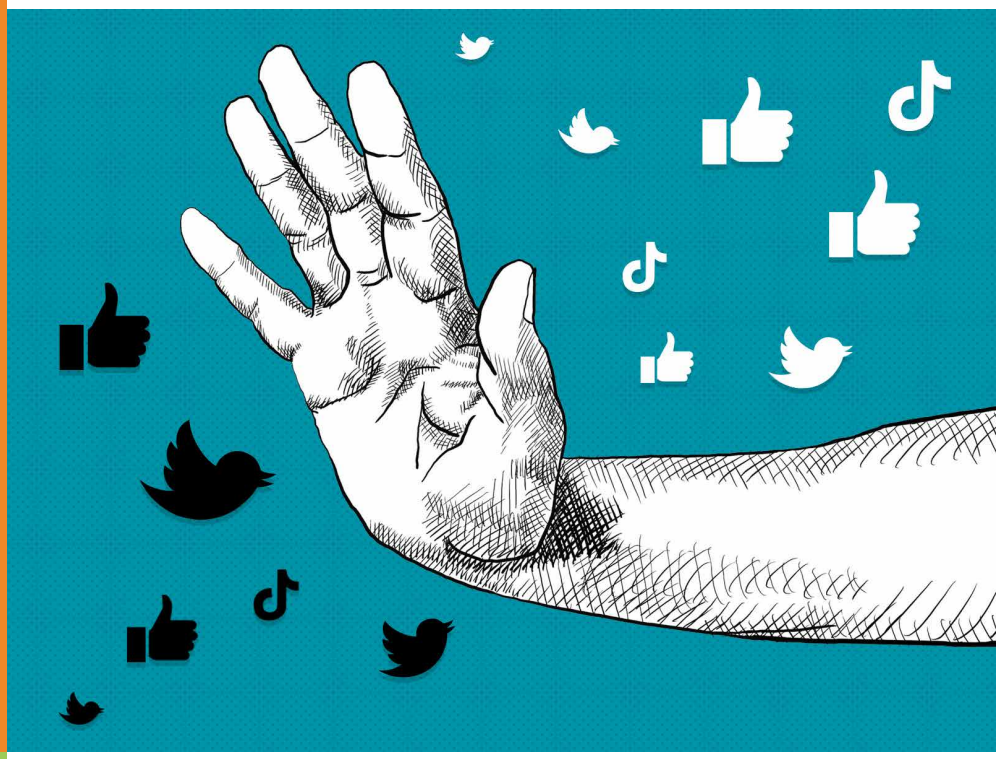
HEINRICH BÖLL STIFTUNG
DEMOCRACY

E-PAPER

# Algorithmic misogynoir in content moderation practice



**BRANDEIS MARSHALL**

Published by Heinrich-Böll-Stiftung European Union
and Heinrich-Böll-Stiftung Washington, DC, June 2021

# The author

**Brandeis Marshall** is a computer science scholar and educator who contributes to the data engineering, data science, and data/computer science education fields. Her interests intersect the racial, gendered, and socioeconomic impact of data in technology.

She is currently a Professor of Computer Science at Spelman College and a practitioner fellow with Stanford University Digital Civil Society Lab. She is also the founder of DataedX, an executive training firm focused on enhancing the workforce's data competencies and supporting career development.

# Abstract

Existing content moderation practices, both algorithmically-driven and people-determined, are rooted in white colonialist culture. Black women's opinions, experiences, and expertise are suppressed and their online communication streams are removed abruptly, silently, and quickly. Studying content moderation online has unearthed layers of algorithmic misogynoir, or racist misogyny directed against Black women. Tech companies, legislators and regulators in the U.S. have long ignored the continual mistreatment, misuse, and abuse of Black women online. This paper explores algorithmic misogynoir in content moderation and makes the case for the regular examination of the impact of content moderation tactics on Black women and other minoritized communities.

# Contents

# 1. Introduction

While we are one human race, social, economic, and political constructs impose a hierarchy that centers white colonialist culture and white patriarchy through manipulation, coercion, or force. These social, economic, and political structures are not designed for equity. Rather, they're designed to de-value everything but white culture. As Gulati-Partee and Potapchuk discuss in their work on advancing racial equity, "[r]acial disparities are driven and maintained by public- and private-sector policies that not only disadvantage communities of color but also over-advantage whites."

