The Impact of Racial Bias in AI, Facial Recognition, and Other Technologies

COLLIN SMITH

CHANEL CORNETT

ADONNE WASHINGTON

EDWARD FULTZ
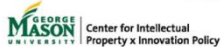
TABLE OF CONTENTS

Collin Smith

Collin Smith is a product attorney at Microsoft, supporting Xbox hardware and accessories, the Applied Sciences Group, Surface Duo and Pen, and Microsoft's international development centers in Tampere, Finland and Tel Aviv, Israel. Before promotion to Principal Corporate Counsel, Collin supported Microsoft 365 engineering, including APIs, AI/ML, organizational analytics and intelligence, and Microsoft's new AI time management assistant Cortana Scheduler.

Before joining Microsoft, Collin worked in-house at Chase (JPMorgan Chase & Co.) in New York, handling a variety of privacy- and technology-impacting consumer banking matters including regulatory inquiries and consumer complaints, matters involving payment technology such as Chase Pay, wire transfers and EFTs, and cryptocurrency. Collin also was one of three attorneys who helped develop an AI-inspired regulatory change management platform for JPMC's legal and compliance departments.

Before becoming a lawyer, Collin served in the United States Army as a logistics officer, including a tour of duty to Iraq in 2010. Collin holds a B.S. in Systems Engineering from the U.S. Military Academy at West Point (Beat Navy!), and a J.D. from The George Washington University Law School.

Chanel Cornett

Chanel Cornett is a Policy Counsel on the Trust & Safety team at Zoom, where she crafts policies to keep users safe on the platform. Prior to joining Zoom, Chanel was a Legal and Policy Officer at Fair Trials, an international NGO where she worked on advancing and protecting criminal procedural rights through legislative drafting, legislative advocacy, and coalition building. Before Fair Trials, she spent her first year post law school graduation working for California Attorney General, Xavier Beccerra's legal and policy advisory team on issues such as criminal justice, immigration, environmental justice, consumer protection, and health care. She is a graduate of Howard University School of Law, where she held internships at the U.S. Securities and Exchange Commission, the U.S. District Court of the District of Columbia, and the U.S. House of Representative among others. She attended undergrad at Marietta College in Marietta, Ohio where she is currently a member of Marietta's Alumni Association Board of Directors. Chanel is a Cleveland, Ohio native based in Washington D.C.

Adonne Washington

Adonne Washington, Esq. is a 2020 graduate of Howard University School of Law. She is currently the Digital Justice Associate Counsel for the national branch of the Lawyers' Committee for Civil Rights Under Law. The Digital Justice Initiative works at the intersection of racial justice, technology, and privacy. Predatory commercial data practices, discriminatory algorithms, invasions of privacy, disinformation, and online hate disproportionately target and harm communities of color, especially Black Americans, immigrants, women of color, and LGBTQ people of color. Adonne is also principal of her own firm, A.R. Washington, PLLC where she offers trademark, copyright, and business law services to small businesses and creatives. Adonne received her BA in social relations and policy with a minor in public relations from Michigan State University.

Edward Fultz

Edward Fultz is a graduate of Morehouse College and Howard University School of Law, where he was President of his class. Currently, Edward is an associate at Reed Smith in the D.C. office, where he practices Intellectual Property, Trademark Law, Data Privacy, Sports, and different sectors of Entertainment and Media. During law school, his thesis was exploring AI interactions with dating applications, the implication on cyber-stalking laws, and the privacy laws that are impacted.

The Impact of Racial Bias in Content Moderation AI

Chanel Cornett, Esq.
Policy Counsel, Trust & Safety
Zoom

AGENDA

I. Content Moderation
   A. The Policies
   B. The Enforcement
II. How is AI Used in Content Moderation?
III. Why is AI Used in Content Moderation?
IV. Where AI Gets it Wrong
V. Platform Racism
VI. Stats
VII. Solutions
VIII. Q&A

Content Moderation

"The term "content moderation" means the intentional removal, labeling, or altering of user-generated content on a covered platform by the covered platform or an automated or human system controlled by the covered platform, including decreasing the algorithmic ranking of user-generated content, removing user-generated content from algorithmic recommendations, or any other action taken in accordance with the covered platform's terms of service, community guidelines, or similar materials governing the content allowed on the covered platform."

