







# SOCIAL MEDIA FAILS WOMEN







## THE REPORT CARD







## Methodology

This report card was conducted in partnership with the Institute for Strategic Dialogue (ISD). UltraViolet and ISD developed this report card over several weeks, analyzing both existing policy and recent or ongoing news coverage of platforms and their policies. First, UltraViolet reviewed existing user policies at Facebook, Instagram, Twitter, Reddit, YouTube, and TikTok. ISD then reviewed these policies against each aspect of UltraViolet's Policy Recommendations, giving a numerical score based on the policies in place at said corporation compared to the recommendations. The numerical score ranges from 0 to 4, 0 being the policy does not exist at said company and 4 being that the platform is in line with policy recommendations. ISD also provided explanations for each score and notes for where an assessment could not be made. Once reviewed, UltraViolet averaged out the score for each company, splitting Facebook into two as Facebook and Instagram are often used separately, and used Harvard University Graduate School of Education's grading rubric to identify a letter grade appropriate to the averaged score.

## SOCIAL MEDIA FAILS WOMEN: THE REPORT CARD

Policy	 Facebook	 Instagram	 Twitter	 YouTube	 TikTok	 Reddit
Includes misogyny, misogynoir, transmisogyny, gendered and racialized disinformation, cyberstalking, sexual harassment, and revenge porn in hate speech rules and transparency reports	<b>D</b>	<b>D</b>	<b>B</b>	<b>D</b>	<b>C</b>	<b>B</b>
Bans and explicitly includes cyberstalking, sexual harassment, revenge porn, deadnaming, misgendering, and other means of virtual sexual exploitation and harassment as forms of misogyny and hate speech	<b>D</b>	<b>D</b>	<b>D</b>	<b>D</b>	<b>D</b>	<b>C+</b>
Creates a clear and enforceable process, such as a weighted system of escalating warnings and strikes based on the level of offense, that leads to permanent deplatforming in the event of frequent or severely abusive violations of policies related to misogyny and health-related disinformation, including disordered eating, extreme dieting, vaccines, abortion, and pregnancy	<b>D</b>	<b>D-</b>	<b>B</b>	<b>D+</b>	<b>C</b>	<b>B</b>

Policy	 Facebook	 Instagram	 Twitter	 YouTube	 TikTok	 Reddit
Establishes a clear and enforceable process for removing content and deplatforming users, user reporting, and closing loopholes for health-related disinformation, including disordered eating, extreme dieting, body shaming, body dysmorphia, transgender people, vaccines, abortion, and pregnancy	<b>F</b>	<b>F</b>	<b>C</b>	<b>D</b>	<b>C</b>	<b>D</b>
Ensures that content that violates these policies can be reported by any user who witnesses it, not just the victims or victims' representatives	<b>F</b>	<b>F</b>	<b>A</b>	<b>D+</b>	<b>B</b>	<b>A</b>
Ensures that users are able to appeal a moderation decision and have the opportunity to provide additional related posts, cultural context, and suggested updates to what content is considered hateful	<b>D</b>	<b>D</b>	<b>D</b>	<b>D</b>	<b>D</b>	<b>D</b>
Directs those who have been exposed to extremist groups and hate content to resources for countering extremism	<b>D</b>	<b>F</b>	<b>F</b>	<b>F</b>	<b>F</b>	<b>F</b>
Creates a human-monitored help desk to support victims of harassment, stalking, and other online sexual abuse; takes immediate action to protect survivors' privacy and safety	<b>F</b>	<b>F</b>	<b>F</b>	<b>F</b>	<b>F</b>	<b>C</b>

Policy	 Facebook	 Instagram	 Twitter	 YouTube	 TikTok	 Reddit
Ensures that algorithms and human moderators do not base moderation and promotion on the political alignment of the content or the user who posted it	<b>D</b>	<b>D</b>	<b>C</b>	<b>D</b>	<b>C</b>	<b>B</b>
Supports and protects victims of harassment, hate disinformation, and abuse; centers the experiences of marginalized people and groups	<b>C</b>	<b>C</b>	<b>C</b>	<b>D</b>	<b>D</b>	<b>C</b>
Ensures that algorithms encourage users to engage more frequently with legitimate, well-researched news and peer-reviewed articles that offer opposing viewpoints, not with disinformation, conspiracies, opinion posts, or claims that are not backed by science	<b>D</b>	<b>D</b>	<b>B</b>	<b>C</b>	<b>D</b>	<b>D</b>
<b>Letter grade</b>	<b>D-</b>	<b>F</b>	<b>C-</b>	<b>D</b>	<b>D+</b>	<b>C</b>

