UC BERKELEY
CENTER FOR LONG-TERM CYBERSECURITY

CLTC WHITE PAPER SERIES

A Secure and Equitable Metaverse

Designing Effective Community Guidelines for Social VR

RAFI LAZERSON

CLTC WHITE PAPER SERIES

A Secure and Equitable Metaverse

Designing Effective Community Guidelines for Social VR

RAFI LAZERSON

NOVEMBER 2022

CLTC
Center for Long-Term Cybersecurity
UC Berkeley

Contents

EXECUTIVE SUMMARY 1

INTRODUCTION 4

METHODOLOGY, SCOPE, AND LIMITATIONS 5

INTERACTIONS AND HARASSMENT IN SOCIAL VR 7

User interactions in VR feel "real" and are conduct-based 7

Harassment in social VR is likened to in-person harassment 8

Harassment in social VR is severe and not uncommon 10

Ensuring inclusive social VR requires immediate action 11

Robust community safety practices are needed for social VR platforms 12

REVIEW OF META'S COMMUNITY GUIDELINES IN SOCIAL VR 12

Robust community guidelines are key to community safety 14

Too much ambiguity regarding which and where Meta policies apply in VR 17

Applying Facebook Community Standards was a positive step for VR safety 20

Ambiguity about whether Facebook Community Standards apply to VR leads to a significant loss of user safety 23

Without comprehensive playbooks, Conduct in VR Policy is too broad to be effective 24

The Conduct in VR Policy is unclear in its delineation of public vs. private VR 26

The Conduct in VR Policy lacks transparency elements that are in the Facebook Community Standards 28

Summary of Findings 28

RECOMMENDATIONS 30

Social VR community guidelines should be accessible **30**

Social VR community guidelines should be specific and comprehensible **30**

Social VR community guidelines should maintain baseline level of thoroughness of 2D social media community guidelines **33**

Social VR platforms should partner across sectors to assess harms and prepare for future threats **35**

Social VR community guidelines should be transparent with a change log, an explanation of values, and policy rationales **37**

Future Research **37**

REFERENCES 39

Appendix A: Horizon Policy (July 2022 Version) **43**

Appendix B: Horizon Policy (June 2022 Version) **44**

Appendix C: Conduct in VR Policy (August 2022 Version) **45**

Appendix D: Conduct in VR Policy (July 2022 Version) **47**

Appendix E: Facebook Community Standards Principles and Policies **49**

Appendix F: Facebook Community Standards Hate Speech Policy **50**

Appendix G: Facebook Community Standards Bullying and Harassment Policy **55**

Appendix H: YouTube Harassment & Cyberbullying Policies **62**

ACKNOWLEDGMENTS 67

ABOUT THE AUTHOR 67

Executive Summary

This paper assesses the novelty of user interactions in social virtual reality (VR); examines hate and harassment in social VR; explores the need for thorough community guidelines; reviews the effectiveness of Meta's community guidelines in social VR; and proposes specific strategies and tactics for how Meta and other platforms can design effective community guidelines in social VR.

User interactions in social VR feel real and present due to the immersive quality of donning a VR headset and the experience of avatar embodiment. User interactions in social VR are also primarily synchronous and conduct-based (integrating speech, avatar gestures, and movements), and are a departure from two-dimensional (2D) social media, where interactions are primarily asynchronous and content-based (using text, images, video).

Harassment in social VR is likened to in-person harassment, is likely pervasive, and impacts individuals from marginalized communities disproportionately. Social VR is surging in use by the public. As investment in the metaverse grows, social VR will increasingly become an essential part of how society plays, connects, and works. Without immediate action to review and design effective community safety practices, including community guidelines, social VR stands to exacerbate inequalities rather than expand opportunities for inclusive positive interaction.

Community guidelines are a crucial component of a platform's community safety practices, providing the public with:

- Injunctive norms prescribing behaviors as acceptable or unacceptable;
- Guidance for reporting harms and seeking remedy;
- Transparency into the platform's internal moderation policies;
- Commitments of responsibility from the platform in ensuring user safety; and
- Reference for the public to hold the platform accountable for moderation commitments.

Community guidelines are only effective if they are:

- Accessible: Guidelines should be easily found and understood, and provide clarity regarding when they apply.
- Comprehensive: Guidelines should delineate broad, high-level *principles*, divided by mid-level *policies* that cover the various categories of online harms.

- Specific: Guidelines should clarify low-level *playbooks* for each policy, giving detailed specifics as to what constitutes violations.
- Transparent: Guidelines should provide rationales to the policies; explanation to the values that inform the guidelines; and a change log of any edits of the guidelines.

An in-depth analysis of Meta's community guidelines relevant to social VR finds four need-based opportunities for growth:

1. Need for increased accessibility:
    * There are two, perhaps three, community guidelines that apply in Meta's social VR:[1] *Horizon Policy,*[2] *Conduct in VR Policy*, and *Facebook Community Standards*. There is an unclear relationship between the three community guidelines and where each applies.

2. Need for increased comprehensiveness:
    * The *Horizon Policy* and the *Conduct in VR Policy* do not thoroughly delineate *principles* or break them down by *policies*. As a result, the *Horizon Policy* and the *Conduct in VR Policy* do not address the breadth of potential online harms.

3. Need for increased specificity:
    * The *Horizon Policy* and the *Conduct in VR Policy* do not provide *playbooks*, leaving ambiguity about what Meta considers violations.
    * The *Conduct in VR Policy* does not provide enough clarity about how Meta distinguishes public from private VR spaces.

4. Need for increased transparency:
    * The *Horizon Policy* and the *Conduct in VR Policy* lack a change log of edits, policy rationales, and explanation of values.

The analysis also finds that clarity is needed on whether the *Facebook Community Standards* continue to inform the *Conduct in VR Policy* and apply to social VR.

1 As of August 2022

2 "Horizon Policy" is used in this paper to refer collectively to the July 2022 versions of the *Horizon Mature Worlds Policy* and the *Horizon Prohibited Worlds Policy*, as well as their predecessor, the June 2022 version of the *Horizon Worlds Prohibited Content Policy*.

