Science and Technology Innovation Program

MALIGN CREATIVITY

HOW GENDER, SEX, AND LIES ARE WEAPONIZED AGAINST WOMEN ONLINE

Authors

Nina Jankowicz
Jillian Hunchak
Alexandra Pavliuc
Celia Davies
Shannon Pierson
Zoë Kaufmann

Wilson Center

Science and Technology
Innovation Program

January 2021

Table of Contents

Acknowledgements

The research team is grateful for the opportunity to explore this critical topic, which would not have been possible without the support of Jane Harman, Meg King, Liz Newbury, and the Wilson Center staff who supported this project and its launch. We are also thankful that eleven women took the time to share their often painful experiences with online gendered abuse and disinformation in order to raise awareness about the human toll of these campaigns.

Executive Summary

This report strives to build awareness of the direct and indirect impacts of gendered and sexualized disinformation on women in public life, as well as its corresponding impacts on national security and democratic participation. In an analysis of online conversations about 13 female politicians across six social media platforms, totaling over 336,000 pieces of abusive content shared by over 190,000 users over a two-month period, the report defines, quantifies, and evaluates the use of online gendered and sexualized disinformation campaigns against women in politics and beyond. It also uses three in-depth interviews and two focus groups to emphasize the impacts gendered abuse and disinformation have on women's daily lives.

Key conclusions:

- ***Gendered and sexualized disinformation is a phenomenon distinct from broad-based gendered abuse*** and should be defined as such to allow social media platforms to develop effective responses. The research team defines it as "a subset of online gendered abuse that uses false or misleading gender and sex-based narratives against women, often with some degree of coordination, aimed at deterring women from participating in the public sphere. It combines three defining characteristics of online disinformation: *falsity, malign intent, and coordination*."

- ***Malign creativity—the use of coded language; iterative, context-based visual and textual memes; and other tactics to avoid detection on social media platforms***—is the greatest obstacle to detecting and enforcing against online gendered abuse and disinformation.

- ***Gendered abuse and disinformation are widespread.***

  » Gendered abuse affected 12 of 13 research subjects, while nine out of 13 subjects were targeted with gendered disinformation narratives.

  » These narratives were racist, transphobic, or sexual in nature. The overwhelming majority of recorded keywords relating to abuse and disinformation were identified on Twitter and directed towards then-Senator, now Vice President Kamala Harris, who accounted for 78% of the total amount of recorded instances.

  » Sexual narratives were by far the most common, accounting for 31% of the total data collected. The majority of these narratives targeted Vice President Kamala Harris. Transphobic and racist narratives only accounted for 1.6% and 0.8%, respectively. While these numbers appear low, these disinformation narratives are also relatively widespread, impacting a number of different research subjects.

  » Online gendered abuse and disinformation is often intersectional in nature, with abusers often engaging with both sex- and race-based narratives, compounding the threat for women of color.

- ***Online gendered abuse and disinformation is a national security issue.*** In interviews with women targeted by Russian, Iranian, or Chinese state media, the research team found that gendered tropes formed the basis of disinformation and smear campaigns against them.

- » This presents a democratic and national security challenge; as adversaries attempt to exploit widespread misogyny, women may be less likely to choose to participate in public life.

- ***A number of challenges complicate detecting and responding to online gendered and sexualized abuse and disinformation:***

  - » Malign creativity is perhaps the greatest challenge to detecting, challenging, and denormalizing online abuse because it is less likely to trigger automated detection and often requires moderate-to-deep situational knowledge to understand.

  - » "Platform policies lack a coherent definition of 'targeted harassment,' meaning" much of the abuse women face is not violative of platforms' terms of service, leaving abusers to continue their activities without facing consequences. There is also a lack of intersectional expertise in content moderation, which results in abuse toward women, people of color (POC), and other marginalized communities going unaddressed.

  - » Targets bear the onus of detection and reporting. Managing an onslaught of abuse on social media requires time to block, report, and mute abusers. These burdens are discounted and affect their daily lives offline.

- ***Social media platforms, lawmakers, and employers have largely ignored this threat to democracy and national security.*** In order to mitigate the threat, the researchers recommend:

  - » **Social media platforms** should introduce incident reports that allow women to report multiple abusive posts at once to provide more context and a more holistic view of the abuse they are experiencing.

  - » They should also regularly update platform classifiers or keywords to reflect and root out malign creativity, improve automated detection methods, and introduce nudges or friction to discourage users from posting abusive content.

  - » Finally, they should create a cross-platform consortium to track and respond to online misogyny, similar to existing consortiums which counter terrorism and extremism.

  - » **Lawmakers** should include content moderation transparency reporting requirements in social media regulation bills to improve understanding of the problem and introduce accountability for women's online protection.

  - » They should create clear standards that prohibit the use of gendered and sexualized insults and disinformation in official business.

  - » Critically, US lawmakers should reauthorize the Violence Against Women Act (VAWA) and include provisions against online gender-based harassment.

- **Employers** should develop robust support policies for those facing online harassment and abuse, including providing mental health services, support for employees' or affiliates' legal fees and other expenses (such as anti-doxxing service subscriptions).

- They should also outline clear mechanisms for targets to report such campaigns against them to official communications and human resources staff.

Gendered and sexualized abuse and disinformation online is sprawling, and even assessing it in the broadest terms presents obstacles in detection and analysis. As this report indicates, abusers' malign creativity means addressing this problem will not be easy. Dedicated, collective efforts by platforms, policymakers, and employers can elevate this from its misidentification as a special interest issue to a question of the right to equal participation in democracy and public life without fear of abuse and harassment.

Introduction

2021 is a year of firsts for the United States: Kamala Harris is the country's first woman Vice President; Avril Haines serves as the first woman Director of National Intelligence; and Janet Yellen is the first woman to lead the U.S. Department of the Treasury. For the first time, women helm the White House communications team. More women—and more women of color—have been sworn in as Members of the United States Congress than ever before.

But for these women, and many others in journalism, activism, academia, and beyond, political aspirations and engagement in public life come with a tacit cost. Social media platforms, lauded for connecting people, for helping protest movements organize, and for giving up-and-comers in a variety of fields the ability to compete, are also vectors for harm. Disproportionately, that harm—in the form of gendered and sexualized harassment—is directed at women, particularly women of color. Whatever the intent, the cumulative global impact of these online behaviors carries a huge risk not only for women's equality, but for national security, and more broadly, the health of democracy itself.

The genesis of this study was the authors' recognition of the nexus between gendered and sexualized harassment and disinformation online. Disinformation, the use of false or misleading information with malign intent, is a weapon of political influence. It is used against nation states, politicians, and racial minorities, as well as systematically deployed against women in public life. Female candidates in the U.S. Democratic presidential primaries were subject to character attacks more often than their male peers.[1] Malign foreign actors, including the Kremlin, are engaging in organized and ad hoc campaigns to silence and discredit women and discourage their participation in democracies worldwide.[2] [3] Deepfake videos are also being used to silence women; 96 percent of all deepfakes depict women in fabricated, non-consensual pornography.[4] A recent investigation identified a bot on the messaging app Telegram that created over 668,000 fabricated, pornographic images of women without their consent.[5]

There have been many studies of the tactics of online influence and disinformation. However, few consider the specificities of gender-based tactics and their effects on women's engagement in the public sphere. In addition to the targeted harassment that women face online, this report investigates the growing trend of gendered or sexualized disinformation campaigns to which women are specifically vulnerable—regardless of whether discrediting women is the end goal, or whether this serves a broader goal of sowing discord and mistrust within a society. These techniques discourage women from being involved in politics and public discourse, sustaining gender imbalances in representation across a variety of industries worldwide.[6]

This research aims to define, quantify, and evaluate the use of online gendered and sexualized disinformation campaigns against women in public life in order to inform policy responses of social media platforms, governments, and employers. Rather than focusing on a single social media platform, it draws upon data collected across six social media platforms from both the "mainstream" and "alternative" spheres to give a broader snapshot of the threats women in public life faced over a two-month period in the last quarter of 2020.

This report grapples with the distinctions between gendered disinformation and gendered online harassment. Through analysis of the volume and characteristics of online harassment and disinformation, the report assesses the coordination and, where possible, the likely intent of the online vitriol directed at female political leaders, journalists, and activists. It pays special attention to abusers' use of malign creativity—coded language, iterative, context-based visual and textual memes, and other tactics to avoid detection—which compounds the problem and makes responding more difficult. In addition to the data gathered from the social media platforms, the analysis also draws upon in-depth interviews with targets of state-sponsored gendered and sexualized disinformation, as well as two focus groups of female scholars and analysts in information operations, internet governance, human rights, and communications, many of whom have experienced online misogyny themselves.

More broadly, this report strives to build awareness of the direct and indirect impacts of gendered and sexualized disinformation upon women engaged in or aspiring to engage in public affairs, as well as its corresponding impacts on national security and democratic participation. Though this research necessarily has a small sample size, it demonstrates the high level of vitriol to which women in public life are subject, with over 336,000 individual pieces of gendered and sexualized abuse posted by over 190,000 users directed at 13 research subjects during the two-month data collection period. Over half of the research subjects were targeted with gendered or sexualized disinformation narratives, with women of color subjected to compounded, intersectional narratives also targeting their race or ethnicity. While the report demonstrates that this problem is diffuse and, thanks to the use of malign creativity, difficult to detect, it is a problem that demands action. The report outlines the most urgently needed solutions within social media platforms, government, and at the employer or organizational level.

Over half of the research subjects were targeted with gendered or sexualized disinformation narratives, with women of color subjected to compounded, intersectional narratives also targeting their race or ethnicity.

This report is by no means easy reading, but it offers a glimpse into the realities of being a woman online, and a step toward making our public spaces more equitable, democratic, and secure.

Existing Scholarship on Gendered and Sexualized Disinformation

There is a burgeoning field of scholarship focused on harassment and abuse in online spaces. However, academic and policy discussions around gendered disinformation as a distinct type of disinformation remain fairly new. In order to learn from previous work and to effectively position our study, the research team selected and reviewed literature from the disinformation, online abuse, and human rights scholarship. This included scholarship on online abuse and harassment directed at women generally, abuse directed at women active in public life, state-backed influence operations with gendered tactics, and disinformation that employs gendered or sexual language. Engaging with the existing body of work helped the team to better understand the unique online threats facing women, in particular, female political leaders, journalists, and activists.

Women in public service use social media as a tool to gain exposure, to connect with constituents, and to advance their messages on their own terms outside the medium of the news media.[7] Activists and journalists use social media as a part of their jobs: to report on evolving stories, to connect with sources, and to publicize their work.[8] While maintaining a social media presence is now necessary for success in these and other careers, an online presence can be a "double-edged sword" that can open the door to online harassment.[9]

Women are uniquely affected by gendered abuse online. In *Credible Threat: Attacks Against Women Online and the Future of Democracy*, a sociological survey of women in public life, Sarah Sobieraj characterizes online digital abuse against women as expressions of "digital misogyny" and "patterned resistance" against women's full and equal participation in public life. She argues that the attacks are "aimed at protecting and reinforcing a gender system in which women exist primarily as bodies for male evaluation and pleasure."[10] Drawing upon qualitative interviews with victims of gender-based attacks online and professionals in content moderation and internet safety, Sobieraj observes that the women who are subject to the harshest forms of abuse are members of marginalized groups, those who speak out in or about male-dominated domains—such as politics, sports, foreign policy, defense, and cybersecurity—as well as women who are "perceived as feminist or non-compliant with traditional gender norms."[11] Further, she illustrates women of color and other marginalized groups experience some of the most intense online attacks, given that their abuse is often intersectional in nature, as this study corroborates.[12] [13]

Online harassment has a silencing effect on women's voices and limits their ability to actively participate in public discourse. A study from the Data & Society research institute shows that 41 percent of women between the ages of 15 and 29 self-censor to avoid online harassment.[14] The National Democratic Institute (NDI) further documented this phenomenon in a report on the prevalence of online violence against young politically-active women in Kenya, Columbia, and Indonesia.[15] The study measured the impact of the attacks on the women's participation in online political discourse by tracking the Twitter engagement behavior of politically-active women before and after they experienced online attacks. The authors documented whether women posted less frequently on the platform, took a break, or left permanently after experiencing gendered online abuse. Comparing Twitter data with the survey responses from victims, NDI found "strong evidence" that online abuse "decreased women's willingness to continue engaging in social media."[16] Another study from NDI looking at the topic of violence against women in politics denoted the distinction between the direct and indirect effects of online violence against women:

> *While acts of violence against women in politics are directed at individual women, they have an intent beyond their specific target: to frighten other women who are already politically-active, to deter women who might consider engaging in politics, and to communicate to society that women should not participate in public life in any capacity.*[17]

Abuse as a mechanism to subdue women in public service--and deter the future participation of other women in the public sphere--is a recurring theme in scholarship on gendered harassment and disinformation. Sobieraj also documented this phenomenon, noting how abuse pushes some women to censor themselves online by avoiding certain topics, softening their opinions, limiting their participation in and even entirely opting-out of political discussions online because of the "unreasonable burdens of participation."[18] This withdrawal from public conversation is another tactic women feel compelled to use in order to protect themselves from online abuse. One of Sobieraj's interviewees noted: "In order to participate [professionally] you have to wade through all this filth and it [...] take[s] longer for women to just do the basic function of participating, and that's even when you take out the fear for one's safety, or the psychological effects it might have on you, or the stress that it causes."[19]

Disinformation is one form of such abuse, but within this body of research, there is no standard definition of disinformation targeted by gender. Some scholars consider gender to be simply one understudied aspect of information operations, rather than a specific kind of disinformation.[20] Others have adopted the term "gendered disinformation," expanding on the broad description of the term set by Nina Jankowicz, the lead author of this study, in her 2017 reporting: "[a mix of] old ingrained sexist attitudes with the anonymity and reach of social media in an effort to destroy women's reputations and push them out of public life."[21] However, follow-on studies employ a distinct definition. Di Meco describes gendered disinformation as "the spread of deceptive or inaccurate information and images against women in politics, following story lines that often draw on misogyny and distrust of women in politics, frequently referring to their sexuality."

NDI's study defines it as "information activities (creating, sharing, disseminating content) which: (a) Attacks or undermines people on the basis of their gender; (b) Weaponizes gendered narratives to promote political, social or economic objectives."[22] These definitions do not establish how gendered harassment and disinformation differ. For example, NDI argues that gendered disinformation "exists at the intersection of disinformation with online violence, such as abuse and harassment: it seeks impact primarily at the political level, though can also cause serious harm at the personal level."[23] We argue that disinformation is a subset of online harassment, a distinction that is critical as social media platforms and governments attempt to determine the broader disinformation phenomenon's impact on society and root out malign content.

Finally, the majority of studies investigating online harassment directed at women in public life, as well as the few looking at gendered disinformation, rely on datasets that are limited to a single platform, Twitter, as Twitter's public API allows easy external data collection. The reality is that women endure online abuse beyond Twitter, and gendered disinformation narratives often thrive on alternative platforms. This study includes mainstream social media outlets like Reddit as well as niche platforms such as Parler, Gab, 4chan, and 8kun (formerly 8chan). This is the first study of this scope and breadth on the topic of gendered disinformation, and yields unique insights into how gendered online abuse occurs in less-moderated settings.

Defining Online Gendered and Sexualized Disinformation

There are a number of unique, discernable narratives and tactics employed as part of online gendered disinformation campaigns. To understand gendered disinformation, it is important to first understand disinformation more broadly. Disinformation is false or misleading information shared with malign intent. When social media companies root out disinformation on their platforms, however, content moderators tend to rely on disinformers' tools and tactics, rather than veracity or intent, to decide whether content should be removed. In the context of foreign influence operations or domestic disinformation campaigns, it is often much quicker and simpler for platforms to determine whether a group of accounts has violated spam policies or has used fake accounts or machine amplification—a few examples of "coordinated inauthentic behavior"—than it is to determine intent or accuracy.[24] As such, coordination is becoming an increasingly important proxy indicator of disinformation campaigns.

Based on our cross-platform data collection, interviews, and focus groups, this study defines gendered and sexualized disinformation as a subset of online gendered abuse that uses false or misleading gender and sex-based narratives against women, often with some degree of coordination, aimed at deterring women from participating in the public sphere. Our definition of gendered and sexualized disinformation combines three defining characteristics of online disinformation: **falsity, malign intent, and coordination**.

Falsity and coordination are relatively straightforward. But how can platforms or policymakers uncover the sometimes nebulous or tacit coordination of these campaigns, or determine the intent of anonymous online accounts? Sobieraj argues that "digital misogyny is an extension of the history of attempts to curtail women's freedom to use public spaces as equals."[25] That is, the attacks themselves telegraph their intent. Sobieraj observes that "aggressors repeatedly draw upon three overlapping strategies—intimidating, shaming, and discrediting—to silence women or to limit their impact in the digital publics."[26] Intimidation tactics include threats to women's physical safety, often in the form of rape threats, death threats, and other threats of violence, DDoS attacks, spam, and doxxing. Shaming tactics "exploit double standards about women's sexual behavior and physical appearance to taint targets," often through leaks of private information or images and peddling of false information. Attacks on women's credibility communicate that they are unfit for public life, especially elected office, and that their views should not be taken seriously. Sobieraj stresses that "gender is [...] at the very center of the attacks themselves. Femininity and femaleness are the weapons of choice used to undermine women's participation and contributions [...] As in the physical publics - the body is the locus of abuse."[27] Further, NDI has identified patterns in the way state-aligned gendered disinformation attacks women and perpetuates gendered narratives in the Philippines and Poland.[28] In particular, the authors observed the use of sexualized online harassment and campaigns amplifying narratives that women are too stupid and untrustworthy to hold public office, among others.

Our definition of gendered and sexualized disinformation combines three defining characteristics of online disinformation: falsity, malign intent, and coordination.

This report expands upon these case studies, examining patterns in gendered disinformation in the US, Canada, the United Kingdom, and New Zealand. Despite definitional nuances, all the studies referenced here agree that gendered and sexualized disinformation is detrimental to women's equal participation in the public sphere and to democracy.

Platforms' Attempts to Shield Women from Online Abuse and Disinformation

Social media companies' Community Guidelines seek to limit the gendered harassment which women face online. However, many of these policies, especially those concerning harassment, showcase the failures of "one-size-fits-all" tech policymaking. In an ideal world, these policies would be the first and best recourse available to women targeted by harassment. Too frequently, however, they are created by cisgender white men for users like themselves. As a result, they often fail to imagine the myriad and unique attacks which abusers employ against women and people of color in the public sphere, as well as the effect these attacks have on their targets. Furthermore, they do not address how gendered and sexualized disinformation threatens women's careers, reputations, and participation in public life.

The most popular social media platforms share a common set of content restrictions. Facebook, Twitter, Youtube, and many other major companies ban hate speech, harassment, promotion of violence, and abuse. Most platforms remove offensive content and remove users who repeatedly violate their Terms of Service or Community Guidelines. However, each company has a unique, platform-specific user code of conduct.[29]

Twitter

Twitter forbids its users from posting threats of violence, abusive behavior, hateful conduct, glorification of violence, promotion of terrorism or violent extremism, endorsement of suicide or self harm, or spam, among other restrictions. The platform's Terms of Service prohibit posting coordinated abuse, as well as false or manipulated content, restrictions which should theoretically limit the spread of harmful content including disinformation. Crucially, Twitter distinguishes between **technical** and **social** coordination for abusive behavior, imposing different punishments for disparate offenses. Users commit technical coordination violations by manipulating Twitter as a platform to artificially spread a message i.e. one person tweeting from multiple accounts. Social coordination, on the other hand, is when users congregate on or off the platform to amplify a specific message. For example, a group might use Twitter DMs to coordinate their tweets, or a single user could encourage their followers to "dogpile," meaning to direct abuse at another user. While Twitter bans all technical coordination, it permits social coordination unless it encourages harm, creating a higher barrier for women to prove that abusive content should be removed.

For example, and as detailed in case studies below, a popular user might direct their followers to "dogpile" a female journalist, flooding her mentions with hate. Twitter's moderation policies would require the journalist to individually report the tweets and prove that she has been harmed in a serious way, rather than merely proving to moderators that the popular instigator initiated social coordination and therefore should be punished. Unfortunately, it is rare for users to employ social coordination against a single person in a positive way. Additionally, tweets containing gendered disinformation often fall under the purview of the harassment policy simply because the platform does not have an overall disinformation policy. Instead, tweets containing false information are labelled or have click-throughs added.[30] Twitter also prohibits tweets which "deceptively promote synthetic or manipulated media that are likely to cause harm," and adds labels to other such content.[31]

Facebook

Facebook maintains similar overall protections to Twitter, prohibiting users from promoting or threatening violence or sexual assault, posting disinformation or inauthentic behavior, using hate speech, harassing others, or spreading disinformation which could cause harm. However, Facebook's content moderation too often fails to protect users. Its Community Standards are opaque and for the most part do not indicate what actions moderators might take against offenses. Even in the few instances where Facebook's policies are clear, they indicate the extent to which the platform's moderation is ad-hoc and uneven. For example, Facebook's prohibition of "cruel and insensitive" content is unclear, while its restrictions on regulated goods are specific, forbidding users from buying or selling guns, drugs, alcohol, tobacco, live or endangered animals, human blood, weight loss products, and/or historical artifacts.[32]

Even where the company's general moderation policies should prevent many forms of gendered harassment and disinformation, conflicting, vague policies compound the situation. For example, in order to facilitate discussion about public figures, the company applies different standards to abuse directed against public versus private individuals. While the platform only removes severe harassment against public figures, its Terms of Service do not protect most degrading content targeting private citizens, including "claims about someone's sexual activity." For most female political candidates, journalists, or activists (who are, by nature of their work, public figures), this distinction is meaningless: discussion of their sexual activity is harmful, intended to undermine their credibility, and dissuade other women from becoming involved in public life. Additionally, Facebook's stance on violent threats against public figures is largely ineffectual, as death threats are theoretically banned under the prohibition on promoting "high-severity violence," but less severe threats are only prohibited when targeting minor public figures or private citizens.[33] The platform thus largely ignores public figures as targets of violent threats, as its rules for the most part only apply to attacks on private citizens or low-level public figures.[34] Worryingly, this excludes a significant amount of gendered abuse directed at female candidates from the jurisdiction of Facebook moderators.