We exist in a global society that exudes anti-Blackness – and this is true for online spaces as well. The tech industry is monopolized by white men in ownership, leadership, and the workforce. When building businesses, Black women receive funding at levels 20 times lower than the median national funding level that includes all demographics of business owners. This demonstrates how much more difficult it is for Black women to raise the capital they need. Men account for 70-80% of the tech workforce and leadership. This industry has built a well-documented culture of toxicity, where men hold all the cards and make all the rules and covert structural racism persists. And this toxic tech culture is perpetuated in the classroom when computer science educators are not committed to changing its culture. These circumstances make the pursuit of equitable practices tenuous at best.

Being a Black woman – offline and online – means experiencing two extremes: 1) no one pays attention to what you say or how you say it, or 2) your words are fodder for scrutiny, surveillance, and judgment. Being simultaneously invisible and hypervisible is a consistent theme (as explored in Dr. Tressie McMillian Cottom's *Thick*) and the result of part of what Moya Bailey calls misogynoir (Crunk Feminist Collective, March 2010), the "anti-Black racist misogyny that Black women experience." The Abuse and Misogynoir Playbook sheds light on a five-phase cycle of disbelieving, devaluing, and discrediting the contributions of Black women as the historical norm. Algorithmic misogynoir builds on Bailey's description to identify how these interactions play out online and in code for Black women.

In digital spaces, Black women's presence, experiences, and interactions include everything from communicating our accomplishments to sharing our trauma. But now, the interactions and responses come extremely quickly, from real, bot, and troll profiles from anywhere on the globe.

Black people and women are more likely to be the targets of online harassment than their white and male counterparts. Black women have been criticized internationally for their scholarship, had posts removed for statements that are met with more scrutiny than the posts of their white counterparts, and been suspended or banned from a platform for speaking out against any form of algorithmic discrimination.

Black women's distinct circumstances aren't recognized and centered. Shireen Mitchell's 2018 Stop Online Violence Against Black Women report showed how online campaigns using Facebook ads were created to disparage Black girls and women with sexualized memes, hashtags, and fake accounts to help spread disinformation ahead of and during the 2016 U.S. Presidential Election. And Charisse C. Levchek's Microaggressions and Modern Racism documents anti-Black racism at the micro and macro level, both via in-person and online interactions. Levcheck points to racial slurs and other forms of hate speech done by micro-aggressors and macro-aggressors on the internet. She specifically calls on organizations to enact policies and procedures to address instances of racism, penalize these micro/macro-aggressors, and support survivors of racism.

All this falls under what Matamoros-Fernandez calls "platformed racism," which is "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." Content moderation can work towards creating inclusive, welcoming spaces for Black women, but current practices embrace misogynoir and then deploy it algorithmically.

# 2. Current content moderation practices

## 2.1. The problem with generalization

Content moderation, according to Grimmelmann, consists of the "governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse."

