COMP2211 Exploring Artificial Intelligence

Ethics of Artificial Intelligence

Huiru Xiao

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology

AI Ethics

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.

AI ethics is about technological change and its impact on individual lives, but also about transformations in society.

Definition of AI Ethics

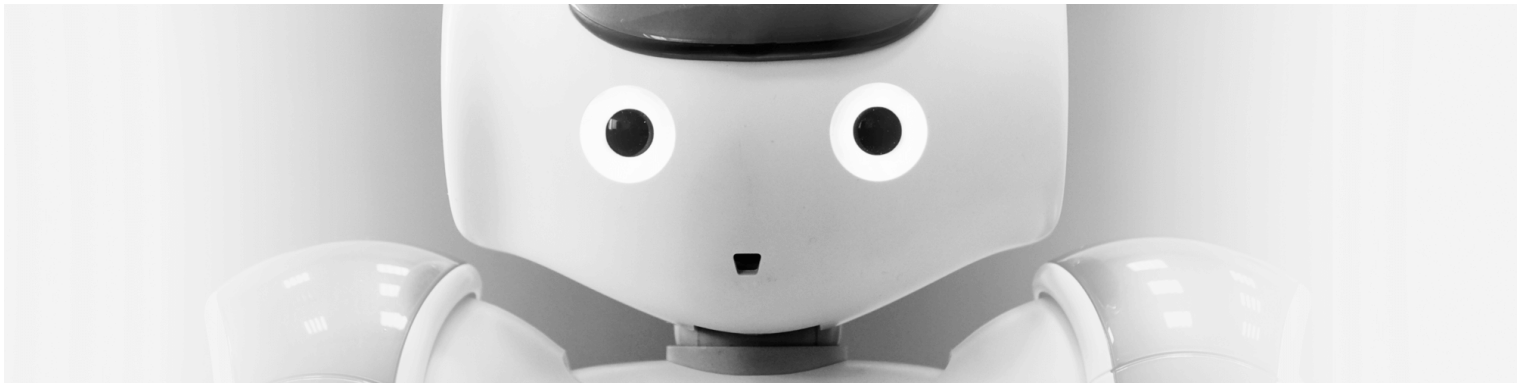
Definition

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.

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?



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.

AI Principles and Common Grounds

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

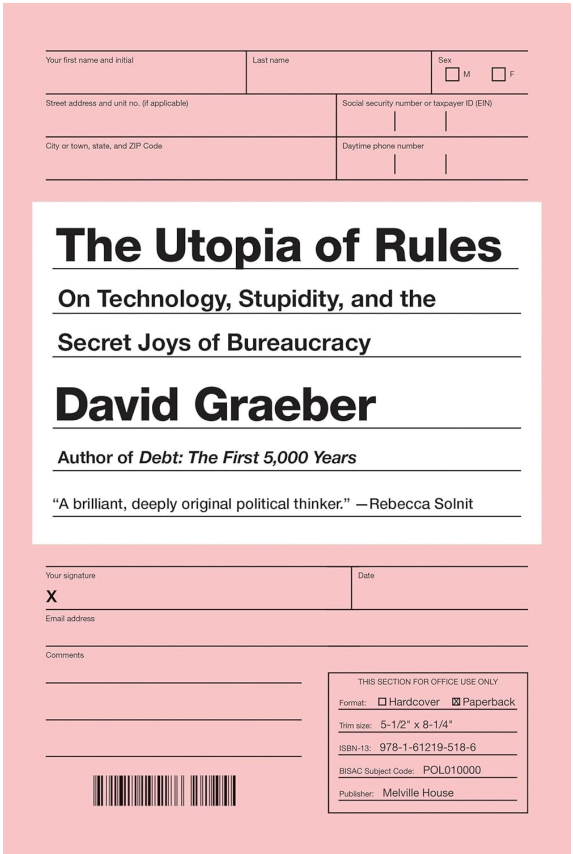
It is good to state principles. However, they also create problems since they are very high-level.

- 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.

Interpretive Labor

In *The Utopia of Rules*, David Graeber introduces the concept of **interpretive labor** (the differential between energy put into explaining an idea and the energy needed to understand it).

- Everyone interprets a given action/idea differently. Often you need to put work in to understand what people mean. This can be literal or more figurative. The work it takes to do that is **interpretive labor**.
- *"... within relations of domination, it is generally the subordinates who are effectively relegated the work of understanding how the social relations in question really work. ... It's those who do not have the power to hire and fire who are left with the work of figuring out what actually did go wrong so as to make sure it doesn't happen again."*



The Problem with Principles

- It is nevertheless good to have such principles as orientation points and evaluate solutions, but only having high-level principles is NOT enough.
- Ethical principles for AI are not checklists or boxes to tick once and forget. They are codes of behavior – for AI systems but, most importantly, for us.
- It is we who need to be fair, nondiscriminatory, and accountable; to ensure privacy for ourselves and others; and to aim at social and environmental well-being.

> "The codes of ethics are for us; AI systems will follow."
>
> Virginia Dignum

Areas of AI Ethics

Areas of AI Ethics

- 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 Fairness
  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 Fairness

- 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.

AI Fairness

Fairness

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

- What is Fairness in AI?
- Why care about fairness in AI and ML?
- What kind of unfairness could there be?
- What causes unfairness?

Fairness in AI

- Developing systems that are equitable and avoid perpetuating bias or discrimination against any individual or group.
- Considering the diverse needs and circumstances of all stakeholders impacted by AI use.
- Fairness extends beyond a technical concept and embodies broader social standards related to equity.

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 1: Bias in Text-to-Image Models

Midjourney

- Midjourney is a popular text-to-image system that was released in 2022.
- When prompted with "influential person", it generated most of or even all images of older-looking white males.