YouTube

Like Facebook and Twitter, YouTube forbids spam, harassment, hate speech, violence, "harmful or dangerous" content, explicitly sexual content, content which promotes violent criminal groups, and inauthentic engagement. Its harassment policy makes exceptions for topical discussion of high-profile issues about public figures, as well as satire and harassment awareness. Regardless of whether a user is a public figure, however, YouTube treats malicious insults against protected identities more harshly. This could theoretically extend to gendered harassment against women in public life, although there is limited evidence to indicate YouTube uses the policy for such purposes.

Unlike other major social media platforms, YouTube follows a clear "Three Strikes" system for content moderation. The first violation results in a warning; repeated violations lead to "strikes." Each strike involves a freeze on certain account features: for example, users with a first strike will not be able to upload videos or engage in certain activities for a week, while users with a second strike cannot post content for two weeks. Accruing three strikes within 90 days results in channel termination. YouTube also will cut off advertising revenue opportunities from videos and channels as a punishment for high-profile creators who violate Community Guidelines. While YouTube's policies are focused on interpersonal conduct, these enforcement mechanisms primarily limit spam: from April to June 2020, only 3.8% of terminated channels were banned for abusive content or harassment, while 92% of removals were because of spam.[35]

Other Platforms

While Facebook, Twitter, Instagram, and other popular sites like TikTok and Twitch have similar user policies, more niche sites like 4chan, 8kun, and Gab have minimal limitations, trading on their broader definition of free speech. Because these sites generally only restrict content which violates the law, they tend to attract users interested in content forbidden on more mainstream platforms, with white supremacist content as a notable example. Such sites tend to employ fewer restrictions on both content and users: 4chan and 8kun do not require users to create accounts, and function on a system of virtual anonymity. Other moderation policies have significantly evolved with time: Reddit's bare-bones user policies were originally more closely aligned with 4chan and 8kun, but its increasing popularity and aggressive content moderation surrounding QAnon has led to an exodus of users who were engaging with harmful content.[36] Additionally, while the site has had serious issues regarding racism and white supremacy, its main user base is not composed of those groups.[37] Finally, Parler's marketing is somewhat misleading in that it is advertised as an alternative site but its Terms of Service are stricter than Twitter's, even as it aims to attract frustrated Twitter users who feel they are being censored.[38]

Cross-Platform Moderation and Enforcement Issues

Measures to protect women from harmful, abusive, gendered and sexualized harassment and disinformation doexist, but barriers at every level prevent their proper enforcement. These issues begin with reporting, the first step for moderation. For example, Facebook only permits targets themselves to report prohibited behavior, putting the onus of reporting on the harassed, and forcing them to re-experience abuse and potentially re-traumatization. Similarly, Twitter's reporting mechanism can be critically cumbersome for attacks that come en masse. Other platforms have little to no moderation at all.

Enforcement processes, which are uneven and opaque across platforms, also present a challenge. To make decisions, content moderators for the individual companies generally rely on internal documents, which are much more specific than publicly available policies, are sometimes contradictory, and change frequently.[39] Further, contract workers making these decisions have competing incentives to flag and act on content, often under difficult conditions. They must attempt to protect users, reconcile contradictory instructions, and make decisions in seconds while working in harsh psychological and physical environments.[40] There is also evidence that final enforcement decisions may sometimes be influenced by politics: Facebook has violated its own policies to protect high-profile conservative leaders from moderation,[41] and Twitter has been criticized for acting more quickly on minor harassment targeted at President Donald Trump than on more severe abuse targeted at progressive female politicians.[42]

Finally, while enforcement mechanisms might remove a share of the abuse, under the current framework, the costs of abuse fall almost entirely on the targets. Abusers might have their accounts temporarily locked or eventually permanently removed, but they can easily create new accounts or attack targets on different platforms. Targets must endure not only the abuse but also its aftermath; reporting processes and protection of targets' own personal information from abusers is practically complex and emotionally exhausting.

Some social media platforms have taken steps to give users greater control over their online experience and to fend off harassment in less burdensome ways by offering customized opt-in moderation features. For example, TikTok now offers its creators a variety of safety controls for their videos. Comment filtering settings are available: creators can switch on a filter that automatically "hides" spam and offensive comments from appearing on their videos. Another filter option allows for creators to specify specific keywords they wish to be filtered out from their videos' comment sections, which they can use to police coded or target-specific language not typically detected by TikTok's moderation systems, effectively customizing their moderation experience on the app. Users also have control over who is allowed to send them direct messages and who can view, comment on, and download their videos--options which include sharing to everyone, friends, or just themselves. Twitter also filters out low-quality content from users' replies and allows users to mute certain keywords from appearing in their timelines. While neither TikTok nor Twitter have released data on how effective these measures are at reducing online abuse on their platform, these features serve as tools that targets can use to create the semblance of a safe personal environment on the platform. These features still place responsibility for moderation on victims, but TikTok in particular, with its mission of being the "last sunny corner of the internet," gives users more tools and agency to manage online harassment in simple and customized ways that lighten their load.

Methodology

This exploratory research takes a sequential mixed methods approach, drawing upon quantitative and qualitative data sources. Key among these were: cross-platform social media data gathered from six social media platforms; existing literature in the field; semi-structured interviews; and focus groups.

Research Questions

This project aims to define and evaluate the use of online gendered abuse and disinformation against women in public life in order to inform policy responses by governments and social media platforms. Three research questions informed this project's scope:

- ***RQ1: Definition.*** How should online gendered and sexualized disinformation against women be defined in practice? What is the difference between gendered or sexualized disinformation and gendered or sexualized harassment?
- ***RQ2: Prevalence.*** How prevalent is gendered and sexualized disinformation against female public figures online?
- ***RQ3: Tactics.*** What are the common tactics used to convey gendered and sexualized disinformation? What are the primary narratives or themes used in gendered and sexualized disinformation? Do these narratives travel across platforms? On which platforms do they gain most traction?

The current analysis cannot and does not seek to attribute intent to users who create or disseminate the disinformation examined below; nor does it seek to evaluate whether or not a user genuinely believes the narrative they are generating or spreading. This research focuses primarily on the content and target of the abuse, with a view to understanding its impact on women engaged in public life.

Time Frame and Geographical Scope

Data collection occurred between September 1 to November 9, 2020, so chosen as to track the abuse towards the subjects throughout the historic 2020 election in the United States. The collection period also captured the 2020 New Zealand election, in which incumbent Prime Minister Jacinda Ardern, a subject in this study, competed and won. As all research was collected online and studied women from four countries, no limitations were placed on the geographical scope of the data collected; however, only English-language data was gathered.

Subjects

Selection Process

The research team chose 13 female politicians from English-speaking countries as research subjects for this project. The subjects were chosen based on their participation in public discourse and likely or known exposure to online disinformation campaigns. Over the course of the project, a small number of candidates were excluded or added based on current events or whether sufficient data was available for analysis.

The team collected data on six US House of Representatives candidates and two Senate candidates—representing three Republicans and five Democrats—as well as Senator Kamala Harris during her successful Vice-Presidential bid. The team also included Michigan Governor Gretchen Whitmer in data collection, given the sustained threats against her during the coronavirus pandemic.[43] Three other international politicians across the political spectrum in English-speaking countries were also included in order to provide a comparative assessment of the gendered disinformation environment.

Subjects were selected to represent diversity in political affiliation, race, ethnicity, and levels of visibility. In order to explore the intersection between gender and race, the research team included women from diverse backgrounds; however, the team recognizes that future research will require a wider look at diversity and integration of an intersectional approach.

The following individuals comprised the final list of subjects for the project:

International Politicians (3)

- Prime Minister Jacinda Ardern (New Zealand)
- Secretary of State for the Home Department Priti Patel (UK)
- Deputy Prime Minister Chrystia Freeland (Canada)

US Politicians (10)

- Senator Susan Collins (R)
- Senator Kirsten Gillibrand (D)
- Senator and Vice-President-Elect Kamala Harris (D)
- Rep. Jaime Herrera Beutler (R)
- Rep. Alexandria Ocasio-Cortez (D)
- Rep. Ilhan Omar (D)
- Rep. Elissa Slotkin (D)
- Rep. Elise Stefanik (R)
- Rep. Lauren Underwood (D)
- Governor Gretchen Whitmer (D)

Quantitative Data Collection

The research team collected data from six social media platforms, selected based on size of user base and ideological variation in users: Twitter, Reddit, Gab, 4chan, 8kun, and Parler. Facebook and Instagram were also considered for collection, but were not included as both platforms employed data collection limitations, which in turn limited our ability to collect data in a structured, sustainable manner.

For all platforms, a list of keywords reflecting abuse or disinformation was built to inform data collection. These lists were grounded in colloquial misogynistic slurs in the English language (i.e., "bitch," "slut," etc.) and tailored to include specific keywords, nicknames, or hashtags that reflected individualized abuse directed at each subject (see Appendix C for the full list of terms used). Keywords reflecting this individualized abuse were largely informed by scoping research via CrowdTangle and incorporated into our retrospective two-month data collection on November 11, 2020. The team also undertook a preliminary data collection and analyzed the results for any slurs, keywords, or disinformation narratives that had initially been missed, and incorporated these into our final collection criteria.

The team undertook a standardized collection approach for Twitter, Reddit, Gab, 4Chan, and 8Kun. In order to be included in the final data set, any posts collected from these platforms were required to contain the subject's name, online handle, or any other known moniker, as well as one or more terms on the subject's tailored list of abusive keywords. If unique terms or hashtags were created to abuse the subject directly, the presence of this term alone qualified the post for inclusion in the final data set (e.g., "camel-toe Harris" or "heels up Harris" for Kamala Harris). The exception to this keyword collection was on Parler, where the platform's interface allowed only for the collection of hashtags, and not keywords. Lists containing abusive hashtags for each subject were also therefore created for collection where possible.

Finally, the data collected was separated into "results" and "data points." **Results** encompass all the posts collected from all platforms. These results were then analyzed against the list of abusive keywords for each subject and separated into categories reflecting those keywords. References to "data points" do not correspond to the number of posts collected, but rather the unique keywords identified within those posts. The number of total data points collected is therefore higher than the total number of results collected, as some results contained multiple keywords.

Limitations

Parler

Data collection for the hashtags "#AOC" and "#KamalaHarris" faced technical issues due to the overwhelming volume of content posted; as a result, no data containing either of these hashtags was collected, which is recognized as a gap in the data.

Twitter

The team found instances of apparent false positives wherein the Tweet body did not appear to match the collection criteria. Further investigation revealed that the keywords were occurring in Quote Tweets--a Tweet written by another user, retweeted with a comment. This dataset suggests a recurring pattern, whereby users quote an abusive Tweet with a mention of the research subject's name or handle. For the purpose of this analysis, Twitter data was categorized as follows:

(i) The primary group consists of Tweets in which the original Tweet body matches the collection criteria; users who posted these Tweets are categorized as primary users engaging in conversations involving potential sexual abuse or disinformation regarding our research subjects. This is the group used in the data analysis.

(ii) The secondary group consists of the Tweets in which the Tweet body does not include the collection criteria, and is dependent on the Quoted Tweets for proper matching. These are secondary users not necessarily directly involved in the conversation, though they may share similar sentiments. While these Tweets have been excluded from the current analysis, the behavior trend is indicative of the wider engagement in these conversations.

Network Visualization Methodology

The visualization of cross-platform networks enables better understanding and representation of the content sharing behaviors. Data in the network visualizations corresponds to data points, posts containing abusive or disinformation narratives according to keyword lists (see Appendix C), as described above. Zooming into the most abusive content or narratives within the networks allowed exploration of user behavior, including whether and how users disseminated their own messages individually and alongside the messages of others. Network visualizations were created to capture two different types of interactions: user-to-abusive keyword relationships (captured on all social media platforms, such as when a number of users publish the word "tranny" on both Twitter and 4chan in Figure 2) and user-to-subreddit and board relationships (captured on Reddit, 4chan, 8kun, and Gab, such as when users post abusive fantasies to multiple subreddits in Figure 7).

Data Ethics

Privacy and Consent

The current analysis draws on publicly available data; all research subjects are high-profile public figures both offline and online, and correspondingly expectations of privacy are reduced.

Collation and Pseudonymization

Following the completion of data collection, data was collated, de-duplicated, and pseudonymized in accordance with the EU's General Data Protection Regulation. All Personally Identifying Information was pseudonymized (username fields, link URLs, and unique identifiers). In the case of Twitter, user handles (i.e.: @username) in Tweet bodies were also pseudonymized. The exception to this were the handles of the research subjects.

Qualitative Data Collection

In this mixed method design, the research team also sought to supplement social media data with a deeper exploration of the lived experiences of women who have been subject to online misogyny at the state-sponsored and domestic levels. This occurred through two means of data collection: interviews and focus groups. Lead researcher Nina Jankowicz conducted three semi-structured interviews with well-known female journalists who have been subject to known, overt, state-sponsored disinformation campaigns emanating from Iran, China, and Russia—Yeganeh Rezaian, Leta Hong Fincher, and Nicole Perlroth, respectively. Given limited access to data attributing online harassment campaigns directly to state entities, the research team chose these journalists due to the direct nature of their targeting by state-sponsored media or government agencies. The sample size was kept small due to limited resources to conduct, transcribe, and analyze the interviews. The interviews were conducted on-the-record, lasted between 45-60 minutes, and were recorded and fully transcribed with interviewee consent. In addition to describing their experiences, each interviewee was asked a series of questions (see Appendix D) about characteristics of gendered disinformation they have experienced and observed. These responses informed the research team's definition of gendered and sexualized disinformation.

Research team members Alexandra Pavliuc and Nina Jankowicz also conducted two focus groups with a total of eight female scholars and analysts who study disinformation, including two women of color and one representative of a marginalized group. The researchers issued a broad invitation to 30 women in media, academia, and policy analysis who have focused their work and engagement on disinformation or issues adjacent to it; the eight respondents participated in one of two 60-minute focus groups. They discussed their personal experiences with online misogyny, attempted to define their experiences, and workshopped solutions to the problem (see Focus Group Guide, Appendix E). Contributions to the focus groups were not for attribution. The focus groups were recorded and fully transcribed, with participant consent.

Data Analysis

Overview

The data collected can be broadly separated into categories of gendered abuse, uncoordinated disinformation, and coordinated disinformation. Gendered abuse involves the often casual use of derogatory terms aimed at degrading or insulting women based on gender. The gendered abuse recorded throughout this project ranged from name-calling to sexually violent threats. One widespread example of such abuse is the frequent reference to Alexandria Ocasio-Cortez's former job as a bartender, which abusers used in attempts to undermine her political qualifications and express misogynistic views. For example, in response to Ocasio-Cortez's attempts to block Trump from picking a new Supreme Court Justice, one user wrote: "Suddenly the slut bartender is now a constitutional scholar." In the data collected, gendered abuse was more widespread across all of the subjects than disinformation. While this type of language is undeniably problematic and deeply harmful, this analysis will focus primarily on coordinated or uncoordinated disinformation directed at the research subjects.

In the context of gender, disinformation involves the spreading of rumors or alleged "facts," often of a sexual nature, in order to humiliate, discredit, or disempower the subjects. These campaigns could be either coordinated or uncoordinated. A coordinated disinformation campaign is one which is intentionally conducted by a person or group of people, who may or may not believe in the narrative. Though not of a sexual nature, one example of coordinated disinformation was a campaign orchestrated in September 2020 by Project Veritas, which claimed to have uncovered evidence that Congresswoman Ilhan Omar was orchestrating a widespread ballot harvesting scheme. Whether or not Project Veritas or their collaborators truly believed the story, their coordinated efforts to spread it had a significant impact on Omar's public reputation and made her a target of increased abuse online. The day the story was released, abuse against Omar rose 1,871 percent.[44]

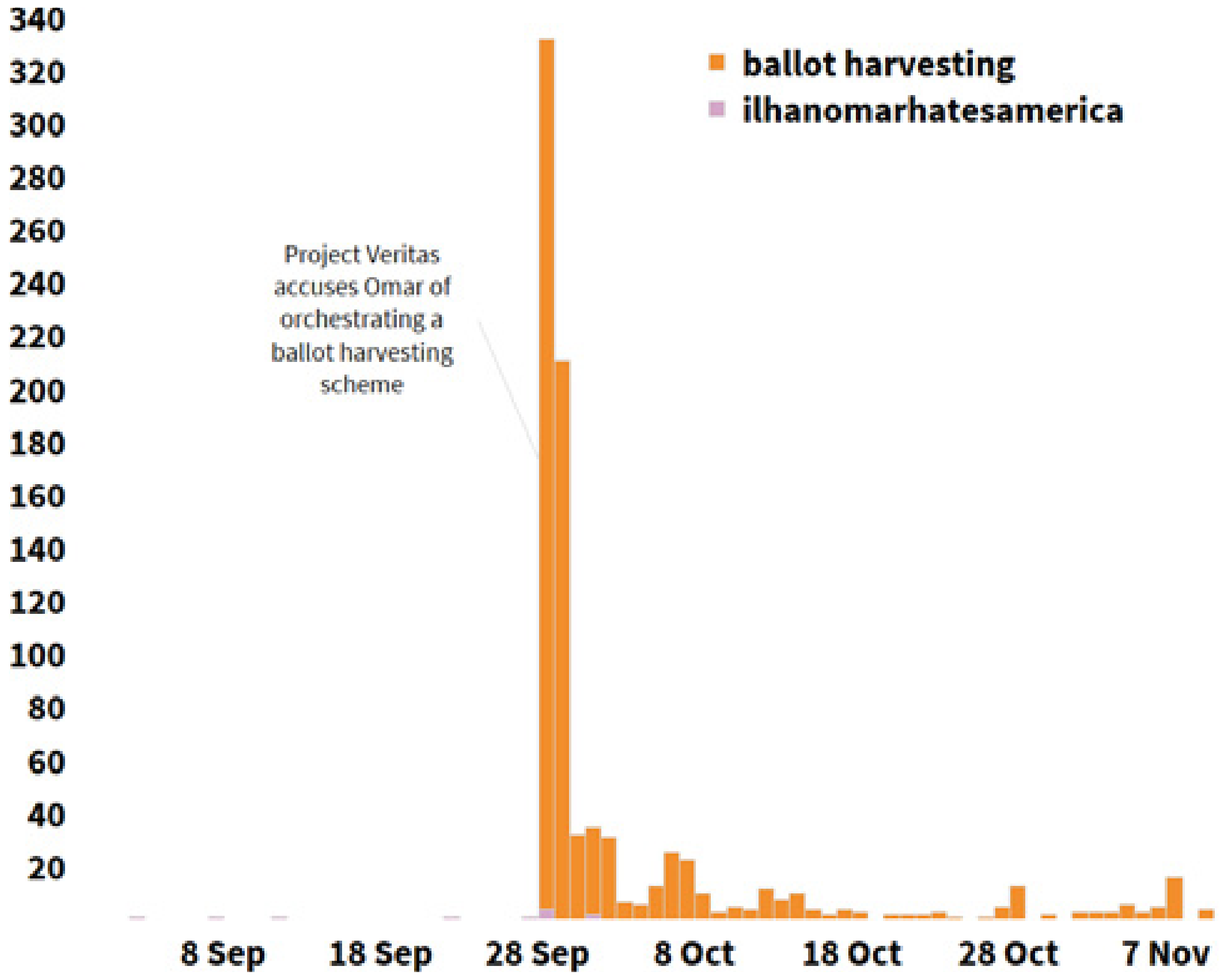


Figure 1: A spike in abuse against Ilhan Omar occurs in response to a ballot harvesting disinformation campaign

By comparison, uncoordinated disinformation is unplanned, but can be equally detrimental. Part of its impact comes from the difficulty of identifying the originating source: whereas coordinated disinformation may have a clear origin (e.g. Project Veritas), uncoordinated disinformation is often difficult to trace and spreads quickly. Uncoordinated disinformation may take the form of "dogpiling," wherein a rumour is spread about a subject by multiple users with various intentions. While some users may spread the narrative out of malice and an intention to harm the subject, others may simply pass it on in the belief that it is reliable information. Governor Gretchen Whitmer recently became the subject of an uncoordinated disinformation narrative when she displayed a pin with the numbers "8645" on her desk during a video interview. The slogan "8645" has been used for several years to refer to removing Donald Trump from office, as 86 means "to get rid of/throw out" in restaurant parlance and 45 refers to Trump's status as the 45th President.[45] Some users, however, misinterpreted 86 as a call for assassination and spread the message that Whitmer was openly calling to have Trump killed. While the source of this narrative is unknown and does not appear to have been coordinated, it was nonetheless extremely powerful: mentions of "8645" reached over 6,000 instances on the day of Whitmer's interview and were echoed by the Trump campaign.[46] Importantly, while it is theoretically possible to define the difference between a coordinated and uncoordinated disinformation campaign, it is not always possible to distinguish between the two in practice. Arguably, the most successful coordinated disinformation campaigns are those which appear organic.

