GLAAD Social Media Safety Index Platform Scorecard — Research Guidance

The 2024 SMSI **Platform Scorecard** consists of 12 indicators that draw on best practices and guidelines from the Ranking Digital Rights (RDR) Big Tech Scorecard, the annual evaluation of the world's most powerful digital platforms on their policies and practices affecting people's rights to freedom of expression and privacy.

In 2022, GLAAD released the inaugural Platform Scorecard evaluating five major social media platforms. The methodology behind the Platform Scorecard was developed in collaboration with Goodwin Simon Strategic Research (GSSR) and Ranking Digital Rights (RDR). After developing a first set of 12 draft indicators, the research team revised and refined the indicators based on feedback from RDR, interviews with five expert stakeholders working at the intersections of technology and human rights, and input from the SMSI advisory committee. Additional methodological considerations were identified during the subsequent policy analysis and company research.

During the 2024 research cycle, we added an evaluation of Threads to the Platform Scorecard. The Scorecard now looks at six major social media platforms: Facebook, Instagram, and Threads (whose parent company is Meta), X/Twitter, YouTube (parent company - Alphabet/Google), and TikTok (parent company - ByteDance).

Guidance for future researchers interested in applying these indicators can be found below.

Indicator 1

The company should disclose a policy commitment to protect LGBTQ users from harm, discrimination, harassment, and hate on the platform.

LGBTQ people and other vulnerable communities are frequent targets of online abuse, hate, discrimination, and harassment. Companies should have a policy in place that protects LGBTQ users from abuse, hate, discrimination, and harassment on the platform (Element 1). This policy should include both sexual orientation (Element 2) and gender identity (Element 3) as protected categories. In order to give users a clear understanding of the types of content and behaviors that are prohibited on the platform, the company should disclose a detailed list of prohibited content and behaviors and provide illustrative examples (Element 4). Companies should also acknowledge the LGBTQ community's history of reappropriating derogatory terms and disclose an explicit acknowledgement and exception of self-expressive usage of otherwise derogatory LGBTQ-related terms by LGBTQ users (Element 5).

One example for a “protected groups” policy can be found in YouTube’s policy disclosures. In its Hate Speech policy, the company says "Hate speech is not allowed on YouTube. We don’t allow content that promotes violence or hatred against individuals or groups based on any of the following attributes, which indicate a protected group status under YouTube’s policy.” The list of attributes includes both “Gender Identity and Expression” and “Sexual Orientation.” In addition, the company has a Harassment & Cyberbullying policy that provides the following: “We don’t allow content that targets someone with prolonged insults or slurs based on their physical traits or protected group status…” This policy also provides that “we take a harder line on content that maliciously insults someone based on their protected group status, regardless of whether or not they are a high-profile individual.”[1]

An example of an explicit acknowledgement and exception of self-expressive usage of otherwise derogatory terms can be found on a page explaining Meta’s Hate Speech policy: “In other cases, speech, including slurs, that might otherwise violate our standards is used self-referentially or in an empowering way…Our policies are designed to allow room for these types of speech but require people to clearly indicate their intent. Where intention is unclear, we may remove content.”

Potential sources:

- Community guidelines
- Hate speech policy

Indicator 2

The company should disclose an option for users to add gender pronouns to user profiles.

On some social media platforms, it has become common practice for users to add their pronouns to their user handles and bios. However, policies such as real name requirements and character limits prevent users from fully expressing their identity by use of pronouns on some platforms.