**Across the board, social media platforms fail the test when it comes to creating an internet experience that is safe and inclusive for Black women, women of color, and LGBTQ people.**

Even on platforms where written policies exist, enforcement is, at best, spotty. The platforms' own research shows the scope of the harm -- from the prevalence and spread of disinformation to the conservative bias of the algorithms.

**Social media corporations, including Facebook, Instagram, Twitter, Reddit, YouTube, and TikTok, must adopt our recommended policy changes to their platforms, in order to curb the real, ongoing harms caused by the spread of hate, disinformation, and harassment online.** However, change cannot depend on just corporations. The government must step in to regulate the platforms, because they are too powerful and too focused on profits to put people first.

# KEY TAKEAWAYS

- **Only one of the platforms currently has a human-monitored help desk** for victims of harassment threats, stalking, or bullying.
- The platforms need to **catch up to the reality and harms of disinformation plaguing their sites**. Some platforms do not have any bans on disinformation at all, while others have limited bans that address COVID-19 and election disinformation.
- Written policies may sound promising, but there is a **lack of accountability** for the fact that there is no enforcement of the policies.
- The platforms claim that they **do not base moderation decisions on the political alignment of the content or the post, but numerous studies, data leaks, and information from whistleblowers have confirmed that is just not true**. At Facebook, Mark Zuckerberg has directly overridden moderation decisions and rules to protect his own interests.
- **Reddit, the platform once known as a haven for white supremacy and misogyny, has set itself apart as an industry leader in the last two years by making robust changes to its hate speech policies**, as well as to moderation and enforcement. However, it is concerning that the platform once known as a home for misogynist Red Pillers is now leading the pack on fighting misogyny.
- Overall, the platforms **must do a lot more to actively fight extremist ideologies currently thriving online**, including connecting users with resources for countering extremism and hate, and by prioritizing factual content to counter extremist ideologies ignited by disinformation; **Twitter is leading the way on this by forming relationships with news organizations, Reuters and the Associated Press**.
- **TikTok has stated on hiring pages they provide support for the physical and psychological health and well-being of content moderators**, however due to lack of transparency on broader platforms' internal policies, worker protections, and contractor resources, we are unable to grade platforms with precision on this matter.
- **Transparency reports provided by the platforms do not highlight how easy or difficult it is for users to report problematic content or how quickly the reports are being handled**. Recent user experience surveys indicate that there is significant need for improvement on this front.
- The results of this report demonstrates **how far all social media platforms have to go before they can claim to be a safe space for women and girls, particularly Black women, women of color, and LGBTQ+ people**.

## FACEBOOK: D-

- Facebook fails to address misogyny, misogynoir, transmisogyny, gendered disinformation, racialized disinformation, cyberstalking, deadnaming, and misgendering. **Facebook is ignoring the ways that harassment and hate can target multiple identities.**
- **Facebook's enforcement leaves much to be desired.** Posts are slow to be removed or may not be removed at all, and exceptions are often made for high-profile accounts.
- Facebook's disinformation policies **fail to address health-related disinformation**, such as disordered eating, abortion, and pregnancy.
- Facebook **fails victims of harassment by putting the burden on them to report each individual post and user**, while other platforms let any user report content that violates their policies.
- Facebook often **exposes users to extremist groups or hate content without providing resources to counter disinformation, hate, and extremism.**
- **Moderation decisions must be able to be appealed**, to both protect victims of harassment, as well as those whose content was removed because of discrimination.