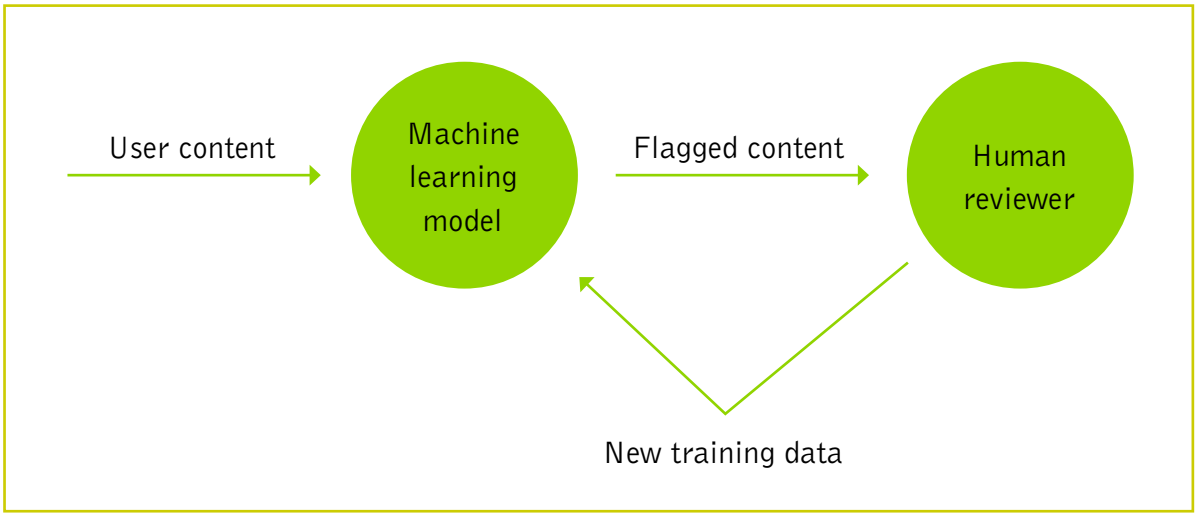



**Figure 1: How Content Moderation Works**

**Figure 1** presents the backbone of content moderation approaches. User-generated content enters the platform or other digital system, where a series of machine learning algorithms are executed to automatically screen and vet the appropriateness of the content. If the content doesn't raise any platform governance issues, the content is posted. If the content violates the platform's governance guidelines, the content is labelled as sensitive content, not posted, or removed. If the content is considered potentially problematic, it's forwarded to human content moderators to review manually and decide whether it can be posted. The pain point for handling content moderation on platforms is how to deal with the fast influx of user-generated content.

We're in a vicious cycle of more data begets more data. Controlling data/content has become a top priority for platforms, which use this combination of human content moderators and automated content moderation practices, as detailed in Gillespie's *Custodians of the Internet*. Human content moderators were once deemed sufficient given the manageable content generation rate of the early internet. However, manual content screening soon couldn't keep up with the rate of user-generated content. Automated content moderation strategies were built to help handle the additional workload and, in theory, conduct the moderation more efficiently. Some particularly offensive content, like child pornography, takes a severe psychological toll on human content moderators. Automated content moderation algorithms can protect people from manually screening that content.

(As a side note: Facebook and Twitter have received frequent attention related to labor issues regarding their content moderation practices. Facebook's content moderators have spoken out about the lack of mental health support and overall unconducive work experience. Paul M Barrett's 2020 report on content moderation and consequences calls for Facebook to make content moderators in-house employees, rather than third-party contract workers, and to triple this workforce, among other recommendations.)

Still, content moderation remains difficult. There is plenty of content that algorithms can't properly categorize because the nuances of our culture aren't as predictable as once believed. Social, economic, political, technical/algorithmic, gender, and racial factors all affect Black women. Society's approach is always to silo these factors and address them one at a time with the mindset that solving the individual factors will help solve the sum of the issues.

Most approaches seek to standardize the pain points, meaning that the newly enacted solutions are supposed to affect all the communities in the same way. Standardizing the identification and handling of perceived and actual problematic content requires establishing a sameness criteria, demonstrating consistent patterns of inappropriate content and a universal effective routine for solving the problems – which doesn't work in our global society, which is more multicultural in our demographic composition.

This isn't an addition problem of disparate impacts. It's more like a multiplication problem of disparate impacts. The result of the simplistic approach is inconsistent content moderation practices that silence Black women.

## 2.2. Double standards

Nominally, content moderation is subject to platform standards and guidelines, such as Facebook's Community Standards. At Facebook, 23 different categories make up their content moderation guidelines, covering violence and criminal behavior, safety, objectionable content, integrity and authenticity, intellectual property, and content-related requests and decisions. Facebook broadly details what content is deemed inappropriate, yet refrains from sharing how their standards are enacted.

For example, Carolyn Wysinger, a high school teacher in Richmond, California, had her Facebook comment deleted within 15 minutes for hate speech even though she was responding to Liam Neeson's anti-Black remarks. While promoting a revenge movie, the Hollywood actor had confessed that decades earlier, after a female friend told him she'd been raped by a Black man she could not identify, he'd "roamed the streets hunting for Black men to harm."

The actor's remarks were not removed. Violence against Black men, for being Black, isn't content designated to be removed from the platform, but calling white men fragile is considered hate speech? According to Facebook's 23 categories, it seems that Liam Neeson's content breached both their Violence and Incitement and Hate Speech policies. It's unclear which community standard policy Wysinfer's post violated.

Women are called fragile, weak, and sensitive regularly, yet that speech isn't marked as hate speech. Or, was it the word "fragile" in conjunction with "white men" that triggered the automated content moderation algorithms? We don't know, because there is a lack of transparency when it comes to content moderation algorithms, processes, and practices. We do have access to Facebook's Community Standards Enforcement Report, which is heavily sanitized, presenting aggregated and summarized data.

