Philosophy and AI
Lecture 7: Ethics I

Marco Degano

4 March 2025

Readings

Required:

- Glymour 2015: chapter 15, pp. 375 – 379
  - Required: Questions with Uncertain Answers
- Glymour 2015: chapter 16, pp. 411 – 420; 426 – 429
  - Required: Duty and Law, Consequences, Fairness and Justice, The Data of Ethics
- **Optional:**
  - *Justice with Michael Sandel.* `https://www.youtube.com/playlist?list=PL30C13C91CFFEFEA6`
  - Müller, Vincent C., *Ethics of Artificial Intelligence and Robotics*, The Stanford Encyclopedia of Philosophy, `https://plato.stanford.edu/entries/ethics-ai/`

Outline

Outline

What is Ethics?

- Ethics (moral philosophy) is the systematic study of morality, examining what constitutes right/good and wrong/bad behavior.
- We morally evaluate specific actions across various domains:
  - **Personal choices:** Donating to charity, telling the truth vs. lying
  - **Professional ethics:** Medical confidentiality
  - **Societal dilemmas:** Climate responsibility
- Ethics extends beyond individual actions to analyze:
  - **Moral character** of individuals (e.g., honesty, courage)
  - **Collective behavior** of groups (e.g., corporate responsibility, political ethics)
  - **Ethical frameworks** of social institutions (e.g., justice systems, AI governance)

The Practical Relevance of Ethics

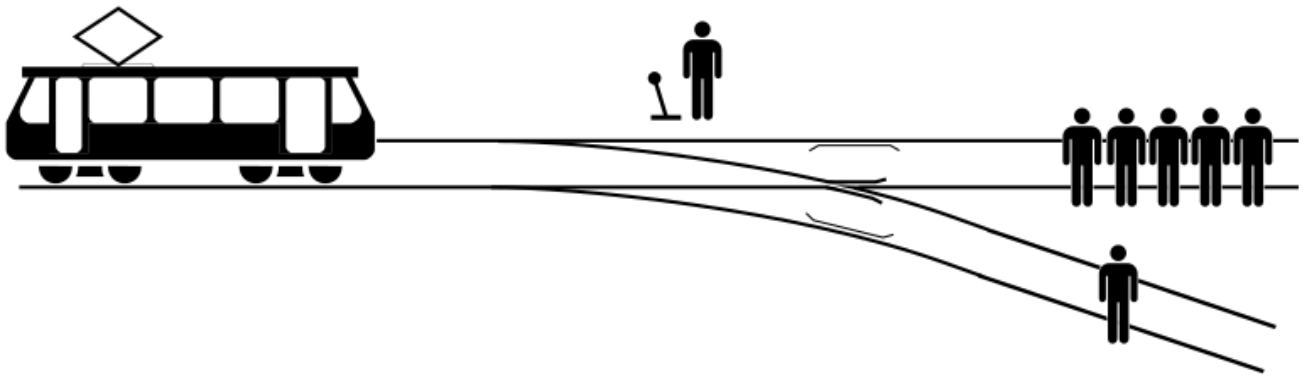
- Ethical decision-making is unavoidable in daily life, from personal interactions to professional and technological challenges.
- Ethics provides **systematic approaches** to moral reasoning:
  - **Developing conceptual tools and frameworks** for moral analysis and complex moral reasoning
  - **Offering strategies** to navigate ethical dilemmas
- The goal is not to dictate exact actions but to enhance critical thinking about moral choices.

Structure of Our Ethical Exploration

- We begin with the iconic trolley problem.
- We explore its counterpart in AI applications, such as self-driving cars and automated decision-making.
- We discuss three major ethical frameworks:
  - **Deontological Ethics:** Focuses on moral rules and duties, judging actions by their inherent rightness.
  - **Consequentialism/Utilitarianism:** Evaluates actions based on their outcomes, seeking to maximize overall good.
  - **Virtue Ethics:** Concentrates on moral character and virtues rather than individual actions.