---

[1] *YouTube’s extension of this protection to “high-profile” individuals is a best practice that other platforms should also implement (e.g., Meta’s policies do not apply to public figures, thereby permitting bullying and harassment and enabling expression of general anti-trans animus via such attacks). While the Platform Scorecard currently does not assess whether platforms extend policy protections against harm, discrimination, harassment, and hate to LGBTQ “high-profile” individuals, GLAAD’s research team will add a corresponding element in future iterations of the SMSI Platform Scorecard methodology.*

Therefore, companies should have a dedicated feature that allows users to add their pronouns to their profiles (Element 1). In order to strike a balance between user expression and privacy and safety, companies should also give users control over the audiences that can see their pronouns (Element 2). For example, Instagram discloses a feature allowing users to add up to four pronouns to their profiles. However, the company falls short of full credit as the disclosure indicates that the feature may not be available for all users. In addition, Instagram discloses only limited options for users to customize who can see their pronouns. While users have the option to show their pronouns to followers only, the company does not disclose more granular options for users to select who can see their pronouns.

Potential sources:

- Terms of service
- Company help pages

Indicator 3

The company should disclose a policy that prohibits targeted deadnaming and misgendering of other users.

Transgender, nonbinary, and gender non-conforming users are among the most vulnerable when it comes to online abuse and harassment.[2] Therefore, companies should have a policy in place that contains a clear prohibition against targeted misgendering (Element 1) and deadnaming (Element 2). Companies should also clearly explain the processes and technologies that they use to identify content and accounts violating this policy (Element 3) and give users clear menu options to report instances of targeted misgendering and deadnaming (Elements 4-5). Prohibiting targeted misgendering and deadnaming is not enough. In order to make this policy effective, companies also should disclose their processes for enforcing this policy once violations to the policy are detected, including providing details of how it decides what may represent violating content, and the actions it may take in response to content and accounts violating this policy (Element 6). Companies should not require self-reporting of potential violations, but should employ technologies, human review, and/or reporting from others to detect violations to the policy.[3]

---

[2] Online Hate and Harassment: The American Experience 2023 | ADL

[3] *While the Platform Scorecard assesses company disclosures regarding the processes and technologies platforms employ to detect violations to targeted misgendering and deadnaming policies, it currently does not address the distinct issue of whether all users can report instances of targeted misgendering and deadnaming, or if violations can only be reported by the targeted individual. (For more information on how self-reporting requirements complicate the enforcement of targeted misgendering and deadnaming policies, please see GLAAD's post "All*

Earlier this year, X/Twitter quietly revived its policy prohibiting targeted misgendering and deadnaming on the platform. This change in policy makes X/Twitter the only platform besides TikTok that prohibits both targeted misgendering and deadnaming. However, the company falls short of fully protecting transgender, nonbinary, and gender non-conforming users from targeted misgendering and deadnaming as it discloses it needs to hear from targeted individuals in order to determine whether a policy violation has occurred, effectively requiring users to self-report violations to the policy. In addition, the company does not disclose whether it also employs human review and/or automated content moderation to identify violations to the policy.

Element language for Elements 3 and 6 directly draw on element language on terms of service enforcement developed by RDR.

Potential sources:

- Community guidelines
- Hate speech policy

Indicator 4

The company should clearly disclose what options users have to control the company's collection, inference, and use of information related to their sexual orientation and gender identity.

Companies collect vast amounts of data that allow them to make inferences about users' sexual orientation and gender identity. Ranking Digital Rights and other civil society groups have called for greater transparency and user control around data collection and processing of this information. Companies should also give users control over the collection and inference of information related to their sexual orientation (Elements 1 and 2) and gender identity (Elements 3 and 4). Users should also have the ability to delete information related to their sexual orientation (Element 5) and gender identity (Element 6), and have control over how this information is used for the development of algorithmic systems (Element 7).

The platforms evaluated in the SMSI continue to provide insufficient transparency about LGBTQ users' control over their own information. However, recent policy changes by TikTok make the platform comparably more transparent than its peers. In April 2024, the company launched a portal that contains policy disclosures and resources relevant to

Social Media Platform Policies Should Recognize Targeted Misgendering and Deadnaming as Hate Speech.") GLAAD's research team will add a corresponding element in future iterations of the SMSI Platform Scorecard methodology.