- August 2022 edits to the *Conduct in VR Policy* removed a statement that the *Facebook Community Standards* apply to VR.
- Applying the *Facebook Community Standards* to social VR set a positive expectation that the thoroughness of 2D community guidelines would be maintained and implemented in VR.
- If the *Facebook Community Standards* no longer apply to VR, there is increased urgency to make the *Conduct in VR Policy* and *Horizon Policy* more thorough and effective.

Social VR platforms should proactively develop community guidelines that maintain the standards of accessibility, comprehensiveness, specificity, and transparency used by 2D social media platforms. Platforms should develop the community guidelines for the unique experience of immersive, conduct-based interactions in social VR. If applying community guidelines from 2D social media to social VR, detailed *playbooks* are needed to clarify how the *principles* and *policies* apply to unique forms of content and conduct in social VR. If delineating public vs. private spaces, those concepts should be clarified. Researchers, civil society organizations, and policymakers should collaborate with social VR platforms, and serve as independent reviewers, in efforts to develop effective community guidelines in social VR. Detailed examples related to these recommendations, as well as future research, are provided in this paper.

Introduction

Social media platforms such as Facebook and YouTube have developed robust user-facing policies regarding content prohibited on their platforms; i.e., the guidelines are (1) accessible, (2) comprehensive, (3) specific, and (4) transparent. Although there is room for further development, the community guidelines of these two major social media platforms now offer meaningful detail that can contribute to positive platform norms and provide the public with a degree of transparency regarding moderation practices. However, the development of these robust community guidelines took years, and many rampant harms went unchecked on the platforms before policies were delineated. Additionally, the shaping of robust guidelines was in many ways a reaction to public exposure of harms occurring on the platforms, public leaks of internal community safety policies,[3] and public pressure campaigns.[4]

With the rise of the "metaverse" — a dual physicalization of digital worlds (e.g., through immersive virtual reality) and digitization of the physical world (e.g., through augmented reality)[5] — we face an uncertainty: will social VR platforms proactively develop clear community guidelines at this early stage of user adoption, or will their process follow the slow, opaque, and reactive trajectories that were typical of 2D social media platforms? An assessment of social VR community guidelines is needed: how do current social VR community guidelines compare to the robust versions offered by 2D social media platforms, e.g., Facebook?

This paper details the need for community guidelines in social VR to match the baseline for thoroughness now established in community guidelines for 2D social media, and conducts a preliminary assessment of the current state of social VR community guidelines. The goal of this paper is to support a digital future in which social VR platforms proactively include robust community guidelines that clearly delineate prohibited online harms, foster inclusive user norms, and provide the public and end-users with transparency into their moderation practices.

3 Simon van Zuylen-Wood, "'Men Are Scum': Inside Facebook's War on Hate Speech | Vanity Fair," Vanity Fair, February 26, 2019, https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech

4 Such as the #StopHateForProfit campaign, see: "One Year After Stop Hate for Profit: Platforms' Progress," Stop Hate for Profit, June 16, 2021, https://www.stophateforprofit.org/platforms-progress-year-later.

5 See Educause on how the term extended reality (an oft-used alternative to "metaverse") "captures both the continuum between and the growing convergence of the physical and virtual worlds we inhabit" as well as for related discussion points on the "physicalization of the digital world" and "digitization of the physical world." "XR (Extended Reality) Community Group," Educause, accessed October 4, 2022, https://www.educause.edu/community/xr-extended-reality-community-group.

Section 1, *Interactions and Harassment in Social VR*, explores the unique quality of user interactions in social VR, in which users are embodied in their avatars and interactions are primarily through conduct, such as speech and avatar gestures. This marks a shift from user interactions in 2D social media, where users are not wholly immersed in their screens and interactions between users are primarily through content, such as posts and messages. The section reviews how immersive *conduct*-based interactions shape positive and negative experiences in social VR; assesses the prevalence of harassment in social VR; and examines the urgent need for effective community safety practices in social VR.

Section 2, *Review of Meta's Community Guidelines in Social VR*, examines the current state of community guidelines in social VR, by analyzing whether Meta, arguably the most influential metaverse company, has maintained the standards for community guidelines it uses for 2D social media, and assessing how effectively the guidelines account for immersive conduct-based interactions in social VR. The section explores both the strengths and weaknesses of Meta's applicable community guidelines and considers where there is need for further development. The section ends with future research opportunities and detailed recommendations for Meta and other platforms to develop robust community guidelines in social VR.

Methodology, Scope, and Limitations

This research examined the current state of harassment in VR through consulting academic and think-tank research as well as media reports. The author conducted an examination of academic literature focused on assessing and coding community guidelines developed for 2D social media platforms, as well as an in-depth analysis of Meta's community guidelines that are applicable to social VR, including the *Horizon Policy*, *Conduct in VR Policy*, and *Facebook Community Standards*. The author attended several topical webinars and the 2022 Augmented World Expo (AWE) in Santa Clara, California.

Although there are many builders of the metaverse, this paper assesses that well-funded corporations have a disproportionate impact on the formation of the metaverse and on norms within social VR, and therefore have a responsibility to lead the industry in developing responsible policies and practices. The paper focuses primarily on the social VR community

guidelines of Meta, the company formerly known as Facebook that was renamed in 2021 to emphasize its strategic focus on the metaverse. Currently, Meta has outsized influence on the metaverse: Meta spent $10 billion on the metaverse in 2021 alone,[6] and its Quest 2 headsets accounted for 78% of all AR/VR headset sales in 2021.[7]

The goals of focusing on Meta, and indeed the broader goal of this paper, are: 1) to provide an initial assessment for the public as to whether and how Meta has developed robust community guidelines for social VR that maintain established standards of thoroughness used in 2D social media community guidelines; 2) to assess whether Meta's community guidelines in VR account for immersive, conduct-based user interactions; and 3) to propose preliminary recommendations for how Meta can further incorporate such considerations to ensure the development of a secure and equitable metaverse.

Meta's social VR policies are an evolving set of documents, and the intention of this paper is to help inform that evolution. This research completed data collection in August 2022. The underlying findings of this research — that there is a need for accessible, comprehensive, specific, and transparent community guidelines in social VR — can continue to inform the evolution of Meta's policies, the development of other VR platforms' policies, and the efforts of researchers and policymakers.

