



March 2025

PREVENTION BY DESIGN

A Roadmap for Tackling TFGBV
at the Source

Lena Slachmuisder & Sofia Bonilla

*A joint project of the Search for Common Ground, Integrity Institute, and Council on
Technology & Social Cohesion*

Table of Contents

01 Acknowledgements

02 Introduction

03 Methodology

04 Section 1: Behavior-Focused Interventions

04 1.1 Nudging Users to Reconsider Harmful Behavior

06 1.2 Filters to Empower Users in Managing Content Exposure

08 1.3 New User Onboarding and Awareness Campaigns

11 Section 2: Upstream Design Solutions

11 2.1 Enhanced Feedback Mechanisms for Reporting and Documentation

13 2.3 Implement Quarantine Systems for Gray-Area Content

18 2.4 Default Privacy Settings to Minimize User Vulnerability

19 2.5 Rate Limits on Interactions from New or Unverified Accounts

21 Conclusion

22 Additional Resources

Acknowledgements

This white paper is the result of a collaborative effort coordinated by the [Council on Tech and Social Cohesion](#), which seeks to influence the design of technology to build trust and collaboration, rather than polarization and violence. We believe that by harnessing the insights of technologists and peacebuilders, we can understand the harms of technology and offer practical recommendations for building technology which fosters social cohesion.

We are deeply grateful to our co-authors for their expertise, dedication, and invaluable contributions. Their insights and commitment to advocating for upstream approaches to tackling technology-facilitated gender-based violence (TFGBV) have been instrumental in shaping the analysis and recommendations presented in this work. We recognize the following individuals as co-authors of this white paper:

Sofia Bonilla

[Integrity Institute](#), USA

Leah Ferentinos

[Integrity Institute](#), USA

Gabe Freeman

[Integrity Institute](#), USA

Wardah Iftikhar

[Shoor Foundation for Education and Awareness](#), Pakistan

Ravi Iyer

[USC Neely Center](#) and [Psychology of Technology Institute](#),

USA

Matt Motyl

[USC Neely Center](#) and the [Psychology of Technology](#)

[Institute](#), USA

Himanshu Panday

[Dignity in Difference](#), Nepal

Dharini Priscilla

Independent gender and tech consultant, Sri Lanka

Durga Pulendran

[Search for Common Ground](#), Sri Lanka

Amrita Sengupta

[Center for Internet and Society](#), India

Nicholas Shen

[Integrity Institute](#), USA

Theodora Skeadas

[Humane Intelligence](#) and [Integrity Institute](#), USA

Lena Slachmuis

[Search for Common Ground](#), Belgium

Elodie Vialle

Independent tech policy consultant, France

Introduction: TFGBV's pervasive impact

The internet offers connection and opportunity, but for many women and marginalized groups, it is also a space where Technology-Facilitated Gender-Based Violence (TFGBV) is pervasive. A [2024 report by Snapchat](#) found that nearly one in four users across six countries—including Australia, India, and the U.S.—were victims of sextortion, a form of TFGBV. Globally, 38% of women report experiencing some form of online violence, [according to the Economist Intelligence Unit](#). Beyond inflicting individual harm, this trend creates a ["chilling effect" that silences women's voices](#), reducing diversity in public discourse and pushing women and girls out of spaces where they could exercise agency and leadership.

The [United Nations Population Fund \(UNFPA\) defines TFGBV](#) as “an act of violence perpetrated by one or more individuals that is committed, assisted, aggravated and amplified in part or fully by the use of information and communication technologies or digital media against a person on the basis of gender.” Common forms of TFGBV include doxxing, cyberstalking, hate speech, and more. The abuse of women did not begin with the rise of social media; yet, [the design choices of social media platforms](#) built to optimize engagement and attention have enabled, amplified, and accelerated this abuse. Platforms reward immediacy, emotional impact, and virality—qualities that cause harmful content to thrive. For example, emotionally charged content designed to provoke outrage, such as targeted harassment campaigns or hate speech, [receives disproportionately high engagement on social media platforms](#) due to their algorithmic ranking systems.

Platforms have historically been slow to adopt safety features, often acting only after significant public outcry. Twitter [introduced its report button in 2013, seven years after its launch](#), while Instagram didn't include a mute button [until 2018](#). This timeline reflects a broader industry issue: in the corporate race to “move fast and break things,” protecting vulnerable users has often been an afterthought for technology executives.