## INSTAGRAM: F

- **Instagram and Facebook largely have the same policies.** This is a problem, because the two platforms have different mediums, play different roles, and the user experience and brands are very different to consumers.
- On top of Facebook's policy flaws, **Instagram has additional flaws related to misogynist algorithms, body shaming, sexist filters, and more.**
- **Instagram needs its own policies that address the specific issues created by a video- and image-centric platform**, such as harms to body image and mental health, and discrimination from users, moderators, and the algorithm because of race, ability, gender expression, and body size.



## TWITTER: C-

- **Twitter's hate speech policy is fairly robust, but has room for improvement**, such as addressing cyberstalking and revenge porn. There needs to be better enforcement of existing policies, particularly with regards to tweets misgendering people, a rampant issue on the platform.
- **Twitter's system of escalating punishment, including deplatforming, and removal of violating content is fairly robust**, though with room for improvement around more consistent enforcement.
- **Twitter recently added a ban on COVID-19 disinformation, but does not ban other health-related disinformation**, including other types of vaccines, abortion, pregnancy, eating disorders, and extreme dieting.
- According to the policies mentioned on its website, Twitter's moderation is based on the content published and not on political alignment. Yet Twitter's own research shows that its **algorithm is biased in favor of conservative content**.
- Twitter is **taking some positive steps to stop the spread of disinformation**, including building relationships with news organizations, the Associated Press and Reuters, and adding a new site prompt that reminds people to read an article before retweeting it solely based on the article's headline.

## YOUTUBE: D

- YouTube **fails to name women and nonbinary people in its policy on discrimination**. It's not enough to only mention gender, because the current policy makes it possible for men to claim "reverse discrimination."
- YouTube **fails to ban content** that promotes or participates in misogyny, misogynoir, transmisogyny, gendered disinformation, racialized disinformation, cyberstalking, deadnaming, misgendering, revenge porn, and sexual harassment.
- YouTube has **allowed extremist content**, such as content which calls for violence against an individual or a group, to thrive, due to insufficient policies and enforcement banning this content.
- YouTube has taken steps to remove COVID-19 disinformation, but is **failing to explicitly ban and remove disinformation and misinformation regarding abortions, health, dieting, disordered eating, and body dysmorphia**.
- YouTube currently allows for strikes or terminations to be appealed if the content creator believes it was in error. **This policy must be applied equitably** to creators of color, women creators, LGBTQ creators, immigrant creators, non-English-speaking creators, and religious minority creators.
- **YouTube's three-strike system resets and isn't cumulative**, making it easy for users to remain on the platform and continue to spread hate, harassment, and disinformation.

## TIKTOK: **D+**

- TikTok **fails to explicitly name**, ban, and commit to removing revenge porn, deadnaming, and misgendering.
- **TikTok must institute a weighted system for suspending and/or banning accounts.** Content which violates community standards must be removed, and steps must be taken to assess suspension or banning in an equitable way. Anecdotal evidence shows that there is preferential treatment for white, able-bodied, cisgender, straight, and English-speaking content creators over BIPOC, disabled, LGBTQ, non-English-speaking content creators.
- Health disinformation is rampant on TikTok, and **TikTok must ban accounts that spread disinformation about COVID-19, vaccines, abortions, dieting, body dysmorphia, and disordered eating.**
- TikTok's algorithm has been found to **favor conservative content**, and to send users down an extremist rabbit hole.
- **TikTok has set the industry standard on transparency reports**, but, despite this, TikTok has failed to remove much of the content related to bullying and harassment.

## REDDIT: **C**

- Reddit has fairly robust hate speech and harassment policies, but there's room for improvement, in part, because it is left up to individual subreddits to moderate and enforce, making it easier for hate-based groups to thrive under their own set of rules. **Reddit should also explicitly address misogyny, misogynoir, transmisogyny, deadnaming, misgendering, gendered disinformation, racialized disinformation, and cyberstalking.**
- Reddit **lacks policies banning disinformation**, especially disinformation related to health, such as COVID-19, vaccines, pregnancy, abortion, disordered eating, and extreme dieting.
- **One way in which Reddit is currently leading is by providing some support for victims of harassment and hate through its partnership with the Crisis Text Line**, which offers free counseling support in the United States. However, Reddit's support must be expanded to help victims protect themselves from their harassers.