Introduction to HICSS-57 Minitrack on AI and Digital Discrimination

Sara Moussawi
Carnegie Mellon University
smoussaw@andrew.cmu.edu

Xuefei (Nancy) Deng
California State University
Dominguez Hills
ndeng@csudh.edu

Jason Kuruzovich
Rensselaer Polytechnic Institute
kuruzj@rpi.edu

1. Introduction

A technology is biased if it unfairly or systematically discriminates against certain individuals or groups by denying them an opportunity or assigning them a different and undesirable outcome [1]. As we delegate more and more decision-making tasks to autonomous systems and algorithms, such as using artificial intelligence (AI) for employee hiring and loan approval, digital discrimination is becoming a serious problem. In her *New York Times* best-seller book "Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy," Cathy O'Neil provides examples presenting instances where algorithms have perpetuated biases, often rooted not in the models themselves but in the data representing the human judgments they are built upon.

However, the digitization of business processes traditionally reliant on human judgment also presents an opportunity to enhance objectivity and equity. By leveraging transparent, well-audited algorithms, businesses can identify and mitigate inherent human biases, turning a potential threat into an instrument of fairness. This underscores the necessity for rigorous, ongoing research aimed at understanding and rectifying bias in technology.

According to Cambridge Dictionaries Online, *Discrimination* is defined as treating a person or particular group of people differently, especially in a worse way from the way in which you treat other people, because of their race, gender, sexuality, etc. [3]. *Digital discrimination* refers to discrimination between individuals or social groups due to lack of access to Internet-based resources or in relation to biased practices in data mining and inherited prejudices in a decision-making context [4]. It is a form of discrimination where users are treated unfairly, unethically, or just differently based on their personal data such as income, education, gender, age, ethnicity, religion, or even political affiliation during the process of automated decision-making [5].

AI-powered decision-making can cause discriminatory harm to vulnerable groups. In a decision-making context, digital discrimination can emerge from inherited prejudices of prior decision-makers, designers, engineers, or reflect widespread societal biases [6]. One approach to address digital discrimination is to increase the transparency of AI systems, and research has called for collaborations with disadvantaged groups whose viewpoints may lead to new insights into fairness and discrimination [7].

Potential ethical concerns also rise in the use of AI that builds on Large Language Models (LLM) such as ChatGPT, the virtual AI chatbot that debuted in November 2022 by the startup OpenAI and reached 100 million monthly active users just two months after its launch. Professor Christian Terwiesch at Wharton found that ChatGPT would pass a final exam in a typical Wharton MBA core curriculum class [8], which sparked a national conversation about ethical implications of using AI in education. While some educators and academics have sounded the alarm over the potential abuse of ChatGPT for cheating and plagiarism, industry practitioners from the legal industry to the travel industry are experimenting with ChatGPT and debating on the impact of the AI on the business and future of the work [9]. In essence, a Large Language Model is a deep learning algorithm that trains on large volumes of text. The bias inherited in the data can lead to emerging instances of digital discrimination especially as various LLM-based models are trained on data from different modalities (e.g., images, videos, etc.). Furthermore, the lack of oversight and regulations over data collection and model training and use can also prove to be problematic and unethical in some cases. Given the rapid developments and penetration of AI chatbots, it is important for us to investigate the boundaries between ethical and unethical use of AI, as well as potential digital discrimination in the use of LLM applications.

URI: https://hdl.handle.net/10125/107204
978-0-9981331-7-1

Addressing the problem of digital discrimination in AI requires a cross-disciplinary effort. For example, researchers have outlined social, legal, and ethical perspectives of digital discrimination in AI [10]. In particular, prior research has called for our attention to research the three key aspects: (1) how discrimination arises in AI systems; (2) how design in AI systems can mitigate such discrimination; and (3) whether our existing laws are adequate to address discrimination in AI [11].