There is evidence that regulatory and human-centered design approaches can lead to safer online environments. Resources such as the Integrity Institute's [Focus on Features](#) lay the groundwork for

such claims by illustrating the extensive connections between platform design and digital harms and providing actionable guidance for reducing abuse by rethinking features at their root. The [United Kingdom's Age-Appropriate Design Code \(AADC\)](#) turned this concept into action. By targeting specific platform features and affordances—such as default privacy settings and restricted data collection—[this legislation has significantly improved the experiences of social media users](#) under the age of 18 in the UK. These initiatives illustrate the evidence for and power of addressing product design to reduce digital harm. Furthermore, as noted in "[No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users](#)", platforms' current approaches to mitigating abuse often fall short because they rely on retroactive measures, placing the burden on victims to protect themselves and report harm. This underscores the opportunity—and necessity—of acknowledging the clear link between digital harms and design features before damage is caused.

In contrast to existing solutions—such as content moderation—that address harm reactively, they place the burden of safety on victims, this paper advocates for a proactive, design-focused approach that embeds safety and user empowerment into social media platform design.

Methodology

This paper explores two complementary strategies for mitigating TFGBV:

- **Behavior-Focused Interventions:** These recommendations aim to influence user interactions by deterring harmful behaviors in real time and empowering individuals to manage their online safety.
- **Upstream Design Solutions:** These proposals address the structural features of platforms, including algorithmic systems and content moderation processes, to tackle harm at the source.

Each section provides specific recommendations, evidence of their effectiveness, and practical steps for implementation.

Section 1: Behavior-Focused Interventions

Behavior-focused interventions aim to address harmful interactions at the user level by:

- Dissuading potential abusers from engaging in harmful behavior.
- Empowering users to create safer personal environments.
- Reducing the reliance on reactive measures, such as reporting or self-censorship.

1.1 Nudging Users to Reconsider Harmful Behavior

Recommendation

Nudges are subtle, real-time prompts designed to encourage users to pause and reconsider potentially harmful language before posting. By fostering reflection, nudges can promote constructive interactions and alignment with community standards and reduce the incidence of TFGBV.

Implementation

- **Adaptive Language Detection:** Platforms can employ machine learning models such to detect harmful language in context and trigger tailored nudges. These models can provide a basis for language detection and nudging that can be customized by platforms based on user feedback and contextual needs.
- **Image detection:** Nudity media classifiers can detect when someone is about to send a NSFW photo containing nudity. While there are a range of classifiers depending on the use-case and product specific needs, many large platforms like Snapchat purportedly utilize models created in-house to detect potential NSFW content on the platform.
- **Contextual Prompts:** When a user attempts to post potentially abusive content, platforms can prompt them with a message encouraging them to reconsider. For instance, [Instagram has used nudges effectively to reduce offensive language](#) by reminding users of community guidelines or prompting empathy. Examples of nudges include:

- “This comment may not align with our community standards. Please review before posting.”
- “Are you sure this reflects the tone you want to convey?”
- **Broad Integration:** Applying nudges across posts, comments, replies, and direct messages can ensure comprehensive coverage.
- **Iterative Refinement:** Using A/B testing to evaluate the effectiveness of nudges can optimize them for diverse user demographics and cultural contexts.

Effectiveness

- **In 2022, Twitter displayed embedded prompts** for users to reconsider potentially harmful tweets before posting. [Matt Katsaros](#) studied the effect of this feature and found that 9% of users decided not to post, and 22% edited their tweet. More importantly, the nudge had a lasting impact, as recipients were less likely to post offensive content in the following weeks. “The nudge changes the behavior in the moment, but more importantly, it has a lasting impact. People are more likely to rethink their approach in future interactions,” [explained Katsaros](#) at a 2024 Symposium on Comment Section Research & Design.
- **Users who were nudged were less likely to receive offensive replies** themselves, he added: “By cutting off one offensive comment at the start, you reduce the likelihood of a toxic back-and-forth.”

Existing Examples

- **Snapchat:** The expanded in-app warning [feature](#) warns teens when they are receiving messages in chat from someone who has been blocked or reported by others, or is from a region where the teen’s network isn’t typically located – signs that the person may be a scammer or otherwise suspicious.
- **X and Instagram:** Both platforms have demonstrated the ability of nudges to reduce harmful interactions and foster healthier digital communities. [Instagram has reported](#) that, over the course of one week it sent approximately one million nudges to users, 50% of the time users deleted or amended their comment as a result. The reduction in hurtful comments posted is

When Instagram sent one million nudges to users over the course of one week, users deleted or amended their comments 50 % of the time



also long lasting, [according to Instagram's research](#) on what it calls “repeat hurtful commenters” — people who leave multiple offensive comments within a window of time. X reported that nudging users to reconsider replies containing harmful language “resulted in people [changing or deleting their replies over 30%](#) of the time when prompted for English users in the U.S. and around [47% of the time](#) for Portuguese users in Brazil.”