I. The Policies

- Typically named, "Community Guidelines", "Community Standards", "Content Policies", "Content Standards", etc.
- The "laws" which define what's allowed and what's not allowed on a platform.
- Examples include: violence, hate speech, misinformation, CSAM, abuse, harassment, doxing, terrorism, trademark/copyright infringement

II. The Enforcement
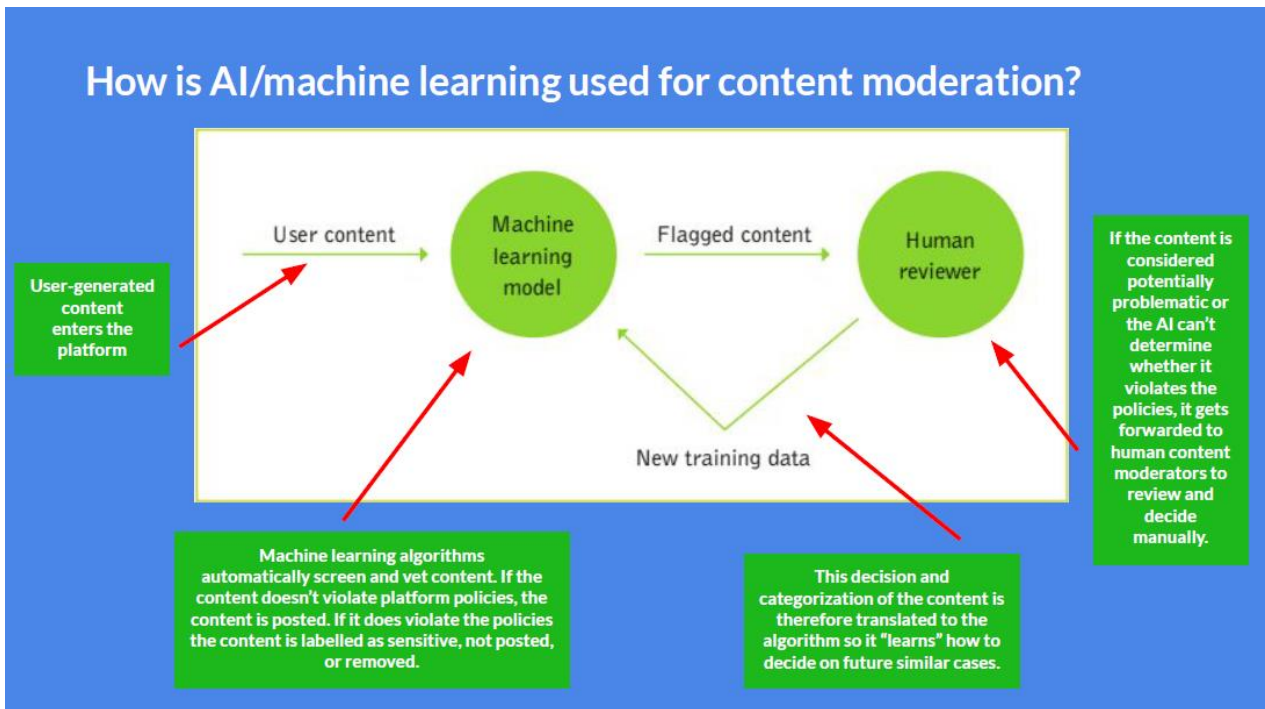
- Human enforcement: platforms either hire a contractor/vendor that specializes in content moderation, has its own team of analysts that make decisions on reported violations, or have a combination of both, with their internal team handling more complex reports.
- AI/machine learning enforcement: AI is used to make decisions on whether content violates the platforms policies and should, thus be removed or flagged as sensitive

How is AI/machine learning used for content moderation?