This paper primarily examines community guidelines because of their significant function in rule-setting and providing platform accountability (as the paper discusses). However, even the most comprehensive community guidelines are only effective if they are part of a platform's broader set of community safety practices, spanning policies, products, and operations. This paper is meant to contribute to wider efforts to review community safety practices in social VR. Similarly, VR is only one component of the metaverse, a term used to refer to a variety of technologies as well as new methods of approaching technologies.[8] This paper focuses on social VR because of the technology's novel form of user interactions and its prominence within discourse surrounding the metaverse. This paper is meant to contribute to wider efforts to review community safety practices across metaverse technologies and developments.

6 Mike Isaac, "Meta Spent $10 Billion on the Metaverse in 2021, Dragging down Profit," The New York Times, February 2, 2022, https://www.nytimes.com/2022/02/02/technology/meta-facebook-earnings-metaverse.html.

7 "AR/VR Headset Shipments Grew Dramatically in 2021, Thanks Largely to Meta's Strong Quest 2 Volumes, with Growth Forecast to Continue, According to IDC," IDC:, March 21, 2022, https://www.idc.com/getdoc.jsp?containerId=prUS48969722.

8 Eric Ravenscraft, "What Is the Metaverse, Exactly?" WIRED, April 25, 2022, https://www.wired.com/story/what-is-the-metaverse/.

Interactions and Harassment in Social VR

USER INTERACTIONS IN VR FEEL "REAL" AND ARE CONDUCT-BASED

Although the metaverse includes various forms of virtual and augmented reality, social VR, i.e., VR that offers interaction between users, has been heralded as a principal benefit of the metaverse over Web 2.0 and two-dimensional social media. For example, Meta, perhaps the company most currently associated with the metaverse, issued the following statement in October 2021 upon changing its name from Facebook to Meta: "The defining quality of the metaverse will be a feeling of presence — like you are right there with another person or in another place. Feeling truly present with another person is the ultimate dream of social technology. That is why we are focused on building this."[9] For Meta, it is the deep immersion of users in social VR environments that epitomizes the possibility of the metaverse and VR technology.

Social VR has been described by researchers as "an embodied way to connect with one another across distances."[10] Embodied in the first-person perspective of an avatar, users can feel a genuine sense of presence and immersion in the virtual environment. This embodiment allows users to interact with others in VR with a shared sense of co-presence,[11] despite being potentially distant in the physical world. Multi-dimensional objects within a VR environment contribute to the user's sense of presence in ways that are impossible with objects a user may interact with on a computer or phone screen.

Significantly, user interactions in social VR are primarily *conduct*-based, using synchronous real-time speech along with avatar movements and gestures. This provides users in social VR with a more fluid and natural world-like experience than is possible with 2D social media, where

9 Mark Zuckerberg, "Founder's Letter, 2021," Meta, October 28, 2021, https://about.fb.com/news/2021/10/founders-letter/.

10 Michelle Cortese and Jessica Outlaw, "Social and Multi-User Spaces in VR: Trolling, Harassment, and Online Safety," *The IEEE Global Initiative on Ethics of Extended Reality (XR) Report* (December 15, 2021): 1–17.

11 Ibid.

interactions are primarily asynchronous[12] and *content*-based, limited to text, image, or video posts and messages.

In a sense, social VR builds on the conduct-based interactions of 2D multiplayer online gaming, such as in sports games, where users often interact with others in synchronous speech and avatar gestures. However, where users of 2D online gaming visually see an avatar distinct from themselves, in social VR, the user is embodied into the avatar through the wearing of a VR headset. This embodiment removes the sense of separation and distinction between the user and the avatar, contributing to interactions between users that feel real and present. Embodiment, immersion, and synchronous conduct-based interactions open the possibility for numerous positive forms of connection between users that were previously impossible in digital technologies. This can include: 1) virtual versions of offline spaces, such as classrooms, live shows, conferences, or parties;[13] or 2) virtual fantastical worlds that do not directly mimic the offline world, such as worlds with zombies or where people have the ability to fly.[14] To the user, any VR world, even the fantastical, can feel real and present due to avatar-embodiment, world-immersion, and synchronous conduct-based interactions.

HARASSMENT IN SOCIAL VR IS LIKENED TO IN-PERSON HARASSMENT

VR's sense of presence, however, can also make negative experiences feel more "real" than on traditional 2D social media sites.[15] In its community guidelines for AltspaceVR, a highly popular social VR platform, parent company Microsoft likens violations of personal space in VR to violations of in-person personal space:

> AltspaceVR is a truly unique environment. The power of the platform is the ability to communicate verbally with fellow users and engage with them non-verbally through body gestures and the spatial positioning of your avatar. Like with real-world

12 Lindsay Blackwell et al., "Harassment in Social Virtual Reality: Challenges for Platform Governance," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 100:1-100:25, https://doi.org/10.1145/3359202.

13 Cortese and Outlaw, "The IEEE Global Initiative on Ethics of Extended Reality (XR) Report–Social and Multi-User Spaces in VR."

14 See PCmag's definition of social VR: "Getting together in a simulated world using a virtual reality (VR) system and social VR app. Participants appear as avatars in environments that can be lifelike or fantasy worlds." Source: "Definition of Social VR," PCMag, accessed June 17, 2022, https://www.pcmag.com/encyclopedia/term/social-vr.

15 Blackwell et al., "Harassment in Social Virtual Reality."

> interaction, users in VR can feel social discomfort if those with who they're interacting are violating real-world personal space.[16]

Indeed, experiences of harassment in VR have been described as comparable to in-person harassment. As far back as 2016, author Jordan Belamire shared a blog post regarding being groped by another user within just three minutes of joining a social VR zombie-shooting game. Belamire described the immersiveness of the fantastical virtual world as feeling realistic, and the groping she was subjected to as feeling real and violating.[17]

Katherine Cross, a University of Washington researcher of online harassment, similarly noted that when virtual reality is immersive and real, toxic behavior in that environment is also real. "At the end of the day, the nature of virtual-reality spaces is such that it is designed to trick the user into thinking they are physically in a certain space, that their every bodily action is occurring in a 3D environment," Cross told *MIT Technology Review*. "It's part of the reason why emotional reactions can be stronger in that space, and why VR triggers the same internal nervous system and psychological responses."[18]

In addition to the immersiveness created by donning VR headsets, other VR hardware can contribute to the realistic experience of harassment in social VR. For example, in a 2022 report, a metaverse researcher described how she was subjected to sexual assault within just one hour of wearing Meta's Oculus headset, and that her controller vibrated when the male avatars assaulted her, creating a physical sensation aligned with what she was experiencing online.[19] As haptic gloves, suits, and other VR immersion hardware become a common part of VR use, experiences of harassment may feel increasingly indistinguishable from in-person harassment.[20]

16 "Community Standards," Microsoft, January 20, 2022, https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/community/community-standards.