LGBTQ users. In the section “Respecting Your Privacy,” the company provides that it does not collect users’ sexual orientation information. Further, TikTok discloses that users who share information related to their sexual orientation and/or gender identity can delete this information.

Potential sources:

- Privacy policy

Indicator 5

The company should disclose that it does not recommend content to users based on their disclosed or inferred sexual orientation or gender identity, unless a user has opted in.

LGBTQ users should have full control over the information they see on their social media feeds, and recommendation of content based on their disclosed or inferred sexual orientation and gender identity should be off by default (Element 1). Companies should also explain how users can opt in (Element 2) and opt out (Elements 3 and 4) of seeing content based on their disclosed or inferred sexual orientation and gender identity.

Companies continue to disclose very little regarding the options users have to control the content they may see on their feeds based on their disclosed or inferred sexual orientation or gender identity. None of the companies evaluated in the 2024 SMSI index disclosed that recommendation of user-generated content based on their disclosed or inferred sexual orientation and gender identity is off by default. Companies also provided only limited information regarding the options that users have to control the content they see on their feeds. For example, X/Twitter’s page explaining its Recommendation Algorithm discloses limited information about the options that users have to control the content they see. However, the company falls short of full credit as it is not clear that users can opt out of all content related to their disclosed or inferred sexual orientation and gender identity.

Potential sources:

- Privacy policy

Indicator 6

The company should disclose that it does not allow third party advertisers to target users with, or exclude them from seeing content or advertising based on

their disclosed or inferred sexual orientation or gender identity, unless the user has opted in.

Ranking Digital Rights and other civil society organizations have long called attention to the harms caused by the targeted advertising-driven business models of social media companies that rely on the collection of vast amounts of user data. Targeted advertising based on sensitive categories raises additional concerns for user privacy and safety, and there is an acute need for users to have full control over how their data is used for targeted advertising.

Companies should not target LGBTQ users with targeted advertising unless they have opted in (Element 1). In order to ensure LGBTQ users are not *excluded* from economic, financial, and other opportunities, companies should also make a commitment not to exclude LGBTQ users from advertising (Element 2). LGBTQ users should also have control over how their user information is used for targeted advertising (Element 3-6). In order to give insight into how companies detect content and accounts violating these rules, they should also disclose the processes and technologies used to identify advertisers who are in violation of these policies (Element 7).

Companies that have a clear disclosure that prohibits advertisers from targeting users with advertising based on their sexual orientation and gender identity receive full credit on Element 1. For these companies, Elements 3-6 are not applicable.

Meta's Business Help Center page "About Meta's advertising policy on discriminatory practices" contains a clear prohibition against both wrongful targeting and exclusion of LGBTQ users from ad services: "Our Advertising Standards don't allow advertisers to run ads that discriminate against individuals or groups of people based on personal attributes such as race, ethnicity, color, national origin, religion, age, sex, sexual orientation, gender identity, family status, disability or medical or genetic condition. This means that advertisers may not (1) use our audience selection tools to (a) wrongfully target specific groups of people for advertising, or (b) wrongfully exclude specific groups of people from seeing their ads; or (2) include discriminatory content in their ads."

Element language for Element 7 directly draws on element language on targeted advertising developed by RDR.

Potential sources:

- Advertising policies

Indicator 7

The company should disclose that it prohibits advertising content that could be harmful and/or discriminatory to LGBTQ individuals.

Companies should also disclose a policy that prohibits advertising content that could be harmful and/or discriminatory to LGBTQ individuals (Element 1). This content includes, but is not limited to, misinformation around gender affirming care, misinformation around PrEP, and content advertising so-called “conversion therapies.” The company should also disclose the processes and technologies it uses to identify advertising content or accounts that publish advertising content harmful or discriminatory to LGBTQ users (Element 2).