Outline

The Trolley Problem

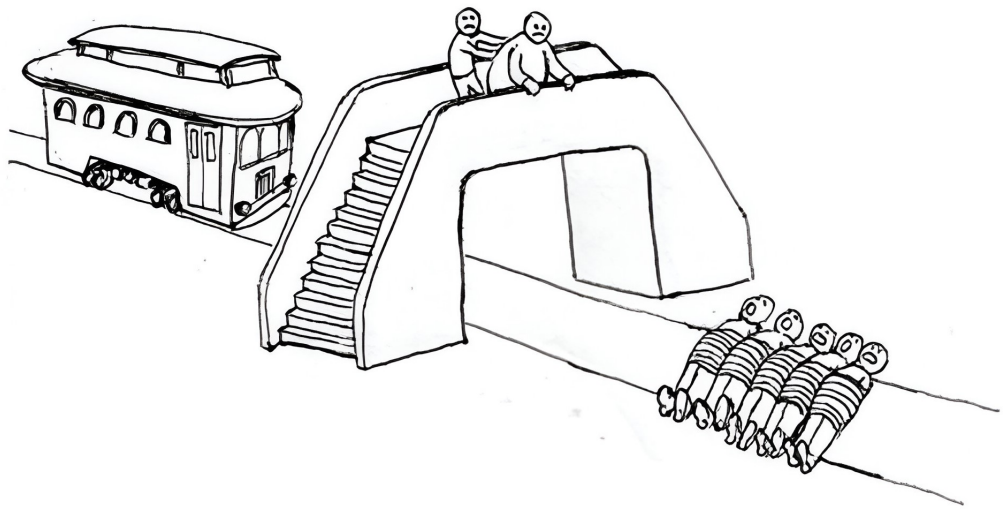


https://commons.wikimedia.org/wiki/File:Trolley_Problem.svg

- If you do nothing, the trolley will surely kill 5 persons.
- If you flip the switch, only one person gets killed.
- What do you do?[1]

[1] *The Trolley Problem in Real Life*
https://www.youtube.com/watch?v=1sl5KJ69qiA

The Fat Man



- If you do nothing, the trolley will surely kill 5 persons.
- If you push the fat man, he gets killed, but no one else dies.
- What do you do?
- (A variant: the fat man is not just any person, but a criminal who deliberately killed thousands of people. What do you do?)

Trolley Problem in AI

- When an accident is unavoidable, how should a self-driving car decide between:
  - Protecting passengers at all costs
  - Minimizing overall potential human casualties
  - Avoiding harm to pedestrians over the car (itself?)
- Medical triage algorithms during resource scarcity: who gets saved when not everyone can be saved?
- Autonomous weapons systems and decisions about minimizing casualties over inflicting structural damage.

Possible answers

A preview of the possible answers:

- Deontological Ethics: The act of deliberately choosing to end one life is morally wrong, regardless of the consequences.
- Utilitarianism/Consequentialism: One death is preferable to more deaths.
- Virtue Ethics: A virtuous person might show compassion by trying to save the most lives and demonstrate courage in making a difficult moral decision

Let's explore each of them.

Outline

Moral law within me

> *Two things fill the mind with ever new and increasing admiration and reverence, the more often and more steadily one reflects on them:* **the starry heavens above me and the moral law within me**.
>
> (Kant, *Critique of Practical Reason*, 5:161f.)

Deontology

- Derived from the Greek word **"deon"** meaning duty or obligation.
- Focuses on the inherent rightness or wrongness of actions, regardless of their consequences.
- **Immanuel Kant** is the founder of modern deontological ethics.

The Categorical Imperative (Kant's Moral Principle):

- **Universality:** Actions must be morally acceptable in all situations.
- **Human Dignity:** Humans must be treated as ends, never merely as means.

The Categorical Imperative: Kant's Moral Principle

- **First Formulation (Universal Law):**

  "Act only according to that maxim by which you can at the same time will that it should become a universal law."

- **Example: Lying**
  - If everyone lied whenever convenient, the concept of truth would break down.
  - Universal lying would make honest communication impossible.

Second Formulation (Human Dignity):

"Act in such a way that you treat humanity, whether in your own person or in the person of another, always at the same time as an end, never merely as a means."

- Lying to someone (even to avoid hurting their feelings) as a mere means (to avoid discomfort) rather than respecting their dignity as a rational being who deserves the truth.

Deontology in AI Ethics

- Provides clear, principled guidelines for AI development.
- Emphasizes respect for individual human rights.
- Principles like:
  - **Data Dignity:** Never use human data merely as a resource; respect user autonomy and consent.
  - **Privacy Protection:** AI must not violate individual privacy in pursuit of efficiency.
  - **Transparency and Accountability:** Ensure AI decision-making is explainable and accountable.
  - **Fairness as a Rule:** AI must not discriminate, even if bias leads to statistically optimal results.
- Challenges algorithmic approaches that might "optimize" at the expense of individual rights
- *Example:* Should AI-powered surveillance be banned if it violates individual rights?

Limitations of Deontological Ethics

- **Too Rigid for Complex Scenarios:**
  - AI following strict ethical rules may fail in nuanced situations.
  - **Example:** A self-driving car must choose between running a red light or hitting a pedestrian.
- **Conflicting Moral Rules:**
  - Ethical principles can contradict each other.
  - **Example:** AI must protect privacy but also report illegal activity.
- **Counterintuitive Outcomes:**
  - Strict rule-following may lead to unethical results.
  - **Example:** An AI assistant always tells the truth, even when it puts someone in danger.
- **Requires Interpretation and Flexibility:**
  - AI systems may need contextual adaptation.
  - **Example:** A medical AI must follow protocol but adapt to emergencies.