17 Jordan Belamire, "My First Virtual Reality Groping," *Athena Talks* (blog), October 22, 2016, https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee.

18 Tanya Basu, "The Metaverse Has a Groping Problem Already," MIT Technology Review, December 16, 2021, https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/.

19 Weilun Soon, "A Researcher's Avatar Was Sexually Assaulted on a Metaverse Platform Owned by Meta, Making Her the Latest Victim of Sexual Abuse on Meta's Platforms, Watchdog Says," Business Insider, May 29, 2022, https://www.businessinsider.com/researcher-claims-her-avatar-was-raped-on-metas-metaverse-platform-2022-5.

20 This can be further underscored by Meta's recent announcement that it aims to make VR experiences as realistic-looking to users as the physical world (passing the so-called "visual Turing test"). See Lisa Brown Jaloza, "Passing the Visual Turing Test: The inside Story of Our Quest for Visual Realism in VR," Tech at Meta, June 21, 2022, https://tech.fb.com/ar-vr/2022/06/passing-the-visual-turing-test-the-inside-story-of-our-quest-for-visual-realism-in-vr/.

HARASSMENT IN SOCIAL VR IS SEVERE AND NOT UNCOMMON

Unfortunately, harassment in social VR is not uncommon. Many forms of harassment in social VR have been described by users, including homophobia,[21] racist chanting,[22] groping,[23] stalking,[24] hitting,[25] death threats,[26] and even gang rape.[27] A 2018 survey of VR users found that 49% of women respondents and 36% of male respondents described experiencing sexual harrassment in VR.[28] The survey results suggest that harassment in social VR may be pervasive regardless of user identity. Additionally, the higher rates of reported harassment by women respondents suggest that there may also be a disproportionate experience of harassment in social VR for individuals of marginalized identities. Considering the growth in social VR since 2018, the Future Research section at the end of this paper calls for updated surveys to provide current assessments of harassment in social VR.

The impacts of harassment online, even outside of the immersion of VR, are severe, with adverse effects on mental health, physical health, and economic well-being.[29] Harassment online has also been shown to impact the ability of individuals of marginalized identities to fully present themselves in online spaces. For example, a 2021 survey of online gamers found that 59% of women gamers conceal their gender when playing games online, including through the use of non-gendered identity or pretending to be male, in order to avoid harassment.[30] The use of names, avatars, voice-changers, and other techniques that conceal being identified as a woman marks a sign of the great lengths many have to go through to avoid harassment in online gaming.

21 Jessica Outlaw, "Harassment in Social VR: Stories from Survey Respondents," *Medium* (blog), May 8, 2018, https://jessica-outlaw.medium.com/harassment-in-social-vr-stories-from-survey-respondents-59c9cde7ac02.

22 Kat Lo [@lolkat], "I Jumped on VR Chat Recently for Research and within 1 Minute of Entering the First Public Room I Witnessed: 1. a Dozen People Chanting the n Word and Laughing for 5 Minutes 2. Several Men Crowding and Making Sexual Comments to a Nervous Girl Who Said She Was 15 in Her Profile," Tweet, *Twitter*, May 31, 2022, https://twitter.com/lolkat/status/1531608673117667328.

23 Basu, "The Metaverse Has a Groping Problem Already."

24 Outlaw, "Harassment in Social VR."

25 Ibid.

26 Ibid.

27 Nina Jane Patel, "Fiction vs. Non-Fiction," *Medium* (blog), December 21, 2021, https://ninajanepatel.medium.com/fiction-vs-non-fiction-d824c6edfbe2.

28 "Virtual Harassment: The Social Experience of 600+ Regular Virtual Reality Users," Slidedeck (The Extended Mind, 2018), https://drive.google.com/file/d/1afFQJN6QAwmeZdGcRj9R40hVr00ZNO4a/view.

29 "Online Hate and Harassment: The American Experience 2021," ADL, May 3, 2022, https://www.adl.org/online-hate-2021.

30 "Reach3 Insights' New Research Reveals 59% of Women Surveyed Use a Non-Gendered/Male Identity to Avoid Harassment While Gaming," Reach3, May 19, 2021, https://www.reach3insights.com/women-gaming-study.

Some have noted a probable influence of online gaming culture on social VR environments.[31] While there are positive aspects of this influence, such as unique forms of community-building in-game (e.g., Warcraft guilds) and on gaming-adjacent platforms (e.g., Discord), there is also growing concern that without immediate action, some of the exclusionary norms that have developed in many online gaming platforms and spaces will also become entrenched in social VR and increasingly difficult to change. Similarly, in absence of meaningful platform moderation, individual users are left with the onus of community safety, and may be adopting strategies to avoid harassment that are similar to those used in online gaming. Considering the pervasiveness of harassment in social VR, it is likely that many individuals from marginalized communities already feel pressure to conceal their identities in social VR in order to avoid harassment. This may be through choosing an avatar with a different skin tone, shape, or gender presentation; selecting not to incorporate religious or traditional garb on their avatar; or opting for a username that does not reference their identity. When interacting in a social VR platform, it may also mean limiting speaking if they feel their marginalized identity may be revealed through their voice. Harassment in social VR may even prompt individuals to avoid a particular platform or social VR altogether, as was anecdotally reported in 2018 research.[32] The Future Research section details the need for updated surveys to assess the prevalence and diverse experiences of users in VR compelled to conceal elements of their identities or avoid VR platforms altogether.

