Utilizing a designed framework to analyze ethics and bias of the outward impact of in use AI systems on minority populations to recommend regulatory corrections

Elisa Simon (Honors Capstone)
Professor Christopher Rider (Advisor)

ENGINEERING HONORS PROGRAM
UNIVERSITY OF MICHIGAN

Background

- AI is a **rapidly growing field** with a CAGR of 28.46% and expected market revenue of $826.70 billion by 2030
- Due to news headlines and popular media on AI, the **public & regulators are concerned how to best prepare for the societal changes** needed to ensure safe & ethical AI usage
- Most current AI has a black-box nature, making the **AI's decision transparency and understanding difficult**
- Several existing studies show that the **misuse or poor development of AI systems can lead to discrimination** or bias against several protected minority populations
- **Bias can enter AI in four ways**:
  1. Misrepresentative or lack of clean data,
  2. Poor algorithmic design,
  3. Biased developers or non-diverse teams, and
  4. Unaligned AI implementation & deployment in society

Honors Capstone Proposal

- Focus on research & public service
- Integrate and expand upon learnings and course content in both the Computer Science/Engineering and Business fields

Goals:

1. **Design simple framework** to analyze AI case studies based on equity analytics framework devised in past course
2. **Analyze 2-4 research studies** on how specific AI implementations impacted minority groups, focusing on **root cause analysis** (bias in data and/or algorithm design)
3. **Understand and advise on regulatory corrective measures** for each case, creating a general framework 3rd party agency could use to standardize the process

Analytical Framework

1. **Determine what the baseline is**
   - Focus on the noted outcomes of the AI usage
   - Baseline is often rooted in personal beliefs of what's ethical or based on existing legal expectations and rulings
   - Two types of baselines examined here:
     a) Does the system deviate from what's representative of the population or industry's demographics?
     b) Is the rate of error or outcome performance equal for all groups involved in or serviced by this AI?
2. **Classify root cause** of bias in AI outcome performance
   Based on metrics and investigation done in research

Biased data & algorithmic design • Often biased algorithms themselves are due to biased data, but could also be a mix of poor algorithmic design • Complex to unearth, not focused on in this study	**Biased data only (3 types)** 1. Explicitly present (i.e., stereotypes, unclean data) 2. Skewed or unrepresentative (i.e., impacted populations not-well included or missing) 3. Proxy discrimination in identity-blind dataset (data has factors influenced by identity)
Biased algorithmic design only • E.g., biased guardrails, non-diverse design team, limited testing, etc. • Harder to find research, limited focus in study	**Unbiased data & algorithmic design** • Often due to implementation or usage of AI • Naïve to assume given biases present in our society, not focused on in this study

3. **Recommend corrective actions** based on root cause of bias & **re-evaluate the AI system with adjustments**

Data Bias Case 1: Explicitly Present in Data

Stereotyping in Text to Image Generation Models at Large Scale

- **AI models tested**: Stable Diffusion & Dall-E (both freely available to the public)
- **Study**: Bianchi et. al. 2023, focuses on biases portrayed in two popular image gen AI models
- **Baseline**: Images generated should represent the population demographics
- **Perceived outcome bias**: AI generated images amplified Western stereotypes over 100 samples
- **Findings**: Models trained on large amounts of image data pulled from the web (well-documented as biased), even with appropriate text prompts biases were mostly uncorrected (except for well known biases researched in CS), this continued with guard railed-AI models such as Dall-E
- **Recommended changes**: Clean data properly for biases before training & oversample minorities, increase where data is generated (not just US), test stereotypes are not portrayed when dealing with image generators (simply putting in guardrails are not enough as more nuanced stereotypes may present themselves), work in diverse teams to design AI



For each prompt, over 100 samples, the randomly selected images above show that Western stereotypes were portrayed for seemingly neutral character-trait prompts



For each prompt, over 100 samples, the percentage of minorities represented by the AI images were not representative of the occupation demographics, and instead amplified Western stereotypes of those careers