1.2 Filters to Empower Users in Managing Content Exposure

Recommendation

Filters give users control over their online experiences by enabling them to block specific keywords, topics, or other triggers. These tools reduce exposure to harmful content and help users create safer, more positive environments. Filters do not just hide harmful content but allow users to define the boundaries of their online interactions, creating a protective buffer against TFGBV and other potential harms. Because filters do not prevent a user from posting, but rather are used to tailor the visibility of content, they do not run afoul of free speech considerations. By creating safer environments, filters also encourage greater participation from users who might otherwise disengage due to harassment.

Implementation

- **Customizable Filters:** Allow users to define their own blocked keywords, phrases, or emojis, with regular reminders to adjust these settings.
- **Sensitive Content Quarantine:** Establish a separate dashboard where filtered content is stored, mirroring email spam folders.
- **Machine Learning Integration:** AI models like [Textgain](#)'s adapt over time to detect evolving abusive language patterns.
- **Awareness Campaigns:** Platforms should educate users about filter tools during onboarding and via periodic prompts to ensure widespread adoption.

Effectiveness

- **Instagram's Hidden Words Feature:** One year after the Hidden Words feature launched, users with large followings (10,000+ followers) saw [40% fewer comments](#) that might be

offensive after turning on the feature. It effectively mitigates harmful interactions by empowering users to define their personal boundaries.

- **Trollwall AI [points out](#)** that moderation through filtering aims to maintain healthy online communities, and leads to more engagement as users feel safer online. In 2024, TrollWall AI reportedly hid [over 3 million hateful comments across Facebook and Instagram](#), averaging more than 9,000 comments in a single day.
- **[Perspective API](#) and [Coral](#)** offer a toxicity filter for comments sections, aimed at fostering a healthier online discourse. [FACEIT](#), one of Europe's largest gaming platforms, experienced a [20% decline in toxic messages](#) since employing Perspective API's filtering.
- **Discord introduced [a new Ignore feature](#)** that allows users to ignore rather than block other users. The feature "allows a person to hide any new messages, DMs, server notifications, profiles and activity from selected users without alerting them...In practice, DMs received from an Ignored person will appear in the inbox with an icon and a grayed-out name, so they are available if the ignorer chooses to look at them."
- **[Matt Katsaros of the Justice Collaboratory at Yale Law School](#) [shared findings](#)** of the effects of such filters. His research on [Nextdoor](#) revealed that, while filtering offensive comments reduced visibility of such toxic comments by 12%, there was little change in how users behaved overall. "Hiding offensive comments reduces their visibility, but it doesn't seem to encourage more productive or positive engagement," Katsaros noted. This suggests that filtering alone may prevent users from seeing toxic content, but it may not change the culture of the platform.



FACEIT, one of Europe's largest gaming platforms, reported a 20% decline in toxic messages since employing Perspective API's filtering

Existing Examples

- **[Instagram's Hidden Words](#) and [Sensitive Content Filters](#)**: Allows users to define their own set of words to filter out, preventing content they deem unwanted from appearing in their DM requests. Once

this feature is activated, Instagram will also filter a list of potentially harmful words, emojis, numbers, or phrases [developed in conjunction](#) with anti-discrimination and anti-bullying organizations. Instagram's [Sensitive Content control](#) allows users to control how much sensitive content shows up on the Explore page.

- **Private companies such as [TrollWall AI](#), [Bodyguard](#) and [Textgain](#)** provide robust filtering mechanisms that report increased user satisfaction and engagement in safer online spaces. Textgain [points out](#) that machine learning models can also refine filters to take context into account.
- **Civil society groups** have also developed filtering tools such as the [Uli browser extension](#).
- **Google's Jigsaw** created an [open-source](#) tool, [Harassment Manager](#), to help people document and manage toxic language targeted at them on social media..
- **[TikTok offers a feature](#)** that allows users to reset their 'For You' feed, providing an opportunity to start afresh and tailor content recommendations to their current preferences. This action resets the feed to display a new set of popular videos, similar to the experience of a new user, and as users interact with these videos, the platform's algorithm begins to personalize the feed based on the new interactions.
- **Sensitive Content Alert:** TikTok flags potentially harmful content and [nudges users so they're aware of potentially harmful searches](#) that result in distressing content and provides users the option to view or avoid it based on individual preference.
- **[X's Safety Mode](#)** helps reduce unwanted interactions by temporarily blocking accounts that use potentially harmful language or repeatedly send uninvited replies and mentions. When activated, the system analyzes Tweets and the relationship between users to assess the risk of negative engagement. If an account is flagged for potentially harmful behavior, it's automatically blocked for seven days, preventing it from following, viewing Tweets, or sending Direct Messages to the user with Safety Mode. However, accounts the user follows or frequently engages with won't be autoblocked. Users can review and undo any blocks in their settings at any time.

