Principles of AI Ethics and Limits of Contemporary AI

AI, Ethics, and Society

Arash Sheikhlar

September 9th 2024

Who am I

- **Arash Sheikhlar, *PhD student in computer science, will soon graduate.***

Background:

- Electrical engineering (BSc in electronics and MSc in control theory),
- Computer Science, PhD in Computer Science and General Machine Intelligence,

Research interests

- Autonomous Generalization: developing mechanisms that allow AI agents to cumulatively learn and reason
- Reasoning architectures: systems that use logic to make inferences

Hobbies

- Watching Sci-Fi movies
- Playing card/video games

Course overview and assessment of my part

Overview of the next two weeks

- General principles of AI ethics (this week)
- Ethical issues with the current AI architectures (this week)
- What Future AI (the next 20-40 years) will look like
- How to make future AI systems more ethical

Assessment

- Assignments and group projects
- Class discussions
- Final exam – written exam in person on Canvas

This week

- ***Main principles of AI ethics***
- **Contemporary AI**
- **Evaluate contemporary AI and its applications based on principles of AI ethics**

Learning goals:

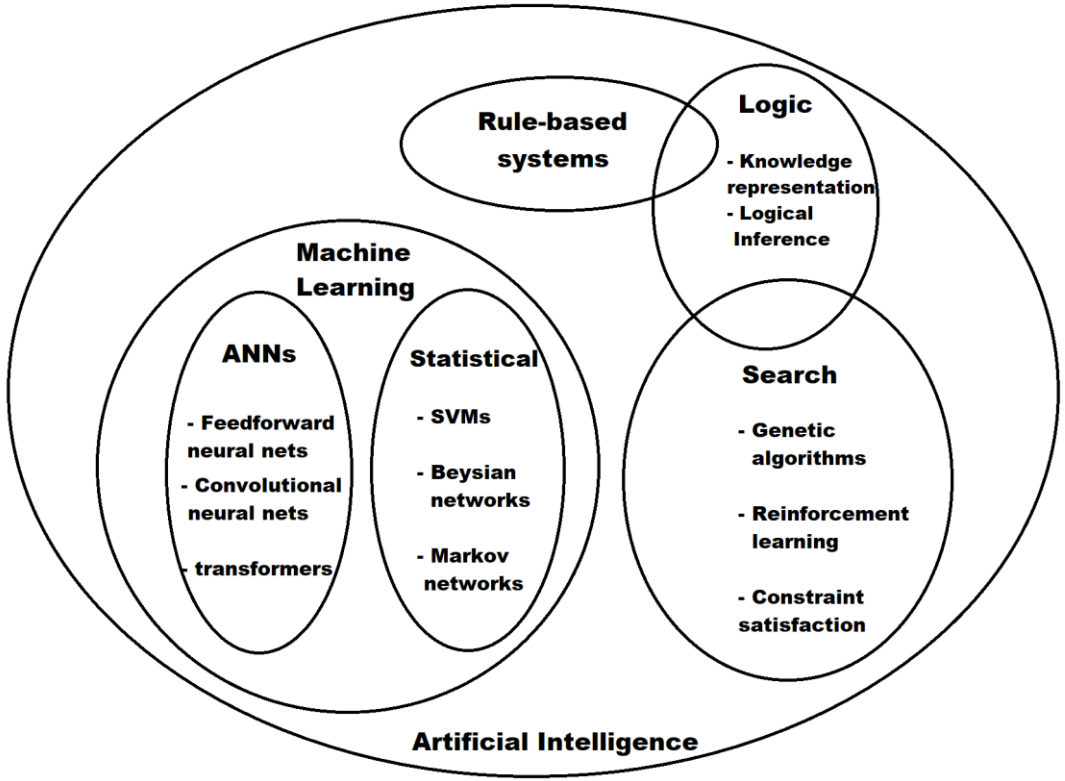
- What the principles are
- Why they are deemed important
- What related issues they are related to
- How principles have to be implemented
- Do contemporary AI systems meet these principles?
- How can contemporary AI be made more ethical?
- Evaluating some large AI projects in relation to AI ethics principles

Main principles of AI ethics*

- Transparency
- Justice & fairness
- Safety
- Responsibility
- Privacy

* Jobin, et al. "The global landscape of AI ethics guidelines." *Nature machine intelligence* 1.9 (2019): pp 389-399