User content → Machine learning model → Flagged content → Human reviewer

New training data

User-generated content enters the platform

Machine learning algorithms automatically screen and vet content. If the content doesn't violate platform policies, the content is posted. If it does violate the policies the content is labelled as sensitive, not posted, or removed.

This decision and categorization of the content is therefore translated to the algorithm so it "learns" how to decide on future similar cases.

If the content is considered potentially problematic or the AI can't determine whether it violates the policies, it gets forwarded to human content moderators to review and decide manually.

Why is AI used for Content Moderation?

1. The emotional toll of human content moderation
2. The amount of online content required to be moderated

Facebook's race-blind practices around hate spe came at the expense of Black users, documents show

How bias creeps into the AI designed to detect toxicity

Black Instagram users were 50% more likely than white users to have their accounts automatically disabled, internal research reportedly showed

TECH • ARTIFICIAL INTELLIGENCE

Google's Hate Speech Detection A.I. Has a Racial Bias Problem

BY JONATHAN VANIAN
August 16, 2019 1:29 PM EDT

The algorithms that detect hate speech online are biased against black people

A new study shows that leading AI models are 1.5 times more likely to flag tweets written by African Americans as "offensive" compared to other tweets.
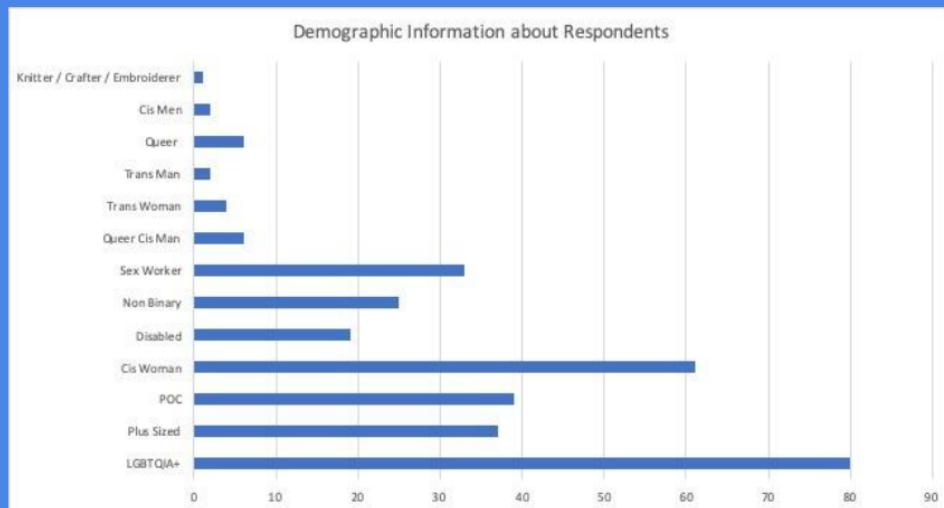
Instagram

Beyond a technical bug: Biased algorithms and moderation are censoring activists on social media

Published: May 16, 2021 1.17pm BST

Platform Racism

"a new form of racism derived from the culture of social media platforms – their design, technical affordances, business models and policies – and the specific cultures of use associated with them." - Matamoros-Fernandez

Demographic Information about Respondents

Category	Value
Knitter / Crafter / Embroiderer	1
Cis Men	2
Queer	6
Trans Man	2
Trans Woman	4
Queer Cis Man	6
Sex Worker	33
Non Binary	25
Disabled	19
Cis Woman	61
POC	39
Plus Sized	37
LGBTQIA+	80

Solutions

1. Use and further development of context sensitive AI models
2. Publish transparency on:
   a. How and what AI models are used
   b. Automated decision statistics
3. Establish equitable platform policies to account for different communities
4. Community centric content moderation
5. End or limit use of automation for content moderation
6. Increased diversity in the content moderation/Trust & Safety field

Sources