1.3 New User Onboarding and Awareness Campaigns

Recommendation

The onboarding process should prioritize user education about safety tools, privacy settings, and

community guidelines. After onboarding, milestone-based reminders and general awareness campaigns ensure users remain informed of the settings and tools at their disposal as their platform activity evolves.

Implementation

- **Structured, Step-by-Step Introductions:** [Platforms can improve user retention by asking users about their goals early on](#) and tailoring onboarding to their specific needs, which fosters a sense of value and security. By introducing safety tools, privacy settings, and community guidelines progressively, onboarding allows users to adopt key features without feeling overwhelmed.
- **Opt-out rather than Opt-in for privacy settings:** Rather than expecting users to be required to opt-in to higher privacy settings, the platforms should [make greater privacy the default](#). When users choose settings with less privacy, they should be prompted to consider the trade-offs before confirming their selection. For example, platforms can follow this recommendation by requiring that location sharing on maps should be turned off by default.



The Uli browser extension redacts slurs and abusive content, and archives problematic content that appears in four languages: English, Tamil, Hindi, and Malayalam

- **Automated, Milestone-Based Awareness Prompts:** Automating reminders and safety prompts at user milestones (e.g., first post, gaining followers, joining a group) is an important enhancement to the onboarding experience. This approach continuously reinforces best practices in user safety and privacy by providing guidance relevant to the user's evolving engagement level. For example, if a user clicks "hide content" on a particular post, the platform could have an automatic prompt that says, "We noticed you've been hiding a lot of messages lately. Do you want to create a filter?" This would

remind the user of the resources at their disposal. Similarly, if a user gains a large following over a short period of time, the platform could automate a prompt that says, "We noticed you went viral for the first time! Do you want to enhance your account security?" Offering users guidance may even retain engagement, as they would be more equipped to stay safe as their account gets more attention.

- **Continuous, Personalized Engagement:** Platforms that practice continuous onboarding, such as progressive checklists, quizzes, and tailored prompts, help users retain safety information and stay engaged. [Reminding new users of norms](#) and the types of posts which comply with the community guidelines are shown to encourage new users to post, by giving them the confidence that their comments will adhere to platform policy.
- **Interactive Engagement:** Gamified checklists and quizzes for social media users can reinforce safety education, increasing adoption of protective measures. In a similar way that the social media or digital applications will prompt the users when an update is needed, similar approaches could be used to prompt users to be more aware of the choices they are or could make to strengthen their online safety. By being transparent about the tools available and how they can enhance the user's experience, platforms can build long-term trust, increasing the likelihood that users will remain engaged and continue using the service with confidence.

Effectiveness

- [Research](#) supports a **structured, multi-step onboarding process** that sets norms and guides users through critical features—especially those related to privacy, content moderation, and reporting tools. [Multiple studies](#) across [different platforms](#) show that displaying a message with community guidelines prior to posting results in new users more likely to post content that adheres to those guidelines and fewer comments reported for abuse.
- **Effective onboarding not only strengthens user safety** but decreases the risk of abuse which could otherwise decrease engagement or prompt the user to exit the platform altogether.

Existing Examples

- [Pinterest's Tailored Goal-Oriented Onboarding](#) helps align user experiences with platform norms, fostering a sense of value and safety.
- **Instagram's [Privacy Prompts for Minors](#):** Defaults new users under 16 or 18 (depending on the country) to private profiles and provides guidance on privacy settings. For existing young users with public profiles, Instagram will “show them a notification highlighting the benefits of a private account and explaining how to change their privacy settings.”
- [TikTok's teen privacy and safety settings](#) default settings for users aged 13-15 so that only people who the user approves can follow them and view their content and any direct messaging is not available.

Section 2: Upstream Design Solutions

Upstream design solutions address structural and algorithmic elements of platforms that amplify harm. By embedding safety into platform design, these solutions prevent abuse at its root by fundamentally altering how content is ranked, promoted, and moderated.