ENSURING INCLUSIVE SOCIAL VR REQUIRES IMMEDIATE ACTION

The significant investment of resources and time into the metaverse from the world's largest corporations and wealthiest investors, as well as from researchers, creatives, and small businesses, suggests that social VR will become a growing part of our day-to-day lives, and could transform domains such as work,[33] education, and tourism.[34] This underscores how deleterious the impacts of exclusionary norms in social VR could be. The inability for some users to present as themselves in or even enter social VR without fear could have severe health and economic ramifications. Without immediate steps taken to ensure the formation of inclusive norms in social VR, the metaverse stands to exacerbate societal inequalities,

31 Blackwell et al., "Harassment in Social Virtual Reality."

32 Outlaw, "Harassment in Social VR."

33 For example, an anticipated Meta VR headset is expected to provide users with high-resolution image quality enabling the reading emails or writing code. See Richard Lawler, "Meta's VR Roadmap Reportedly Plans Four New Headsets for Release through 2024 - The Verge," The Verge, May 2, 2022, https://www.theverge.com/2022/5/2/23053888/meta-virtual-reality-headset-cambria-quest-vr-mr.

34 David Heaney, "Meta Vision Of The Metaverse Shows Futuristic Headset Design," UploadVR, June 16, 2022, https://uploadvr.com/meta-concept-future-headset-design/.

benefitting the most privileged and causing harm to already disadvantaged and marginalized communities.

Efforts to build an inclusive metaverse cannot wait; the metaverse already exists, and norms are becoming established. It may be that the metaverse is not yet in its idealized form, and the base of users may still be small compared to traditional social media. However, this nascent stage is precisely when intervention can be effective, before norms become too entrenched. We cannot afford to make the same mistakes as occurred with 2D social media, where exclusionary user behavior and opaque platform policies became increasingly difficult to reverse with each passing year. Indeed, the metaverse market worldwide is already growing rapidly. Augmented reality is already in widespread use (e.g., in applications such as TikTok filters and Pokemon Go), and virtual reality is gaining momentum. In 2021, shipments of augmented reality and virtual reality headsets grew 92.1% year over year, reaching 11.2 million units.[35] In February 2022, Meta's social VR platform Horizon reached 300,000 monthly users, a 10x increase in about three months.[36] There are numerous other social VR platforms, some of which have even more active users than Horizon. For example, Rec Room, an online video game that can be accessed in 2D or VR, grew from about one million active VR monthly users in January 2021 to surpassing three million in April 2022.[37]

Technology improvements will make the metaverse more visually appealing and user-friendly, and hopefully will offer more interoperability between platforms. But that does not negate that a version of the metaverse already exists in which norms are being established.

ROBUST COMMUNITY SAFETY PRACTICES ARE NEEDED FOR SOCIAL VR PLATFORMS

On a given platform, social norms such as the prevalence of harassment are significantly influenced by the community safety practices that the platform implements.[38] Community safety

35 "AR/VR Headset Shipments Grew Dramatically in 2021, Thanks Largely to Meta's Strong Quest 2 Volumes, with Growth Forecast to Continue, According to IDC."

36 Alex Heath, "Meta's Social VR Platform Horizon Worlds Hits 300,000 Users - The Verge," The Verge, February 17, 2022, https://www.theverge.com/2022/2/17/22939297/meta-social-vr-platform-horizon-300000-users.

37 Jamie Feltham, "Rec Room Passes 3 Million Monthly Active VR Users," UploadVR, April 14, 2022, https://uploadvr.com/rec-room-3-million-vr-users/.

38 There are many other factors that contribute to the development of norms in online platforms, such as platform product design, government regulation, user demographics, broader online and offline culture, and broader societal changes. This paper focuses on community safety practices, because they are most directly associated with harms on a platform, and they are within the capacity of each platform to develop.

practices include three primary elements: policy, product, and operations.[39] Effective community safety practices incorporate comprehensive and intentional design of all three elements.

- **Policy:** Sets the ground rules for what is allowed and not allowed. Policy includes:
  - External-facing communications, e.g., community guidelines, legal policies and terms of service, or public commitments by the platform to moderate particular content/conduct; and
  - Internal guidelines provided to moderators regarding reviewing and taking action in response to user actions.[40]

- **Product:** Features within a platform that provide users and moderators with the ability to ensure a safe experience. Such features include:
  - User tools, e.g., for blocking, muting, recording, and reporting;
  - Moderator tools, e.g., for deleting content, and detection of violations;
  - Educational tools, e.g., tutorials, videos, and signage regarding community guidelines; and
  - Invisible safety tools, e.g., age-gating, recommendations, and downranking.

- **Operations:** Give teeth to the policy via enforcement and penalties. Operations include:
  - How the platform enforces the policy through human and AI moderation. This enforcement may result in user warnings, demotion or removal of content, account suspension/termination, or limiting account access to platform features.

In this early period of social VR adoption, when norms and culture are still being solidified, platforms need to create comprehensive community safety practices to set standards for the future of social VR. This should be done by platform developers working alongside external researchers, policymakers, and civil society to ensure the community practices developed are equitable, effective, and transparent.

39 See "Promoting Safety with Policy, Product and Operations," Meta, November 15, 2018, https://about.fb.com/news/2018/11/inside-feed-community-integrity-keeping-people-safe/.

40 Internal guidelines could also be categorized under community safety operations.

Review of Meta's Community Guidelines in Social VR

In support of reviewing the community safety practices of social VR platforms, this next section provides a preliminary analysis of Meta's community guidelines in social VR. Meta was chosen as the focus because the company has written about its goal to work with outside researchers to develop an ethical and responsible metaverse,[41] and because the company has an outsized influence on the development of standards in the metaverse and thus has a greater responsibility to get it right.

ROBUST COMMUNITY GUIDELINES ARE KEY TO COMMUNITY SAFETY

User-facing community guidelines were chosen as a point of analysis because they accomplish five crucial community safety functions:

1. Rule-setting: Community guidelines serve as the primary set of injunctive norms prescribed from the platform to users regarding what behaviors are acceptable and unacceptable on the platform.
2. Guidance: Community guidelines serve as the primary resource for users to seek remedy and moderation from the platform, including to appeal platform actions on their content or account.
3. Clarity: Community guidelines serve as the only public-facing insight into a platform's internal moderation policies.
4. Responsibility: Community guidelines serve as the primary means by which a platform can make public commitments regarding what is unacceptable on the platform. The guidelines are also a public commitment that the platform pledges to play an active role in ensuring community safety and that it will not place the onus of safety solely on the user, such as through platform features (e.g., blocking or personal boundaries).
5. Accountability: Community guidelines serve as the primary tool for the public to hold companies accountable for harmful behavior on their platforms.

