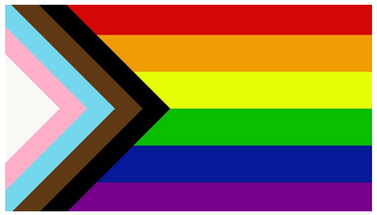

Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities

Nenad Tomasev, Kevin R. McKee, Jackie Kay, Shakir Mohamed

DeepMind

UCL

Introduction

Algorithms have moral consequences for queer communities, yet their impact remains critically underexplored from a fairness perspective.

Sexual orientation and gender identity are prototypical examples of unobserved characteristics, presenting challenging obstacles for fairness research.

There is a pressing need to develop and improve fairness frameworks applicable to the case of unobserved and sensitive characteristics.

Privacy

Sexual orientation and gender identity are highly private aspects of personal identity.

Attempts at using AI to develop 'gaydar' tools infringe on the privacy of queer individuals and pose significant safety risks.

Recent advancements in facial recognition undermine the safety of queer gatherings or clubs. Chatbots based on language models could be used to infiltrate online spaces and solicit private and sensitive information.

Privacy-enhancing solutions are needed in order to deal with the increasing risks in this domain.

Censorship

Multiple groups and institutions around the world impose unjust restrictions on the freedom of expression and speech of queer communities.

Language processing tools can be abused to erase queer voices and enforce heteronormative views.

Automated content moderation systems pose the (unintentional) risk of censoring queer expression.

Language

Researchers face an opportunity to develop language-based AI systems inclusively, overcoming human biases and establishing inclusive norms.

AI systems must be able to adapt to the evolution of natural language and avoid censoring content as a result of its adjacency to the queer community.

Fighting Online Abuse

AI systems could potentially help human moderators flag abusive online content.

Yet, recent work shows that AI may moderate queer content in unexpected and inappropriate ways: one toxicity detection system indicated drag queens were as highly offensive as white supremacists.

Health

Machine learning presents key opportunities to improve medical treatment decisions for queer communities, who are challenged by extraordinarily high incidence of HIV, STIs, and substance abuse.

At the same time, the routine omission of sensitive characteristics from medical ML research datasets poses significant risks.

Mental Health

AI systems have the potential to help address the alarming prevalence of mental health-related issues in the queer community, including the high incidence of suicide..

Reinforcement learning may present both opportunities and risks when it comes to behavioral interventions.

Employment

Research indicates that human evaluators rate resumes significantly lower if they contain queer-related content.

Big data approaches to recruitment and hiring could infringe on job candidates' privacy by outing them to prospective employers without consent.

Considerations for Queer ML Fairness Research

Queer fairness needs to be considered through an intersectional lens, as the impact of AI systems may vary between queer, racial, class, and other subcommunities.

Fairness measures that do not rely on group membership information—including individual, contrastive and counterfactual fairness—seem especially promising.

Promising, recently developed methods optimizing fairness across all plausible yet unobserved groups include *Adversarially reweighted learning* and *distributionally robust optimisation*.

It will be necessary to work towards understanding and **resolving the tension between privacy and fairness**.

Model explainability may prove crucial for ensuring ethical and fair AI applications in cases when fairness metrics are hard or impossible to reliably compute.

Researchers must **move beyond the unquestionable cisnormativity of sex and gender** categories traditionally used in the AI research.

It is crucial that the AI community **involves more queer voices** in the development of AI systems, algorithmic fairness, and ethics research.