be explained.

It is nevertheless good to have such principles as orientation points and evaluate solutions.

Fairness

The topic of enforcing fairness has become important, in particular in machine learning.

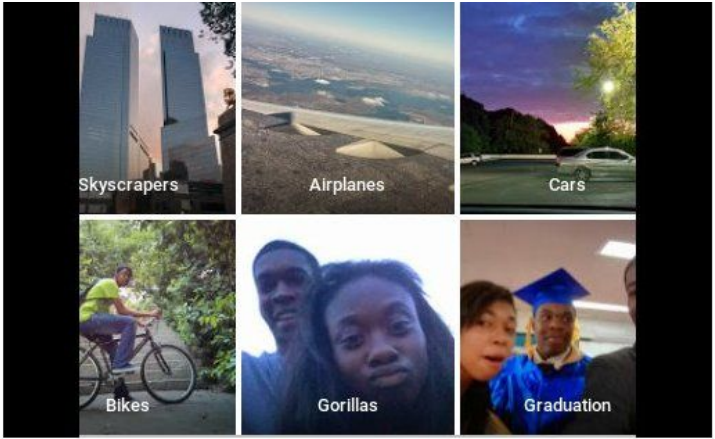
- Why care about fairness in ML?
- What kind of unfairness could there be?
- What causes unfairness?
- What concepts of fairness are there?

Why Care?

- Many things become automated by machines:
  - Employers select candidates by using ML systems.
  - Linkedin and XING use ML systems to rank candidates.
  - Courts in the US use ML systems to predict recidivism.
  - Banks use credit rating systems, which use ML.
  - Amazon and Netflix use recommendation systems.
- If these systems act unfair, groups and individuals may suffer.

Unfairness: Example

- Face recognition in Google Photo mis-classifiers black people.



diri noir avec banan @jackyalcine · Jun 29

Google Photos, y'all [redacted] My friend's not a gorilla.

Unfairness: Example

- A search query in XING orders less qualified male candidates higher than the more qualified female candidate.

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query "Brand Strategist".

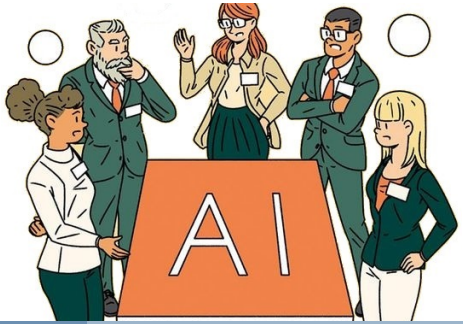
Possible Reasons for Unfairness

- Skewed sample: If some initial bias happens, such bias may compound over time: future observations confirm prediction and fewer opportunities to make observations that contradict prediction.
- Tainted examples: For example, word embeddings may lead to gender stereotypes if they are present in the text one learns from.
- Limited features: Some features may be less informative for a minority group.
- Sample size disparity: Training data from a minority group is sparse.

Can Machines Make Moral Decisions?

- Philosophers usually consider machines as not capable of making moral decisions.
- However, one can try to find properties such that machines could act morally.
- Machines need to have at least:
  - Beliefs about the world
  - Pro-attitudes (intentions)
  - Moral knowledge
  - The possibility to compute what consequences one's own action can have

  in which case they can be considered as moral agents.



Moral Decision: Example

- Self-driving cars will come into situations where they have to choose between bad alternatives (e.g., killing the passenger or a pedestrian).
- How should such a car choose in such a situation?
- Note that because of its much faster reactivity, a car might be able to make decisions where a human cannot at all.

Ask what ordinary people think a car should do in such moral dilemmas.

Practice Problem

Match the European Union Principles with the given Common Grounds by completing the following table with numbers.

Seven European Union Principles

(A) Human agency and oversight
(B) Technical robustness and safety
(C) Privacy and data governance
(D) Transparency
(E) Diversity, non-discrimination and fairness
(F) Societal and environmental well-being
(G) Accountability

Common Grounds

(1) Autonomy
(2) Beneficence
(3) Non-maleficence
(4) Justice
(5) Explicability

A	B	C	D	E	F	G

Practice Problem

A	B	C	D	E	F	G
1	3	1	5	4	2	1

Acknowledgment

- Most of the materials in this lecture notes were obtained from an online source:

  Joschka Boedecker and Wolfram Burgard and Frank Hutter and Bernhard Nebel and Michael Tangermann, 16. AI & Ethics, Ethical Consideration about AI & Machine Ethics, Foundations of Artificial Intelligence, lecture notes, Albert-Ludwigs-Universität Freiburg, delivered on 24 July, 2019:

  `http://ais.informatik.uni-freiburg.de/teaching/ss19/ki/slides/ai16_ethics.pdf`.

That's all!

Any questions?