Overview of AI systems

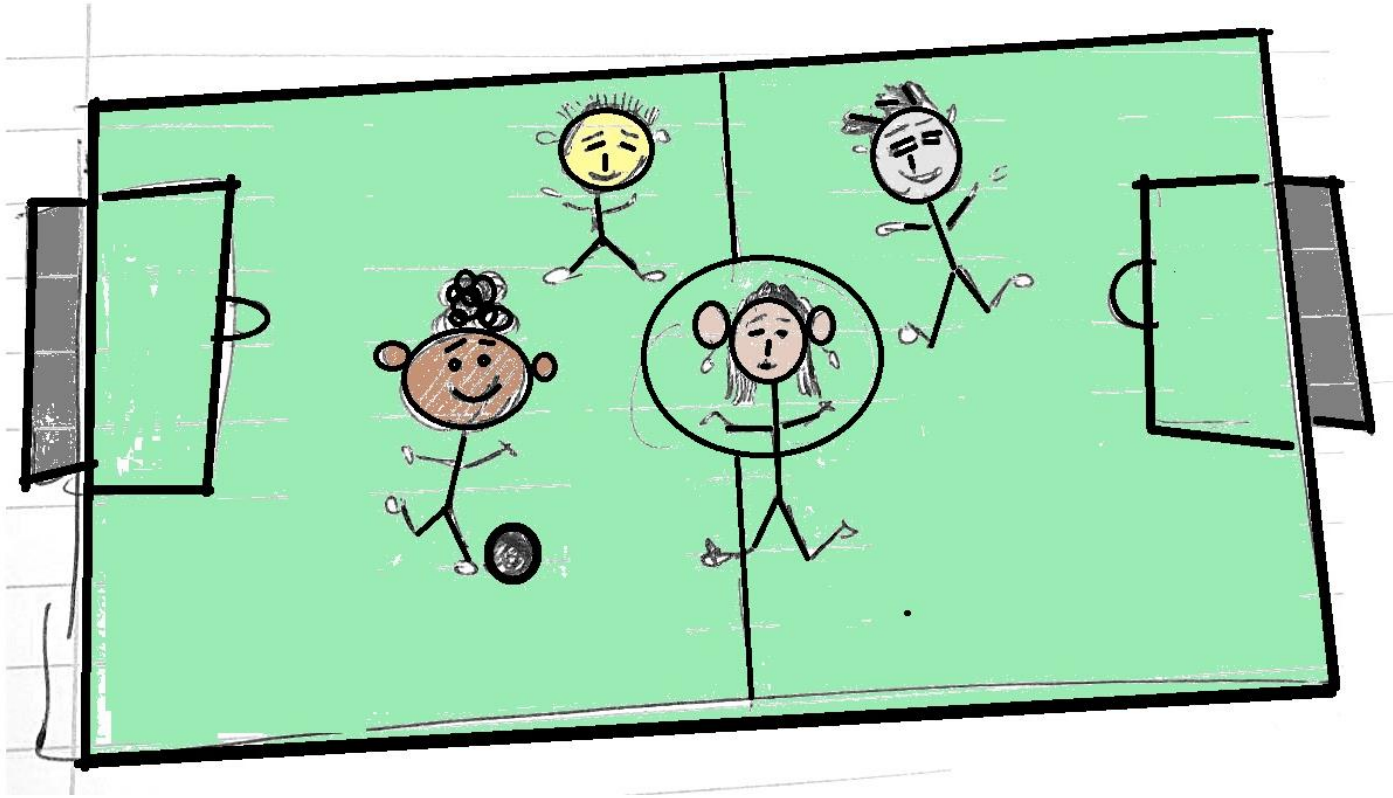


Transparency

Transparency: understandability of decisions/operations/plans of systems

- *Importance*: building trust, scrutinizing the issues, and minimizing harm
- *Issues*:
  - Accountability issues of the black-box systems, e.g., Large Language Models for medicine
  - Safety issues: no transparency leads to no improvement, leading to less safety
  - (What other issues?)
- *Implementation*:
  - Knowledge representation and reasoning using explicit causal networks, decision trees, etc.
  - Explainability techniques: Post-Hoc and real-time.
  - Documentation:
    - (ANNs) open training (and test) data, model types
    - (Symbolic AI) reasoning types

Justice and Fairness: Everyone's Game!



Justice and Fairness

- **Justice and fairness:** Promoting equality by avoiding bias and discrimination
    - Equality in use/access: algorithms must be open-source
    - Equality in training/opportunity: no one must be excluded from AI training (What else?)
    - Equality in AI-based judgment: decision-making based on specific factors, gender & race, is not allowed
  - *Importance*: technology must be accessible/useful to everyone
  - *Issues*:
    - conflict of interest between technology providers & users in the competitive AI market,
    - biases in the datasets and data-sensitive algorithms that use these datasets, leading to unequal impacts
  - *Implementation*:
    - following guidelines for data collection
    - reducing the reliance on data-dependent methods,
    - bias detection and correction algorithms (Any other ideas ?)

