2. INSTAGRAM

- **Launch date:** October 6, 2010
- **Number of Users/Visitors:** 2 billion monthly active users [132]
- **Short Overview of Content Moderation Process:** Content moderators review posts that have been flagged by AI, reported by users or non-users. Non-users can file a report available on Instagram's Help Centre. Most of this work is outsourced to third-party vendors.
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online?** Yes

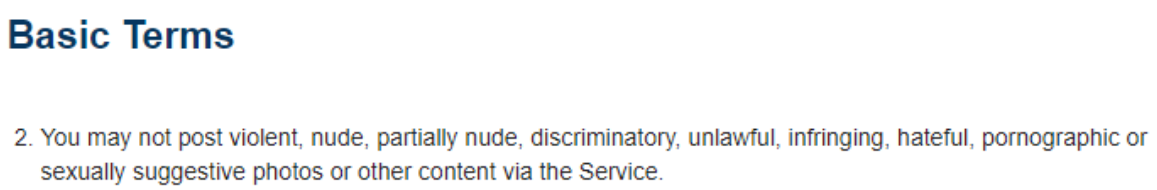
[132] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," *Statista*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on April 30, 2023).

Key Developments

Terms of Use

While the first traceable Terms of Use for Instagram are from 2012, the document did not include a relevant provision on hate speech until 2013 (see Figure 6). The Terms did not define hateful content, however, nor reference any characteristics that Instagram protected from hateful content. By 2018, the provision was deleted from the Terms of Use.

Figure 6[133]



Community Guidelines

The first traceable Instagram Community Guidelines date to 2012, but the Guidelines did not include a hate speech provision until 2015. Under the subheading "respect other members of the Instagram community," the 2015 guidelines noted:

"We want to foster a positive, diverse community. We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages. We do generally allow stronger conversation around people who are featured in the news or have a large public audience due to their profession or chosen activities.

It's never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. When hate speech is being shared to challenge it or to raise awareness, we may allow it. In those instances, we ask that you express your intent clearly."[134]

The above provision is an amalgamation of various types of objectionable content, ranging from harassment to hate speech, but the provision lacks a clear definition of any of these terms. The second paragraph, however, implies that Instagram considers hate speech to involve encouraging violence against or attacking individuals on the basis of protected characteristics. It is unclear, however, if the company intends to treat credible threats, hate speech, degrading content, blackmail, and harassment as separate types of content – or if the company considers all of these

[133] https://web.archive.org/web/20130123212202/http://instagram.com/about/legal/terms/updated/
[134] https://web.archive.org/web/20150825000805/https://www.facebook.com/help/instagram/477434105621119/

forms of speech to be hate speech, given the second paragraph focuses on hate speech specifically. In July 2020, however, Instagram added hyperlinks to this provision, which implied that the company considers these types of objectionable content separately.[135] The company added a link to the phrase "hate speech" that directed people to the hate speech policy in the Facebook Community Standards. The phrase "credible threats" linked to the Violence & Incitement policy in the Facebook Community Standards, while the phrase "degrade or shame them" linked to the Bullying & Harassment policy in the Facebook Community Standards. Thus, it appears that this provision of the Instagram Community Guidelines corresponds to several different policies within the Facebook Community Standards.

By adding these hyperlinks to the Instagram Community Guidelines, the company implied that Facebook's Community Standards apply to content on Instagram. The scope of the hate speech provision in Instagram's Community Guidelines differs from the hate speech policy in the Facebook Community Standards, however. Instagram's policy references ten protected characteristics, compared to the 16 that Facebook's policy covers. Thus, it is not clear whether Instagram prohibits hate speech against the ten protected characteristics listed in the Instagram Community Guidelines or the 16 listed in the Facebook Community Standards.

The Community Standards Enforcement Report page of the Transparency Center, however, states "Facebook and Instagram share content policies. What is violating on Facebook is violating on Instagram. Throughout this report, we link to our Community Standards, which include the most comprehensive description of these policies."[136] This statement suggests there is no need for two sets of policies and the Community Standards reflects the policies enforced on Instagram. Why then does Instagram continue to list the Community Guidelines on its website? Why are the Community Guidelines listed under "Other Policies" on the Meta Transparency Center? When did the Community Standards become the default rules for both platforms? There have been no updates to the relevant provision in Instagram's Community Guidelines since 2020, so perhaps sometime after that date? It is not clear.

This confusion is problematic in terms of the legality requirement in Article 19 (3) of the ICCPR. Users have no way of knowing what exactly the rules are for Instagram. Thus, even if the Community Standards are, in fact, the rules for both platforms, we still feel it is important to analyze Instagram's Community Guidelines in this report, since the Community Standards Enforcement Report is the only place where Meta clearly states that the Community Standards apply to both.

[135] https://web.archive.org/web/20200730024324/https://help.instagram.com/477434105621119
[136] https://transparency.fb.com/data/community-standards-enforcement/?source=https%3A%2F%2Ftransparency.faceb

Analysis of Policy Scope

As described in the previous section, the scope of content covered by Instagram's hate speech policy is somewhat unclear. The most basic interpretation of the Instagram policy guidance, however, suggests the company has defined hate speech as incitement to violence or attacks based on protected characteristics since 2015, as reflected in Table 3. This scope of covered content aligns with Article 20(2) of the ICCPR. Table 4 illustrates that the scope of characteristics covered by Instagram's policy also has not changed since 2015, though it is much broader than the list of characteristics covered by Article 20 (2).