2.1 Enhanced Feedback Mechanisms for Reporting and Documentation

Recommendation

Platforms should build or integrate existing user-friendly reporting systems that provide feedback on abuse reports and enable users to document and store evidence easily. These measures improve trust and accountability while empowering victims to seek resolution.

Implementation

- **Streamlined Reporting Interface:** Provide a central dashboard where users can report abuse, track submission status, and view outcomes.
- **Automatic Evidence Capture:** Save flagged content in encrypted formats, including metadata and timestamps, for use in legal or institutional actions.
- **Collaborations with Advocacy Groups:** Collaborate with organizations such as the National Democratic Institute, Pirth.org, and Freedom House, which jointly developed the [Rapid Online Support and Assistance Mechanism \(ROSA\)](#) to aid those targeted by TFGVB.
- **Existing protocols like the Christchurch Call Crisis Response Protocol and the EU's Digital Services Act (DSA) [could be expanded](#)** for their scope to include TFGVB .
 - For the Christchurch Call, this means developing rapid response mechanisms to address coordinated harassment campaigns that disproportionately target women.

- For the DSA, it entails mandatory reporting on OGBV incidents, gender-disaggregated statistics, and compliance with gender-sensitive design standards. These expansions would position OGBV as a priority issue, ensuring platforms address it with the same urgency as extremist content.
- **Additionally, platforms should adopt interoperable documentation tools**, enabling users to collect and share evidence of abuse—such as screenshots and metadata—across multiple platforms. These measures reduce retraumatization, empower victims to take action, and encourage platforms to improve cross-platform accountability for abuse.
- **Platforms can collaborate with or integrate existing reporting mechanisms** developed by civil society to track instances of TFGVB across various platforms. This would lighten the burden on platform engineers and indicate a willingness to cooperate with civil society to tackle this important issue.



In 2023, the Revenge Porn Helpline reported a 90% removal rate, meaning it removed over 200,000 individual non-consensual intimate images from the internet



Pirth.org allows users to report online threats across any social media platform—their team reviews reports and escalates threats to platforms, easing the burden on victims

Effectiveness

- **Block Party and BodyGuard**: Demonstrate the utility of combining transparent reporting with automated documentation tools, reducing the burden on victims.
- **StopNCII.org, a project of the Revenge Porn Helpline (RPH), uses hashing technology** across platforms to stop the spread of unwanted intimate photos online across multiple

platforms. In 2023, the RPH revealed that [women have 28 times more intimate images shared than men](#). The RPH has an over [90% removal rate](#), meaning it has removed over 200,000 individual non-consensual intimate images from the internet.

Existing Examples

- **Block Party:** Allows bulk reporting and organizes abusive content for easier review.
- **Pirth.org:** Allows users to report online threats across any social media platform. Upon submission, it instantly generates a personalized action and resource list, including digital safety tools, helplines, legal aid, mental health support, and more. With user consent, the [Pirth.org](#) team reviews reports and escalates threats to platforms, easing the burden on victims.
- **Instagram's Reporting System:** Offers contextual guidance and [feedback](#) for flagged content.
- **Instagram report status:** Allows for users to see the status of the reported content or account that they submitted to the platform and also allows users to see the full history of their past reports.
- **Instagram disabling screenshots and screen recording** for ephemeral images or videos sent in private messages in efforts to combat sextortion and revenge porn that commonly happens when a user sends an intimate image and the receiving user screenshots or screen records it for ransom.

2.2 Move Away from Engagement-Based Content Ranking

Summary

Engagement-based ranking systems prioritize content that drives high interaction, such as likes, shares, and comments. While effective for increasing user retention, these algorithms frequently amplify harmful content, including sensationalist, divisive, and abusive material. Shifting to trust-based or quality-oriented ranking systems can mitigate these harms while maintaining user engagement.

Harm Amplification through Engagement-Based Ranking

- **Sensationalism and Divisiveness:** Content designed to provoke strong emotional reactions —especially outrage— is rewarded by engagement algorithms. Meta CEO Mark Zuckerberg,

describing the ‘natural engagement platform’ [wrote](#): “One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content.”