2. Scope

This minitrack focuses on understanding and addressing the discrimination problems arising in the design, deployment, and use of artificial intelligent systems. It welcomes a diverse set of methodologies including empirical studies, design research, theoretical frameworks, case studies, etc., from scholars across various disciplines such as information systems, computer science, library science, sociology, and law, etc. Topics presented in this minitrack include, but are not limited to:

- AI-based Assistants: Opportunities and Threats
- AI Explainability and Digital Discrimination
- AI Systems Design and Digital Discrimination
- AI Use Experience of Disadvantaged / Marginalized Groups
- Biases in AI Development and Use
- ChatGPT and Ethical Use
- Digital Discrimination in Online Marketplaces
- Digital Discrimination and the Sharing Economy
- Digital Discrimination with Various AI Systems (LLM based AI, AI assistants, etc.)
- Effects of Digital Discrimination in AI Contexts
- Ethical Use/ Challenges/ Considerations and Applications of AI systems
- Literacy of AI Users
- Responsible AI practices to Minimize Digital Discrimination
- Responsible AI Use Guideline and Policy
- Societal Values and Needs in AI Development and Use
- Sensitive Data and AI Algorithms
- Social Perspective of Digital Discrimination
- Trusted AI Applications and Digital Discrimination
- User Experience and Digital Discrimination

3. Summary of Articles

This mini-track presents eight papers in HICSS-57. We introduce them briefly below.

Three papers explore the ethical issues associated with AI. First, in their paper titled: "What Is Ethical AI? – Design Guidelines and Principles in the Light of Different Regions, Countries, and Cultures", Lier et al. [12] derive key topics, design requirements, and design principles for ethical AI. The authors adopt cultural dimensions, text mining and topic modeling analysis to examine the issue of ethical AI in different regions, countries, and cultures. Second, focusing on developing ethical AI systems, Han et al. [13] propose an actionable ethics-aware guideline for AI developers consisting of five recommendations to help developers construct high-quality AI systems. The work also shows how to implement the guideline by using AI predictive models constructed on a national big data set that estimates children's risk of experiencing abuse and neglect in the United States. Third, Sengupta et al. [14] generate useful insights by examining public perceptions and community discourse related to AI ethics on the Reddit platform using a multi-methodological and multi-level approach.

Investigating the extent to which algorithms replicate and amplify data and human biases was the focus of two studies. In their paper titled "Does a Fair Model Produce Fair Explanations? Relating Distributive and Procedural Fairness," Yang and Howe [15] investigate interactions between fairness and explanations in neural networks showing that there are circumstances where controlling for one can impact the other. The authors then explore this relationship experimentally. They design a loss term for explanations called GWAD, Groupwise Attribution Divergence, and compare its effect to existing loss terms for fairness. They show how including this term reduces explanation differences with no significant loss of accuracy. Also, aiming to explore whether algorithms replicate human biases, Arhin and Treku [16] explain the contexts around accuracy-fairness trade-offs and make the empirical case for why, when, and how the trade-offs manifest in AI systems. The authors then use Python-generated synthetic data and propose a classification framework to help better understand the algorithmic accuracy-fairness trade-off.

Leveraging data to help improve model performance in scenarios where obtaining sufficient amounts of data, e.g., hate speech, is a challenge is the focus of Svetasheva and Lee [17]'s work. In their research, the authors propose leveraging synthetic data generation to improve hate speech detection. This path is possible following the recent development of

generative pre-trained transformers which can dramatically improve the quality of synthetic data. The paper establishes that large language models can serve as both data generators and annotators, exhibiting performance comparable to and even surpassing that of humans.

This minitrack also presents work that aims to better understand AI literacy in adult education. Wolters et al. [18] conducted a structured literature review on AI literacy with a focus on adult education and make recommendations for future research areas on this topic.

Finally, the large adoption of AI in discriminatory pricing among companies is not well received by consumers who wish to be informed about AI algorithms monitoring their online activities and have negative reactions and aversion towards AI. To that effect, Peng et al.'s work [19] utilizes a lab experiment to better understand the role of price sensitivity and explanation in the relationship between AI disclosure and consumers' revenge behavior.