Table 3

Content Explicitly Covered by Instagram's Hate Speech Policies		2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Hate(ful) speech/ content		X	X	X	X	X	X	X	X	X	X	X
Promotion of Hatred												
Support for Organized Hate (Including Symbols)				X	X	X	X	X	X	X	X	X
On the basis of protected characteristics	Incitement to or Threats of Violence			X	X	X	X	X	X	X	X	X
	Attacks			X	X	X	X	X	X	X	X	X
	Statements of inferiority or content that demeans											
	Dehumanization											
	Expressions of contempt or disgust											
	Calls for exclusion or segregation											
	Discrimination											
	Denying or mocking historical atrocities, or valorizing the perpetrators											
	Slurs											
	Harmful Stereotypes											
	Conspiracy Theories											
	Cursing											

Table 4

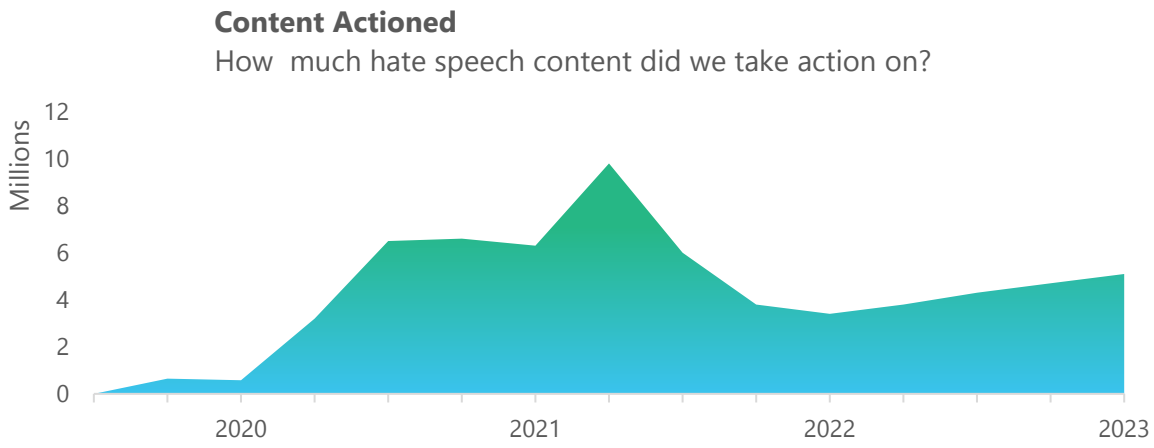
Characteristics Protected in Instagram's Hate Speech Policies											
	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	**0**	**0**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**
Race			X	X	X	X	X	X	X	X	X
Ethnicity			X	X	X	X	X	X	X	X	X
National Origin			X	X	X	X	X	X	X	X	X
Religion			X	X	X	X	X	X	X	X	X
Gender			X	X	X	X	X	X	X	X	X
Color											
Immigration Status											
Sex			X	X	X	X	X	X	X	X	X
Gender Identity			X	X	X	X	X	X	X	X	X
Sexual Orientation			X	X	X	X	X	X	X	X	X
Age											
Disability			X	X	X	X	X	X	X	X	X
Disease/ Medical Condition			X	X	X	X	X	X	X	X	X
Veteran Status											
Occupation											
Weight											
Pregnancy											
Caste											
Victims of a Major Event											
Socio-Economic Status											
Culture											
Tribe											

Notes: An X indicates the company's hate speech policies covered that protected characteristics for at least one month during the given year.

Changes in Enforcement Volume

As it does for Facebook, Meta provides a Community Standards Enforcement Report for Instagram. However, these reports do not exist for years prior to 2015, the year Instagram made its only major changes to the scope of its hate speech provisions.[137] Thus, it would be difficult to use this data to assess how changes in the scope of Instagram's Community Guidelines impacted enforcement volumes. Nevertheless, Figure 7, which reports the amount of Instagram content actioned due to violations of the hate speech prohibition, shows substantial changes in this metric over time. Because this report is part of the Community Standards Enforcement Report, it raises questions about whether Instagram's Community Guidelines, or Facebook's Community Standards, represent the final word on what content is and is not allowed on Instagram. Thus, changes in the Community Standards, as documented in the previous section, could possibly drive the changes in enforcement volume depicted in Figure 7. However, Figure 7 shows a large increase in the amount of content actioned in both Q2 and Q3 2020, and there was only a noticeable change in the scope of Facebook's hate speech policy in August 2020. In fact, Meta attributed these 2020 increases in Instagram content actioned for hate speech violations to improvements in proactive detection technology for the English, Spanish, and Arabic languages, and noted that they expected continued fluctuations in these numbers as the company adjusted to COVID-19 related workforce disruptions.[138]

Figure 7[139]



[137] "Hate Speech, Community Standards Enforcement Report," *Meta Transparency Center,* https://transparency.fb.com/data/community-standards-enforcement/hate-speech/instagram/ .

[138] See Guy Rosen, "Community Standards Enforcement Report, August 2020," *Meta Newsroom,* August 11, 2020, https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/, and Guy Rosen, "Community Standards Enforcement Report, November 2020," *Meta Newsroom*, November 19, 2020, https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/.

[139] "Community Standards Enforcement Report," Meta Transparency Center, https://transparency.fb.com/data/community-standards-enforcement/ .