41 Andrew Bosworth and Nick Clegg, "Building the Metaverse Responsibly," Meta, September 27, 2021, https://about.fb.com/news/2021/09/building-the-metaverse-responsibly/.

Even if many users do not currently read the community guidelines in a thorough manner (which would be required to make the sharing of rule-setting injunctive norms highly effective), these functions collectively show the importance of community guidelines in providing transparency and accountability to users and the public regarding a platform's moderation practices. Additionally, as will be discussed in the Recommendations section, platforms should continue to make efforts to introduce the community guidelines to users to increase their norm-building effectiveness.

Although this paper's review focuses on social VR community guidelines, it does not mean that a review of other community safety practices is not needed. On the contrary, even the most comprehensive community guidelines are only effective if they are part of a platform's broader set of meaningful and effective community safety practices, spanning policies, products, and operations.

Community guidelines are only effective if they are:

- Accessible: Guidelines are easily found and understood by the end-user, and provide clarity regarding when they apply.
- Comprehensive: Delineates broad high-level *principles*, divided by mid-level *policies* that cover the various categories of online harms.
- Specific: Clarify low-level *playbooks* for each policy, giving detailed specifics as to what constitutes violations.
- Transparent: Provide rationales to the policies; explanation to the values that inform the guidelines; and a change log of any edits of the guidelines.

Underlying this analysis of Meta's community guidelines are the following questions:

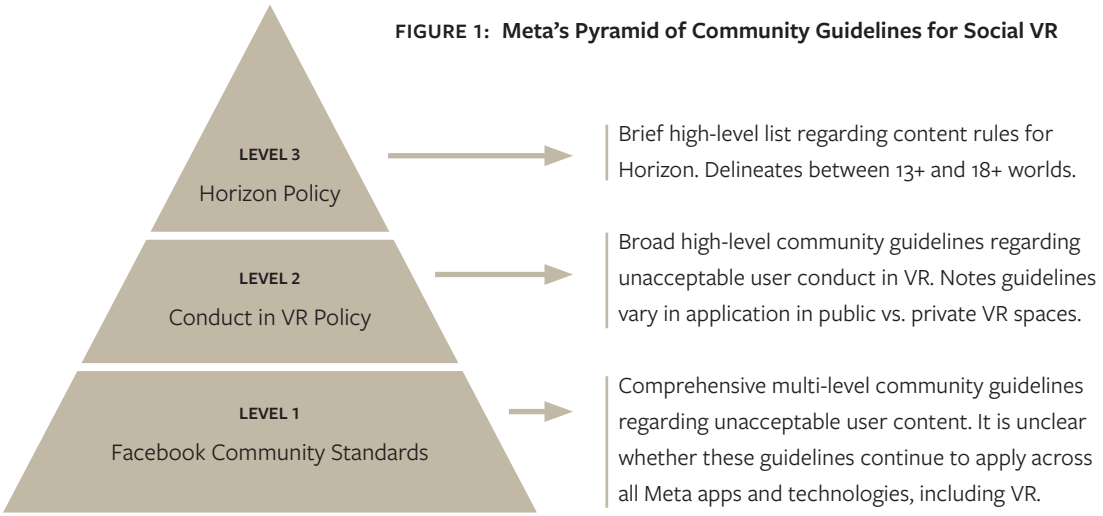
1. Are Meta's community guidelines for social VR accessible, comprehensive, specific, and transparent?
2. How has Meta developed community guidelines for social VR, considering the technology's unique enablement of *conduct*-based user interactions?
3. How do Meta's community guidelines for social VR compare with its community guidelines in 2D social media?
4. In August 2022, Meta updated its community guidelines related to social VR; how have the updates impacted the previous three questions?

An analysis of Meta's community guidelines that apply in social VR yields five primary findings:

1. There are two, perhaps three, community guidelines that apply in Meta's social VR:[42] *Horizon Policy*, *Conduct in VR Policy*, and *Facebook Community Standards*. August 2022 updates to the *Horizon Policy* and the *Conduct in VR Policy* suggest that the *Facebook Community Standards* no longer apply to VR. The ambiguity regarding where and how each policy applies weakens the accessibility of the policies.
2. If the *Facebook Community Standards* no longer apply to VR, that would be a significant setback for community safety, and for transparency and accountability in how Meta moderates its platforms. This is particularly so in the absence of comprehensive community guidelines for VR.
3. The *Conduct in VR Policy* lacks the level of specificity, and therefore effectiveness, of the *Facebook Community Standards*.
4. The *Conduct in VR Policy* lacks a change log, policy rationales, and explanation of values.

There is no single list of public-facing community guidelines for users to follow in Meta's social VR. There are at least two, and perhaps three: the *Horizon Policy*, the *Conduct in VR Policy*, and possibly the *Facebook Community Standards*. Figure 1 (below) shows the three layers of user-facing community guidelines.[43]

FIGURE 1: Meta's Pyramid of Community Guidelines for Social VR



42 As of August 2022.

43 The pyramid in Figure 1 depicts only the external community standards that are user-facing and norm-building. Meta has internal documents that give moderators more detail on how to interpret the user-facing community standards. Additionally, users of Meta's products have various terms of service and policies that must be adhered to.

Level 1, the *Facebook Community Standards* (often referred to by Meta as simply the *Community Standards*), is an extensive list of content-focused guidelines for what is (and is not) allowed on the Facebook social media platform. Prior to a July 2022 update of the *Horizon Policy*, and an August 2022 update of the *Conduct in VR Policy*, it was clear that the extensive *Facebook Community Standards* also applied to VR. The recent updates have left that ambiguous. (More information about this can be found in the next subsection.)

Level 2, the *Conduct in VR Policy*, provides a short list of conduct and content that may not be allowed in VR, depending on the context an individual is in (e.g., in public vs. private VR spaces).

Level 3, the *Horizon Policy,* notes five types of content that are prohibited across Horizon (e.g., promotion of illegal drugs), and three types of content that are allowed in worlds marked for mature audiences at least 18 years old (e.g., promotion of marijuana and tobacco).