- [Studies about Facebook and Twitter](#) reveal that divisive content is shared twice as often as neutral content.
- **Hate Speech and Misinformation:** Research shows that hate speech and extremist ideologies thrive under engagement-driven models, as seen in incidents such as the [2016 U.S. election](#) and [health misinformation campaigns](#).
- **Gendered Harm:** Equimundo’s [Manosphere-Rewired](#) report highlights how misogynistic content gains visibility through these algorithms, exacerbating TFGBV by amplifying toxic narratives



“One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content”

- Mark Zuckerberg

Effectiveness of Non-Engagement Signals

[A review of evidence from 20 approaches to content ranking](#) that prioritize quality and safety over engagement include:


1. **Quality Metrics:** Platforms like YouTube have improved retention by [prioritizing watch time over clicks](#), encouraging the promotion of higher-quality videos.
2. **User Surveys:** Facebook’s [“worth your time” surveys](#) collect feedback on the value of content, aligning ranking with user satisfaction.
3. **“Bridging Content” & Diverse Metrics:** Algorithms that prioritize content appealing to a broad audience [reduce the visibility of polarizing material](#).

These approaches demonstrate that [non-engagement ranking models can reduce harm without compromising engagement](#): “In 2012 YouTube switched from maximizing clicks to maximizing a combination of clicks and watch-time. They reported a short-term drop in clicks but a long-term increase

retention,” write the report authors. “Similarly, Facebook uses quality proxies such as misinformation labels and click-gap scores to reduce misinformation in content ranking, balancing short-term engagement with long-term user satisfaction.”

Implementation

- **Trustworthiness and Quality Metrics:** Integrate factors such as user credibility, source reliability, and content quality into ranking algorithms.
- **Algorithm Transparency:** Share details of ranking criteria to foster user trust and accountability.
- **User-Curated Rankings:** Allow users to provide feedback on the relevance and quality of content, tailoring their feeds to their preferences.



In 2024, YouTube “trained [its systems] to elevate authoritative sources higher in search results, particularly in sensitive contexts,” which mitigates risks while optimizing high quality information in its ranking framework

Existing Examples

- **Pinterest:** Pinterest does not rely solely on engagement signals for ranking and recommendations.
- **LinkedIn:** Prioritizes content relevance and professional value, using quality metrics to rank posts.
- **YouTube & Google Search:** YouTube [promotes watch time over clicks](#), incentivizing creators to focus on informative, engaging content. In 2024, YouTube reported that its systems are “trained to [elevate authoritative sources](#) higher in search results, particularly in sensitive contexts” – mitigating risks while optimizing high quality information in its ranking framework. Google Search predicts quality using a wide variety of signals, including long established information retrieval signals (the most famous being PageRank, Google’s founding algorithm). As a result, users get results from trusted medical organizations and other authoritative sources when using Google Search, [especially around sensitive topics](#). Replacing engagement-based ranking with trust-based models offers a clear pathway to reducing the amplification of TFGBV while maintaining platform integrity.

2.3 Implement Quarantine Systems for Gray-Area Content

Recommendation

Quarantine systems temporarily isolate flagged content in a separate review area, allowing for thoughtful moderation before public exposure. This approach reduces immediate harm while balancing safety and free expression. Gray-area content—material that skirts the edges of platform policies—poses a significant challenge. Over-censorship risks alienating users and undermining free expression, while under-moderation exposes users to harm. Quarantine systems mitigate this tension by creating a buffer for review.

Implementation

- **Dedicated Quarantine Dashboard:** Develop a centralized interface where users and moderators can review flagged posts. Content should remain obscured by default, with the option for users to reveal or act on it.
- **AI-Assisted Moderation:** Employ algorithms to flag gray-area content, prioritizing nuanced analysis over blanket removals. Natural Language Processing (NLP) models are commonly used for detection and moderation of hate-speech, harassment, threats, identity-based attacks, and all varieties of toxicity. NSFW image and video classifiers are often used in the same strategy as NLPs for abusive media related material like revenge porn, child porn, and sextortion. The specific type of media classifiers used are typically based on the cloud service the platform utilizes, but a couple of the most popular are Amazon's [Rekognition](#) and Google's [Cloud Vision](#). Tools like [PhotoDNA](#) by Microsoft are also used to support content moderation for known images of child exploitation and sexual exploitation through a type of image hashing technology.
- **User Collaboration:** Involve users in the review process by enabling them to confirm, dispute, or clarify moderation decisions.

Effectiveness

- **Instagram's Restrict Feature:** Empowers users to manage and curate their online experience without blocking or otherwise engaging with the abuser.
- **Sensitivity Screens:** Flags potentially harmful content and allows users to opt in or out of viewing it. Anecdotal evidence suggests that, once a sensitive content screen is applied to a video, engagement drops drastically.

- **Outlook’s Spam Quarantine:** Demonstrates how quarantining can balance protection with transparency by allowing users to review flagged content.
- **Shadowbanning:** Partially or completely hiding a user’s content from being visible to other users. This is commonly done by social media platforms as a punishment or a temporary hold for additional review when it is believed that a user has violated particular safety guidelines. While not many companies publicly state that they do this, it is a fairly common moderation and engineering tactic.