For example, Alphabet’s advertising policies contain a page titled “Inappropriate content" that prohibits: "Content that incites hatred against, promotes discrimination of, or disparages an individual or group on the basis of their race or ethnic origin, religion, disability, age, nationality, veteran status, sexual orientation, gender, gender identity, or any other characteristic that is associated with systemic discrimination or marginalization." And Meta’s Unrealistic Outcomes advertising standards policy expressly prohibits the promotion of: “Conversion therapy products or services.”

Potential sources:

- Advertising policies

Indicator 8

The company should regularly publish data about the actions it has taken to restrict content and accounts that violate policies protecting LGBTQ individuals.

In order to provide insight into how company policies are enforced, the company’s transparency report should disclose the number of pieces of content restricted for violating the company’s policies protecting LGBTQ users (Element 1). This includes content removals, but also other types of enforcement actions the company may take (e.g., hiding content, labeling content with a warning to the user). This data should be broken out by different types of policy violations—for example, hate speech against LGBTQ users, and targeted misgendering and deadnaming (Element 2). The company should also disclose the number of accounts restricted for violations of policies protecting LGBTQ users (Element 3) and break out this LGBTQ-specific data by different types of policy violations (Element 4).

Wrongful removal of content and accounts can have significant implications for freedom of expression and human rights. Therefore, companies should be committed to reinstate wrongfully removed content and accounts in a timely manner. Hence, companies should

also disclose the number of pieces of content (Element 5) and accounts (Element 6) reinstated after they were wrongfully removed. Drawing on RDR best practices, this data should be disclosed four times a year (Element 7).

The platforms evaluated in the 2024 SMSI Platform Scorecard continue to fall short of providing comprehensive data on content and account restrictions for violations to policies protecting LGBTQ users. For example, TikTok's "Community Guidelines Enforcement" report discloses data on content and account removals for violations to its Community Guidelines, which prohibit different forms of hate, discrimination, and harassment against LGBTQ users. However, the report does not break out this data for different types of policies protecting LGBTQ users—for example, hate speech against LGBTQ users, and targeted deadnaming and misgendering.

Element language for Elements 7 directly draws on element language on transparency reporting developed by RDR.

Potential sources:

- Transparency report

Indicator 9

The company should take proactive steps to stop demonetizing and/or wrongfully removing legitimate content related to LGBTQ issues in ad services.

LGBTQ creators and other underrepresented groups are frequent targets of wrongful demonetization and removal from ad services on social media platforms, depriving them not only of tools for expression, but also creating economic and financial inequities. Companies should disclose the concrete steps they take to address wrongful removal and demonetization of LGBTQ creators (Element 1) and disclose that they initiate or participate in meetings with stakeholders that represent on behalf of or are content creators who have been demonetized and/or had their legitimate content related to LGBTQ issues removed from ad services (Element 2). In order to provide insight into content and account removals impacting LGBTQ creators, the company should publish data on the number of pieces of content and accounts related to LGBTQ issues removed, filtered, demoted, or demonetized in ad services for violating the company's policies (Elements 3 and 4). Transparency is also needed in regards to the number of pieces of legitimate content and accounts related to LGBTQ issues that were reinstated after they were wrongfully removed, filtered, demoted, or demonetized in ad services for violation to the company's policies (Elements 5 and 6). The company should publish this data at least once a year (Element 7).

Despite advocates and LGBTQ creators raising concern over the removal and demonetization of LGBTQ-related content from ad services on YouTube, Alphabet continues to provide limited transparency on the state of demonetization and removal of LGBTQ creators and their content. The company discloses piecemeal solutions rather than a comprehensive plan outlining concrete steps to address demonetization, filtering, and removal of LGBTQ creators. The company's transparency reports provide no data giving insights into removal and demonetization of LGBTQ creators and LGBTQ-related content from ad services.

Potential sources:

- Company blog
- Transparency report

Indicator 10

The company should disclose a training for content moderators, including those employed by contractors, that trains them on the needs of vulnerable users, including LGBTQ users.