Safety

- **Safety:** Making sure systems do not cause harm to individuals (or society).
  - *Importance*: “do not harm” is more important than “doing good”,
  - *Issues*: privacy violations, physical harm, or psychological damage, examples
    - Misuse of private information via companies, e.g., facial recognition
    - physical harm due to malfunctioning or improper use, e.g., accidents in autonomous cars (What else?)
    - Stress and social anxiety due to interaction with AI systems, AI companion chatbots
  - *Implementation*:
    - using testing and monitoring techniques
    - anomaly detection
    - Safety guidelines for the use/creation of AI systems for specific domains

- **Question**? Can safety, justice & fairness, and transparency principles be conflicting?

Responsibility

- **Responsibility:** Accountability of actions and decisions
  - *Importance*: The need for ethical conduct by AI developers and users.
  - *Issues*: it is usually not clear who is responsible for AI mistakes. AI, designer, or user? (Why?)
  - *Implementation*:
    - ethical training regarding the use and development of AI
    - promoting a culture of integrity within AI development teams
    - proper documentation for the systems

Privacy

- **Privacy:** Protecting personal data
  - *Importance*: protecting individual rights.
  - *Issues*: It is challenging to find the balance between privacy and the need for large datasets for data-driven AI development
  - *Implementation*: technical methods like
    - data minimization techniques
    - privacy-by-design approaches

    And
    - data protection regulations
    - increased public awareness about the privacy rights

When we talk about ethics in AI, what type of AI do we really mean?

AI from a historical perspective

Idea: In 1950, Turing proposed the idea of building child machines, which are machines that can mimic human children's learning and reasoning behavior. Minsky (1952) built the first Artificial Neural Net (ANN), called SNARC. In 1959, Rosenblatt creates the Perceptron - which signaled the coming of sub-symbolic systems over half a century before contemporary ANNs. Their creation has been inspired by biological neural networks.

Birth of AI: In 1956, McCarthy and Minsky, propose ideas for building reasoning systems based on logic rules and symbolic representations. In the 1960s and 1970s, heavy emphasis was put on symbolic AI and rule-based systems such as chess.

Fathers of AI:

- McCarthy presents situation calculus, a formal framework for planning and reasoning for robotic agents.
- Minsky introduces the idea of a “society of mind” (1986) where intelligence is considered a phenomenon unifying multiple entities/processes that work together. Those entities process information at multiple layers of abstraction, realizing consciousness.

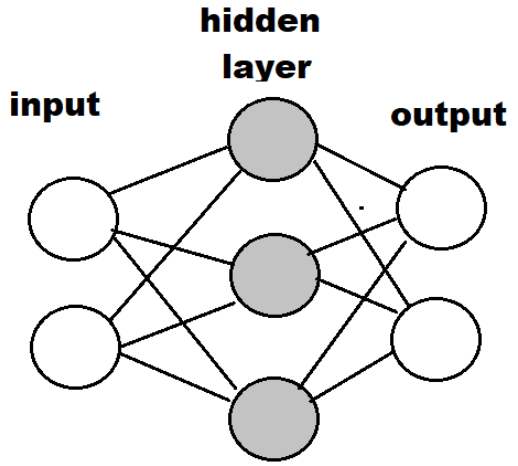
AI Winter: The 1970s and 1980s marked the AI winter, where the progress in AI research slowed down.

Machine learning boom: In the 1990s and 2000s, AI matures in the form of machine learning (ML) algorithms, which are data driven methods, as opposed to rule-based symbolic systems.

Contemporary AI: Practical Artificial Neural Nets

In the 2010s, Artificial Neural Networks (ANNs) requires enormous amounts of data and computing power. ANN led to extensive automation in the domains of

- Image recognition (object detection): inputs are image data.
- Natural language processing: text and audio data



In recent years, ANNs are used in building large language models (LLMs), e.g., GPT models.

- Training data : texts on the internet.
- How do ANNs get training:
  - Training: feed input data (examples) with known outputs, which tunes the weights (edges between nodes) until it predicts well.
  - Backpropagation: adjust the weights when a wrong prediction is made.