At Twitter, 15 rule categories are covered under the content moderation umbrella. These are divided into four groups: safety, privacy, authenticity and enforcement, and appeals. The content moderation decisions are swift and harsh for Black women.

Consider the case of Shana V White, currently the Senior Associate, CS Equity and Justice Initiatives at the Kapor Center. As a computer science educator with 16 years of experience, she engages her 25,000+ Twitter followers as a vital advocate for teachers and marginalized communities. White was permanently banned from the platform on April 26, 2021 for her comments to someone who supported former Senator Rick Santorum's disinformation narrative about Native Americans. She appealed and her account was restored that evening. Days later, White once again issued comments in defense of a marginalized community and was permanently banned again. She announced her return to Twitter on May 27, 2021.

## Jason S Campbell's April 26th tweet


## Shanas V. White's 26th April response tweet


So a post sharing disinformation about the United States' history remains viewable to the public (~9.8M views), but a snarky and sarcastic comment results in the profile's permanent ban? The initial post seems to violate the Platform Manipulation and Spam policy.

White's response likely violated Twitter's Suicide/Self-Harm policy. However, again, we don't know because of the lack of transparency. Like Facebook, Twitter's Transparency Report is also heavily sanitized, presenting aggregated and summarized data.

In these content moderation snafus, it's the Black women who are responding to incendiary posts and then de-voiced as a consistent outcome. Black women who defend themselves or others are tagged as the agitators who need to be tone policed by white colonialists in powerful positions.

We don't know how much of this algorithmic misogynoir is ignited by content moderation algorithms or by fellow users reporting Black women's content. Still, the final decision of their content's appropriateness, regardless of the outcome, becomes part of the platform's online portfolio of a user's suspected and attributed violations (see Twitter's consequences section in its hateful conduct policy). This is reminiscent of an online version of the U.S. criminal system where, once a person is tagged as an agitator, it is very difficult to reduce or remove that perception. Twitter lays out a ladder of content moderation enforcement options – including tweet-level, direct message-level, and account-level enforcement with permanent suspension being the most severe consequence – but the escalation process for content moderation enforcement remains confusing.

Here are a few more high-profile content moderation and online harassment instances:

- Dr. Safiya Noble's *Algorithms of Oppression: How Search Engines Reinforce Racism* was publicly criticized on Twitter by a staff historian from the Institute of Electrical and Electronics Engineers (IEEE), an international professional society. The staff historian hadn't read the book prior to posting his commentary under the IEEE umbrella. The post was taken down only after public outrage.

- Joy Buolamwini, Dr. Timnit Gebru, Dr. Helen Raynham, and Deborah Raji's Gender Shades paper on racial and gender discrimination in commercial facial recognition software drew sizable international critique from Amazon, which developed one of the technologies evaluated in the paper. Amazon went so far as to post their disagreement with the paper's findings on their blog. Buolamwini responded using Medium articles like this to share their prior communications with each company about the racial and gender inequities in their facial analysis findings.

- Dr. Timit Gebru contended with online harassers when discussing Duke University's PULSE AI tool, a human face creation technology. Her pioneering work in algorithmic inequities in artificial intelligence were minimized, she was addressed as if she were too emotional or hysterical, she endured mansplaining, and experienced being treated like an angry Black woman, which is a racist stereotype.

- Months later, Dr. Gebru advocated for more transparency about internal Google processes regarding research paper publication criteria. One of her papers, accepted for publication at a prominent international conference, was scrutinized by those inside of Google, but she did not receive clarification on the process or issues.

She was abruptly terminated by the organization. As Dr. Gebru shared this experience in real time on Twitter, she was attacked again and again on Twitter after she was ousted by Google, yet her harassers retained their online profiles for months.

# 3. Proposals and suggestions

## 3.1. Clarifying the role of social media companies

In the U.S., attempts at creating content moderation legislation continue to overlook Black women online. Sometimes, this is explicit: the January 2021 Congressional Research Service Report on Social Media mentioned neither Black communities, Black people, nor Black women. Of particular note in the conversation around moderation is Section 230 of the Communications Decency Act of 1996, which states that, "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." (47 U.S.C. § 230). This means that interactive computer services, a descriptor used to define websites in the 1990s, aren't liable for third-party content posted to their websites. In general, this protection ensures that websites can't be sued if a user posts illegal content. Copyright violations, pornographic materials, and federal crime violations have been specifically listed in Section 230 exceptions.