The following analysis examines the primary disinformation narratives identified by the research team, key findings from the platforms examined, and unanticipated results.

Top-Level Quantitative Findings

- Between September 1 and November 9 2020, over 336,000 data points of gendered abuse and disinformation were posted on the platforms monitored by the research team. Of the 13 research subjects, the overwhelming majority of recorded keywords relating to abuse and disinformation were directed toward Kamala Harris, accounting for 78% of the total number of recorded instances.

- The research team identified three overarching types of disinformation narratives that impacted multiple subjects: **sexual, transphobic, and racist**. Sexual disinformation narratives were the most common, accounting for approximately 5% of the total data collected, while transphobic and racist narratives accounted for 1.6% and 0.8%, respectively. While the frequency of these narratives appear low compared to generalized gendered abuse, it is critical to recognize the potential for emotional and psychological harm caused by the spread of disinformation.

- Generalized gendered abuse online was far more widespread than disinformation. Terms indicating gender-based abuse accounted for 50.4% of the total data collected, with "bitch", "witch", and "ugly" as the top three keywords recorded. Additionally, while 12 of the 13 subjects received gendered abuse online, the research team identified active disinformation narratives employed against 9 subjects.

- "Sex," "Bitch," "Sexy," "Witch," and "Ugly" were the top five recorded keywords by volume across all platforms and subjects. The research team found in some cases that certain keywords did not correspond directly to gendered abuse of the subjects. For example, in the case of Elise Stefanik, the majority of posts including the keyword "sex" were in reference to her support of Donald Trump despite the accusations of sexual abuse leveled against him. These terms were therefore excluded from subject analysis where necessary.

Overarching Disinformation Themes

The research team identified disinformation narratives for 9 of the 13 subjects. Eight of these subjects were the targets of specific and personal disinformation narratives. The research team identified three overarching themes across the narratives.

Transphobic Narratives

Four subjects were targeted with disinformation narratives in which users asserted that they were secretly transgender women. This narrative targeted Kamala Harris, Gretchen Whitmer, Alexandria Ocasio-Cortez, and Jacinda Ardern. It was frequently supported by a photograph or video as "proof" of the subject's gender deception.

- A photo of **Kamala Harris** was circulated which positioned her side by side with her alleged former identity, a man named Kamal Aroush. "Aroush," in reality a photoshopped image of Harris, was given a backstory to increase the legitimacy of the narrative. This image appears to have originated as part of a QAnon campaign. The narrative was promoted by one user on Gab in 33 separate posts (see large orange arrow in the network visualization below).

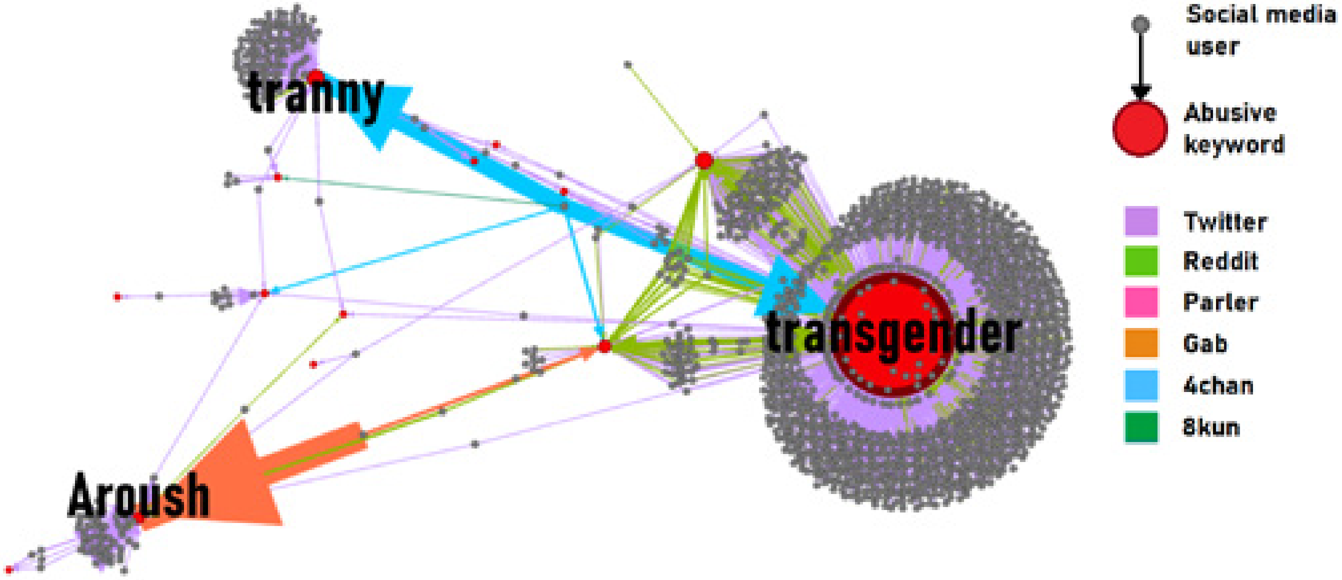


Figure 2: Keyword network of users interacting with transphobic terms towards Kamala Harris, including the "Kamal Aroush" disinformation narrative

- This narrative also targeted **Jacinda Ardern** after a video showed a pleat in her dress interpreted as evidence of male genitalia.
- **Gretchen Whitmer** has had her facial features compared to transgender celebrity Caitlyn Jenner, with users asserting that her bone structure is "proof" of this identity.
- No purported "evidence" has been spread alleging that **Alexandria Ocasio-Cortez** is transgender, but users online frequently target her with derogatory terms including "tranny" or "transsexual" as accusations or insults.

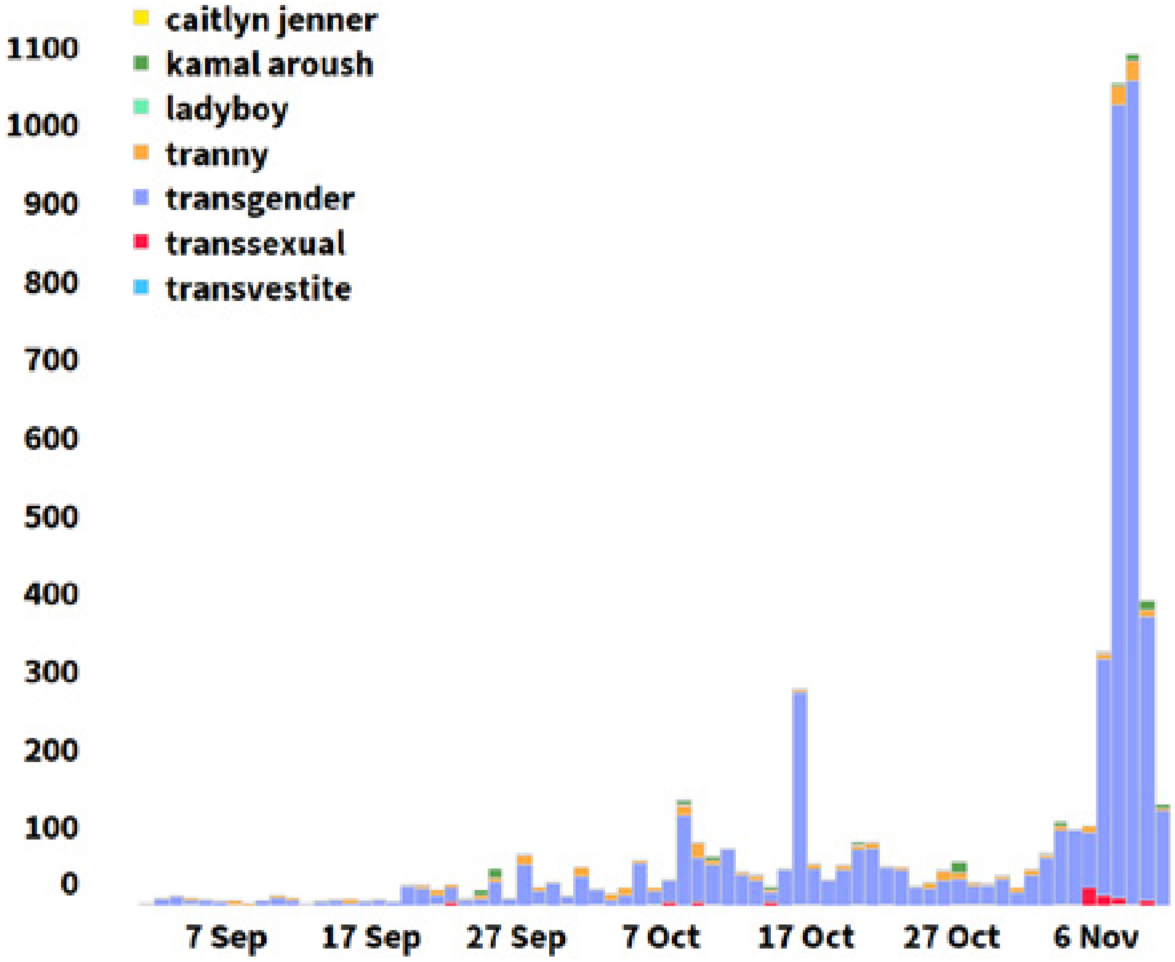


Figure 3: Frequency of terms related to transgender disinformation narratives. While a spike was observed following the US election, it should be noted that a portion of these are false positives and relate to discussion around Joe Biden mentioning transgender people in his victory speech.

The "secretly transgender" narrative is a longstanding fixture of gendered online abuse. This rumor targeted Michelle Obama throughout and beyond the Obama Administration, asserting that she was formerly a man named Michael. At their foundation, these narratives tap into the trope of the duplicitous woman, implying that not only are transgender individuals inherently deceptive, but that this deception is responsible for the power and influence that these women hold. To this end, the narrative is also deeply misogynistic in its assumption that women cannot gain power without trickery. Proponents of these disinformation campaigns appear to assume that transgender identities, especially "hidden" ones, are so abhorrent that once the truth is revealed these women will lose all credibility and power.

The data collected suggests that this transphobic narrative is targeted primarily at higher-profile political figures. Harris, Ocasio-Cortez, Whitmer, and Ardern ranked first, second, fifth, and ninth respectively in the number of data points collected related to transphobic narratives. All of these subjects have publicly challenged traditional political spheres: Harris as the first Black, South Asian, woman Vice President; Ocasio-Cortez as a young Latina Congresswoman, Ardern as a young female Prime Minister who gave birth while in office, and Whitmer as a female Governor who imposed significant restrictions during the COVID-19 crisis. Considering these trends, the "secretly transgender" narrative may serve two possible goals. The first is to strip these women of their power and attractiveness, since the transphobia inherent in this narrative dictates that transgender people can be neither attractive nor powerful. The second is to justify their political success, as the misogyny inherent in the narrative dictates that women, particularly young and attractive women, cannot rise to power without deception or male characteristics. It is also possible that this narrative is a byproduct of the sheer volume of abuse received by high-profile political women.

Racialized and Racist Narratives

Gendered abuse online also manifested in racist or racialized disinformation. Of the 13 subjects examined, five are women of color who were subjected to racist abuse. Three of these women were targeted with racist and racialized disinformation.

- This was most clearly observed in reference to **Ilhan Omar**, whose Black and Muslim identities were weaponized to portray her as a dangerous foreign "other." Abuse targeting Omar's ethnicity and religion has manifested in a multitude of ways since she entered the political sphere. In 2019, a photoshopped image circulated purporting to show her without her hijab, revealing a balding head with unkempt hair. The photo aimed to humiliate Omar based on her appearance, religion, and ethnicity. Narratives identified in the data collected, however, appear to have shifted from attempts at humiliation towards a portrayal of Omar as a terrorist and political saboteur. These are grounded in Omar's identity as a refugee and her connections with other Somali immigrants and attempt to cultivate and capitalize on mistrust of Black and Muslim communities. One such disinformation campaign asserted that Omar was orchestrating widespread ballot fraud in Minnesota with the help of the Somali community, while another more sexualized narrative claimed that she immigrated illegally by marrying her biological brother. The

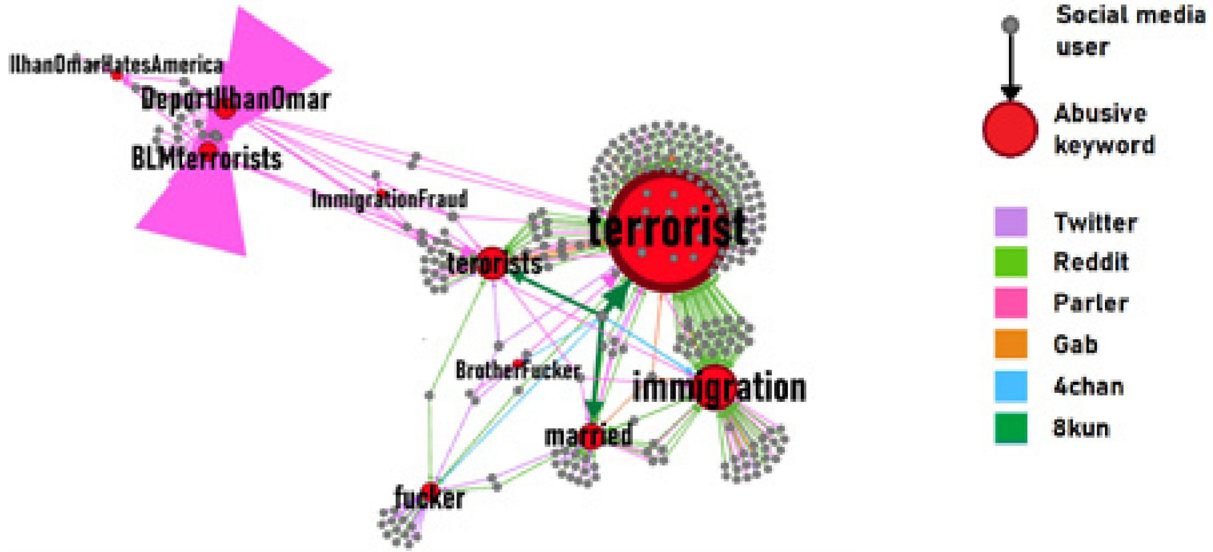


Figure 4: Keyword network of users interacting with racist terms about Ilhan Omar, including disinformation narratives about her allegedly "illegal immigration," related to false accusations that she had married her brother.

latter narrative plays on the taboo of incest to portray Omar as a foreigner who does not adhere to American cultural norms and by extension, cannot represent the American people. As depicted in the network visualization below, a small but vocal group of users on Parler posted a number of racist hashtags such as #DeportIlhan-Omar (with one user responsible for 91 uses of the hashtag). A wider array of users posted keywords relating to the false narrative about Omar marrying her brother on Twitter, Reddit, 4chan,and 8kun (with one user engaging extensively with this narrative on 8kun).

- A similar narrative targeted **Kamala Harris**, whose citizenship status was called into question throughout her political campaigns. Users focused on the fact that Harris' parents are both immigrants, incorrectly arguing that Harris is not a natural-born citizen and is therefore ineligible to run for office. Additionally, the minority identities of both Harris and Ocasio-Cortez were used as sources of criticism and delegitimization in several instances. One narrative insisted that as a Black and South Asian woman, Harris could speak for neither group, while others online accused her of exaggerating her racial identities in order to further her political goals. This resulted in the hashtag #KamalaAintBlack, which was recorded 657 times throughout the collection period. **Alexandria Ocasio-Cortez** was targeted with similar criticism when it was revealed that she used to go by the nickname "Sandy." The hashtag "Sandy Cortez" subsequently spread with the aim of delegitimizing Ocasio-Cortez's identity as a working-class woman of color and reframing her as a privileged politician.

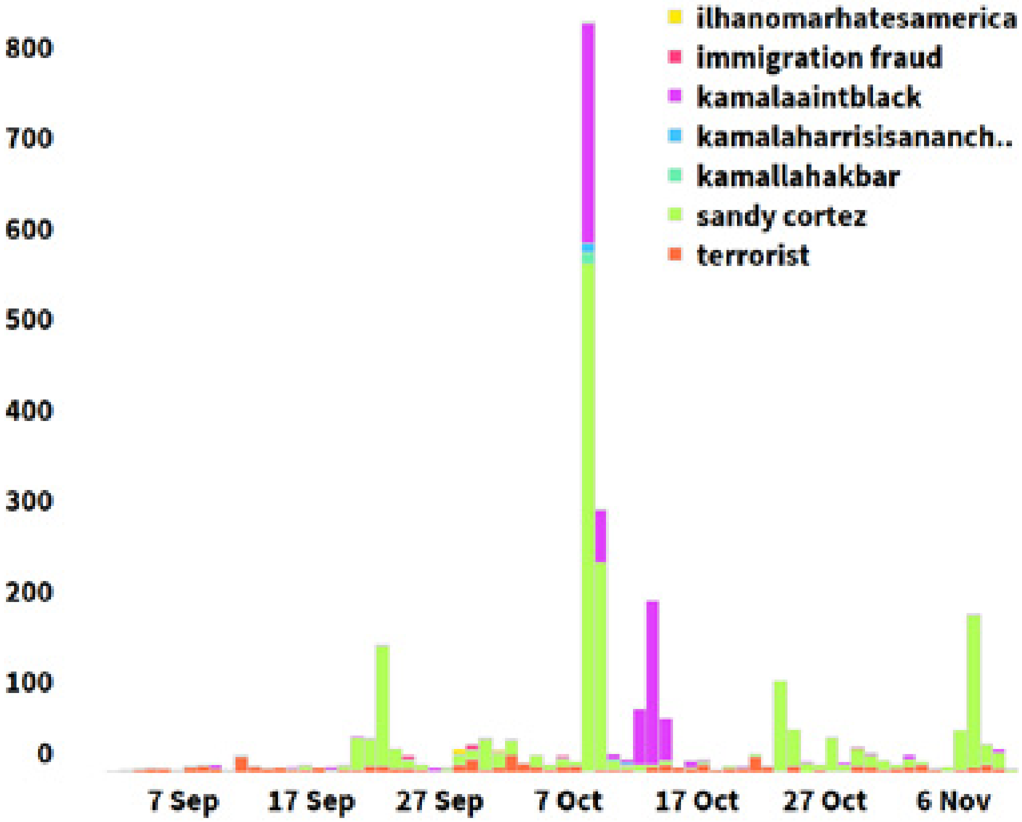


Figure 5: A timeline of the racist and racialized terms used against subjects throughout the collection period. The spike on October 7 reflects the Vice Presidential debate for which Ocasio-Cortez's Tweets received backlash.

As comparative data was not collected on male politicians of color, it is difficult to know whether these racialized narratives occur with more frequency when directed at women leaders. However, racialized disinformation provides an additional avenue by which female politicians of color receive abuse online and is often compounded with other modes of harassment.

Sexualized Narratives

While a majority of the subjects received abuse in the form of sexualization, four of the research subjects faced sexualized disinformation: Kamala Harris, Gretchen Whitmer, Kirsten Gillibrand, and Alexandria Ocasio-Cortez. Harris, Whitmer and Ocasio-Cortez have all been targeted with disinformation relating to an allegedly scandalous sexual past in attempts to discredit or humiliate them.

- In the 1990s, **Kamala Harris** began a romantic relationship with San Francisco mayor Willie Brown, who was legally married but separated from his wife. This history has been used to frame Harris as a "homewrecker" and an individual who uses sex to further her career. The rumor that Harris "slept her way to the top" is one of the most widespread disinformation campaigns targeting her, and has led to a swath of derogatory nicknames including "Heels-up Harris," "Headboard Harris," "Super Spreader" and "Joe and the Ho." These nicknames, as well as other abusive terms, were employed by users who also engaged with the Willie Brown narrative, that is, many users engaged with multiple abusive keywords and narratives about Harris. This narrative has also been used to spread the rumor that Harris is planning on taking the seat of President for herself once established in the White House.

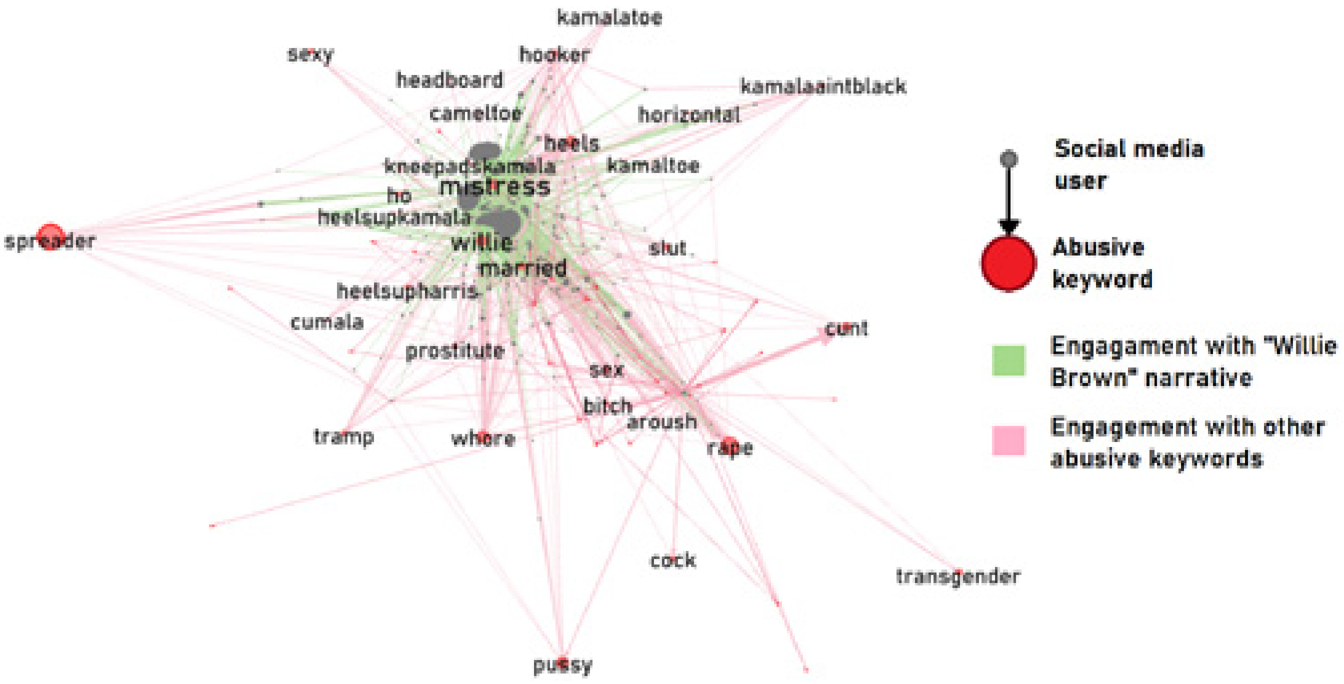


Figure 6: Keyword network of users who discussed both the Willie Brown narrative (green) and wider abusive keywords and narratives (pink). Dark grey shapes in the center of the network are clusters of multiple social media users.