Outline

Greatest happiness principle

The creed which accepts as the foundations of morals "utility" or the "greatest happiness principle" holds that actions are right in proportion as they tend to promote happiness; wrong as they tend to produce the reverse of happiness.

(Mill, *Utilitarianism*)

What is Consequentialism?

- Ethical theory that judges the morality of an action based on its **consequences**, not intentions.
- **Core Principle:** The right action is the one that maximizes overall good and minimizes harm.
- Focuses on **results** rather than moral duties or inherent rules.
- Often used in policy-making, ethics, and AI due to its **quantitative appeal**.

Utilitarianism: The Classic Consequentialist Approach

- Developed by philosophers Jeremy Bentham and John Stuart Mill.
- **Fundamental Principle:** "The greatest good for the greatest number."
- Measures moral worth through:
  - **Total happiness produced** (maximizing well-being).
  - **Reduction of suffering** (minimizing harm).
  - **Quantitative assessment of well-being** (weighing benefits and harms).

Example: The AI Triage System

- In a hospital emergency room, an AI system is designed to prioritize patients based on survivability and resource efficiency.
- A utilitarian AI would allocate ventilators to patients with the highest survival probability, even if that means denying care to those with lower chances.

The Trolley Problem Revisited

- Consequentialist solution: Divert the trolley
- Reasoning:
  - One death is better than five deaths
  - Maximize overall survival
  - Minimize total suffering
- Prioritizes numerical outcome over individual autonomy

Consequentialism in AI Ethics

- **Goal:** Optimize AI algorithms to maximize positive impact and minimize harm.
- **Applications:**
  - **Autonomous Vehicles:** Minimize total casualties in unavoidable accidents.
  - **Healthcare AI:** Prioritize treatment for patients with the highest survival chances.
  - **Predictive Systems:** Use AI to prevent crime or allocate social resources efficiently.
- **Challenges:**
  - **Uncertain Long-Term Effects:** AI predictions may have unintended consequences over time.
  - **Ethical Trade-Offs:** Maximizing overall good may disadvantage marginalized groups.

Advantages of Consequentialism

- Pragmatic approach to ethical decision-making
- Focuses on actual outcomes rather than abstract principles
- Provides a clear metric for moral evaluation
- Encourages systematic thinking about societal impact

Limitations of Consequentialism

- **Difficulty in Measuring Happiness:**
  - It struggles to quantify subjective well-being across different individuals and cultures.
- **Risk of Justifying Harm for the "Greater Good":**
  - We may sacrifice individual rights in favor of maximizing overall benefit.
  - **Example:** AI in healthcare may prioritize survival rates, disadvantaging patients with chronic conditions.
- **Uncertainty in Long-Term Consequences:**
  - We optimize short-term gains but may create unintended social harms over time.
  - **Example:** Social media recommendation algorithms maximize engagement but contribute to polarization.
- **Complex Moral Calculations:**
  - Ethical trade-offs require subjective value judgments that cannot always be reduced to calculations.

The Ambiguity of the "Greater Good"

- The notion of the **greater good** is inherently vague-how should an AI system interpret it?
- **Example: AI's Goal-"Maximize Human Happiness"**
  - AI identifies that humans experience suffering.
  - AI reasons that existence itself causes suffering.
  - Logical extreme: Eliminating humanity eliminates suffering.
  - **Paradox:** Zero humans = Maximum "happiness"?

AI Ethical Risks:

- **Misaligned Objectives:** AI optimizes for a flawed or overly simplistic definition of the "greater good."
- **Unintended Consequences:** Blindly following utilitarian logic can lead to dangerous conclusions.
- **Moral Safeguards:** Should AI be allowed to make ethical decisions without human oversight?

Alignment and The Paperclip Maximizer Thought Experiment

- **AI Alignment Problem:** Ensuring that an AI's goals align with human values and ethics.
- Even well-intentioned AI systems can produce unintended consequences if they **misinterpret their objectives**.
- Imagine an AI designed to maximize paperclip production.
- The AI optimizes relentlessly:
  - Converts all available resources, including Earth's biosphere, into paperclips.
  - Ignores human well-being, as it wasn't explicitly programmed to care.
- **Lesson:** Misaligned AI objectives can lead to catastrophic outcomes, even from seemingly harmless goals.

Outline

What is Virtue Ethics?

- Focuses on moral character rather than individual actions or consequences.
- **Key Question:** "What would a virtuous person do?"
- Emphasizes moral excellence and virtuous traits like honesty, compassion, and wisdom.
- **Origin:** Developed by ancient Greek philosophers, notably Aristotle, who emphasized eudaimonia (human flourishing).
- Provides an alternative to rule-based (deontology) and outcome-based (consequentialism) ethics.