TOO MUCH AMBIGUITY REGARDING WHICH AND WHERE META POLICIES APPLY IN VR

There is a significant amount of ambiguity regarding where each of the aforementioned Meta community guidelines applies in VR. As a result, the community guidelines are largely ineffective in being accessible to users.

Horizon Policy: There is some ambiguity about whether the *Horizon Policy* applies only to content created within the Horizon app or whether it includes content created in non-Horizon apps that can be accessed with a Meta VR headset. On the surface it would seem that the guidelines apply only to content or worlds created in the Horizon app. However, Meta recently launched the Meta Horizon profile,[44] a new social VR profile that is required for all users of a Meta VR headset. A user's Meta Horizon profile is active even when accessing a non-Horizon app via a Meta VR headset. As such, some ambiguity remains regarding whether the *Horizon Policy* also applies to the Horizon profile and not just the Horizon app. If it also applies to the Horizon profile, then even a user creating content in a non-Horizon app (but who is using their Meta Horizon profile) would be subject to the guidelines. The ambiguity regarding what Horizon products are included within the *Horizon Policy* leads to opacity for users about how and where Meta may moderate content. Additionally, the ambiguity diminishes accountability

44 "Introducing Meta Accounts and Meta Horizon Profiles for VR," Meta, July 7, 2022, https://about.fb.com/news/2022/07/meta-accounts-and-horizon-profiles-for-vr/.

for Meta, leaving it unclear to the public where in VR the company promises to apply the *Horizon Policy*.

Conduct in VR Policy: There is ambiguity about whether the *Conduct in VR Policy* applies only within Meta VR apps or whether it also applies within third-party VR apps that are accessed via a Meta VR headset. The August 2022 *Conduct in VR Policy* states: "Consistent with these values, developers, including Meta, may take action, such as limiting functionality or restricting features in their respective apps and products."[45] By stating apps *and* products, is Meta implying that the company intends to moderate conduct on first-party apps *and* on third-party apps that are accessed via the Meta headset? Similar questions arise from the following sentence in the *Conduct in VR Policy*: "Meta may also take action on our platform, such as suspending accounts." What does "our platform" mean in this sentence? Does it refer to Meta-created apps, Meta headsets, or perhaps to a broader "Meta," at the company-wide level, meaning that a user violating the *Conduct in VR Policy* within a VR world may also be penalized across Meta, for example having their Facebook account suspended? The ambiguity regarding where in VR the *Conduct in VR Policy* applies creates opacity for users about how and where Meta may moderate content. Additionally, the ambiguity diminishes accountability for Meta, leaving it unclear to the public where in VR Meta commits to moderate with the *Conduct in VR Policy*.

Facebook Community Standards: There is ambiguity about whether the *Facebook Community Standards* continue to apply in VR. Prior to a July 2022 update of the *Horizon Policy*, and an August 2022 update of the *Conduct in VR Policy*, it was clear that the *Facebook Community Standards* apply to VR. Both the *Horizon Policy* and the *Conduct in VR Policy* directly stated that the *Facebook Community Standards* apply.[46] For example, consider the following passage from the former *Conduct in VR Policy*:

> We want everyone to feel safe while they enjoy an immersive virtual experience. The [Facebook] Community Standards outline what is and is not allowed on Facebook[47] apps and technologies,[48] and apply to both content and conduct in VR. To help you

45 See Appendix C.

46 See Appendix B and Appendix D for the prior versions of the *Horizon Policy* and the *Conduct in VR Policy*, which reference the *Facebook Community Standards*.

47 In this context, it is apparent that Facebook refers to the company now called Meta rather than Facebook the social media platform.

48 The sweeping clarification that the *Facebook Community Standards* apply to all Meta apps and technologies is significant and arguably positive. It connotes that the *Facebook Community Standards* apply even on non-Meta social VR platforms if accessed through a Meta headset. This would suggest that Meta would treat its headsets, and associated VR app store, as something closer to an Apple App Store than a web browser, with community safety standards that apps must

better understand how the Community Standards apply to conduct in a virtual space, we highlighted areas of our policy and how they apply to conduct in VR.

The former *Conduct in VR Policy* thus appeared to serve primarily as a secondary clarification document, i.e., clarifying that the *Facebook Community Standards*, the primary document, applies to both content and conduct in VR. The former *Conduct in VR Policy* provided a few brief examples of how the *Facebook Community Standards*, written for 2D content, can be understood for VR conduct. Indeed, in the examples it provided, the *Conduct in VR Policy* continued to link back to the *Facebook Community Standards*.

However, the updated *Conduct in VR Policy* no longer references the *Facebook Community Standards* or links back to it. Instead of serving as a secondary clarification document, the updated *Conduct in VR Policy* appears to be a primary document, serving as *the* source rather than a clarification of the guidelines for conduct in VR. This leaves users and the public with ambiguity. Do the *Facebook Community Standards* still apply to VR? As will be noted in the next subsection, the *Facebook Community Standards* are extensive and specific; whether or not they apply in VR is of enormous consequence. If the *Facebook Community Standards* do apply in VR, do they apply only within Meta-created apps or do they also apply when using third-party apps accessed via a Meta VR headset? Of related significance, do the *Facebook Community Standards* no longer apply across Meta's "apps and technologies," as was declared in the former *Conduct in VR Policy*? In other words, is Meta limiting the coverage of the *Facebook Community Standards* to only a select few apps? If so, which apps are still covered by the *Facebook Community Standards*?

In summary, there is ambiguity about where in VR the *Horizon Policy* and the *Conduct in VR Policy* apply, and there is ambiguity about whether the *Facebook Community Standards* apply in VR at all, and if they do, which apps they apply to. Collectively, this ambiguity largely renders ineffective the five functions of community guidelines (discussed on page 14).

1. Users are not provided with a clear set of injunctive norms prescribed from Meta regarding what behaviors in VR are acceptable and unacceptable on Meta apps and technologies.
2. Users are not provided with a clear and transparent resource to seek remedy and moderation from Meta, including to appeal actions taken by Meta on their account.
3. The public is not provided with insight into Meta's internal policies for moderation in VR.

---

uphold. The removal of this line from the August 2022 *Conduct in VR Policy* leaves users with uncertainty about whether the *Facebook Community Standards* still apply across all Meta apps and technologies.