Image detection services—like Amazon’s Rekognition, Google’s Cloud Vision, and Microsoft’s PhotoDNA—support content moderation for violent imagery and known images of child exploitation and sexual exploitation

Existing Examples

- **Instagram’s [Restrict Feature](#):** The “restrict” feature prevents flagged comments from appearing publicly until approved by the user, reducing exposure to harassment. Direct messages from restricted accounts move to Message Requests and do not trigger notifications. Users can view them, but the sender won’t see read receipts or activity. This empowers users to manage exposure to other specific users without alerting the restricted party.
- **[Instagram](#) and [TikTok](#) Sensitive Content Flags:** Offers users greater control over what they see while maintaining freedom of expression.
- **[TikTok Shadowbanning](#):** According to the TikTok Safety Center, accounts that repeatedly share content deemed unsuitable for the For You feed may have their posts removed from the feed and become less visible in search results.

Quarantine systems are a practical solution for moderating ambiguous content without over-censoring or ignoring potential harm.

2.4 Default Privacy Settings to Minimize User Vulnerability

Summary

Platforms should implement privacy-first defaults, such as limited profile visibility and restricted discoverability, to reduce user vulnerability. These settings protect users from unsolicited interactions and targeted harassment, particularly for women and marginalized groups. These should be accompanied by clear guidance to users about how to change these defaults, and understanding the consequences of less privacy.

Harm Addressed by Default Privacy Settings

Public-facing profiles expose users to increased risks, including stalking, doxxing, and other forms of TFGVBV. Privacy defaults provide a safer starting point for all users while offering flexibility for those who wish to adjust their settings.

Effectiveness

- **India's Locked Profiles Initiative:** When Facebook activated locked profiles for women in India by restricting access to photos and posts, it resulted in [significantly reduced harassment](#). One account illustrates the feature's impact: Next to an image of an Indian woman covering her face with a saree, a common tradition, the woman shared that she received 367 friend requests and comments like "very beautiful" and "where do you live." After enabling Facebook's "locked profile" feature, introduced in India in 2020, the unsolicited messages stopped. "By June 2021, the feature had been adopted by [34% of women users in India](#), said the internal report."
- **Instagram's Privacy Defaults for Minors:** Automatically sets accounts for new users [under 18 to private](#), ensuring safer interactions.



Facebook's Locked Profiles Initiative, which restricts access to photos and posts, significantly reduced online harassment. In fact, it was adopted by 34% of women users in India

Implementation

- **Privacy-First Defaults**
 - New accounts should default to high-privacy settings, such as hidden contact details, limited searchability, and restricted message requests.
 - Privacy settings, like location sharing, should require explicit opt-in consent.
- **Real-Time Visibility Snapshots:** Provide users with tools to see how their information appears to others, similar to Facebook's "View As" feature.

Existing Examples

- **Facebook's Locked Profiles in India:** Successfully protected users in regions with heightened risks of harassment.
- **Instagram's Privacy Defaults for Minors:** Demonstrates the efficacy of defaulting to safer settings for vulnerable users.
- **TikTok's Video Interaction Controls:** With Duet and Stitch set to "off" by default, users can choose to enable these features each time they post a video.

By establishing privacy-first defaults, platforms can protect users from TFGBV while maintaining flexibility for those who prefer greater visibility.

2.5 Rate Limits on Interactions from New or Unverified Accounts

Recommendation

Platforms should impose rate limits on actions such as friend requests, comments, and messages for new or unverified accounts. These limits serve as preventive friction and reduce misuse and abuse by bad actors and bots while creating safer online environments.

Harm Addressed by Rate Limits

Newly created or unverified accounts are often used for spam, harassment, and coordinated abuse campaigns. Without restrictions, these accounts can overwhelm victims with harmful interactions.

Effectiveness

- **Twitter’s Rate Limits:** Control the frequency of tweets, follows, and messages for new users, reducing spam and abuse.
- **USC Neely Center’s Design Code:** Reinforces the importance of rate limits in preventing the exploitation of platform features for harm. “Small groups of motivated actors have been instrumental in promoting content with [societal risk](#), [targeting others](#), and misusing other people’s [images](#) and [information](#). Most people have no need to reach the rate limits that platforms have for such functionality, and so limiting high usage to those who are trusted and have a demonstrated need creates a safer ecosystem for both individuals and society.”
- **Implementation Structured Rate Limits:** Restrict actions for new accounts until they demonstrate authentic behavior through verified activity or longevity.
- **Threshold Escalation:** Flag accounts that exceed activity thresholds for additional review.
- **Gradual Unlocking:** Increase interaction limits as accounts gain credibility.