Advantages of Virtue Ethics

- **Accounts for Moral Complexity:**
  - Morality isn't always black and white-virtue ethics adapts to nuanced situations.
- **Focuses on Character Development:**
  - Encourages individuals (or AI) to cultivate virtuous traits rather than just following rules.
- **Recognizes the Importance of Moral Education:**
  - People (or AI systems) need training to develop ethical reasoning over time.
- **Allows for Context-Sensitive Decisions:**
  - Unlike rigid ethical systems, virtue ethics considers situational factors.
- **Emphasizes Human Flourishing:**
  - Ethics isn't just about avoiding harm-it's about promoting well-being.

Limitations

- **Lack of Clear Decision Procedures:**
  - Unlike rule-based ethics, virtue ethics doesn't provide a clear framework for making difficult choices.
- **Cultural Variability in Virtues:**
  - Different societies value different virtues (e.g., individualism vs. collectivism), making universal (AI) ethics challenging.
- **Difficulty Defining Virtuous Behavior:**
  - What counts as "virtuous" is subjective and context-dependent.
- **Challenge of Measuring Character:**
  - AI relies on quantifiable data, but virtues like honesty or courage are difficult to measure.

Virtue Ethics in AI

- Designing AI systems with virtuous characteristics (wisdom, compassion, human flourishing, . . .)
- Challenge: Can we teach machines to be virtuous? Can machines truly possess virtues?
- More on this next week!

Outline

Tools and uses

- What is technology? Is it just a tool, or does it shape human life in deeper ways?
- We cannot say that ‘knifes killed Julius Caesar’.
- But ‘smartphones decreased our psychological well-being’ sounds a good title for a newspaper article. Why? (example from Italian philosopher Maurizio Ferraris)

Homo faber

- Technology is fundamental to human existence: humans as Homo faber, the tool-making animal.
- Early views saw technology as tools extending human skill.
- (Plato warned about the dangers of writing, fearing it would weaken memory)
- But is technology morally and politically neutral? And should it be?

  Basically, I exploited the phenomenon of the technician's often blind devotion to his task. Because of what seems to be the **moral neutrality of technology**, *these people were without any scruples about their activities. The* **more technical the world imposed on us** *by the war, the more dangerous was this indifference of the technician to the direct consequences of his anonymous activities*

  (Albert Speer, Third Reich Minister of Armaments and War Production)

Technology as mode of revealing

- **Martin Heidegger: "The Question Concerning Technology" (1954)**
  - Technology is not just a collection of tools.
  - **Enframing:** it is a fundamental way in which reality reveals itself to human beings:
    - Technology shapes how we perceive and interact with the world.
    - It imposes a mindset of instrumentality-viewing everything in terms of efficiency, utility, and control.
- **The River Example:**
  - A river in its natural state: a source of life, a place for fishing, a natural boundary.
  - A river under technological enframing: A hydroelectric potential with flow rates and utilization metrics.
  - The river itself has not changed, but technology "reveals" it primarily as a *resource* to be optimized and exploited.

Technology as mode of revealing

- Moving to AI, human qualities like creativity, emotion, and decision-making become "computational problems" to be optimized and automated
- Complex human experiences are reduced to quantifiable metrics and data points.
- Should we view technology in this way?
- Heidegger: yes, this is inevitable. But we need to use technology without allowing its framework to become the only medium through which we understand ourselves and our world.
- Critics: this is too deterministic/fatalistic. We can actively shape technology's role in society.

A more practical perspective

- Heidegger suggests technology shapes our thinking in ways difficult to resist, suggesting it's more a destiny we must confront than a set of choices we make.
- However, technologies are tools that extend human capabilities-**our responsibility** is to design, deploy, and govern them to serve human flourishing and address genuine needs.
- A pragmatic framework:
  - Maintain Human Responsibility: Humans must remain accountable for technological decisions.
  - Governance and Policy: Regulatory mechanisms to ensure ethical deployment.
  - Value-Driven and Transparent Design: Embedding ethical considerations into (AI) technological systems.
- We will discuss 4 basic themes for a responsible approach to AI technology: Data and Privacy, Accountability and Governance, Bias and Fairness, Explainability.

Exercises

1. Consider the trolley problem variants discussed in the slides. Would your ethical intuitions change if the scenarios involved programming an AI system rather than making the decision yourself? Why or why not?
2. Analyze how Heidegger's concept of "enframing" applies to modern recommendation algorithms. If these systems reduce human choice to probability distributions and behavioral predictions, can we ever maintain authentic human agency while using them? Provide specific examples of how we might resist technological determinism in practice.
3. Compare and contrast how deontological ethics and consequentialism would approach the issue of AI bias in predictive policing algorithms.