Lecture 3: DEI Principles in AI Ethics: Navigating the Challenges of Bias

Exploring the Intersection of Diversity, Equity, Inclusion, and Artificial Intelligence

Hongshan Guo

2024-09-10

Section 1

Recap from Last Week

Lectures: Sustainability (3x3) and Leadership

- People, Planet, Profit
- Identifying leadership behavior - Act like a leader?

Labs: ‘A’ and ‘B’

- Identify Apple Inc. Related Cases
- Sustainability and Leadership in Action: What do you see

Section 2

DEI Principles in AI Ethics: Navigating the Challenges of Bias

2: Introduction

- What is DEI?
  - Diversity: Variety in human differences
  - Equity: Fair treatment and access
  - Inclusion: Environment where all can thrive
- AI Ethics: Ensuring AI systems are fair and unbiased
- Today's Focus: How DEI principles apply to AI development and deployment
- Why are we having this session: DEI are being compressed into datasets, which then gets used to derive 'insights' for leaders to act upon.

2.1: Introduction (Continued)

- Directly impacts an organization's long-term viability and social responsibility.
- Sustainable leaders must ensure that AI systems to be:
  - fair,
  - unbiased, and
  - beneficial to all stakeholders.
- Ignoring DEI in AI can lead to reputational damage, legal risks, and erosion of trust, undermining sustainability efforts.
- From a triple-bottom line perspective, this relates to:
  - people (ensuring AI benefits all segments of society),
  - planet (using AI ethically to address environmental challenges), and
  - profit (developing trustworthy AI for long-term business sustainability).
- Key competency for leaders aiming to build resilient, ethical, and sustainable organizations in our increasingly AI-driven world.

3: The Problem of Built-in Biases

- Sources of AI Bias:
  1. Data Collection
  2. Algorithm Design
  3. Interpretation of Results
- Teaser: “Did you know? An AI system once classified a turtle as a rifle. Let’s explore why!”

3.1 Turtle as Rifle, a built-in bias example

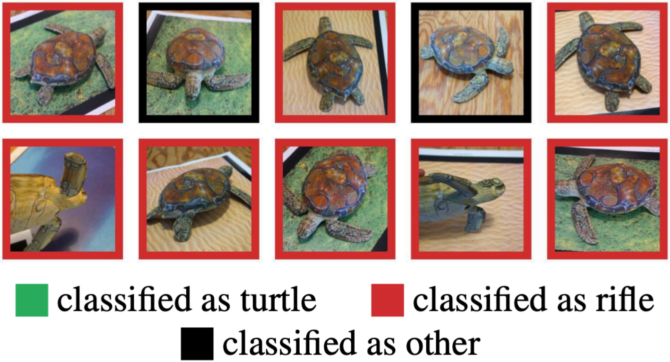

Figure 1. Randomly sampled poses of a 3D-printed turtle adversarially perturbed to classify as a rifle at every viewpoint[2]. An unperturbed model is classified correctly as a turtle nearly 100%

4: Interactive - Spot the Bias

- Instructions: Identify potential biases in these scenarios
    1. A speech recognition system struggling with accents
    2. An image dataset with 80% male CEOs
    3. A predictive policing algorithm focusing on certain neighborhoods
- Discussion: What biases did you spot? How might they impact real-world applications?

5: Case Study 1 - Twitter's Image Cropping Algorithm

- The Issue: Algorithm favored white faces over black faces
- Root Cause: Focus on high contrast areas and facial features
- Impact: Reinforced racial biases in social media representation
- Twitter's Response:
  - Made algorithm choices more transparent
  - Eventually removed automatic cropping

6: Case Study 2 - Amazon's AI Recruiting Tool

- The Problem: AI tool discriminated against women
- Key Issues:
  - Trained on 10 years of resumes, mostly from men
  - Penalized resumes including "women's"
  - Downgraded graduates of women's colleges
- Amazon's Action: Discontinued the tool
- Teaser: "Imagine applying for your dream job, but the AI says no. Ever heard of the 'this is an examplary candidate' trick?"

7: Leadership in DEI and AI Ethics

- Leaders' Roles:
  - Champion diversity in AI teams
  - Mandate thorough bias testing
  - Foster a culture of ethical AI development
- Why It Matters:
  - Business: Broader market appeal, avoid PR disasters
  - Ethical: Fair opportunities for all
  - Social: Build trust in AI technologies
- Reflection: "As a future leader, how will you promote DEI in tech?"

8: Scenarios - Possible

Here're some very straight forward examples of AI-driven solutions that you could develop without having good DEI awareness:

- 1 "The Resume Screener": AI favors certain universities
- 2 "The Loan Approver": Lower approval rates for specific postcodes
- 3 "The Employee Performance Predictor": Bias against flexible workers
- Discussion: Identify DEI breaches and propose solutions

9. Scenarios - Real

Now let's look at some real debacle of DEI failures by big corporates and highly-paid execs:

- Google Image Recognition Controversy (2015) : Tags
- Amazon's AI Hiring Tool Bias (2018): Data
- Microsoft's *Tay* Chatbot Incident (2019): Guardrails/24h
- Uber's Facial Recognition System Failure (2021): Facial Rec Color

10: Critical Thinking - The Complexities of “Fair AI”

Title: “Fair for Whom? Navigating the Complexities of DEI in AI”

- The Dilemma of Fairness:
  - What does “fair” mean in different contexts?
  - Can optimizing for one group inadvertently disadvantage others?
- Potential Trade-offs:
  - Accuracy vs. Inclusivity
  - Generalizability vs. Specificity
  - Speed of development vs. Thorough bias checking
- Discussion Points:
  - 1 “Is ‘fair AI’ truly fair for everyone? Who might be left out?”
  - 2 “How do we balance addressing historical inequities without creating new ones?”
  - 3 “What are the risks of over-correcting in AI design?”

10.1: Critical Thinking - The Complexities of “Fair AI” (Continued)

- Case Example:
  - An AI healthcare diagnostic tool is adjusted to be more accurate for underrepresented groups
  - Result: Slight decrease in accuracy for the majority group
  - Question: “Is this ethical? How do we decide?”
- Activity: “Ethical AI Design Spectrum” (if we have enough time)
  - Position yourselves on a spectrum in the room
  - One end: “Maximum DEI considerations in AI, even if it slows development”
  - Other end: “Rapid AI development with basic fairness checks”
  - Discuss reasons for their positions

11: Balancing Act - Doing Enough, But Not Too Much

- Key Considerations:
  1. Regulatory Compliance: Meeting legal requirements
  2. Ethical Imperatives: Going beyond the law to ensure fairness
  3. Practical Constraints: Time, resources, and technological limitations
- Finding the Balance:
  - Continuous testing and iteration
  - Diverse team input in all stages of development
  - Regular ethical audits and transparency reports
- Discussion Question: “How can we determine if we’ve done ‘enough’ to ensure fairness in AI systems?”
- Thought Experiment: “The AI Fairness Meter”
  - If you could design a tool to measure AI fairness, what metrics would you include?
  - How would you weigh different aspects of fairness against each other?

12: Conclusion and Next Steps

- Key Takeaways:
  - AI bias is often subtle but impactful
  - Proactive DEI integration is crucial in AI development
  - Ongoing vigilance and testing are necessary
- Your Action Item: “One thing I’ll do to promote DEI in tech...”
- Further Reading:
  1. Open for a surprise: The endearing results of Twitter’s new image crop, VOX
  2. Student proves Twitter algorithm ‘bias’ toward lighter, slimmer, younger faces, The Guardian
  3. NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles, arxiv.org