In order to ensure that content moderators are aware of the unique challenges that LGBTQ communities and other vulnerable users face online, companies should disclose required training for moderators that trains them on the needs of vulnerable users in protected categories (Element 1), including LGBTQ users (Element 2).

In the 2024 SMSI, Meta received partial credit based on disclosure in its "Gender Identity Policy and User Tools" policy. According to the policy, Meta's human reviewers around the world "have undertaken specific training on gender identity policy enforcement in 2022. We give reviewers more explicit and detailed internal guidance about when to consider a trans, non-binary or genderfluid person to be attacked on the basis of gender identity. This helps us better enforce our policy at scale for the 2.8 billion people who use our technologies, across every country and language where we operate. It involves providing guidance on the language used by the LGBTQ+ community to identify indicators for gender identity for trans, genderfluid, non-binary and gender nonconforming people (such as the Trans Pride flag)." However, it is not clear from the company's disclosure whether Meta has conducted similar training since 2022.

Potential sources:

- Company blog
- Annual report

Indicator 11

The company should have internal structures in place to implement its commitments to protect LGBTQ users from harm, discrimination, harassment, and hate within the company.

The company should disclose that it has an LGBTQ policy lead who advises policy and product teams on how companies' policies, products, and services may impact the online rights, safety, and privacy of LGBTQ users (Element 1). The potential risks that LGBTQ users may face online are constantly evolving. Therefore, the company should also disclose that it engages with organizations representing the needs of LGBTQ and other vulnerable users to ensure they are up to date on any challenges that LGBTQ users may face (Element 2). The company should also disclose that it has a formal training in place that trains employees at different levels of the company about the needs of LGBTQ users (Element 3).

Notably, TikTok was the only platform evaluated in the 2024 Platform Scorecard disclosing that it has an LGBTQ policy lead. In this context, the page "Combating hate and violent extremism" provides the following: "...we have a dedicated team who champions fairness considerations across our products and policies to help ensure representation and inclusion across different communities. This cross-disciplinary team is staffed with policy and program leads focused on specific communities (such as LGBTQ+, BIPOC, Persons with Disabilities, and more) as well those working holistically on embedding human rights frameworks."

Potential sources:

- Company annual report
- Company blog

Indicator 12

The company should make a public commitment to continuously diversifying its workforce, and ensure accountability by periodically publishing voluntarily self-disclosed data on the number of LGBTQ employees across all levels of the company.

In order to ensure a company's commitment to diversity is implemented internally, companies need to build diverse teams across different levels of the company, including policy teams and engineering and product teams.

The company should make a public commitment to taking proactive steps to diversify its workforce (Element 1). The company should also disclose an internal reporting mechanism that allows employees to voluntarily self-disclose their sexual orientation and gender identity (Element 2). This voluntarily disclosed data should be published in the company's workforce numbers (Element 3) and should be broken out by different teams (Element 4). The company can only receive full credit on this indicator if it publishes this data at least once a year (Element 5).

X/Twitter is the only company evaluated in the 2024 Platform Scorecard that fails to make a renewed commitment to diversifying its workforce. TikTok makes a commitment to a diverse workforce in its Code of Conduct: "Born to be global, we span our operation across geographies and house a workforce with diverse backgrounds. We champion diversity and inclusion, as we understand only by sticking to this principle, are we able to attract and maintain a robust workforce, which is essential for achieving our mission of 'Inspire Creativity, Enrich Life.' Moreover, it is also our commitment to the whole society that we cherish and respect uniqueness, and we encourage people to be their true and creative selves." However, the company provides comparably little transparency as it does not disclose an internal reporting mechanism that allows employees to voluntarily self-disclose their sexual orientation and gender identity. Further, the company does not disclose any workforce diversity numbers.

Potential sources:

- Diversity report
- Company blog