- **Kirsten Gillibrand** has been accused of involvement in the sex cult NXIVM after it was revealed that her father briefly did some legal work for the organization several years ago. Our analysis found that this was the main narrative that users promoted around Gillibrand, with 98% of these interactions occurring on Twitter.

- In response to **Gretchen Whitmer's** COVID-19 lockdown measures, Whitmer's detractors spread the sexualized narrative that she had earned the nickname "Stretchin' Gretchen" during college in relation to a sex act. This narrative, which does not address Whitmer's state policies in any way, shifts focus from her position as a political leader to her alleged sexual history.

- Similarly, users online have spread the rumor that **Alexandria Ocasio-Cortez** filmed a sex tape several years ago, leading some to create photoshopped images that are presented as evidence of this. Additionally, while not explicitly disinformation, Ocasio-Cortez is also a popular subject for online sex fantasies and fetishisation. For example, discussions about Ocasio-Cortez and "feet pics" or an alleged OnlyFans account were recorded over 1,200 times during the collection period.[47] On Reddit, the research team observed that a number of users cross-posted calls for sexual fantasies about Ocasio-Cortez on multiple erotic and roleplay subreddits. Some of these calls included users seeking political roleplays about conservative politicians or Trump supporters having rough sex with Ocasio-Cortez, as well as roleplays about liberals degrading Ivanka Trump. One user, who posted similar calls to eight subreddits, went as far as fantasizing about Trump raping Ocasio-Cortez and having his aides film the rape to use as blackmail in the future. This content demonstrates the degree to which a young female politician can be sexualized and fetishized online, a trend which demeans her credibility as a serious politician.

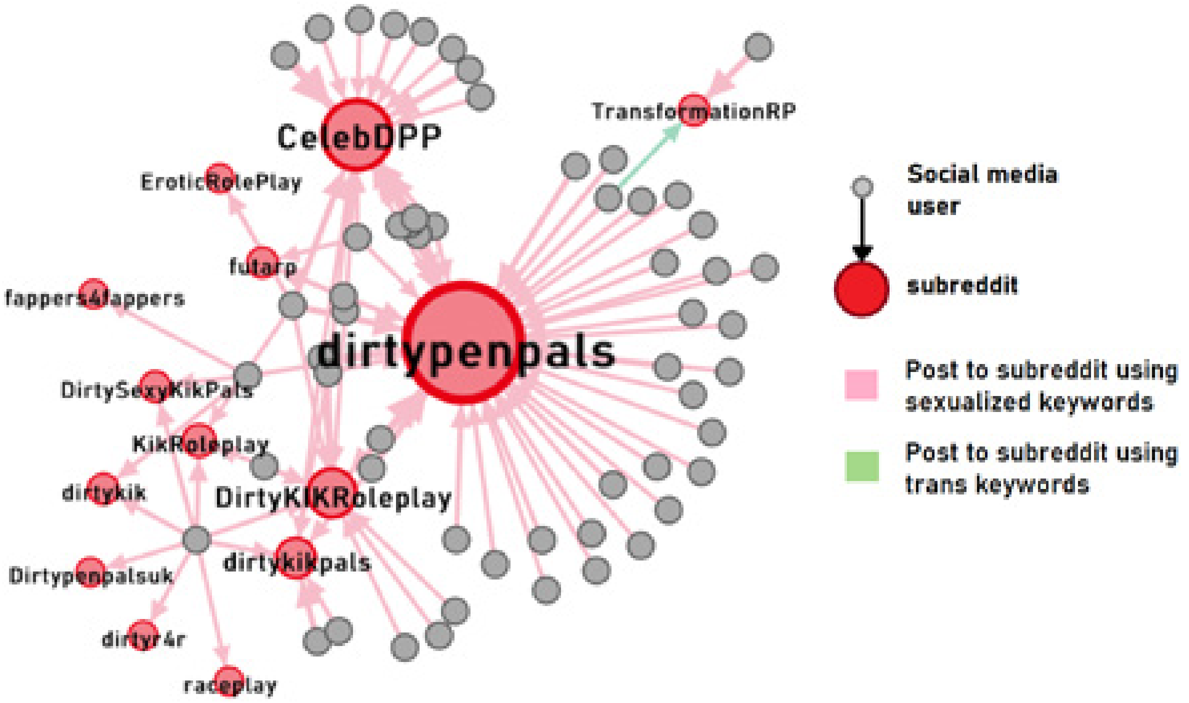


Figure 7: Subreddit network of Reddit users posting calls for sexual content (pink) and transphobic content (green) centered around AOC to multiple sexual and roleplay subreddits.

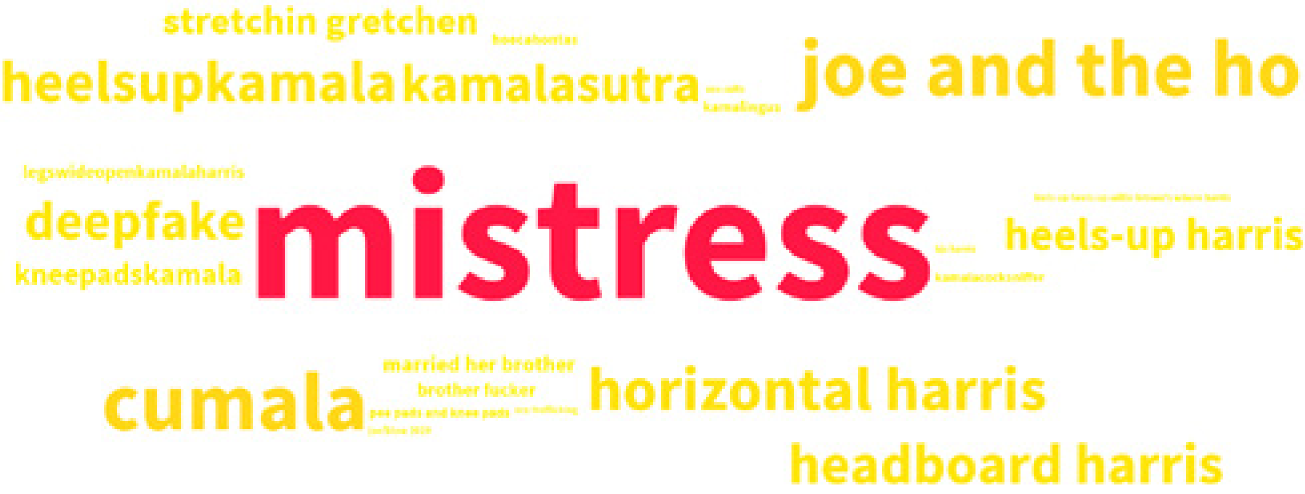


Figure 8: Frequency of terms used in sexualized disinformation campaigns, the majority of which target Kamala Harris. The term "Super Spreader," most commonly related to the COVID-19 pandemic, was also co-opted by abusers to refer to Kamala Harris in a sexual manner. However, given the false positives that were collected through the scraping of this term, it is not included in the above graphic.

Equally of note are the subjects that abusers chose not to sexualize. One hypothesis is that older politicians were less likely to encounter sexualized or sexually violent abuse: Chrystia Freeland, Kirsten Gillibrand (other than the short-lived narrative relating to her father's legal work for NXIVM), Susan Collins, and Priti Patel. The exception to this hypothesis is Ilhan Omar. The sexualization of younger subjects appears to be a method of undermining their influence; Omar's abusers, however, have largely chosen to focus on framing her as a foreign threat that should be expelled rather than fetishized. This may be because abusers see delegitimization using racist narratives as a more direct route to removing Omar from power. It is also important to note that abusive narratives are by no means coherent. While Whitmer was the subject of a sexualized narrative (stemming from "Stretchin' Gretchen"), a competing narrative characterized her as a Hitler-esque figure, with keywords including "Whitler" and "Gestapo Gretchen" gaining traction. This narrative simultaneously accorded her political power and undermined it.

Platform Analysis

Twitter accounted for 95% of the total data collected, indicating that this is the platform on which the greatest volume of abuse takes place. The research team hypothesizes that this is due to the structure of the platform, which allows users to address the subjects directly. Twitter is the only platform on which all the subjects had personal accounts and could engage with other users. While all six platforms were monitored for data related to all 13 research subjects, the data collection did not record instances of abuse of each subject on every platform. Additionally, in every case except one, Twitter produced the most data points for the subjects.[48] The ability of users to directly address the subjects on Twitter allows in many ways for "higher impact" abuse, allowing users to yell at targets rather than simply about them.

The research team observed the highest number of sexually explicit and violent keywords on 4chan and 8kun, with Twitter and Reddit ranked third and fourth respectively. Parler, however, was unique in its lack of highly vitriolic language. The keywords "sex" and "bitch" were within the top three recorded terms for Twitter, Reddit, 4Chan, Gab and 8kun, while Parler's top three keywords were #PhonyKamala, #RecallWhitmer, and #RecallGovWhitmer. Additionally, keywords indicating gendered abuse were among the lowest recorded terms on the platform. It is unclear why user behavior on Parler was different, particularly given that Parler brands itself as a "free speech" social media alternative. This could be the consequence of the echo chamber created by a relatively small user base.

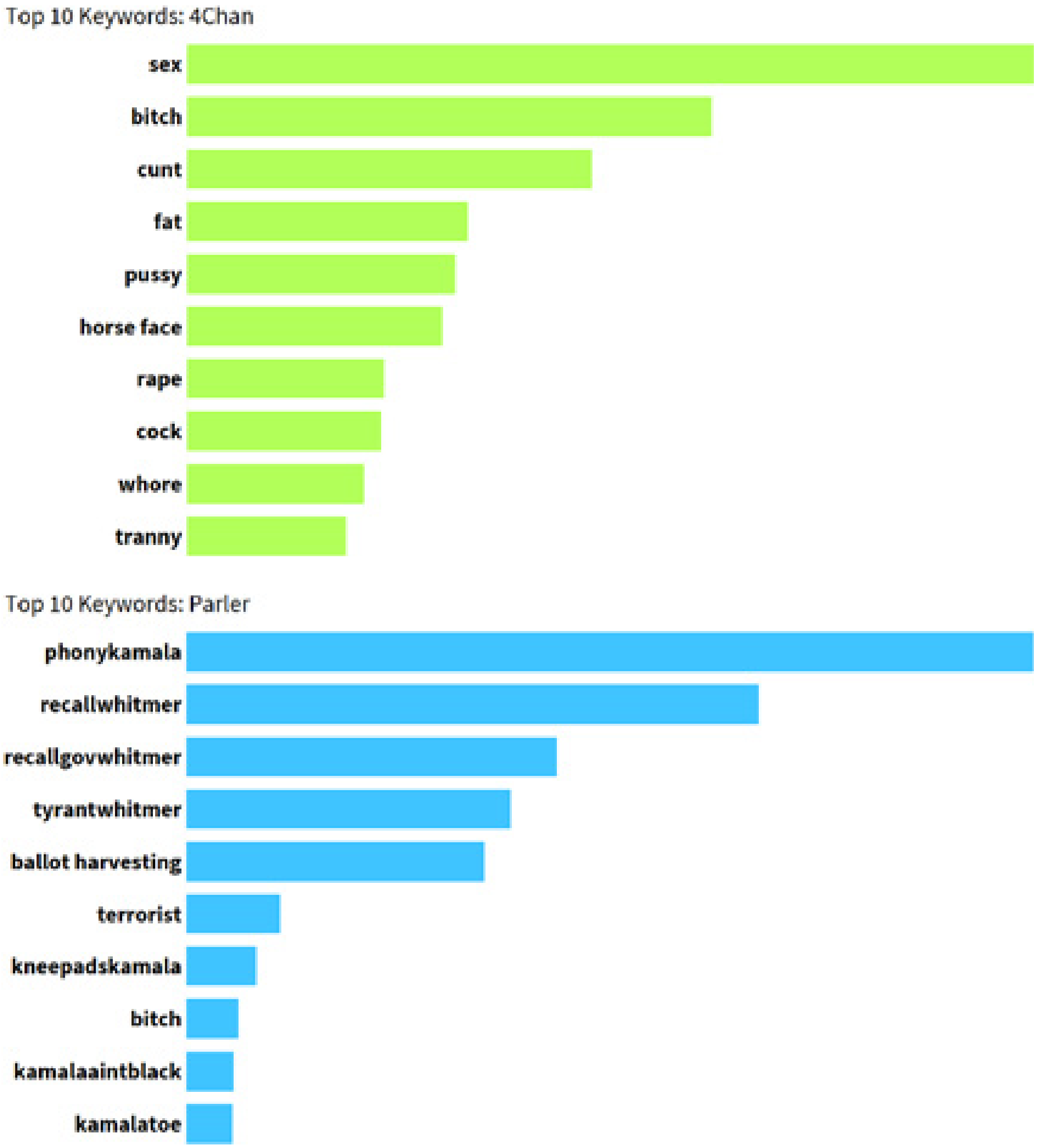


Figure 9: Differences in top 10 keyword use between Parler and 4Chan.

Exploring Behavior and Coordination

In order to evaluate user behavior and assess potential coordination, the research team visualized user relationships (i) between the abusive keywords users employed across social media platforms (captured on all monitored social media platforms), and (ii) between subreddits and boards users posted to (captured on Reddit, 4Chan, 8kun, and Gab).This overview of user interactions with keywords across platforms illustrates how the themes of sexual, racist, and transphobic disinformation narratives and broader abuse overlapped across platforms. Overall, **we observed patterns of intersectionality between generally abusive and sexually abusive keywords or narratives with which users engaged.We also observed instances of individual users who exhibited repetitive abusive posting patterns**.

While the three themes of gendered disinformation recorded during this data collection period - transphobic, racist, and sexual narratives - were presented separately in this report, it is important to emphasize that they do not operate in silos. Keyword networks of user interaction with #KamalaAintBlack and "Sandy Cortez" showed that some users who interacted with these racist narratives also interacted with sexualized narratives and abusive keywords. This duality was most visible in the two largest datasets we collected (Kamala Harris and Alexandra Ocasio-Cortez, see Figure 10 below).

We also observed users engaging in repetitive and abusive posting patterns in multiple networks which targeted Kamala Harris, Alexandra Ocasio-Cortez, Gretchen Whitmer, and Ilhan Omar. This pattern of behavior occurred on all social media platforms, where a few users engaged with the same keyword hundreds of times, or repeatedly posted sexually explicit content to multiple subreddits and boards during the two month collection period. The former pattern was seen when one individual posted the "Kamal Aroush" transphobic narrative about Kamala Harris to Gab 33 times. The latter behavior was observed in Figure 10, where one user (whose pseudonymized username is the same on 4chan and 8kun) posted to as many as 27 4chan and 8kun boards, and repeatedly posted to the Politically Incorrect, Q Research, and Random boards up to 551 times on 4chan and 8kun (indicated by the large arrows in Figure 10). These forms of *individual campaigning* and repeated behaviors may violate some platforms' spam policies. It also demonstrates the repetitive, dedicated behavior present among some users in their amplification of gendered disinformation narratives.

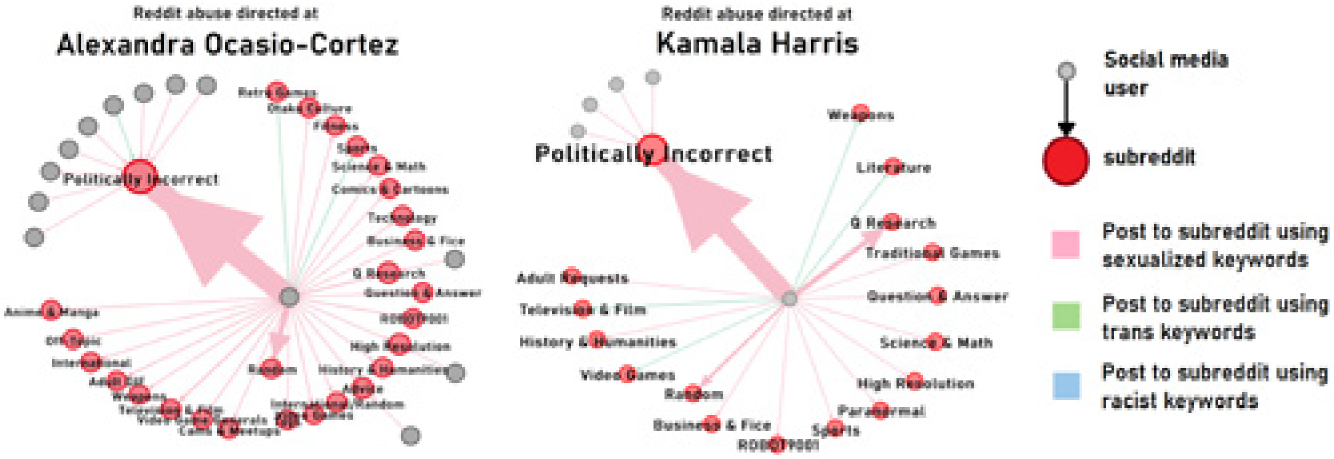


Figure 10.1: One individual (center, grey) posts sexualized (pink), transphobic (green), and racist (blue) content to both 4chan and 8kun about Alexandra Ocasio-Cortez, Kamala Harris (this page) and Ilhan Omar (next page). Gretchen Whitmer was the subject of similar interactions.

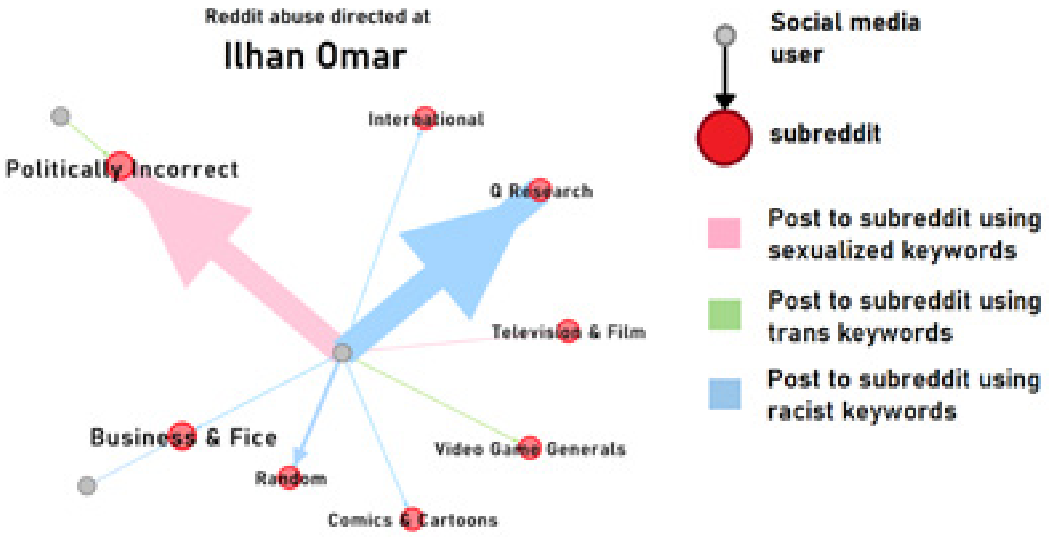


Figure 10.2: One individual (center, grey) posts sexualized (pink), transphobic (green), and racist (blue) content to both 4chan and 8kun about Alexandra Ocasio-Cortez, Kamala Harris (previous page) and Ilhan Omar (this page). Gretchen Whitmer was the subject of similar interactions.

Unique Findings

Abusive Keywords: A War of Intent

Certain keywords recorded were found to have both positive and negative connotations, depending on the intent and context of the post. Two notable examples of this were "Big Gretch," a nickname developed for Gretchen Whitmer, and the use of the word "bitch." Whitmer's critics initially employed "Big Gretch" in relation to Whitmer's COVID-19 policies. Subsequently, however, the nickname was also adopted as a term of endearment, becoming the subject of a rap song publicly embraced by Whitmer herself.[49] A similarly ambiguous term used in relation to other research subjects is "bitch." Initially considered an explicitly abusive term, the research team found that it is frequently used in admiration of the subjects. For example, one user writes "AOC is a that bitch [sic]. I love this woman! ".These findings indicate that a certain percentage of false positives exist among terms that would otherwise be considered clearly abusive. The duality of meaning, as well as the ambiguity of intent, can also lead to challenges in automatic detection of abusive and false narratives, a critical consideration for platform content moderation responses.