Existing Examples

- **Reddit’s Post Restriction Tools:** Communities on Reddit can limit posting frequency for [new users](#) to reduce abuse and spam. Similarly, under Reddit’s Karma System, certain subreddits can choose to automatically remove new posts from users who haven’t met specific [engagement criteria](#), even if the content isn’t spam.
- **Twitter’s [Rate Control Systems](#):** Prevents misuse of features like direct messaging and follows.
- **Instagram’s [Comment and DM Limits](#):** This feature enables users to restrict comments and DM requests during periods of heightened attention. It helps protect individuals from potential abuse by automatically hiding comments and messages from users who don’t follow them or have only recently started following them.

Rate limits are a simple yet effective mechanism for curbing misuse while maintaining accessibility for legitimate users.



According to the USC Neely Center, most people have no need for rate limits, so limiting high platform usage to those who are trusted and have a demonstrated need creates a safer online environment

Conclusion

Preventing tech-facilitated gender-based violence (TFGBV) requires a multifaceted approach that goes beyond reactive measures. By integrating behavior-focused interventions with upstream design solutions, platforms can not only mitigate harm but also address the underlying structures that enable abuse. Thoughtful design choices—such as default privacy settings, revised ranking models, and proactive nudges—play a crucial role in empowering users to manage their safety and ultimately reducing the prevalence of TFGBV.

This proactive, design-driven approach prioritizes inclusivity and safety to ensure that digital spaces remain accessible and empowering for all users. By embedding safety into platform architecture from the outset, companies can shift from a model of damage control to one of prevention, which reduces the burden on individuals to protect themselves from harm.

Additionally, stronger and more meaningful transparency initiatives from platforms will drive continuous improvement in platform design and allow platforms to proactively adapt to emerging threats and evolving patterns of online abuse.

As highlighted by Columbia University's [Institute for Global Politics](#), TFGBV is not just a personal issue—it is a systemic problem that undermines free expression, economic participation, and democratic engagement. Tackling this issue requires platforms to prioritize safety and accountability from the earliest stages of development. Through a combination of user empowerment and responsible design, platforms can create a more equitable and secure digital future.

Additional Resources



1. Center for Internet and Society. (2023). [Technology and Women's Political Participation in India: A Position Paper.](#)
2. Commonplace. (2023). [Trust Through Trickery.](#)
3. Cunningham, Tom, et. al. (2023). [What We Know About Using Non-Engagement Signals in Content Ranking.](#)
4. Equimundo. (2024, June). [The Manosphere, Rewired: Understanding Masculinities Online.](#)
5. eSafety Commissioner. (2023). [Safety by Design: Technology-Facilitated Gender-Based Violence Industry Guide.](#)
6. Integrity Institute. (2023). [Focus on Features: Prevent Harm Through Design.](#)
7. Institute for Strategic Dialogue. (2023). [Misogynistic Pathways Towards Radicalization: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence.](#)
8. Jaurisch, J., & Bahro, J., with Allen, A., Pershan, C., & Szymielewicz, K. (2023, December; last updated 2024, September). [DSA Risk Mitigation: Current Practices, Ideas, and Open Questions. Interface.](#)
9. National Democratic Institute (NDI). (2023). [Conference on Preventing and Disrupting the Spread of Gendered Disinformation in the Context of Electoral Processes and Democratic Rollback.](#)
10. National Democratic Institute (NDI). (2022). [Interventions for Ending Online Violence Against Women in Politics.](#)
11. National Democratic Institute (NDI). [Landscape Tracker: TFGBV Interventions.](#)
12. Neely Center for Ethical Leadership and Decision Making. (2023). [Neely Center Design Code for Social Media.](#)
13. PEN America. (2021, March 1). [No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users.](#)
14. PEN America. (2023, June 29). [Shouting into the Void: Why Reporting Abuse to Social Media Platforms Is So Hard and How to Fix It.](#)
15. UNESCO. (2023). ["Your Opinion Doesn't Matter, Anyway": Exposing Technology-Facilitated Gender-Based Violence in an Era of Generative AI.](#)
16. University of Cambridge. (2023). [Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media.](#)
17. University College London and University of Kent. (2024). [Safer Scrolling: How Algorithms Popularise and Gamify Online Hate and Misogyny for Young People.](#)