Social media platforms are hosting, publishing, and moderating content that may not meet the criterion of being exempt from the liability protections, but that may still create harm by discriminating against some users. The rules of moderation are determined by economically and politically influential private corporations, unregulated. The crux of the debate is how to classify these companies – as either "content pass-through companies" or "responsible content companies" – and whether social media platforms like Facebook and Twitter are protected from liabilities and lawsuits for their moderation conduct under the vague language of "interactive computer service."

If they are classified as content pass-through companies, then they simply present user-generated content (and therefore are not liable for the content shared on their platforms). If they are responsible content companies, then they're in the business of controlling what users view (and therefore culpable for outcomes).

Social media platforms are operating as responsible content companies and should be treated as such. Automated and manual content moderation practices are happening. This means that platforms themselves deploy content management protocols that reach beyond copyright, pornographic materials, and federal crime violations.

In addition, the "interactive computer services" definition needs revisions to capture both sides of the online interactivity. Section 230 defines an "interactive computer service" as "any information service, system, or access software provider that provides or enables

computer access by multiple users to a computer server, including specifically a service or system that provides access to the Internet and such systems operated or services offered by libraries or educational institutions." U.S.C. §230(f)(2).

This definition describes the outgoing online communication or notification aspects of computer services e.g., sharing an article. However it doesn't address the *incoming* online communication processing and interpretations, e.g., posting a reply post or the ensuing direct back-and-forth communications among users in reference to an initial post. The outgoing online communication is similar to a broadcast signal as it intends to be seen by as many people as possible. An incoming online communication, on the other hand, responds to an individual or group of people as a way to engage in conversations publicly.

The interactive computer service description, in being applied to platforms, is woefully insufficient. The applications and uses of online platforms have evolved many times since Section 230 was crafted in the mid-1990s. Platforms are handling both outgoing and incoming streams of communications, and this management operates differently depending on the direction of the communication stream. As this paper will show later, platforms use automated content moderation algorithms to monitor reply messages at a higher level of scrutiny than outgoing communications.

## 3.2. Addressing structural inequalities

Current content moderation practices signal a trajectory that codifies algorithmic misogynoir. For example, the Global Internet Forum to Counter Terrorism (GIFCT) is a collaborative working to "foster collaboration and information-sharing to counter terrorist and violent extremist activity online." The founding members are titans in the social media and internet community: YouTube (Google), Twitter, Facebook, and Microsoft.

Gorwa et al eloquently describe how GIFCT's interlaced technical and political issues are grounded in opacity. The specific engagements of the working groups, partnerships, and collaborations that are part of GIFCT aren't shared. Their transparency reports offer little detail on which algorithmic tools are being used in automated content moderation, or how individual users are the benefactors of this collective. What's missing in these approaches is contextual understanding. GIFCT's mission, vision and core values lack race, gender, class and ableism considerations.

An alternative to this approach is to use the existing public value failure (PVF) framework, which articulates nine categories that delineate society's failure in providing a public value, such as rights, benefits, or privileges of citizens, by governments and policies. The public values framework helps us tease out proposals that could make the internet safer for Black women.

The framework tells us that an "equal playing field" (broad, standardized policies) is less desirable than collective actions and public policies addressing structural inequalities and historical differences in opportunity structures. It is thus important to regularly examine content moderation practices for Black women and other minorities.

The ranking and prioritization protocols in content moderation benefit white communities more favorably than non-white communities. Black women, in contrast, experience punitive actions for vocalizing inequities on these same platforms. The benefits and harms of existing content moderation practices need to be documented.

It would be useful to disaggregate data by race/ethnicity and gender in content moderation guidelines, standards, practices, enforcement, and legal requests. Sanitized and aggregated data hides the demographics of who flagged content, whose content was flagged, which guideline was breached by the content, and how the automated content moderation algorithms decided on their outcomes. It may also be helpful to have equal representation of Black women and other minorities in human content moderator teams. Much more work needs to be done to make transparent the automated content moderation tools, policies, and practices.

