Ethical Frameworks to Navigate Dilemmas in Data-Driven Technologies

Lecture 12

Introduction

We will explore real-world scenarios that require the application of different ethical frameworks

Understanding which frameworks are suitable for specific situations is essential

Failure to select appropriate framework can lead to negative consequences

Different frameworks serve different individuals/organizations

What Actions to Take in Emergency and Disaster Scenarios?

United Nations

Universal Guidelines for AI (UGAI)

Emphasizes the use of AI to enhance public safety, holding organizations accountable for their actions, and includes a termination clause for when human control is no longer possible.

The Toronto Declaration

Toronto Declaration (TD)

Centers on non-discrimination and international human rights principles, with universal applicability, but lacks robust guidance on privacy concerns.

UGAI focuses more on public safety and accountability, while the TD centers on human rights principles.

What Actions to Take in Emergency and Disaster Scenarios?

- The use of AI in emergency and disaster scenarios can raise ethical concerns related to **privacy, autonomy, and human rights.**

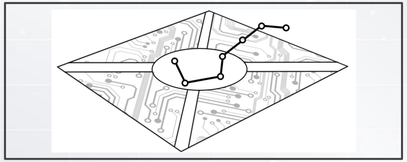
- The **Universal Guidelines for AI (UGAI)** and the **Toronto Declaration (TD)** are ethical frameworks that can provide guidance in decision-making.

- Both frameworks can be valuable tools for **decision-makers, providing guidance** in using AI in an **ethical and responsible manner** in emergency and disaster scenarios.

What if a Biased System Is Still Better than Humans?

Montreal Declaration for responsible AI (MDR AI)

Provides guidance through its principles of prudence, human autonomy, and responsibility. These ensure that negative consequences are anticipated, human control is present, and decisions align with organizational values.



Beijing AI Principles

Emphasize informed consent and educational training, ensuring patients understand the risks of automated analysis and doctors have knowledge to answer questions about limitations and potential failures.

What if a Biased System Is Still Better than Humans?

- AI in medical imaging scans can provide high accuracy in detecting diseases and markers, but ethical concerns arise from their inability to **explain decisions made**.

- Both sets of principles have strengths and weaknesses, with the **MDR AI** providing a **comprehensive** approach, but may require **regulatory instruments**, and the **Beijing AI Principles** offering guidance on **specific issues** but may lack concreteness in **implementation**.

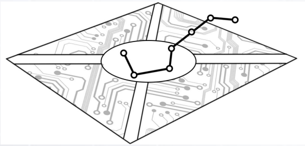
- It is crucial to ensure AI is used ethically and responsibly in medicine, **prioritizing patient autonomy, accountability, and transparency.**

Should AI Enable Self-Destructive Behavior?

United Nations

Universal Guidelines for AI (UGAI)

Prioritize data quality, prohibition of secret profiling, and public safety but lack concrete guidelines for practitioners to follow.



Beijing AI Principles

Emphasize informed consent and ethics by design to prevent addictive interfaces and reward systems, but lack specificity and consistency across products and services.

Should AI Enable Self-Destructive Behavior?

- AI can lead to **self-destructive behavior**, such as compulsive media consumption, due to design patterns like **endless scrolling** and **personalized content**.

- The ethical dilemma is **balancing profit motives** against providing **meaningful control** for users.

- The key is to leverage these principles in AI development to **prioritize the benefit of people** and promote **responsible use of AI**.

What Limits on Persuasive Technology Are Fair?

OECD.AI
Policy Observatory

OECD AI principles

Provide guidelines for responsible disclosure, transparency, privacy, and ensuring well-being

However, OECD principles lack demonstrated use cases for evaluation

European
Commission

Guidelines for trustworthy AI

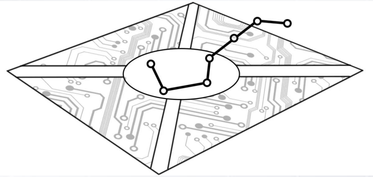
Provide granular guidance for respecting human autonomy, privacy, data governance, and well-being

But, they are written from the perspective of the European legal and regulatory ecosystem

What Limits on Persuasive Technology Are Fair?

- AI integrated in smart toys, can be used to **draw minors** into conversations and **persuade them** in the privacy of their homes.
- To develop beliefs and **behave in specific patterns** that might help an organization **meet their goals**.
- For example, convincing children that they should ask their parents to **buy particular brands** of toys. These technologies can be **highly persuasive**.

What If an AI System Is "Emancipated" from Human Oversight?



Beijing AI Principles

Beijing AI Principles highlight control risks and openness and sharing as particularly relevant in this scenario

Openness and sharing can provide an accountability mechanism

OECD.AI
Policy Observatory

OECD AI principles

OECD AI Principles advocate for appropriate human intervention when needed

Simpler models to limit the emancipation of the AI system from human control under the OECD AI Principles

What If an AI System Is "Emancipated" from Human Oversight?

- **Emancipation** of algorithmic systems from human oversight presents a **control risk**

- **Complexity of interactions** between the platform and a **large user base** might be further complicated with the use of opaque **black box AI systems**

- Frameworks like the **Beijing AI Principles** and **OECD AI Principles** provide guidance on ethical concerns in this scenario

- These frameworks still **lack granular advice** and **high-level nature** creates an additional risk of **ethical concerns** being raised and open to interpretation.

How to Protect Against AI-Driven Harassment?

The Toronto Declaration

Toronto Declaration (TD)

Based on international human rights, emphasizing the need for deeper consideration of the second-order effects of open publication, but lacks prescriptive guidance for alternate publishing models and incentives.

Asilomar AI principles (AAIP)

Provide normative guidance on research goals, funding, culture, and shared prosperity, addressing ethical concerns related to the pursuit of research and balancing the costs and benefits of downstream research.

How to Protect Against AI-Driven Harassment?

- **Deep fakes** demonstrate the **power of AI** in generating convincing fake audio and video segments with someone else's resemblance, leading to **potential harassment and defamation**. Example: link

- **TD and AAIP** highlights the importance of considering the **potential consequences** of AI systems, ensuring they align with i**nternational human rights**, and promoting **open research**.

Additional Resources