Disinformation Exempt

Among the research subjects, Priti Patel is the only high-profile politician against whom coordinated disinformation campaigns were not recorded. This is particularly interesting given Patel's long history of controversial policies. Online abusers appeared to target her directly with misogynistic abuse; "bitch," "fat" and "witch" were the most frequent abusive keywords directed towards Patel. It is unclear what accounts for this apparent lack of targeted disinformation during the data collection period. As the project's only British research subject, it is possible that the difference is cultural; however, this conclusion cannot be drawn without further research to enable a comparison to other British political figures. It is also possible that the lack of a major electoral event in the UK during the project's collection period can account for this. This finding raises questions regarding the frequency and intensity of disinformation cross-culturally.[50]

A "Copypasta" Passage Targets Kamala Harris

While examining the data collected, the research team noted a unique passage that appeared 31 times in reference to Kamala Harris. The passage is an example of a "copypasta," internet slang for a body of text that is copied and pasted across different websites. It begins with the phrase: "I see Kamala Harris as a challenge, more than anything. Here is a woman who, in every single aspect, is absolutely revolting - her exterior AND her personality - yet I can't help but wonder what would be like, to plunge balls-deep into her repeatedly." The passage continues by describing different sexual acts in explicit detail, referring to Harris as a "thing" that the user allegedly despises but wishes to conquer sexually. Upon investigation, the research team found that this exact passage has been circulating online across different websites since at least 2012, and has been used for several different women and girls including Amy Schumer, Chloe Moretz, Emma Gonzalez, Greta Thunberg, and Maisie Williams.[51] Given that the passage is too long to be posted on Twitter where the user could tag the subject directly, it appears to be a way for male or male-identifying users online to assert their masculinity over influential women, using the fantasy of violent sex as a proxy for domination. Kamala Harris was the only research subject targeted by this passage during the data collection period.

Harassment and Political Party Affiliation

The data collected precludes any definitive conclusions regarding differences in abusive language directed at Democratic and Republican subjects. Subjects including Kamala Harris or Alexandria Ocasio-Cortez do not have Republican equivalents that are of comparable political stature, preventing a reliable comparison of keywords and sentiments. Additionally, the time frame during which data was collected saw primarily Democratic or Liberal women centered in historic events, leading to their increased prominence online. While some data suggests that the abuse directed towards Democratic women contained more violent language, this cannot be confirmed within the scope of the data collected.

Qualitative Analysis

The research team conducted in-depth interviews and focus group discussions in order to better understand how women experience gendered and sexualized harassment versus disinformation campaigns. In particular, the qualitative research focused on how nation states use gendered tropes to undermine democracy and how gendered and sexualized campaigns affect subjects' engagement in public life on- and offline. Consistent with Sobieraj's findings in *Credible Threat*, these interviews revealed that online harassment campaigns cause women to reconsider and often adjust their public engagement. Some women take precautions to protect their physical safety after experiencing such campaigns. Women of color have increased concerns, as they receive particularly graphic, vitriolic abuse. Finally, all interviewees and focus group participants reported that they felt social media platforms and governments left the onus on targets to report and respond to harassment campaigns, and that neither were taking adequate measures to protect the online and offline safety of half the world's population.

State-sponsored Gendered and Sexualized Disinformation

Gender and sexuality has long been used to argue that women are ill-suited to positions of power. The domestic deployment and amplification of these narratives is undoubtedly a threat to democracy, but when this societal vulnerability is weaponized by adversaries, it becomes a threat to national security. This section details three overt state-sponsored harassment campaigns aimed at female journalists targeted by Iran, China, and Russia. In all three case studies, state-sponsored entities deployed false or misleading narratives that played on gendered or sexualized narratives against prominent journalists, unleashing what interviewees called a "tsunami" of online "terrorism" and abuse against them.

Iran's targeting of Yeganeh Rezaian

In July 2014, Iranian journalist Yeganeh Rezaian had just married her husband, Jason, who served as The Washington Post's Tehran bureau chief.[52] Not long after her wedding, her Facebook account was hacked and she received a frightening email; the senders threatened to blackmail her by publishing "dirty photos" on social media and sharing them with her husband. Though Rezaian knew the email was a fraud, she was disturbed: "as a woman working in a very traditional society, I was worried about like, 'What is in there? I need to make sure.'" Not long after receiving the email, Rezaian was locked out of her account, the Iranian security services raided her home, and she and her husband were arrested and placed in solitary confinement. Rezaian was released on bail in October 2014; her husband was convicted of espionage and remained in prison for 544 days until the State Department secured his release in early 2016.

Since then, the couple has moved to the United States, but gendered and sexualized disinformation campaigns against Rezaian have not stopped, and neither have their ongoing effects on her offline life. She and her husband were the subjects of an Iranian Revolutionary Guard (IRG) television documentary in 2019 that led to an onslaught of online abuse. Advancing the IRG's false assertions about Rezaian's character, the documentary uses the trope that women are naive and unable to make decisions without the guidance of a man. "It portrays me as a young, uneducated woman...who is so fascinated with living abroad," she said. "This older, smart, CIA agent comes to Iran and he's able to trick her and make her fall in love, even more, with living in America." The documentary asserts that the Rezaians married to make the bride's family more comfortable with the espionage scheme, which built in her a sense of loyalty to Jason and to the United States. "The fact that they show me as a very naive, completely uneducated" woman bothered Rezaian, she said.

> *I had a Master's degree before I met Jason, and it's not that I have never been outside of the country...They try to go back to that super traditional, naive girl who doesn't know anything about the world and the foreign guy who was more experienced, and, obviously, he's a CIA agent so he is well-trained, is easily able to deceive her. It's misogynistic, right?*

After the documentary aired, Rezaian's Instagram account was inundated with abusive comments. She told the research team they were "all about my gender and sexuality," and that they usually "had something to do with my private life and being a woman." In addition to criticizing her career and her choice to marry a foreigner, Rezaian noted she is often ridiculed online for "the way I put makeup on or [the way] I dress." This has changed the type of content she posts publicly: "I intentionally try to never publish any photos of me...wearing a strapless dress because I know that will [provoke] more people," she noted as an example.

She deals with the ongoing harassment by deleting the abusive comments as they come in and never reactivated her Facebook account after it was hacked in 2014. "I hate Facebook because I feel like it's terrible that such a vastly worldwide-used social media platform was so weak." The Rezaians have also already made plans about how they will protect their newborn baby's privacy and safety by not posting pictures of the child online. In the context of the offline harms perpetuated by the online gendered and sexualized disinformation campaigns against her, Rezaian refers to the efforts as "a form of terrorism" with the aim of "making sure that, first and foremost, as a woman, you lose your credibility publicly, and then privately [perpatrators] destroy your self-confidence, because they know that affects your public persona."

Rezaian's experience—as a target of state actors that used a false, gendered narrative to defame her and her husband—is a classic example of state-sponsored gendered disinformation, and a clear demonstration of its implications for women's equal participation in society.

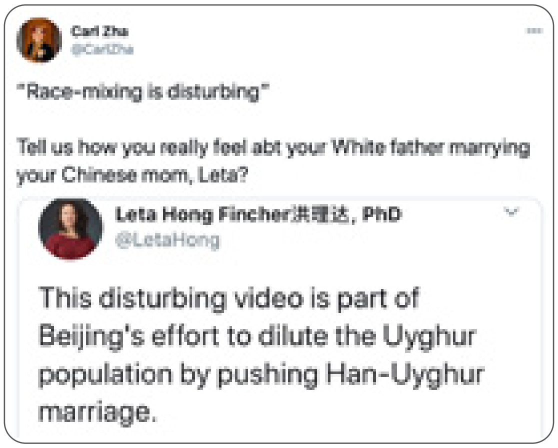
Abusers want to make "sure that first and foremost, as a woman, you lose your credibility publicly, and then privately [perpatrators] destroy your self-confidence, because they know that affects your public persona."

China's targeting of Leta Hong Fincher

Leta Hong Fincher is an American journalist and scholar whose work focuses on feminism in contemporary China. In summer 2020, as more evidence emerged of human rights abuses against Uyghur Muslims in Xinjiang province, the *People's Daily*, a large Chinese state-owned newspaper group, tweeted a video about mixed-race marriage in the province.[53] "Alimjan is one of numerous young people in NW China's #Xinjiang who pursue love earnestly. Take a look at his story!" the text of the tweet read, accompanying a video that depicted an allegedly happy mixed-race marriage.

Hong Fincher was disturbed; the propaganda video was attempting to rewrite the troubling history of the dilution of the Uyghur ethnicity through aggressive promotion of marriage with Han Chinese partners. She posted a Twitter thread drawing attention to the troubling video and connecting it to her research. In one Tweet she wrote: "As I write in my book #BetrayingBigBrother, Xinjiang officials have for years offered bonuses to inter-ethnic couples with one Han Chinese partner marrying a member of an ethnic minority."[54] The thread also goes on to connect Hong Fincher's own background to the thread: "Many people seem to think that the *People's Daily* video of a Uyghur man dating a Han Chinese woman is just a sweet, interracial love story. I myself am mixed race and of course if a young Uyghur Muslim chooses on their own to marry a Han Chinese partner, that is perfectly fine." But, she wrote, that was not what the video depicted; instead it was covering up China's eugenics policies against "undesirable" births from Uyghur and Kazakh populations through forced sterilizations and cash bonuses for interracial marriages.

Her thread generated both high levels of engagement and unprecedented criticism. "I just started getting...bombarded with these accusations of being against mixed race relationships. There were a lot of sexualized insults that I get," Hong Fincher said.[55] She added that the campaign centered on a "coordinated, false narrative" that she did not approve of mixed-race marriage. There were also a number of fake Twitter accounts set up to impersonate her. Eventually she noticed her thread had earned the ire of Carl Zha, a prominent pro-China blogger and podcaster who supports himself through Patreon, a crowdsourcing platform. Isobelle Cockerell, a reporter at Coda Story, describes Zha as "devoted to attacking Western reports of human rights abuses in the region and painting coverage of Uyghur oppression as an influence operation designed to incite tension between the U.S. and China."[56] Hong Fincher said she had "blocked [Zha] a long time ago because he had harassed me in the past more than once." But that did not stop Zha from taking a screenshot of her Tweet and posting the image with the comment:



Carl Zha
@CarlZha

"Race-mixing is disturbing"

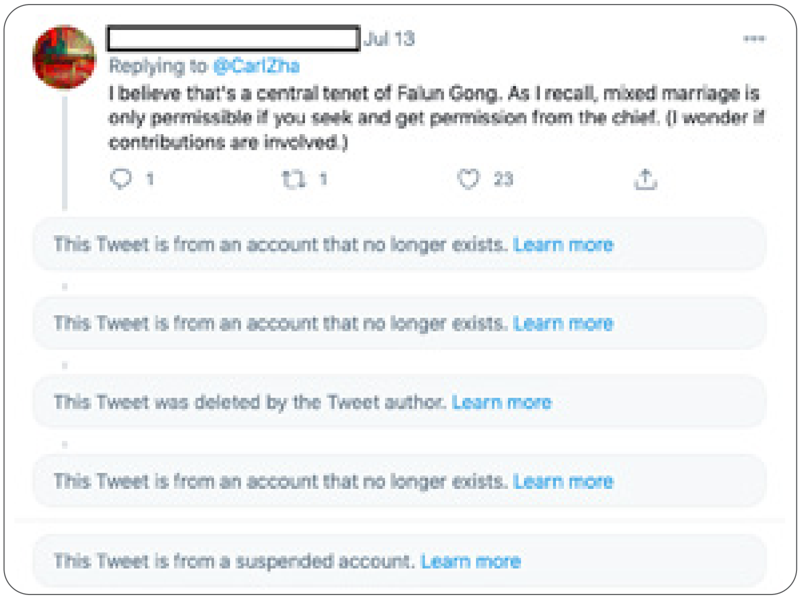
Tell us how you really feel abt your White father marrying your Chinese mom, Leta?

Leta Hong Fincher洪理达, PhD
@LetaHong

This disturbing video is part of Beijing's effort to dilute the Uyghur population by pushing Han-Uyghur marriage.

Figure 11: A screenshot of Carl Zha's Tweet about Leta Hong Fincher. The Tweet reads: "'Race-mixing is disturbing' Tell us how you really feel abt your White father marrying your Chinese mom, Leta?"

Hong Fincher believes it was Zha's Tweet that was the trigger for the harassment campaign against her. "It was something very personalized, but it wasn't over the line harassment," she noted, referencing platform policies about "targeted harassment" that rarely lead to enforcement. Hong Fincher's experience is a classic example of dogpiling, which many of this project's focus group participants underlined as a loose type of coordination between online users, with devastating effects. Hong Fincher blocked and reported hundreds of accounts in the aftermath of Zha's Tweet, including the accounts impersonating her. Cockerell, the journalist from Coda Story, also uncovered hundreds of seemingly-automated pro-China accounts that amplify Zha and other pro-China voices like him.[57] Some of these accounts, as well as those impersonating Hong Fincher, were removed, as evidenced by notes about deleted accounts or deleted Tweets among the replies to Zha's Tweet and Hong Fincher's original thread:



[redacted] Jul 13
Replying to @CarlZha
I believe that's a central tenet of Falun Gong. As I recall, mixed marriage is only permissible if you seek and get permission from the chief. (I wonder if contributions are involved.)

This Tweet is from an account that no longer exists. Learn more

This Tweet is from an account that no longer exists. Learn more

This Tweet was deleted by the Tweet author. Learn more

This Tweet is from an account that no longer exists. Learn more

This Tweet is from a suspended account. Learn more

Figure 12: A screenshot from replies to Carl Zha's original Tweet about Leta Hong Fincher, including five Tweets from accounts that no longer exist, were suspended, or were deleted by the Tweet author.

The harassment targeting Hong Fincher did not stop at inauthentic amplification, however. Others associated with state-run Chinese media expanded on the false narrative that she was "against mixed-race marriage" despite being the product of and party to a mixed race marriage. For instance, Tom Fowdy, a blogger who is a freelance author for the Chinese official state media outlet CGTN, equated an effort to raise awareness about the campaign against Hong Fincher with efforts to ban Carl Zha from Twitter over "allegations of [her] targeted harassment." His blog did not address the substance of the "allegations," but suggested criticism of Zha was a "part of the broader culture amongst China analysts and opponents of the country to discredit everyone who contravenes their views." "Shortly after that," she recalled, "there was a Xinhua correspondent who was based in Belgium," who suddenly began to criticize Hong Fincher's PhD research, conducted at Tsinghua University in Beijing. "I have never been [criticized] by Chinese state media officially," Hong Fincher noted, nor had her work in China been the target of Chinese state derision. At the same time, she noticed that "all these Chinese ambassadors around the world who are on Twitter started dismissing the report about mass sterilization of Muslim women."

Finally, trolls took to other platforms, including Hong Fincher's Amazon page, where they left inauthentic reviews of her work and promoted the false narrative about her views on mixed-race marriages. Hong Fincher says "there were people calling me a multitude of sexualized insults, misogynistic insults...there have been people threatening to gang rape me and rape me and referring to my children." She wondered on Twitter, "Is it any wonder that most women prefer not to call out harassers publicly?".[58] She does credit Twitter with some response; after an email exchange with a Twitter employee and a public awareness campaign led by the Coalition for Women in Journalism, they began taking action on some of the content in the campaign and verified Hong Fincher's account. But women without the profile, resources, or volition to escalate evidence of abuse may not have been able to achieve this result.

Despite Twitter's action, when asked to describe the disinformation campaign against her in one word, Hong Fincher called it a "tsunami." But Hong Fincher weathered that storm. "It was clearly an attempt to intimidate me and shut me up and exhaust me," she said. "I didn't want to engage, obviously. But I just thought 'I can't let them get the upper hand.' I pinned my thread for quite a few weeks just out of defiance and I didn't want to let all these trolls know that their intimidation was working. But it was utterly exhausting and extremely unpleasant."

This case study demonstrates all three characteristics of gendered and sexualized disinformation: malicious state and non-state actors drove a false, gender-based narrative about Hong Fincher's alleged beliefs about mixed-race marriages, used sexualized threats and insults, created fake accounts impersonating her, and unleashed state media employees with some degree of coordination to criticize her previous research in order to protect the reputation of the Chinese state and denigrate Hong Fincher.

Russia's use of gendered narratives

Among malign state actors, the Russian Federation's targeting of women in public life has become a well-established pattern. Whether attacking journalists, activists, politicians, or others engaged in public discourse, the Kremlin and its online influence operations amplify sexist, misogynistic, and gendered narratives. They are aimed at undermining women's credibility and participation in public life and the political process, thereby undermining democracy itself.

In a 2019 study, Dr. Samantha Bradshaw identifies the ways foreign influence operations rely on traditional gender stereotypes in their messaging strategies through an analysis of over 300,000 English Tweets about gender and politics culled from Twitter's Election Integrity Initiative dataset.[59] While discussions around gender identity amounted to only 13 percent of all discussions by observed foreign state-operated accounts—including a large proportion from Russia's Internet Research Agency—Bradshaw concluded that "gender was a cross-cutting theme that intersected" with discussions pushed by foreign state-operated accounts on other topics like race and religion, employing gendered stereotypes "to engender fear, spread skepticism, and foment distrust."[60]

Finnish journalist Jessikka Aro, who identified and reported on pro-Kremlin trolls long before they were a household topic, was a target of their gendered disinformation herself. "They released her medical history and her home address. They created a music video mocking her as a 'Bond girl.' They claimed, without basis, that she was a prostitute soliciting male bigwigs from the CIA and NATO, who fed her lies about Russia."[61] Similarly, the face of Svitlana Zalishchuk, a young Ukrainian parliamentarian, was superimposed onto pornographic images after Russian-backed disinformation spread about her in the wake of a bloody battle in the Russian-sponsored war in Ukraine's Donbas region. In the Republic of Georgia, high-profile women (including the editor-in-chief of a popular magazine, Tamara Chergoleishvili, who is married to a prominent opposition politician) were targeted with fake sex tapes that aimed to destroy the reputations of women in Georgia's highly patriarchal, conservative society.[62]

> I got rape threats, many through DMs, and I would report and just blocked, and blocked, and blocked.

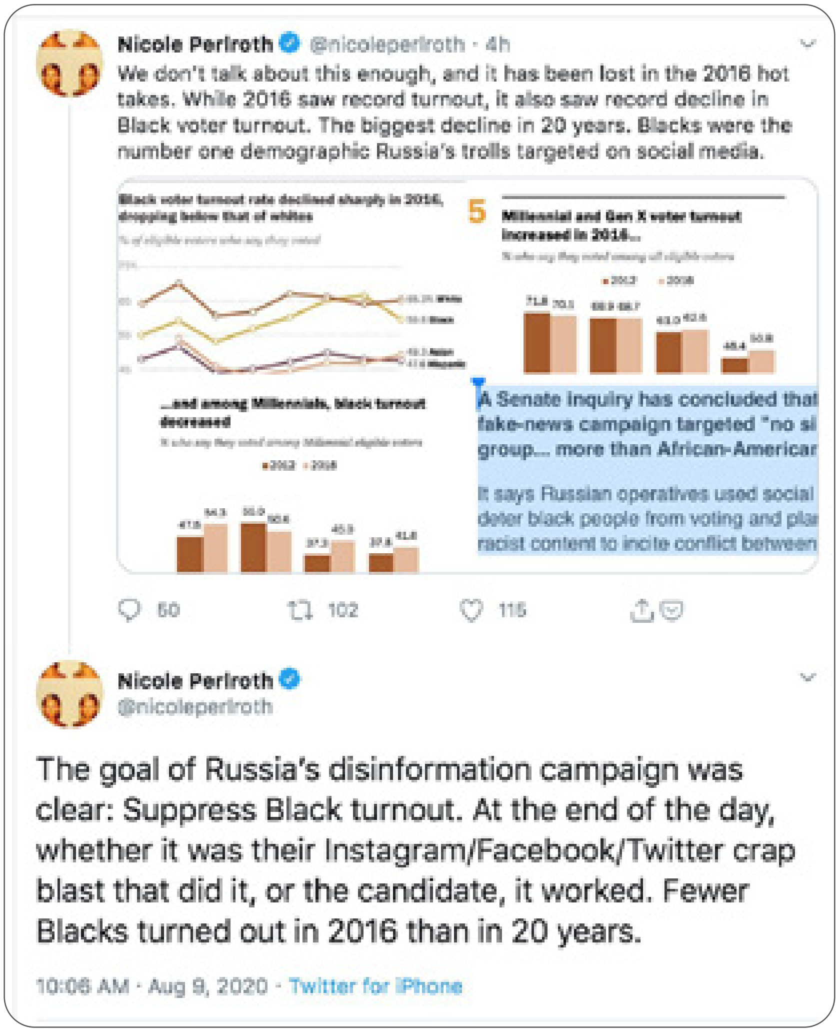
These are brazen examples of state-sponsored gendered and sexualized disinformation, but the phenomenon can also take on more insidious, covert forms, playing broadly on tropes that women are less intelligent than male counterparts, less qualified to participate in public life than men, or creatures only suitable for breeding. These nuanced attacks by state-sponsored entities often encourage the dogpiling effect described by many of this study's focus group participants and other targets of online gendered harassment.