As platform engagement is an integral part of business and marketing activities, potential uneven content moderation enforcements can affect people's livelihood as well as damage their dignity and reputation online. Therefore, it is also important to have independent oversight and public accountability reports and audits. This helps in establishing transparency as a central tenet of policy-making. Sharing and implementing consistent content moderation protocols are required for true transparency across sociopolitical systems. Accessible public discourse and responsive government actions can build trust with more effective communication streams.

## 3.3. Balancing power asymmetries between originator and commenter

Content moderation operates simultaneously in two directions: (1) ways in which a user interacts with the platform, e.g., a user's ability to block another user and (2) ways in which the platform interacts with the user, e.g., the treatment of reply tweets make them notable candidates for online harassment violations. Policy proposals tend to address the former and not the latter, but need to intentionally address both.

When content moderation practices lean on heavily automated decision systems and protocols, social structures and technical processes collide. Notably, the tweet originator is part of an implicit protected class over the tweet commenter. (See Twitter's offensive content documentation.) When someone comments, or replies, to the original tweet, the tweet commenter is notified that their content has been reviewed and potentially deemed to have violated the platform's rules. However, the tweet originator doesn't receive the same treatment – as in the case of Rick Santorum's tweet that remained posted while

Shana White's tweet was removed and her account suspended. In essence, we assess that platforms weren't designed to pre-process and review messages for potentially offensive content. In addition, uncomfortable conversations are being discouraged by the platform.

An alternative is to deploy backtracking and propagation strategies. The backtracking in content moderation strategy operates such that if flagged content is in response to another piece of content and moderated off the platform, then it behooves the content moderation protocols to re-review to the original source for its compliance with handling of harmful and illegal content. In other words, if the reply content is removed, then original source content may be eligible for removal as well.

The propagation in content moderation strategy operates in the opposite way: If the original source content is removed from the platform, then the reply content may be eligible for removal also. Content commenters should receive an alert stating which content moderation guideline the original content violated, as well as their own, if applicable. In the case of backtracking or propagation strategies in content moderation, we recommend more consistent handling of enforcement practices. The approach used to resolve one content moderation issue needs to apply equitably to all issues of the same classification. Right now, we can't be sure exactly what's happening due to tech's opacity and under-reporting when content is examined using automated decision algorithms. All signs are leaning toward an imbalance that needs to be urgently corrected.

# References


47 U.S. Code § 230 – Protection for private blocking and screening of offensive material. (n.d.). *Cornell Law School – Legal Information Institute.* https://www.law.cornell.edu/uscode/text/47/230 (accessed 2021, June 16)

About. (n.d.). Shana V. White. "Illuminate others and purposefully disrupt the status quo." *Shana V. White Blog.* https://shanavwhite.com (accessed 2021, June 16)

About offensive content. (2021). *Twitter.* https://help.twitter.com/en/safety-and-security/offensive-tweets-and-content (accessed 2021, June 16)

Angry black woman. (2021, June 12). *Wikipedia. The Free Encyclopedia.* https://en.wikipedia.org/wiki/Angry_black_woman (accessed 2021, June 16)

Apple. (n.d.). *Apple.* https://www.apple.com (accessed 2021, June 16)

Barrett, P. M. (2020). Who Moderates the Social Media Giants? A Call to End Outsourcing. *NYU Stern.* https://bhr.stern.nyu.edu/tech-content-moderation-june-2020 (accessed 2021, June 16)

Bell, T. A. (2020, June 5). It's Time We Dealt With White Supremacy in Tech. *Marker.* https://marker.medium.com/its-time-we-dealt-with-white-supremacy-in-tech-8f7816fe809 (accessed 2021, June 16)

Bozeman, B., Johnson, J. (2014, May 26). The Political Economy of Public Values: A Case for the Public Sphere and Progressive Opportunity. *SAGE Journals – The American Review of Public Administration.* Vol 45, Issue 1. https://journals.sagepub.com/doi/abs/10.1177/0275074014532826 (accessed 2021, June 16)

Brown, D'S. (2021, February 5). Male Colleagues Harass Black Female Former Googler Timnit Gebru Amid Google Ouster. *Blavity News.* https://blavity.com/male-colleagues-harass-black-female-former-googler-timnit-gebru-amid-google-ouster?category1=news (accessed 2021, June 16)

Buolamwini, J. (2019, January 25). Response: Racial and Gender bias in Amazon Rekognition – Commercial AI System for Analyzing Faces. *Medium.* https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced (accessed 2021, June 16)