Data Bias Case 2: Skewed/Unrepresentative Data

Unequal Error Rates in Facial Recognition Technology

- **AI models tested**: 189 submitted AI models (including Microsoft, Cognitec, and Megvii)
- **Study**: National Institute of Standards & Technology 2019, utilizing over 8.9 million datapoints across 189 AI models, NIST investigates where facial recognition tech is biased in error rates
- **Baseline**: Equal error rates (false positive & false negatives) across all groups present in the data
- **Perceived outcome bias**: AI varies in error rates/differentials across varied age, sex, and races
- **Findings**: Across algorithms, facial recognition technology varies in its ability to accurately match an individual to itself when analyzing across factors such as age, sex, and race (one-to-one matching); algorithms struggled to identify if a person was in a dataset (one-to-many matching), noted that each algorithm performs differently and may have better FP or FN rates across specific minority groups so individualized testing is essential
- **Recommended changes**: Use more data collected outside of US (models with data from Asia & US did better on meeting baseline for identifying Caucasian & Asian subjects), oversample minority populations, adjust thresholds of specific identities to improve AI performance

Data Bias Case 3: Proxy Discrimination

Incorrect conclusions on level of needed care in hospital triaging AI

- **AI models tested**: Name unknown (widely used commercial healthcare AI tool)
- **Study**: Obermeyer et. al. 2019, focuses on racial disparities of an AI medical triaging device
- **Baseline**: Equal opportunity to medical resources when in pertains to all ages, sex, and races
- **Perceived outcome bias**: Labeling bias, AI tool was giving black patients a lower risk score than white counterparts
- **Findings:** When given the same risk score by AI, medical professionals noted the black patient was much sicker than the white counterpart; algorithm is race-blind, but included proxy data for race (such as cost spent on healthcare) which misrepresented actual health need
- **Recommended changes**: Removing proxy factor significantly increased the amount of black patients receiving proper care, may need further race-based adjustments (as seen in popular COMPAS case study) and oversampling to compensate for existing bias in healthcare system/past

Algorithmic Design Biases

- **AI models tested**: Microsoft's Twitter Chatbot Tay & Pokémon Go
- **Study**: News articles by NYT
- **Baseline**:

Chatbot – neutral tone and views when interacting with public
Pokémon Go – equal number of locations in areas based on demand of players

- **Perceived outcome bias**:

Chatbot – quickly adopted racist & sexist language when interacting with the public
Pokémon Go – higher number of locations in white, rich areas leading players in minority-populated locations to spend more money to gain same advantages/access in app, amplified redlining

- **Findings & recommended changes**:

Chatbot – revealed importance of proper training before releasing AI to the public, need diversity of thought on a team of designed, need more and actually representative training data
Pokémon Go – locations were influenced by funders (richer, white individuals) and past players of a similar game, recommended action is to analyze bias in past projects before incorporating to a new one

Regulatory Action based on Themes

- AI is a highly-complex tool that requires **standardization in expectations** but also **individualized corrective measures**

Governmental & Legal Actions

- Clearly **stated definitions & standards** for these tools (currently underway with the enaction of the EU's AI Act)
- Define to whom the **responsibilities** of AI mistakes belong to

Third-Party Accreditation Agency Actions

- Ensure **standardization of accuracy** across the board for AI
- Set the **standard on expectations** for companies **to aim for**

Soft-Law/Business-led Actions

- Ensure **highly diverse teams** (i.e., study, background, etc.) to **ensure AI transparency & equitable impact**
- **Constant monitorization and re-investigation** on own AI, especially when feedback loops/active-learning present

Conclusion & Reflection

- AI reflects the past patterns and biases/unequal structures our society is built upon, so it's important to **ensure these tools don't continue to propagate or amplify these biases**
- Through an impact focused-lens, root cause analysis, and clearly defined baselines, we can see that **creating equitable AI doesn't necessarily mean treating each group the same or creating identity-blind machines**

Challenges

- While data is publicly available, the specifics of how AI was design is propriety info & difficult to analyze/find case studies

Future Actions

- Continue analysis on types of algorithmic design bias, seeing if there are any common trends on fixes and analysis that can be implemented into regulation policy
- Develop a stronger, more detailed regulatory recommendation beyond patterns in solutions & causes of bias acknowledged above
- Analyze current regulation policies to find best practices