RT's Targeting of New York Times Reporter Nicole Perlroth

Nicole Perlroth, a cybersecurity reporter for The New York Times, is no stranger to gendered attacks. When Perlroth was covering Russian interference in the 2016 election and the Edward Snowden leaks, users would "grab unflattering screenshots of media appearances to stoke comments about my appearance. I got rape threats, many through DMs, and I would report and just blocked, and blocked, and blocked."[63] Much of the abuse she receives focuses on her journalistic background. Perlroth said her abusers often ask "'How dare she write about [cybersecurity] when she's not one of us?' I think in my particular position that means not being a computer security researcher, or having a technical computer science background. Much of it has the condescending element of 'How dare you even have an opinion about this when you can't hack?'"

She often shares bylines with male reporters at the Times, who are not subject to the same gendered harassment. Perlroth described the experience of relentless online abuse as if "someone put me in a dryer and just like left it on high for two days."

Russian state media seized on this trend in August 2020. In the leadup to the American presidential election, she tweeted about the Russian Internet Research Agency's well-documented strategy of targeting Black voters in an attempt to suppress Black turnout in the 2016 election. In her Tweet, Perlroth "remind[ed] people that Black turnout was really low, historically low in 2016." She wanted to explore "[why] we don't examine the fact that Black turnout dipped substantially in 2016, and what affect, if any, Russian influence campaigns had on the African American vote—when Russian trolls pretended to be Black Lives Matter activists and spread anti-Clinton memes and narratives like 'Killary,'" she said, citing findings by the Senate Intelligence Committee, which concluded the majority of Russian influence operations in 2016



Nicole Perlroth @nicoleperlroth · 4h

We don't talk about this enough, and it has been lost in the 2016 hot takes. While 2016 saw record turnout, it also saw record decline in Black voter turnout. The biggest decline in 20 years. Blacks were the number one demographic Russia's trolls targeted on social media.

Black voter turnout rate declined sharply in 2016, dropping below that of whites
White
Black
Asian
Hispanic
5
Millennial and Gen X voter turnout increased in 2016...
2012
2016
71.8
70.1
69.9
68.7
63.0
62.6
48.4
50.8
...and among Millennials, black turnout decreased
2012
2016
47.6
54.3
55.0
50.6
37.3
45.9
37.6
41.8

A Senate inquiry has concluded that fake-news campaign targeted "no si group... more than African-Americar

It says Russian operatives used social deter black people from voting and pla racist content to incite conflict between

Nicole Perlroth @nicoleperlroth

The goal of Russia's disinformation campaign was clear: Suppress Black turnout. At the end of the day, whether it was their Instagram/Facebook/Twitter crap blast that did it, or the candidate, it worked. Fewer Blacks turned out in 2016 than in 20 years.

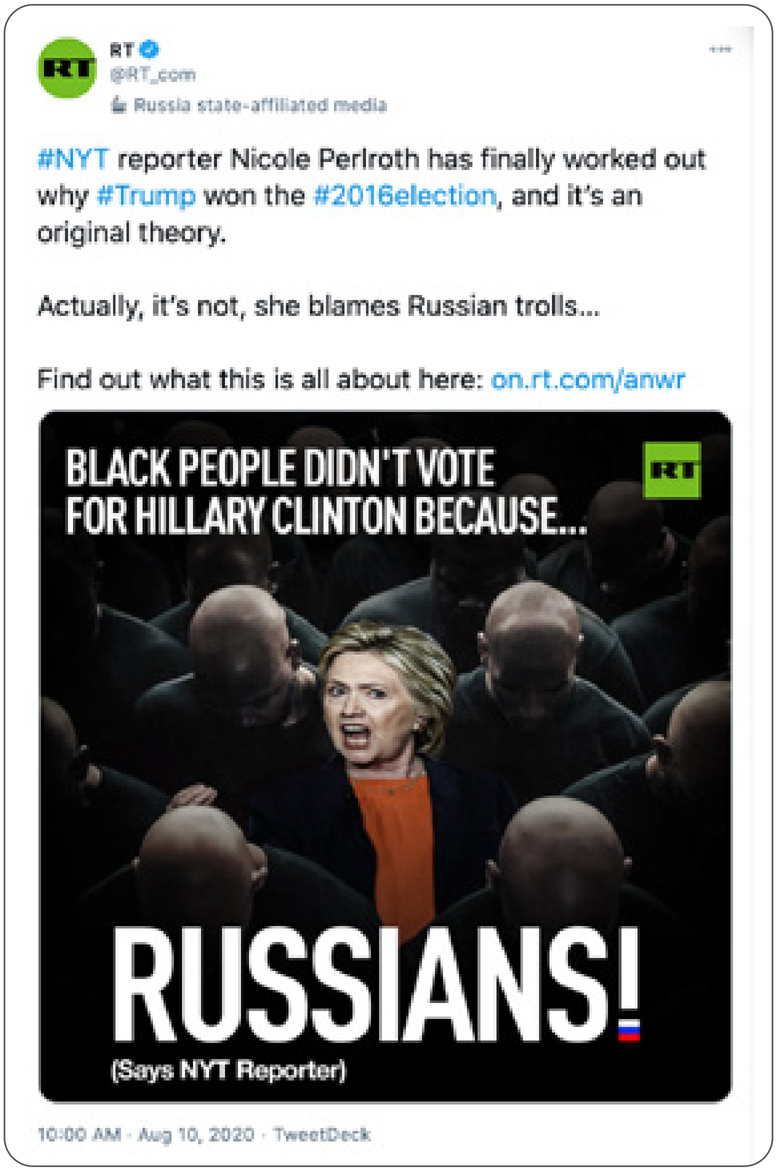
10:06 AM · Aug 9, 2020 · Twitter for iPhone

Figure 13: Perlroth's now-deleted Tweet discussing Russia's strategy targeting Black voters in the 2016 election.

were aimed at suppressing the Black vote. "I come back to this question a lot: Did those influence campaigns tamp enthusiasm from Black voters for Hillary? And how do you measure that? So I re-raised that question and Russia Today," now known as RT, the Russian state-sponsored foreign broadcaster, "picked it up."

RT published an article entitled "White NYT reporter tells black people they didn't vote for Hillary in sufficient numbers because Russia duped them," with corresponding Twitter, Instagram, and Facebook posts.

The social media content promoting the article, as well as the text of the article itself, all played on the gendered trope Perlroth has encountered many times in her career: that she is not intelligent enough to understand the topics on which she is reporting. The article begins: "It only took almost four years, but New York Times reporter Nicole Perlroth



RT
@RT_com
Russia state-affiliated media

#NYT reporter Nicole Perlroth has finally worked out why #Trump won the #2016election, and it's an original theory.

Actually, it's not, she blames Russian trolls...

Find out what this is all about here: on.rt.com/anwr

10:00 AM · Aug 10, 2020 · TweetDeck

Figure 14: Screenshot of RT Tweet targeting Nicole Perlroth.

has finally gotten to the bottom of why black Americans didn't get excited about Hillary Clinton's 2016 candidacy for president: Russian trolls."[64] On Twitter, the broadcaster wrote: "#NYT reporter Nicole Perlroth has finally worked out why #Trump won the #2016Election, and it's an original theory. Actually it's not..."[65]

Additionally, RT described Perlroth's Tweet as "whitesplaining," playing on the racial themes that run through both modern and historical Russian operations.[66] During a summer of racial unrest in the United States, throughout which RT fanned the flames of American polarization and endemic racism, the Russian broadcaster alleged that Perlroth's legitimate questions about the impact of 2016 Russian information operations on their target audiences was racist. RT chose to run unattractive pictures of Hillary Clinton in what appear to be aggressive poses with the article and its corresponding social media content, playing into narratives that Clinton was "nasty" or "shrill," and that older women do not have a place in society. Eventually, the Perlroth narrative was laundered into the American fringe media; in content repackaged from the original RT article, an American website used an unflattering image of Perlroth alongside text that claimed she was "whining" about Hillary Clinton's 2016 election loss.

Even for Perlroth, who is no stranger to online harassment, "the volume of the Twitter blowback and online chatter around my Tweet was way higher than anything I had tweeted in the last two years. So I imagine it was because of RT." In response, Perlroth deleted the tweet—a strategy used by many women experiencing online abuse to lessen the amount of harassment directed at them—which caused her abusers to insinuate that she tacitly admitted that she was wrong. Perlroth's experience demonstrates how women experiencing online abuse have no good choices; by speaking their mind and standing their ground, they may subject themselves to further abuse. The decision to delete the content engendering abuse or lock down accounts can also lead to harassment.

RT's targeting of Perlroth does not include explicitly gendered or sexualized tropes; the elements of this campaign would not be unearthed by an automated tool searching for gendered terms among the state-sponsored broadcaster's articles and social media properties. But using the Russian Federation's online media ecosystem, it advanced a far more insidious narrative—that women are stupid, "whiny," or, in the case of Hillary Clinton, unnecessarily power-hungry—in order to drive online engagement and further societal division that was later amplified and replicated within homegrown American publications and audiences.

Challenges of Detecting and Enforcing Against Gendered and Sexualized Abuse and Disinformation

There are a number of challenges in detecting and responding to online gendered and sexualized abuse and disinformation. Based on the data collected for this study and the responses of interview and focus group participants, the research team has identified the following difficulties in order to aid future responses.

Malign Creativity in Online Misogyny

While this research identified a list of keywords that might signal or be used in conjunction with gendered abuse or disinformation, we found that "malign creativity"—the use of coded language, iterative, context-based visual and textual memes, and other tactics to avoid detection—is perhaps the greatest challenge to detecting, challenging, and denormalizing online abuse, whether gendered and sexualized disinformation or more broadly. The most insidious gendered disinformation narratives used coded language, less likely to trigger automated detection and which often requires moderate-to-deep situational knowledge to understand.

For example, the word "bitch" may be represented using spaces or special characters. This makes these abusive terms and the narratives they support difficult to detect for automatically. Furthermore, in the context of specialized narratives or nicknames, such as "Stretchin' Gretchen" or "Heels Up Harris," a human content moderator may lack the context to understand and take action against abusive content when faced with a one-off, target-generated report about one of these coded narratives. Finally, malign creativity can take the form of visual content, including images and videos.

Memes targeting Kamala Harris as part of the sexualized disinformation campaign alleging she "slept her way to the top" were edited, cropped, animated, or otherwise altered in order to evade detection by platforms' automated and human content moderation efforts, and significantly slowed the process of taking action against the content.

Inadequate Definitions of "Targeted Harassment"

As they stand, platforms' policies on "targeted harassment" often do not adequately address the online abuse and gendered disinformation campaigns to which women in public life are subject. For example, Twitter does not have a clear definition of targeted harassment. The platform forbids "hateful conduct" and abusive behavior, as well as social coordination which encourages harm, as discussed earlier in this paper.[67] Unfortunately, the platform has not yet developed either the policy or moderation infrastructure to deal with users inciting attacks which fall below the hate speech threshold, as is often the case with dogpiling. Frequently, targets will deal with a flood of harassment from many users, who take implicit instruction from one leader (as Leta Hong Fincher experienced with Carl Zha). While this is technically social coordination, Twitter's moderation infrastructure requires that targets report each abusive Tweet or instance of harassment separately, making it difficult for moderators to consolidate reports into one issue. Additionally, the most harmful component of dogpiling is the sheer volume of hate, rather than the actual Tweets or messages themselves, and thus individual attacks frequently do not meet Twitter's threshold for action.

Facebook does not have a clear definition of targeted harassment either; applicable policies are spread across several sections of its Community Standards. The platform specifically prohibits hate speech which might cause violence, harassment, disinformation, manipulated content, and "Coordinated Inauthentic Behavior," which the platform defines as coordinated use of fake accounts. Publishing false information results mostly in demotion in the Newsfeed, rather than uniform content removal, a position which has permitted the rampant spread of mis- and disinformation on the platform.[68] While the platform's policies should in theory deal with targeted harassment if taken generously, in reality they are so vague as to be unenforceable for content moderators; Instagram, which is owned by Facebook, faces similar enforcement obstacles.[69] Finally, in order to facilitate discussion about public figures, Facebook applies different standards to abuse directed against public versus private individuals. The company largely exempts public political figures as targets of violent threats; these rules for the most part only apply to attacks on private citizens or low-level public figures. Worryingly, this excludes gendered abuse directed at female candidates from the jurisdiction of Facebook moderators.[70]

YouTube prohibits hate speech, specific harassment of individuals, and misleading or manipulated content.[71] The platform's misleading content policy seemingly only applies to very specific forms of electoral disinformation (such as false claims about candidate eligibility), and a clause about other forms of manipulated media is too vague to be reasonably actionable.[72] In cases of outright harassment, the company has only taken action in the most extreme circumstances and applied minimal sanctions. For example, right-wing provocateur and YouTube creator Stephen Crowder harassed Vox journalist Carlos Maza with explicitly racist and homophobic slurs for two years; it took massive public and internal outcry for the platform to even demonetize Crowder's channel.[73] With YouTube's three-strikes policy, unless harassment or abuse is clearly severe, violators will receive a warning or a strike. Users who violate the Terms of Service will be prevented from uploading new videos for a set period of time, depending on how many prior strikes they have incurred. This sanction, however, is only meaningful for video creators, a fraction of YouTube's user base. There is no real consequence for those who merely use YouTube to post abusive comments.[74]

Several focus group members and interviewees noted that action is rarely taken when they report targeted harassment, likely because the content does not include a direct threat or direction to harass. Instead, a high-follower account will often post a critical or belligerent comment about the target (sometimes referred to as a "dogwhistle"), which incites the account's followers to engage in much worse abuse, known as dogpiling. As Leta Hong Fincher noted: "It was something very personalized, but not over the line harassment," thereby allowing much of the abuse and disinformation to slip through Twitter's enforcement. Similarly, one focus group participant said that even though there might not be direct incitement to abuse, or clear evidence of coordination, "it is coordinated in the sense that one target is chosen and that target is going to get hell."

Nicole Perlroth described a similar phenomenon: "I was basically getting punched in the face on Twitter for four days. And getting DMs that were saying the word 'bitch' in them and 'you don't know shit,' and sort of Tweets that were subtweets, but were very clearly aimed at me, and then caused their own virus storm of more Tweets and DMs." These networks of abuse must be considered when crafting platform policy.

I've had instances in which I reported disability related harassment. And I know that the language is bad because I'm a disabled woman who faces this on a daily basis online...

Lack of Intersectional Expertise in Content Moderation

Similarly, focus group participants highlighted the need for greater investment by platforms in moderation expertise. Content moderators, frequently working under difficult conditions and without proper training, lack the required expertise in gender, race, and other marginalized communities - and yet they are tasked with making split-second decisions that may directly affect the physical and/or psychological safety of users. One participant noted: "I've had instances in which I reported disability related harassment. And I know that the language is bad because I'm a disabled woman who faces this on a daily basis online... and I get the reply that they didn't find it violative of their harassment policies against a marginalized community. So for me, I'm like 'Do you have the right disability experts in place that are helping you understand the language and the nature of campaigns and why this is actually harassment and not just a normal comment to make?'".

Targets Bear the Onus Of Detection and Reporting

As Leta Hong Fincher noted, the act of shielding oneself from abuse as "exhausting" for targets of gendered disinformation campaigns. This trend is echoed by other research; in *Credible Threat,* Sobieraj writes:

> *Hours and days are lost weeding through comments, Tweets, and messages. Many women invested time documenting the abuse. They organized screen shots, printed and filed materials, and otherwise worked to create a paper trail at the request of law enforcement or employers—or simply to have evidence on hand in the event of escalation. Going to court, filing reports, blocking and reporting—all these strategies sap time.*[75]

A focus group participant noted: "It is largely a content moderation problem that really puts the burden on the individual being attacked to report the harassment. And then they're in wait mode and you don't know if you're going to be waiting 24 hours at maybe the best case scenario, or several days, or maybe there will be no action whatsoever." These testimonials explain the drain that self-reporting–as well as responding to abuse more broadly–can have on targets.

In the data collected for this study, the highest volumes of abuse directed at the research subjects were recorded on Twitter. It is heartening to note that Twitter can sometimes successfully intercept this abuse, as occurred after Leta Hong Fincher reached out to the platform. In many cases, the spread of disinformation and abuse occurs indirectly via retweets, allowing the narratives to gather momentum "under the radar," without the need for a great number of direct abusers. With less visibility and a smaller pool of potential abusers, subjects such as Fincher were able to harness the power of blocking/reporting features to cut off their abusers' access. Focus group participants also echoed that blocking and reporting features have been helpful to them in protecting their online presence, offline safety, and their psychological health. One participant noted that "the limiting comments function on Twitter," which allows Tweet authors to decide which users can reply to their Tweets, "while not perfect, feels like a step in the right direction." Despite these successes, the burden of reporting this abuse and advocating for its removal still falls on the subjects.

Abuse Occurs Outside of Highly Visible Areas

Many women who are the subject of online abuse or disinformation campaigns find that abusive content can spread in ways that are not easily monitored outside the scope of dedicated research, such as the current study. For example, on Twitter, rather than abuse being sent solely in reply to a target's Tweet, or as a Quote Tweet or screenshot of the Tweet, abuse can be sent in reply to other content that may or may not tag the target. On Facebook, abuse occurs in the comments of posts, often on a page or group that the target does not administer or may not even know exists. Perpetrators of abuse may also refer to targets by nicknames or employ malign creativity to make it difficult to track the campaigns and narratives, often causing targets to feel overwhelmed when attempting to assess the abuse against them.

The architecture of other platforms also contributes to the effect of hard-to-detect abuse. On TikTok, abuse comes in the forms of video clips in 60 seconds or less, which replay the harassment on loop, and can also occur as text in comment sections. Abusive clips on the platform layer different forms of multimedia to produce content that compounds the severity of abuse. Harassment of public figures is not typically sent directly to them if they do not have a presence on the platform, but the videos are recommended to like-minded users by TikTok's algorithm, helping them reach wider audiences.[76]

Offline Burden on Targets Discounted

Blocking, reporting, muting, and restricting one's account are ways to manage during an abusive episode or disinformation campaign. However, these mitigating features offered by platforms do not account for the psychological and physical effects on targets. One focus group participant noted that these campaigns are "obviously...designed to push [you] down and it has the effect of this grinding away at your resistance, your ability to get through something psychologically. So, I almost see it as a psychological warfare technique."

These efforts also often have effects on women's physical security as well as the ways they participate in public discourse. Like Yeganeh Rezaian, women can be targets of hacking attempts. One focus group participant noted that in addition to hacking, women, particularly women of color, have to worry about doxxing—the publication and malicious use of an individual's personal details, such as her phone number, address, or children's names—online. The participant explained that in her experience and research: "The first thing that happens to women of color, even if they have children or not, is to threaten their children. Threats to their children get them to be quiet." Further, the same participant noted that offline "practical jokes" are also directed toward women of color and have a silencing effect: "Someone sending 100 pizzas to someone's house is just a warning that they know where you live."

Other participants echoed the sentiment that abuse and disinformation campaigns against them drove them from their platforms—effectively forcing them to disengage from their work in the public sphere—for periods of time: "Oftentimes, my solution is to lock down my account and I don't... I either lock it down or I completely go offline and I don't post for days. And that has the silencing effect that we've been talking about, because you don't feel safe to continue speaking, so you don't speak."

Nicole Perlroth, the *New York Times* cybersecurity reporter, noted that the offline effect of such harassment is "very real."

> *It affects our relationships, it affects our mental health. It is horrible, and I don't think that the New York Times ... I think they know that their reporters get harassed on Twitter, and occasionally we try and do something about that. I don't think people understand the viciousness of it, and how much of it women get, and that it just doesn't end. And there's no clear way to respond to it, except silence.*

When online threats generate offline effects, targets have limited recourse with both platforms and law enforcement to ensure abusers face consequences. There is no easily accessible way for women to escalate online attacks generating offline harm to platform review teams. As noted above, the burden of proof is on the target of the campaign to connect the dots between the online and offline abuse, often a time consuming and retraumatizing process that may or may not lead to greater protection. Oftentimes, rather than pursuing consequences for abusers, women choose to remain silent and instead moderate their own behavior.