Ethical issues of ANNs

Transparency: ANNs designed for real-world application domains have a large number of inputs and outputs, hidden nodes, and layers, making them **black-box** predictors/classifiers. (Some examples of domains?)

Justice and fairness: ANNs are overly sensitive to training data. If the training data is biased, they amplify the biases through their architecture, making decisions/classifications/predictions that reflect horrifying discrimination (e.g., racial and gender biases in what domains?).

Safety and security:

- ANNs can not deal with adversarial attacks, where the input test data is designed in such a way that fools the ANN-based AI systems. (Some examples?)
- ANNs can be used for the creation of deepfakes, which can be used for unethical purposes.

Responsibility: A lack of transparency makes it difficult to understand how the system makes decisions. This allows some AI system designers to avoid the consequences of their systems' malfunctions.

Privacy: ANNs call for large amounts of training data, allowing the ANN developers to collect and gain access to a lot of people's private data.

Can ANNs become more ethical?

Explainable ANNs:

- **Using explainability techniques to** help users understand the decisions **only to some extent!**

Bias mitigation: collecting more training data, bias detection algorithms

- issues: 1) lack of training data in many domains
- 2) privacy issues with collecting more and more data

Safety improvement via adding the human-in-the-loop component to retrain models,

- Issue: high human labor costs

Responsibility: It will still be not clear who is responsible for the decisions made

Privacy: anonymizing the training data, which is challenging for large datasets

Some controversial large AI projects

- We will evaluate the potential ethical issues with the following AI projects
  - GPT-4 by OpenAI
  - AI Surveillance and Social Credit System
  - Tesla’s Full Self-Driving

Some large AI projects – GPT models

GPT-4 by OpenAI

- Generative Pre-trained Transformer (GPT4) is the largest LLM
- 45 gigabytes of training data (3X larger than GPT-3)
- 1.7 trillion parameters (10X larger than GPT-3)
- 8,192 tokens prompt length (2X larger than GPT-3)

- Except for some moderation filtering to avoid generating offensive or violent responses, little improvement has occurred in terms of ethics.

Ethical issues that still exist:

- Transparency: transparency is very limited, and it is not clear how some responses are generated.
- *Justice* and Fairness: the use of GPT may violate the rules of copyright.
- *Safety*: generating harmful content, misinformation, or being misused for unethical purposes.
- *(Discussion 5min) Responsibility*: creators, then users, then models?
- (Discussion *5min*) Privacy: It uses even larger datasets, both from the internet and users and issues about consent.

AI Surveillance and Social Credit System

AI Surveillance: Facial recognition via CCTV camera and collecting large data from social media and financial transactions to predict criminal activities and violent behaviors.
Technology: Convolutional neural nets and database integration of ID databases and social media profiles.
Social Credit System: A scoring mechanism that tracks financial credits and gives rewards or punishments in the form of access to loans, job opportunities, and government services.
(businesses use their own social credit system)
Technology used: Rule-based systems and machine learning.

- Transparency: Non-transparent, limited information about the systems' inner workings
- Justice and Fairness: There exist substantial concerns about fairness of the systems, as they may disproportionately target specific groups of people
- *(Discussion)* Safety: High potential for different types of social damages, e.g., oppression, loss of freedom, and social exclusion
- *(Discussion)* Responsibility: Not clear who is responsible for the high potential misuse of these systems
- *(Discussion)* Privacy: Significant privacy violations due to data collection on individuals without their consent.

Tesla's Full Self-Driving (FSD)

Takes (almost) full control of the car's navigation (minimum assistance from the driver). Its features are: autopilot (cruise control and lane keeping), auto lane changes, auto park, traffic light and stop sign auto control, and full self-driving beta (fully automatic city driving).

- Technology:
  - Sensors: cameras, radar (for measuring the distance of objects), GPS, and ultrasonic sensors (for close objects)
  - ANNs are used to map sensor data to car control reactions.

- Transparency: Limited transparency.
- *(Discussion)* Justice and Fairness: concerns about the impact on jobs (replacing human drivers with AI)
- *(Discussion)* Safety: accidents may occur due to wrong interpretation of data, which has caused several deaths so far.
- Responsibility: Tesla has been testing its self-driving beta on public roads; who is responsible for potential accidents?
- *(Discussion)* Privacy: Tesla cars collect large amount of data about driver and passengers, leading to privacy concerns.

What do you think?

- Write a one-page analysis about one or more of the large AI projects discussed in relation to the five ethical principles introduced in this session.
- Present your analysis in the class for 2 minutes and get feedback about your analysis from classmates at the end.

Thank you!

ANY QUESTIONS?