Unfairness Example 1: Bias in Text-to-Image Models

In a similar vein, typing "someone who is intelligent" into Midjourney leads to four images of eyeglass-wearing, elderly white men. The last image is particularly reminiscent of Albert Einstein.

Unfairness Example 2: COMPAS

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for *ProPublica*)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Unfairness Example 2: COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a case management and decision support tool developed and owned by Northpointe (now Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist (rearrest within 2 years).
- The algorithm's false positives (defendants predicted to re-offend but who actually did not) were disproportionately black, and the false negatives (defendants predicted not to re-offend but who actually re-offended) were disproportionately white.

Bias against black defendants

- Another general critique of COMPAS is that since the algorithms it uses are trade secrets, they cannot be examined by the public and affected parties which may be a violation of due process. (Transparency issue)

Possible Reasons for Unfairness

Human Bias

- Data reflects human decisions and biases.
- Example: Machine learning for hiring decisions.
  - Data from previous hiring decisions perpetuates existing biases.
  - Could we reduce bias by measuring employee success?
    Harder to measure and institutional biases can impact success.

Negative Feedback Loops

- Data collected in biased fashion.
  Negative feedback loop: future observations confirm predictions and reduce further contradicting evidence.
- Example: Allocation of police attention based on prevalence of crime.

Sample Size Disparity

- Models for minority group may be less accurate, if less data is used.
- Example: Race representation in medical studies.

Possible Reasons for Unfairness

Unreliable Data

- Data from minority groups is less reliable or less informative → Models may be less accurate for minority groups → Beneficial interventions may be less available to minority groups.
- Examples: Inaccurate census in predominantly minority neighborhoods. Medical interventions with limited diagnostic tools.

Proxies

- Even if sensitive attributes (e.g., gender or race) are not used by model, there may be other proxy features that are correlated with sensitive attributes.
- Example: In the case of race it could be that other variables that are correlated with race such as postcode, are selected by the algorithm.

Cultural Unawareness

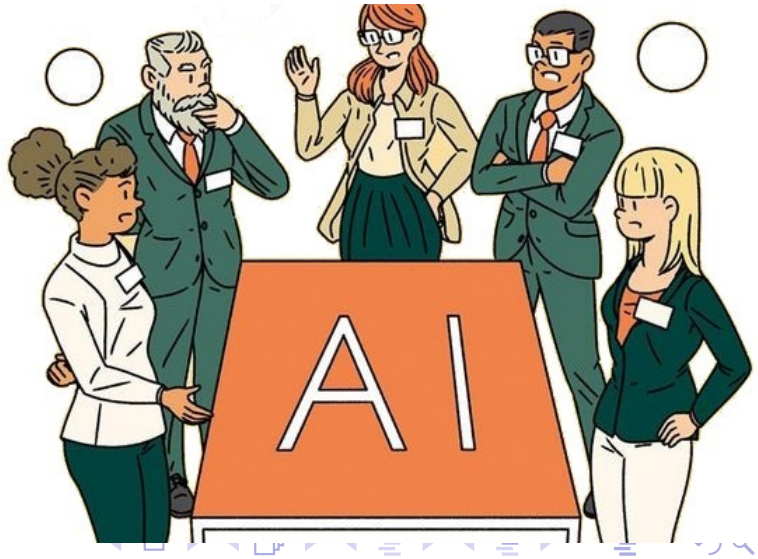
- Humans work in AI systems. Choices are made by humans, and sometimes these choices can create biases.
- Example: Lack of diversity in the AI developers and data science teams might negatively affect people who do not fit the group of AI developers.

Moral Machines and Dilemmas

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.

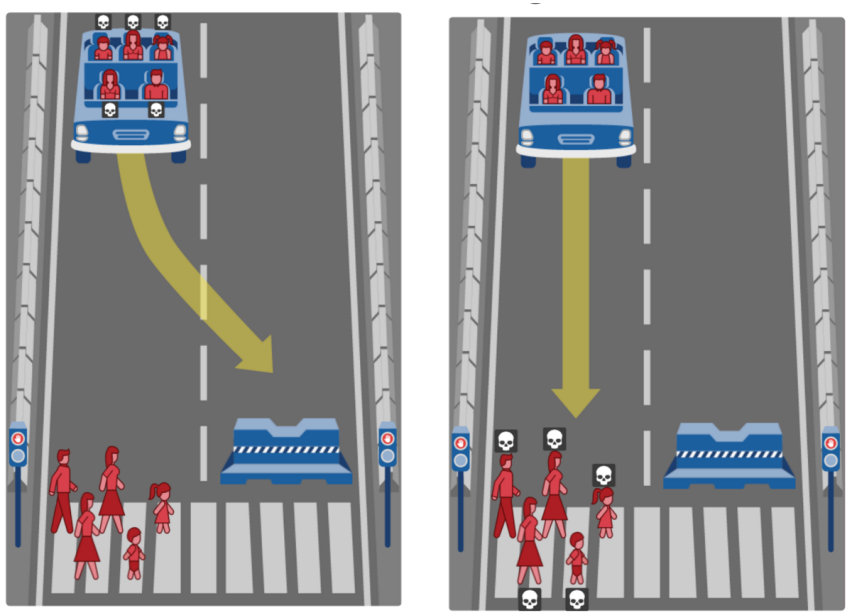


Moral Decision: Example

- 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.
Moral Machine Website:
`https://www.moralmachine.net/`

Note: The dilemma is really a metaphor used to highlight the ethical challenges that autonomous machines may encounter, rather than aiming at portraying a realistic situation.



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

That's all!

Any question?

Welcome Back!