Few Consequences for Perpetrators of Abuse

In the eyes of those who experience abuse and gendered disinformation, platforms rarely seem to exact meaningful consequences on perpetrators. One focus group participant noted that after a violent threat, her abuser was able to continue posting:

> *I always get rape or death threats. On Twitter, one comment that I got, which I actually reported to Twitter, and they did nothing about it, the man said, 'I'm going to rape your dead body.' And Twitter said it didn't violate their norms, and I was just like, 'Okay. Tell me when it does violate your norms. How about that? Why don't you let me know what violates your norms.'*

Another participant noted that even "successful" reports are often myopic, focusing on a single piece of content rather than a broader trend of behavior from an account:

> *On Twitter you'll flag one Tweet or whatever, but you don't want to have to comb through all the Tweets that that person has made to show additional examples. And they might take action on that one Tweet when*

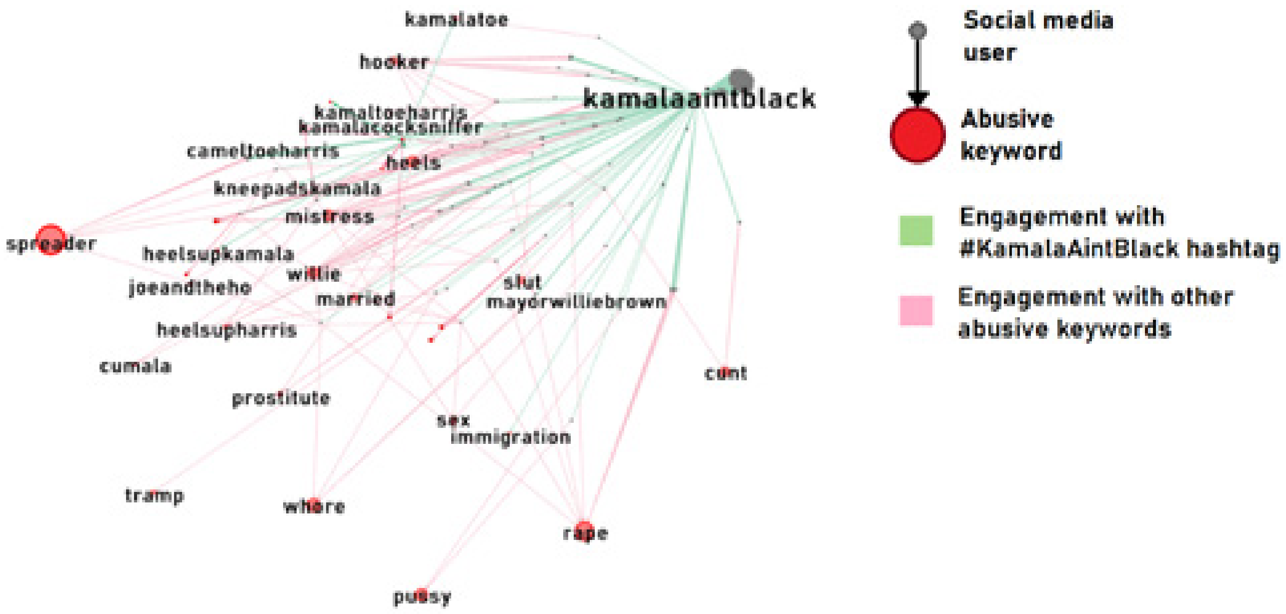
> *really that whole account is problematic. They're not just posting gender-based harassment, they're posting anti-Semitic comments and other attacks on people from other marginalized communities. And it shouldn't take you going through and looking at all of their vile content. That should be something that the platform investigates themselves and takes action on the larger account, not just the one Tweet or post attacking you.*

Nicole Perlroth described an incident in which a colleague had been subject to gendered harassment on Twitter but received no response from the platform. It took her personal intervention to get the content removed, per Twitter's policy:

> *[She] had tweeted out also a screenshot of her reporting it to Twitter, Twitter saying it did not find the report worthy of [being taken down]. And I reported it, and got the same message. And it took me finding the right person at Twitter, sending them an email saying, 'Hi, I'm a New York Times cybersecurity journalist, this is what someone just posted on your site, here is the report saying you did not deem this harassment. This is blatant harassment.' And then Twitter took it down. But it took...urging from a New York Times journalist to do anything. That's ridiculous.*

Women of Color Face Far Greater Threats

Both the quantitative and qualitative data collected in this sample underscore a trend established by other studies: gendered harassment and disinformation campaigns against women of color online are greater in volume and more serious in tone than those that white women face.[77] [78] Several of the women of color in the sample faced multiple gendered or sexualized disinformation narratives in addition to a high level of gendered abuse. White research subjects were the targets of fewer disinformation narratives and received less harassment during the collection period. The



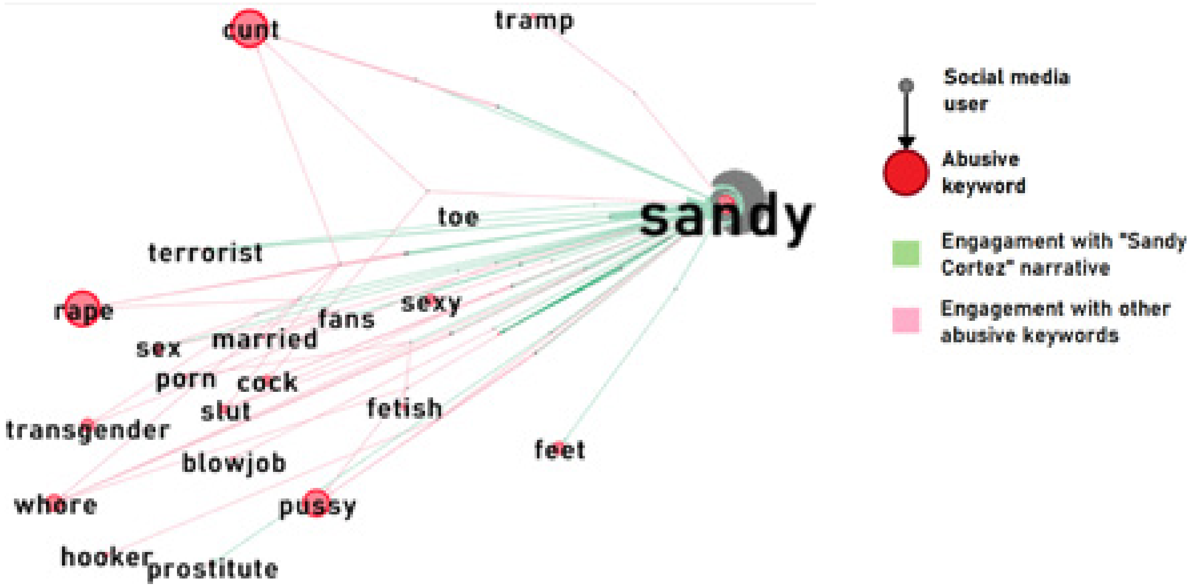


Figure 15: Filtered keyword networks that show which terms users interacted with (pink) alongside #KamalaAintBlack and "Sandy" (green).

network visualizations below demonstrate the intersectionality of the attacks against Kamala Harris and Alexandra Ocasio-Cotrez. Abusers who used the hashtag #KamalaAintBlack across social media also engaged with sexualized narratives when discussing Kamala Harris, such as her relationship with Willie Brown and false narratives around her sexuality with terms such as "HeelsUpHarris," "whore," and "superspreader." Similarly, users who demeaned Alexandra Ocasio-Cortez by calling her "Sandy Cotrez" also used sexualized terms such as "rape," "only fans," and "feet pics."

This trend was echoed in focus group discussions. One focus group participant noted that "women of color deal with both the gender harassment perspective in addition to the racial aspects that come at us. And, sometimes, they're right on top of each other, and layered on top of each other, without people understanding what's happening." Another participant noted that the very reporting tools designed to root out abuse on platforms is also a detriment to women of color's equal participation in the online environment: "The idea of being able to report Tweets and accounts is weaponized against women, and especially women of color," she said. As noted in the Policy Review section above, when platforms design their infrastructure they are not only designed with men in mind, but with white men in mind. Women of color are doubly endangered by these oversights.

In *Credible Threat*, Sobieraj writes that women who belong to marginalized groups "receive more and qualitatively different digital pushback."[79] This is also a challenge for misogyny detection as women of color, and women in non-English-speaking cultural contexts are targeted with different forms of digital pushback than their white, English-speaking counterparts who exist in contexts more familiar to engineers. Some scholars are beginning to research automated misogyny detection in different languages.[80] With the continued rise of malign creativity, platforms, users, and policymakers would be well served by an increase in robust studies in this field that explore both wholly- and semi-automated methods for detecting and blocking misogyny to women in marginalized, and multi-language groups.

Policy Recommendations

The online misogyny women face—and with it, the gendered and sexualized disinformation campaigns against them—is not a problem any one group or sector can solve alone. Currently, there are weaknesses in every adjacent function, from platform policies and content moderation, to political recognition and employer support. The recommendations in this paper focus on critical changes that could have an immediate impact on women's experiences on social media platforms and in public life.

Social Media Platforms

- ***Introduce incident reports.*** Given the pervasive use of malign creativity, coded language, dogwhistles, and dogpiling, all of which may not be evident in user-generated reports that comprise a single piece of abusive content, social media platforms should allow users to compile incident reports that provide more context and a more holistic view of the abuse they are experiencing. Twitter's reporting infrastructure allows users a scaled-down version of this feature, in which users are able to categorize the abuse to which they have been subject, select up to five offending Tweets from the account they are reporting, and write a short explanation of the situation to provide context. However, this feature would benefit from being expanded, so that rather than individual users and pieces of content, it reflects broader patterns and networks of abuse.

  We suggest platforms consider allowing users to create and update broader incident reports that could highlight the harassment they are subject to from a more holistic perspective. Such reports might allow users to add multiple pieces or streams of content from multiple accounts over time, documenting not just a single moment of a campaign against them, but its broader architecture and impact. These reports could be coupled with network visualization and analysis, similar to the type done in this report, to understand the origins and impetuses behind waves of abuse. Finally, platforms should ensure that users have a way to follow, escalate, and appeal the status of incident reports. While this feature would be difficult to immediately roll out at scale, platforms could consider allowing verified users or accounts with a certain number of followers first access to such features, with the goal of widespread distribution in the future, given that many women without large audiences or verified status are also subject to abuse. As a final benefit, taking responsibility for network analysis would indicate that platforms are starting to create a historical record of abuse against women.

- ***Regularly update platform classifiers or keywords to reflect and root out malign creativity.*** Lists of predictably offensive words only account for a fraction of the abuse that women receive online, particularly when dealing with disinformation narratives, which are often contextual and convoluted. For high-profile political candidates, journalists, and academics, social media companies should consider working with individual targets (or their staff) to identify the unique narratives, nicknames, and memes employed against them and reflect such malign creativity in classifiers and keywords that drive platform content moderation, whether automated or human. Where on-platform reporting structures help inform content moderation decisions, platforms should ensure that targets and their staff are aware of this connection and provide adequate training for those managing the public presence of women under attack. Incident reports, as described above, can also help inform updates to classifiers and keywords.

It is important to recognize, however, that updating classifiers and lists of abusive keywords is to some degree a never-ending task; abusers will continue to find ways to employ malign creativity in their campaigns. As with traditional disinformation campaigns, removing the offending content and playing "whack-a-troll" will only go so far toward creating a more equitable, democratic online environment.[81]

- ***Improve automated detection.*** Automated and machine learning models are only as accurate as their training data. This is particularly true for gendered abuse and disinformation; this research faced challenges with collecting the many forms of misogyny, hate speech, and gendered disinformation women face online due to the varied nature of language and malign creativity of perpetrators. Machine learning techniques require carefully gathered and processed training data for the unique racial, cultural, and linguistic environments in which they operate. Any automated method must also incorporate a feedback loop, as recommended by Cathy O'Neal in *Weapons of Math Destruction*.[82] A feedback loop should also continuously update machine learning models using successful reports that individual users have submitted to the platform, as these are ideal examples of human-identified problematic content. Platforms might share these systems in order to eradicate cross-platform abuse, perhaps through a global consortium, as described below.

- ***Together with subject matter experts, update definitions of targeted harassment to explicitly include gender-based abuse and disinformation.*** As discussed above, current definitions of targeted harassment are inadequate, and do not offer subjects of gender-based abuse and disinformation campaigns protection or recourse. After updating their Terms of Service to reflect this widespread problem, platforms should train content moderators to recognize and act against gender-based abuse and employ and consult with more subject-matter experts with intersectional background to assist in policy development, training for content moderations, and oversight of policy enforcement.

- ***Create and enforce transparent and meaningful consequences for persistent abusers.*** Not a single focus group participant was satisfied with social media platforms' enforcement against the online misogyny they had experienced and observed. Much of this dissatisfaction stems from clear incidents of abuse and targeted harassment that meet no or little consequence. The changes outlined above would help highlight persistent abusers; for these individuals, platforms should consider heftier consequences than removal of an isolated piece of content or locking of account features until the offending content is deleted. Platforms might consider imposing—and importantly, enforcing—a tiered escalation system on repeat offenders targeting marginalized groups, including women. Consequences might include locking and placing a public notice on abusive accounts for a period of time; zeroing out the followers/members/subscribers of an abusive account or group, and deplatforming. As with other content moderation decisions, the escalation and enforcement of these policies should be clearly communicated to users, with opportunity for adjudication and appeal.

- ***Introduce nudges or friction to discourage users from posting abusive content.*** Instagram recently introduced a feature designed "to give users pause before posting offensive comments."[83] The platform preemptively identifies such sentiments and "trigger[s] in-app prompts informing people that, if they repeatedly violate the rules, their user accounts could be disabled."[84] During the 2020 U.S. election, Twitter also introduced a new feature nudging users to read articles before sharing them.[85] Using the expanded classifiers described above, platforms should consider similar measures to discourage the posting of abusive and harassing language in the first instance.

- ***Create a cross-platform consortium to track and respond to online misogyny.*** Rather than duplicate threat investigation and enforcement work within individual platforms, social media platforms should consider creating a consortium to address gendered online abuse, similar to the Global Internet Forum to Counter Terrorism (GIFCT). The GIFCT plays an important coordination role in identifying, tracking, and taking action against terrorist content online, but also allows platforms a forum to cooperate on anti-terrorism policymaking and responses, creating industry standards. In the context of gendered abuse, such a body would also serve as a cross-sector convening space and point of contact for civil society groups and government entities working on these issues. Finally, and crucially, it would advance a global view of the problem of online misogyny and gendered disinformation, as it has huge ramifications for women outside of English-speaking, Western countries.

Lawmakers

- ***Include content moderation transparency reporting requirements in social media regulation bills.*** Currently, researchers rely on incomplete data to understand malign activity online, and policymakers lack a complete picture when introducing legislation to mitigate these problems. Lawmakers should mandate that social media companies publicly report about their content moderation activity, with a special emphasis on efforts undertaken to protect marginalized groups, including women. Metrics for reporting should be as consistent across platforms as possible while taking into account each platform's unique infrastructure.[86] Specific metrics for reporting on gendered and sexualized harassment and disinformation might include: number and type of user-generated reports received, number of reports that resulted in action, information about the types and levels of responses, and performance metrics about on-platform mitigation techniques (such as nudges, described above), descriptions of the support and training that staff members and content moderators receive to handle complaints, and description of the personnel and expertise responding to harassment complaints.

- ***Through relevant bicameral committees, create clear standards that prohibit the use of gendered and sexualized insults on official Congressional social media properties, and adopt codes of conduct censuring gender-based abuse by Members.*** On social media properties such as Facebook, female politicians often receive a great deal of often gendered and sexualized abuse in the comments sections of their posts. Developments in First Amendment law, including *Knight First Amendment Institute v. Trump*, in which the Southern District of New York ruled that the President could not block his critics on social media because of their political views, have led to cautious content moderation on official social media accounts.[87] Relevant committees—including the Senate Select Committee on Ethics, the Senate Rules Committee, the House Ethics Committee, and the Committee on House Administration—should develop guidance for Congressional offices on dealing with abusive behavior, in particular gendered and sexualized harassment. Allowing such content—often violative of platforms' Terms of Service—to stand unchallenged in perpetuity on official pages tacitly legitimizes this content and allows it to gain a greater audience. Harassment based on gender or characteristics specific to other marginalized communities should be viewed as distinct from legitimate political criticism.

Similar standards should apply to Members in their official behavior, both on and offline. Elected officials should lead by example, calling out gender-based abuse and harassment when they see it, as well as not engaging in it themselves. Members should not share or employ gendered disinformation or gender-based slurs. Members of the House of Representatives are already prohibited from sharing "visual misrepresentations of other people, including but not limited to deep fake technology," and in official communications may not "disparage" other Members, including through ad hominem attacks.[88] Given the widespread nature of attacks against women in politics, Congress should develop more detailed standards for decorum around gender issues and their intersection with trends such as disinformation.[89]

- ***Reauthorize the Violence Against Women Act (VAWA) and include provisions against online gender-based harassment.*** The 2019 VAWA Reauthorization Act has not received a vote in the Senate, leaving crucial protections for victims of gender-based violence lapsed. When the next Congress considers VAWA reauthorization, lawmakers should add provisions to support targets of online gender-based harassment, including budgetary allocations to build law enforcement awareness about and increase investigation of online gender-based threats. We concur with Sarah Sobieraj that "even in contexts such as the United States, where much of the [online] abuse will not be legally actionable, responding [law enforcement] officers can play an important role in supporting victims. Helping officers become well versed on attackers' tactics and the primary venues for online hostility is an important first step."[90] Lawmakers could also consider including platform transparency and disclosure requirements about gender-based abuse and disinformation (described above) in the VAWA reauthorization bill.

Employers

- ***Develop support policies for employees facing online harassment and abuse.*** For many public-facing industries, including the media, academia, think tanks, and government, employee engagement on social media is now critical to both brand and individual success. Many employers have policies relating to employees' or affiliates' use of social media, but far fewer have support mechanisms for those undergoing online abuse as a result of their work-related social media engagement. Employers should consider providing mental health services, support for employees or affiliates' legal fees and other expenses (such as anti-doxxing service subscriptions), as well as outlining clear mechanisms for targets to report such campaigns against them to official communications and human resources staff. Organizations with non traditional arrangements with affiliates—such as think tanks that associate with fellows who are not technically employees of the institutes they represent—must recognize that these individuals may also require support and benefits beyond institutional affiliation.

Organizations can also engage with social media platforms on behalf of targets of abuse, adding additional legitimacy and urgency to targets' reports; in instances in which professional organizations highlight the campaigns against their members, such as when the Coalition for Women in Journalism released a statement in support of Leta Hong Fincher, platforms seem to take action more quickly. Above all, employers should recognize that the experience of online gendered abuse is extremely isolating, often frightening, and is aimed to have a silencing effect; they should adopt practices and policies to ensure their employees and affiliates feel comfortable speaking out and seeking support.

Conclusion

Lawmakers, technology policymakers, and social media users may be tempted to discount or ignore the problem of gendered and sexualized abuse and disinformation online. It is sprawling, and, as this study and others have found, even assessing it in the broadest terms presents obstacles in detection and analysis. Some might point to the number of female elected officials and key political appointments in the incoming Biden-Harris administration as a sign that gendered abuse and disinformation is not a significant impediment to women's participation in political and public life. The volume of abuse gathered for this research is a testament to the dedication of these women, rather than evidence of a more equitable path.

Beyond the individual and systemic repercussions of online gendered and sexualized abuse, the phenomenon also has implications for national security. These narratives can be readily exploited by malign foreign actors, as demonstrated by the case studies from Iran, China, and Russia in this report. In its current form, the infrastructure of social media platforms facilitates the spread of such narratives and enables their weaponization.

As this report indicates, abusers' malign creativity means addressing this problem will not be easy, but dedicated, collective efforts by platforms, policymakers, and employers can elevate this from its misidentification as a special interest issue to a question of the right to equal participation in democracy and public life without fear of abuse and harassment. An aspiring female politician, journalist, or activist should not need to reconcile her ambition with a lifetime of online abuse. Furthermore, it should not be up to women themselves to identify and enforce against such abuse.

As this study has demonstrated, even the women achieving historic "firsts" in American life have been subject to a variety of gendered and sexualized attacks, or worse, widespread gendered and sexualized disinformation campaigns fueled by malign creativity. They have endured and succeeded in spite of this vitriolic, demeaning, and silencing climate. Others interviewed in this study have admitted the silencing effect that online misogyny has had on them, pushing them to lock down their accounts, to reconsider what to write, say, or share, or moving them to delete content that has generated abuse. The effects of such campaigns are broader than just the target in question; for every incident in which gendered and sexualized narratives against a high-profile female target are allowed to proliferate, influencing the target's public presence, thousands of other women see those narratives and consider whether to engage at all.

We have the tools to reverse this trend. By employing the creativity and technological prowess democracies engender—changing how we report abuse, how we respond to it, and how we support those who experience it—we can make a pariah of malign creativity and the misogyny that inspires it.

The initial publication of this report did not provide specific commentary on Tom Fowdy's blog post or detail his relationship to Chinese official state media outlet CGTN. Fowdy is a freelance author for CGTN. The Wilson Center regrets any misunderstandings and has amended the text to provide greater specificity.

About the Authors

Nina Jankowicz serves as the Disinformation Fellow within the Wilson Center's Science and Technology Innovation Program. She is the author of *How to Lose the Information War: Russia, Fake News, and the Future of Conflict* (Bloomsbury/IBTauris 2020). She has worked as a strategic communications adviser to the Ukrainian Foreign Ministry under the auspices of the Fulbright Public Policy Program and managed democracy assistance programs to Russia and Belarus at the National Democratic Institute. Ms. Jankowicz received her MA in Russian, Eurasian and East European Studies from Georgetown University and her BA in Russian and Political Science from Bryn Mawr College.

Jillian Hunchak is a research analyst who specializes in far-right extremism in French- and English-language contexts. She currently carries out research at Moonshot CVE and Insight Threat Intelligence, and has completed internships at the Organization for the Prevention of Violence (OPV) and the International Centre for the Study of Radicalisation (ICSR). She holds an MA from King's College London in Terrorism, Security and Society.

Alexandra Pavliuc is a PhD student at the Oxford Internet Institute who studies the spread of disinformation online. She is a researcher on the Alternative News Networks Project (part of the Computational Propaganda Project at the OII). Her work has been published by *Defence Strategic Communications* and featured on CBC News. Ms. Pavliuc holds degrees in Data Science and Professional Communication from City, University of London and Ryerson University.

Celia Davies works on interventions programming at Moonshot CVE, with specialisms in gender-based violence and disinformation. She also heads up organizational security and chairs the ethics review committee. Prior to joining Moonshot, Celia led a series of human rights and media development projects in Russia, Ukraine, and the South Caucasus. She holds degrees from the University of Cambridge (BA English Literature) and the University of Edinburgh (LLM International Law). She was also a Winston Churchill Memorial Trust Fellow (2014), conducting research on democratic participation.