4. References

[1] Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6), 16-23.

[2] O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[3] Cambridge Dictionaries Online. Cambridge University. Retrieved February 25, 2023. https://dictionary.cambridge.org/dictionary/english/discrimination

[4] Cheng, M. & Foley, C. (2018). The sharing economy and digital discrimination: The case of Airbnb. *International Journal of Hospitality Management*, 70, 95–98.

[5] Such, J. M. (2017). Privacy and autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 4761–4767).

[6] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671-732.

[7] Van Nuenen, T., Ferrer, X., Such, J. M., & Coté, M. (2020). Transparency for whom? Assessing discriminatory artificial intelligence. *Computer*, *53*(11), 36-44.

[8] Terwiesch, C. (2023). Would Chat GPT Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course. *Mack Institute for Innovation Management at the Wharton School*, University of Pennsylvania, 2023.

[9] Shendruk, A. (2023, February 14). Here's how 10 industries are experimenting with ChatGPT. *Quartz*. https://www.yahoo.com/finance/news/heres-10-industries-experimenting-chatgpt-164600575.html

[10] Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and Discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *40*(2), 72-80.

[11] Miller, K. (2020). A matter of perspective: Discrimination, bias, and inequality in AI. In *Legal Regulations, Implications, and Issues Surrounding Digital Data* (pp. 182-202). IGI Global.

[12] Lier, S. K., Gerlach, J., & Breitner, M. H. (2024, January 3). What Is Ethical AI? – Design Guidelines and Principles in the Light of Different Regions, Countries, and Cultures. In *Proceedings of the 57th Hawaii International Conference on Systems Science (HICSS -57)*, Honolulu, Hawaii.

[13] Han, Y., Landau, A., Kulkari, P., Modaresnezhad, M., & Nemati, H. (2024, January 3). An Implementable Guideline for Developing Ethical AI Systems: The Evaluation of Child Abuse and Neglect Prediction. In *Proceedings of the 57th Hawaii International Conference on Systems Science (HICSS -57)*, Honolulu, Hawaii.

[14] Sengupta, S., Srivastava, S., & McNeese, N. (2024, January 3). Public Perceptions, Critical Awareness, and Community Discourse on AI Ethics: Evidence from an Online Discussion Forum. In *Proceedings of the 57th Hawaii International Conference on Systems Science (HICSS -57),* Honolulu, Hawaii.

[15] Yang, Y., & Howe, B. (2024, January 3). Does a Fair Model Produce Fair Explanations? Relating Distributive and Procedural Fairness. In *Proceedings of the 57th Hawaii International Conference on Systems Science (HICSS -57),* Honolulu, Hawaii.

[16] Arhin, K., & Treku, D. (2024, January 3). Contextualizing the Accuracy-Fairness Trade-off in Algorithmic Prediction Outcomes. In *Proceedings of the 57th Hawaii International Conference on Systems Science (HICSS -57),* Honolulu, Hawaii.

[17] Svetasheva, A., & Lee, K. (2024, January 3). Harnessing Large Language Models for Effective and Efficient Hate Speech Detection. In *Proceedings of the 57th Hawaii International Conference on Systems Science (HICSS -57)*, Honolulu, Hawaii.

[18] Wolters, A., von Straussenburg, A. F. A., & Riehle, D. M. (2024, January 3). AI Literacy in Adult Education—A Literature Review. In *Proceedings of the 57th Hawaii International Conference on Systems Science (HICSS -57),* Honolulu, Hawaii.

[19] Peng, X., Peng, X., & Xu, D. J. (2024, January 3). Does AI Disclosure in Discriminatory Pricing Backfire? The Moderating Role of Price Sensitivity and Explanation for Price Differences. In *Proceedings of the 57th Hawaii International Conference on Systems Science (HICSS -57),* Honolulu, Hawaii.