Chang, E. (2019, March 5). Brotopia. *Penguin Random House.* https://www.penguinrandomhouse.com/books/547571/brotopia-by-emily-chang/ (accessed 2021, June 16)

Community Standard Enforcement Report. (2021). *Facebook.* https://transparency.fb.com/data/community-standards-enforcement/?from=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement (accessed 2021, June 16)

Community Standards. (2021). *Facebook.* https://www.facebook.com/communitystandards/introduction (accessed 2021, June 16)

Deerwester, J. (n.d.). "Liam Neeson is canceled:" Fans react to actor's story of urge for racist revenge. *USA Today News.* https://eu.usatoday.com/story/life/people/2019/02/04/liam-neeson-reveals-shocking-racially-charged-past/2766111002/ (accessed 2021, June 16)

Flaherty, C. (2018, February 6). Questioning "Algorithms of Oppression." *Inside Higher ED.* https://www.insidehighered.com/news/2018/02/06/scholar-sets-twitter-furor-critiquing-book-he-hasnt-read (accessed 2021, June 16)

Gallo, J. A., Cho, C. Y. (2021, January 27). Social Media: Misinformation and Content Moderation Issues for Congress. *Congressional Research Service.* https://crsreports.congress.gov/product/pdf/R/R46662 (accessed 2021, June 16)

Gorwa, R., Binns, R., Katzenbach, C. (2020, February 28). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *SAGE Journals – Big Data & Society.* https://journals.sagepub.com/doi/full/10.1177/2053951719897945 (accessed 2021, June 16)

Grimmelmann, J. (2015). The Virtues of Moderation. *17 YALE J.L. & TECH.* 42 (2015). https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1110&context=yjolt (accessed 2021, June 16)

Gulati-Partee, G., Potapchuk, M. (2014). Paying Attention to White Culture and Privilege: A Missing Link to Advancing Racial Equity. *The Foundation Review*, Vol. 6:1. http://www.mpassociates.us/uploads/3/7/1/0/37103967/paying_attention_to_white_culture_and_privilege-_a_missi.pdf (accessed 2021, June 16)

Guynn, J. (n.d.). Facebook while black: Users call it getting "Zucked," say talking about racism is censored as hate speech. *USA Today News.* https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/ (accessed 2021, June 16)

Hateful conduct policy. (2021). *Twitter.* https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (accessed 2021, June 16)

How The Facebook Ads that Targeted Voters Centered on Black American Culture: Voter Suppression was the End Game. (n.d.). *SOVAW. Stop Online Violence Against Women.* https://stoponlinevaw.com/wp-content/uploads/2018/10/Black-ID-Target-by-Russia-Report-SOVAW.pdf (accessed 2021, June 16)

How well do IBM, Microsoft, and Face++ AI services guess the gender of a face? (2018). *Gender Shades.* http://gendershades.org (accessed 2021, June 16)

Jason Campbell @JasonSCampbell. (2021, April 26). *Twitter.* https://twitter.com/JasonSCampbell/status/1386685340522536961?s=20 (accessed 2021, June 16)

Kleinman, Z. (2019, February 4). Amazon: Facial recognition bias claims are 'misleading'. *BBC News.* https://www.bbc.com/news/technology-47117299 (accessed 2021, June 16)

Kurenkov, A. (2020, June 24). Lessons from the PULSE Model and Discussion. *The Gradient.* https://thegradient.pub/pulse-lessons/ (accessed 2021, June 16)

Levchak, C. C. (2018). Microaggressions and Modern Racism. *Springer.* https://link.springer.com/book/10.1007/978-3-319-70332-9#about (accessed 2021, June 16)

Matamoros-Fernández, A. (2017, February 21). Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Taylor & Francis Online.* https://www.tandfonline.com/doi/abs/10.1080/1369118X.2017.1293130?journalCode=rics20 (accessed 2021, June 16)

McClintock, E. A. (2016, March 31). The Psychology of Mansplaining. *Psychology Today.* https://www.psychologytoday.com/us/blog/it-s-man-s-and-woman-s-world/201603/the-psychology-mansplaining (accessed 2021, June 16)

McMillan Cottom, T. (2019, October). Thick. And other Essays. *The New Press.* https://thenewpress.com/books/thick (accessed 2021, June 16)

Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C. (2020, July 20). PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. *Duke University.* http://pulse.cs.duke.edu (accessed 2021, June 16)

Messenger, H., Simmons, K. (2021, May 10). Facebook content moderators say they receive little support, despite company promises. *NBC News.* https://www.nbcnews.com/business/business-news/facebook-content-moderators-say-they-receive-little-support-despite-company-n1266891 (accessed 2021, June 16)

Molla, R., Lightner, R. (2016, April 10). Diversity in Tech. *The Wall Street Journal.* http://graphics.wsj.com/diversity-in-tech-companies/ (accessed 2021, June 16)

Núñez, A-M., Mayhew, M. J., Shaheen, M., Dahl, L. S. (2021, March 15). Let's Teach Computer Science Majors to Be Good Citizens. The Whole World Depends on It. *EdSurge.* https://www.edsurge.com/news/2021-03-15-let-s-teach-computer-science-majors-to-be-good-citizens-the-whole-world-depends-on-it (accessed 2021, June 16)

Our range of enforcement options. (2021). *Twitter*.
https://help.twitter.com/en/rules-and-policies/enforcement-options (accessed 2021, June 16)

Rooney, K., Khorram, Y. (2020, June 12).
Tech companies say they value diversity, but reports show little change in last six years. *CNBC.*
https://www.cnbc.com/2020/06/12/six-years-into-diversity-reports-big-tech-has-made-little-progress.html
(accessed 2021, June 16)

Ryan Mac @RMac18. (n.d.). *Twitter.*
https://twitter.com/RMac18/status/1382366931307565057?s=20 (accessed 2021, June 16)

Schiffer, Z. (2021, March 5). Timnit Gebru was fired from Google – then the harassers arrived. *The Verge.*
https://www.theverge.com/platform/amp/22309962/timnit-gebru-google-harassment-campaign-jeff-dean
(accessed 2021, June 16)

Shana V. White @ShanaVWhite. (2021). *Twitter.*
https://twitter.com/ShanaVWhite/status/1397873437197078530?s=20 (accessed 2021, June 16)

Shea Wesley Martin @sheathescholar. (2021). *Twitter.*
https://twitter.com/sheathescholar/status/1386744704176431104?s=20 (accessed 2021, June 16)

The Twitter Rules. (2021). *Twitter.*
https://help.twitter.com/en/rules-and-policies/twitter-rules (accessed 2021, June 16)

The State of Black & Latinx Women Founders. (2021). *Digitalundivided.*
https://www.projectdiane.com (accessed 2021, June 16)

They aren't talking about me... (2010, March 14). *The Crunk Feminist Collection.*
http://www.crunkfeministcollective.com/2010/03/14/they-arent-talking-about-me/ (accessed 2021, June 16)

Turner, K., Wood, D., D'Ignazio, C. (2021, January 27). The Abuse and Misogynoir Playbook. *mit media lab.* https://www.media.mit.edu/articles/danielle-wood-and-katlyn-turner-co-author-article-the-abuse-and-misogynoir-playbook-for/ (accessed 2021, June 16)

Vogels, E. A. (2021, January 13). The State of Online Harassment. *Pew Research Center.*
https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (accessed 2021, June 16)

Wood, M. (2019, January 26). Thoughts on Recent Research Paper and Associated Article on Amazon Rekognition. *AWS Machine Learning Blog.* https://aws.amazon.com/de/blogs/machine-learning/thoughts-on-recent-research-paper-and-associated-article-on-amazon-rekognition/ (accessed 2021, June 16)

**Imprint**

Heinrich-Böll-Stiftung European Union, Brussels,
Rue du Luxembourg 47-51, 1050 Brussels, Belgium

Heinrich-Böll-Stiftung Washington, DC, 1432 K St NW, Washington, DC 20005, USA

**Contact, Heinrich-Böll-Stiftung European Union, Brussels**
Zora Siebert, Head of Program, EU Policy
**E** Zora.Siebert@eu.boell.org

**Contact, Heinrich-Böll-Stiftung Washington, DC**
Sabine Muscat, Program Director, Technology and Digital Policy
**E** Sabine.Muscat@us.boell.org

Place of publication: http://eu.boell.org | https://us.boell.org

Release date: June 2021

Layout: Micheline Gutman, Brussels

Illustrations: Pia Danner, p*zwe, Hannover

Editor: Angela Chen