Shannon Pierson studies the impact of disinformation on democracies and social media regulation as a Research Assistant Intern within the Wilson Center's Science and Technology Innovation Program. She has consulted on multiple projects for the Microsoft Corporation's Defending Democracy Program on international election security and internet governance legislation. Ms. Pierson holds a degree in International Studies, with a focus in cybersecurity, from the University of Washington.

Zoë Kaufmann is a Research Assistant Intern for the Wilson Center's Science and Technology Innovation Program, where she researches political disinformation and the implications of social media user policies. She is pursuing a BA in History at Bryn Mawr College and has worked on several political campaigns.

About Moonshot CVE

Moonshot CVE is a London-based social enterprise specializing in countering violent extremism and other online harms. Moonshot designs new methodologies and technologies to support more effective, scalable, and sustainable responses to the threats posed by harms such as violent extremism, disinformation, and gender-based violence, online and offline. Moonshot's work ranges from software development and digital capacity building to leading global counter-messaging campaigns. Since its founding in 2015, Moonshot has worked with a range of governments and private sector organizations around the world, including local and national governments, multilateral organizations, and major technology companies. Recent collaborations include a multi-year media literacy project with USAID, a transnational project on violent extremism with the International Organization for Migration, elections integrity projects with the Anti-Defamation League, and emergency COVID-19 response initiatives with USAID.

Appendix A: Platform Policy Quick Reference

Key:

- 🟥 Forbidden
- 🟦 Unclear
- ⬜ Permitted where not otherwise forbidden (i.e. an action violates another policy like breaking the law)

	Threatening violence	Abuse	Harassment	Targeted harassment against protected groups	Coordinated abuse	Notes
Twitter	Forbidden, including threats of death or sexual assault.	Forbidden under abusive behavior policy.	Included in Hateful Conduct / abuse policy.	Forbidden under hateful conduct policy.	Forbidden under coordinated abuse policy.	"Public interest" exceptions.
Facebook/ Instagram	Forbidden.	Forbidden under Community Standards.	Forbidden.	Covered under Hate speech policy.	Dogpiling vaguely referenced in Group Restrictions.	Disinformation is forbidden.
YouTube	Forbidden under Community Guidelines.	Forbidden under Community Guidelines.	Forbidden under Community Guidelines.	Forbidden under Hate speech policy. Distinguishes between hate speech and harassment by target.	🟦	
Reddit	Forbidden under Harassment and bullying policy, as well as Violent content policy.	Forbidden under Harassment and bullying policy.	Forbidden under Harassment and bullying policy.	Forbidden under Hate content policy.	🟦	
TikTok	Forbidden under Terms of Service and Community Guidelines.	Forbidden under Terms of Service and Community Guidelines.	Forbidden under Terms of Service and Community Guidelines.	Forbidden under Terms of Service and Community Guidelines.	🟦	Mis/ Disinformation is forbidden.
Twitch	Forbidden under Community Guidelines.	Forbidden under Harassment and abuse policy.	Forbidden under Harassment and abuse policy.	Forbidden under Harassment and abuse policy.	🟦	
4chan	Nothing which might violate the law.	🟦	🟦	🟦	🟦	
8chan/ 8kun	Nothing which might violate the law.	🟦	🟦	🟦	🟦	
Gab	Nothing which might violate the law.	🟦	🟦	🟦	🟦	
Parler	Nothing which might violate the law. Forbidden under Community Guidelines / "Elaboration on Guidelines".	🟦	🟦	🟦	🟦	

Appendix B: Platform Architecture Quick Reference

Platform	Structure	Messaging	Anonymity	Privacy options
Twitter	Shortform text, images, videos, links displayed in a Newsfeed. Fleets (stories). Users can retweet or Quote Tweet others' content.	Direct Messages. Default setting permits any user to DM any other user; individuals can restrict this.	Users must create accounts which they register with a phone number or email.	Users can make tweets visible only to approved followers.
Facebook	Text, images, videos, links displayed in a Newsfeed from friends or group members. Stories. Sharing function.	Facebook Messenger. Users can message their friends directly; users who they are not friends with will send message requests.	Users must create accounts which they register with a phone number or email.	Users can set their profile to private; the majority of content will only be visible to friends.
Instagram	Images and videos from followed users displayed in a feed. Stories.	Users can message public or followed accounts directly; this feature can be turned off. Otherwise, users who are not followed will send message requests.	Users must create accounts which they register with a phone number or email.	Users can set their profile to private; content will only be visible to approved followers.
YouTube	Video streaming, including recommendation algorithm. Likes, dislikes.	None.	Users must create accounts which they register with a phone number or email.	Users can set their videos or channels to private.
Reddit	Community-based forums. Threads on specific topics, up-votes and down-votes.	None.	Users must create accounts which they register with a phone number or email.	None. User profiles show activity.
TikTok	Video-sharing app. Users' "For You" pages show videos chosen by algorithm and videos from followed accounts.	Private messages. Users can permit messages from anyone or only mutual followers.	Users must create accounts which they register with a phone number or email.	Users can set their account to private, making their videos visible only to followers. Likes are private unless changed in settings.
4chan	Imageboards organized by topic.	None.	All users are anonymous. Within each thread, posters are assigned random numerical IDs using a combination of cookies and IP tracking.	Not necessary.
	Imageboards organized by topic. Difficult to access; the only public point of entry is through TOR (on the Dark Web).	None.	All users are anonymous. Within each thread, posters are assigned random numerical IDs using a combination of cookies and IP tracking.	Not necessary.
Gab	Text, images, videos, links displayed in a Newsfeed.	Chat app similar to Discord. Both private and public chat rooms; messages in private rooms are encrypted. Messages are deleted after 30 days.	Users must create accounts which they register with a phone number or email.	None.
Parler	Text, images, videos, links displayed in a Newsfeed.	Private messaging between users who follow one another. Otherwise, messages are sent as requests.	Users must create accounts which they register with a phone number or email.	Users can set their profiles to private; content will only be visible to approved followers.

Appendix C: Keywords Used in Data Collection and Analysis

The following keywords were used to identify social media posts that contained transphobic, racist, and sexualized content that might lead to the discovery of disinformation narratives. Variations of some of these keywords were used in parts of the analysis (for example, only the word "Aroush" was used to identify all posts containing the "Kamal Aroush" narrative). Many of these subject specific keywords were discovered throughout the analysis process and added to the final data collection (such as "married her brother" for Ilhan Omar, and "HeelsUpHarris" for Kamala Harris). This process further underscores evidence of malign creativity and coded language to target and abuse women online.

Generalized Gender-Based Abuse	Sexual Disinformation	Transphobic Disinformation	Racist Disinformation
Bartender	BJs Harris	Caitlyn Jenner	kamalaharrisisananchorbaby
Bitch	Brother Fucker	kamal aroush	ilhanomarhatesamerica
Bitchigan	Cumala	transsexual	kamallahakbar
Bitchmer	Deepfake	ladyboy	terrorist
Camel Toe Harris	Feels Up Heels Up	tranny	kamalaaintblack
Cameltoe Harris	Headboard Harris	transgender	immigration fraud
Chubby	Heels up Harris	transvestite	sandy cortez
Cow	Heelsupkamala		
Cunt	Hoecahontas		
Dyke	Horizontal Harris		
Fat	joe and the ho		
Fugly	joe/blow 2020		
Hooker	kamalacocksniffer		
Horse Face	kamalasutra		
Jacinderella	kamalingus		
kamalatoe	kneepadskamala		
kamalatoes harris	legswideopenkamalaharris		
kamaltoe harris	married her brother		
kumaltoe harris	pee pads and knee pads		
kunt kamala	stretchin gretchen		
mistress	super spreader		
nudes	sex cults		
onlyfans	sex trafficking		
overweight	willie brown's whore harris		
priti ugly	NXIVM		
prostitute			
pussy			
rape			
roastie			
occasional cortex			
slut			
tramp			
ugly			
whore			
witch			
witchmer			
blowjob			
camel-face harris			
cock			
Coom			
Feet pics			
Fetish			
waitress			

Appendix D: Semi-Structured Interview Questions

- What have been the main characteristics of gendered disinformation that has been targeted at you online?
- What are other characteristics that you feel women generally face, but that you have not necessarily faced?
- How would you define what you face online? Please provide a term, and a longer statement.
- Have you felt that there was a level of coordination between your attackers when they attacked you online?
- Do you feel that there is a difference between the disinformation targeted against you, and other disinformation you are aware of?

Appendix E: Focus Group Script

We have invited you here today because we are interested in your unique perspectives on the challenges women in public life face online. Each of you study Internet trends in some capacity and as women, you may have faced harassment and misogyny online yourselves. We are grateful for the opportunity to tap into the deep knowledge in this virtual room and workshop policy solutions for platforms and governments.

We will be recording this focus group for note taking purposes. But, in the final report, no comments will be personally attributed to anyone. We will begin the recording now.

PT 1: Discussion on the topic at hand

1. What are the main characteristics of the misogyny that has been targeted at you, or that you have seen be targeted at other women online? (we're aiming for types of content, ways that women are approached, etc)
    a. Do you think there is a level of coordination in these attacks? If yes, how so?
    b. Do you think these types of attacks have an element of disinformation (or manipulated information) to them?
    c. How would you define what you're seeing?

PT 2: Workshopping policy solutions for platforms and governments to these types of attacks (start with social media platforms, then move onto governments)

1. Can you think of an example of a social media platform's regulation to counter misogyny on their platform that you think was successful? Let's go around and each give an example, if we have one.
    a. Let's discuss the main characteristics of your successful examples.
2. Can you think of an example of a social media platform's regulation to counter misogyny on their platform that you think was unsuccessful? Let's go around and each give an example, if we have one.
    a. Let's discuss the main characteristics of your unsuccessful examples.

3. Based on our above discussion, and from what you've seen happening or have experienced online, what can platforms do to improve this problem?

4. *(Ask questions 3-5 again but about successful/unsuccessful government policies)*

5. If you could redesign the internet, a social media platform, or social media itself from scratch, how would you design it to ensure that women feel safe, and are able to participate freely?

6. If you could redesign the internet, a social media platform, or social media itself from scratch, how would you design it to ensure that women feel safe, and are able to participate freely?

Endnotes

[1] Lucina Di Meco, "#SHEPERSISTED: Women, Politics & Power in the New Media World," The Wilson Center, Fall 2019.

[2] "Naked untruth; sexualised disinformation." *The Economist*, November 7, 2019, 47 (US).

[3] Nina Jankowicz, "How Disinformation Became a New Threat to Women," *Coda Story*, December 11, 2017.

[4] Aja Romano, "Deepfakes are a real political threat. For now, though, they're mainly used to degrade women," *Vox*, October 7, 2019.

[5] Jane Lytvynenko and Scott Lucas, "Thousands Of Women Have No Idea A Telegram Network Is Sharing Fake Nude Images Of Them," *BuzzFeed News*, October 20, 2020.

[6] Frances Perraudin, "Alarm Over Number of Female MPs Stepping Down After Abuse," *The Guardian,* October 31, 2019.

[7] Di Meco, 2019.

[8] Sarah Sobieraj, *Credible Threat*: *Attacks Against Women Online and the Future of Democracy* (Oxford University Press, 2020).

[9] Di Meco, 2019.

[10] Sobieraj 2020, 5.

[11] Ibid., 10.

[12] Ibid.

[13] Amnesty International, "Toxic Twitter," March 2018.

[14] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney, "Online Harassment, Digital Abuse, and Cyberstalking in America," Data & Society, 2016.

[15] Kirsten Zeiter, Sandra Pepera, and Molly Middlehurst, "Tweets that Chill," National Democratic Institute for International Affairs (NDI), 2019.

[16] Ibid.

[17] "#NotTheCost: A Call to Action," National Democratic Institute, 2016.

[18] Sobieraj 2020, 119.

[19] Ibid., 112.

[20] Samantha Bradshaw, "The Gender Dimensions of Foreign Influence Operations," Global Affairs Canada, 2019.

[21] Jankowicz 2017.

22 Ellen Judson, Asli Atay, Alex Krasodomski-Jones, Rose Lasko-Skinner, and Josh Smith, "Engendering Hate: The contours of state-aligned gendered disinformation online," Demos, October 2020.

23 Ibid.

24 Evelyn Douek, "What is Coordinated Inauthentic Behavior on Social Media?," *Slate*, July 2, 2020.

25 Sobieraj 2020, 4.

26 Sarah Sobieraj, "Bitch, slut, skank, cunt: patterned resistance to women's visibility in digital publics," *Information, Communication & Society* 21, Volume 11 (2018).

27 Ibid.

28 Judson, Atay, Krasodomski-Jones, Lasko-Skinner, and Smith, "Engendering Hate," 2020.

29 See Appendix A for more detail on different platforms' user policies.

30 Twitter Support, "Notices on Twitter and what they mean," Twitter Help.

31 Twitter Support, "Synthetic and manipulated media policy," Twitter Help.

32 Facebook, "Regulated Goods," Facebook Community Standards.

33 Platforms are less likely to enforce against harassment of public citizens than harassment of private citizens. For example, users cannot "purposefully expose" public citizens to calls for their own death or threaten "severe violence", but public figures are not protected by Facebook's Violence and Incitement policy, which only applies to "minor public figures" and private citizens. Facebook, "Bullying," Facebook Community Standards. Facebook, "Violence and Incitement," Facebook Community Standards.

34 Facebook, "Bullying," Facebook Community Standards.

35 "YouTube Community Guidelines enforcement," Google Transparency Report, Google, September 2020.

36 Kaitlyn Tiffany, "Reddit Squashed QAnon by Accident," *The Atlantic,* September 23, 2020.

37 Kaitlyn Tiffany, "Reddit Is Finally Facing Its Legacy of Racism," *The Atlantic*, June 12, 2020.

38 Rachel Lerman, "The conservative alternative to Twitter wants to be a place for free speech for all. It turns out, rules still apply," *The Washington Post*, July 15, 2020.

39 Nick Hopkins, "Revealed: Facebook's internal rulebook on sex, terrorism and violence," *The Guardian*, January 3, 2018.

40 Andrew Marantz, "Why Facebook Can't Fix Itself," *The New Yorker*, October 12, 2020. Casey Newton, "The Trauma Floor: The secret lives of Facebook moderators in America," *The Verge,* February 25, 2019. Zach Whittaker, "Facebook to pay $52 million to content moderators suffering from PTSD," *TechCrunch,* May 12, 2020. Casey Newton, "The Terror Queue: These moderators help keep Google and YouTube free of violent extremism—and now some of them have PTSD," *The Verge,* December 16, 2019.

[41] Isaac Stanley-Becker and Elizabeth Dwoskin, "Trump allies, largely unconstrained by Facebook's rules against repeated falsehoods, cement pre-election dominance," *The Washington Post*, November 1, 2020.

[42] Donie O'Sullivan and Alaa Elassar, "Twitter bans posts wishing for Trump death. The Squad wonders where that policy was for them," CNN, October 3, 2020.

[43] Nicholas Bogel-Burroughs, "What We Know About the Alleged Plot to Kidnap Michigan's Governor," *The New York Times*, October 9, 2020.

[44] It is important to note that while some coordinated disinformation campaigns may be orchestrated by foreign countries, this project's quantitative analysis did not examine state-sponsored disinformation.

[45] Clara Hendrickson, "What is '8645'? Whitmer's pin an anti-Trump message using restaurant industry slang," *Detroit Free Press*, October 19, 2020.

[46] Ibid.

[47] OnlyFans is a subscription service through which creators can monetize their content, charging subscription fees to "fans." It is popular with sex workers, who share explicit content with users of the site.

[48] Gretchen Whitmer was the only subject for whom Twitter did not return the highest number of data points, with Parler producing the most results. This is likely because abuse towards Whitmer originated on Parler and only later migrated towards Twitter.

[49] Andrea Perez-Balderrama, "Gov. Whitmer responds to 'Big Gretch' rap song made about her," *Detroit Free Press*, May 4, 2020.

[50] The authors note that data collection for Priti Patel contained a high percentage of false positives due to an investigation that broke during the collection period relating to human trafficking in the United Kingdom.

[51] The copypasta message's use in reference to Emma Gonzalez was quoted in the introduction of Sobieraj's *Credible Threat* to illustrate the sorts of vitriol women face online, though its repeated use towards multiple women was not discussed.

[52] Nina Jankowicz, interview with Yeganeh Rezaian, October 2, 2020, via Zoom.

[53] People's Daily China, Twitter Post, July 12, 2020.

[54] Leta Hong Fincher, Twitter Post, July 12, 2020.

[55] Nina Jankowicz, interview with Leta Hong Fincher, October 23, 2020, via phone.

[56] Isobel Cockerell, "Pro-Beijing influencers and their rose-tinted view of life in Xinjiang," *Coda Story*, August 7, 2020.

[57] Cockerell 2020.

[58] Leta Hong Fincher, Twitter Post, August 17, 2020.

[59] Bradshaw 2019.

[60] Ibid.

[61] *The Economist* 2019.

[62] Jankowicz 2017.

[63] Nina Jankowicz, Interview with Nicole Perlroth, October 21, 2020, via Zoom.

[64] "*White NYT reporter tells black people they didn't vote for Hillary in sufficient numbers because Russia duped them*," RT, August 10, 2020.

[65] RT, *Twitter Post*, August 10, 2020.

[66] "Russian trolls' chief target was 'black US voters' in 2016," BBC, October 9, 2019.

[67] Twitter Support, "Hateful conduct policy," Twitter Help Center. Twitter Support, "Abusive behavior," Twitter Help Center. Twitter Support, "Coordinated harmful activity," Twitter Help Center.

[68] Facebook Community Standards, "False News," Facebook.

[69] Taylor Lorenz, "Instagram Has a Massive Harassment Problem," *The Atlantic,* October 15, 2018.

[70] Facebook Community Standards, "Hate speech," Facebook. Facebook Community Standards, "Harassment and Bullying," Facebook.

[71] YouTube, "How does YouTube protect the community from hate and harassment?," Google. YouTube Help, "Spam, deceptive practices, & scams policies," Google Support.

[72] YouTube, "Spam, deceptive practices, & scams policies," Google Support.

[73] Daisuke Wakabayashi, "YouTube takes tougher stance on harassment," *The New York Times*, December 11, 2019. Emily Stewart, "We don't want to be knee-jerk": YouTube responds to Vox on its harassment policies," *Vox*, June 10, 2019.

[74] Those who repeatedly comment abuse may have their channel terminated, although this is reserved for the most severe cases. YouTube Help, "Community Guidelines strike basics," Google. YouTube Help, "Channel or account terminations," Google.

[75] Sobieraj 2020, 110.

[76] Julia Alexander, "TikTok reveals some of the secrets, and blind spots, of its recommendation algorithm," *The Verge*, June 18, 2020.

[77] Amnesty International UK, "UK: online abuse against black women MPs 'chilling," 9 June 2020.

[78] Amnesty International 2018.

[79] Sobieraj 2020, 97.

[80] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti, "Misogyny Detection in Twitter: A Multilingual and Cross-Domain Study," *Information Processing & Management,* 57, no. 6 (November 1, 2020).

[81] Nina Jankowicz, "The Only Way to Defend Against Russia's Information War," *The New York Times*, September 25, 2017.

[82] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* (New York: Crown, 2016).

[83] Adriana Lee, "Instagram Discourages Bullying, Brings Back Photo Maps, More," *Yahoo News*, October 6, 2020.

[84] Ibid.

[85] Taylor Hatmaker, "Twitter plans to bring prompts to 'read before you retweet' to all users," *TechCrunch*, September 24, 2020.

[86] Nina Jankowicz and Shannon Pierson, "Freedom and Fakes: A Comparative Exploration of Countering Disinformation and Protecting Free Expression," The Wilson Center, December 2020. 31.

[87] Knight First Amendment Institute v. Trump, 928 F.3d 226 (2d Cir. 2019).

[88] US House of Representatives Committee on House Administration, "The House of Representatives Communications Standards Manual," January 2020.

[89] Cécile Guerin and Eisha Maharasingam-Shah, "Public Figures, Public Rage: Candidate abuse on social media," Institute for Strategic Dialogue, October 5, 2020.

[90] Sobieraj 2020, 150.

WOODROW WILSON INTERNATIONAL CENTER FOR SCHOLARS

The Woodrow Wilson International Center for Scholars, established by Congress in 1968 and headquartered in Washington, D.C., is a living national memorial to President Wilson. The Center's mission is to commemorate the ideals and concerns of Woodrow Wilson by providing a link between the worlds of ideas and policy, while fostering research, study, discussion, and collaboration among a broad spectrum of individuals concerned with policy and scholarship in national and international affairs. Supported by public and private funds, the Center is a nonpartisan institution engaged in the study of national and world affairs. It establishes and maintains a neutral forum for free, open, and informed dialogue. Conclusions or opinions expressed in Center publications and programs are those of the authors and speakers and do not necessarily reflect the views of the Center staff, fellows, trustees, advisory groups, or any individuals or organizations that provide financial support to the Center.

Woodrow Wilson International Center for Scholars
One Woodrow Wilson Plaza
1300 Pennsylvania Avenue NW
Washington, DC 20004-3027

The Wilson Center

🌐 www.wilsoncenter.org
✉ wwics@wilsoncenter.org
f facebook.com/woodrowwilsoncenter
🐦 @thewilsoncenter
📳 202.691.4000

Science and Technology Innovation Program

🌐 www.wilsoncenter.org/program/science-and-technology-innovation-program
✉ stip@wilsoncenter.org
🐦 @WilsonSTIP
📳 202.691.4321

Wilson Center

Science and Technology Innovation Program