4. Meta has made only ambiguous public commitments regarding the content and conduct it allows and does not allow on its apps and technologies in VR.
5. The public has no clear guidelines to reference to hold Meta accountable for harmful behavior on its apps and technologies in VR.

APPLYING FACEBOOK COMMUNITY STANDARDS WAS A POSITIVE STEP FOR VR SAFETY

The former *Conduct in VR Policy*[49] clarified that the *Facebook Community Standards* applied to VR. Through this policy, Meta had made a significant contribution to the future of community safety practices in social VR. With the update to the *Conduct in VR Policy*,[50] and ambiguity about whether the *Facebook Community Standards* still apply in VR, that significant contribution is in doubt.

The *Facebook Community Standards* are among the most extensive community guidelines of any social media platform. They are thorough in the policies included, specific in how those policies are interpreted, and transparent in delineating the values and rationales behind the policies. Therefore, in carrying over all of the *Facebook Community Standards* to VR, Meta had established an expectation that they would moderate VR with at least the extensiveness and depth of their moderation of Facebook. Since Meta is dominant in the sale of its VR headsets, and is the largest current investor in the metaverse, establishing extensive community standards in its own VR technologies contributes to the development of a baseline for platform moderation policies across major social VR platforms and VR technology-makers. Conversely, by no longer applying the *Facebook Community Standards* to VR, Meta may be contributing to a regressed and diminished baseline for platform moderation transparency and accountability across social VR. Meta may also be setting a dangerous norm that platforms can suddenly reduce the comprehensiveness, transparency, and accountability of their community guidelines.

The *Facebook Community Standards* are specific and detailed, with information sorted into categories and sub-categories with increasing detail and transparency. The cascading detail in the *Facebook Community Standards* can be described as following the well-respected trust and safety framework of the "Three P's:" principles, policies, and playbooks.[51] The *Facebook*

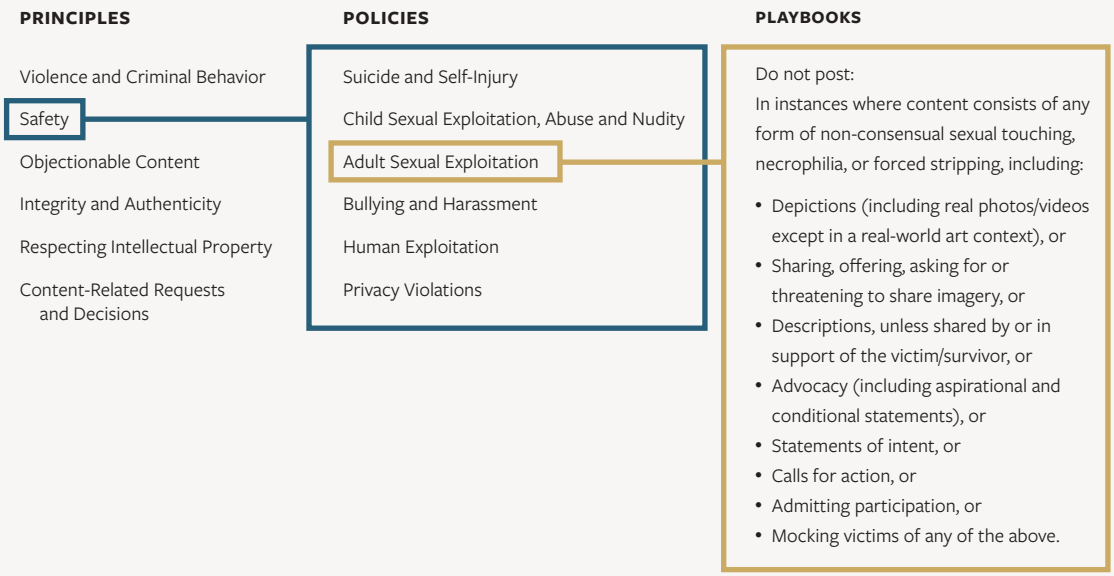
49 Meaning, the July 2022 version. See Appendix D for full text of July 2022 Conduct in VR Policy.

50 Meaning, the August 2022 version. See Appendix C for full text of August 2022 Conduct in VR Policy.

51 During a panel at the May 2022 Safety Matters Summit, Rey Allie described (not verbatim) the "Three P's" as: 1) Principles: at a macro level what the company/organization stands for (e.g., user expression); 2) Policies: the rules that

Community Standards include six high-level principles subdivided into 24 policies, each of which includes numerous playbooks, i.e., detailed and specific forms of content prohibited. Figure 2 shows how one principle, Safety, is subdivided into six policies. It further depicts how the policy, Adult Sexual Exploitation, is further broken down into a number of detailed playbooks.[52]

FIGURE 2: Example of *Facebook Community Standards* Cascading Specificity



Just as "Safety" is subdivided into a number of policies (as shown in Figure 2), each of the other principles in the *Facebook Community Standards* is subdivided into a number of policies.[53] Similarly, just as the policy "Adult Sexual Exploitation" includes detailed playbooks, other policies in the *Facebook Community Standards* also include detailed playbooks explaining in clear terms what content is prohibited on the Facebook platform. It is the specificity of playbooks that make the broader goals of principles and policies practical and effective. In applying the *Facebook Community Standards* to VR, Meta established that community guidelines in social VR should incorporate the baseline of specificity established in its 2D social media platform, Facebook.

---

provide more specificity to corresponding principles (e.g., no hate speech); 3) Playbooks: detailed specifications of how a policy is operationalized, and in unique contexts (e.g., in India, hate speech may include reference to caste). Source: Rey Allie, "Developing and Operationalizing Policy" (Panel, Safety Matters Summit, May 19, 2022).

52 The policy category, Adult Sexual Exploitation, includes other specifics that were not included in Figure 2 for the sake of space.

53 For the full list of *Facebook Community Standards* principles and policies, see Appendix E.

In addition to their level of detail, the *Facebook Community Standards* are also comprehensive in the number of harms they address. In 2021, the Partnership for Countering Influence Operations (PCIO) analyzed the community guidelines of Facebook and 12 other social media and messaging platforms,[54] and found Facebook to have the second most extensive community guidelines (in word count), following only Twitter. Since not all platforms categorize online harms the same way, PCIO utilized a coding system of 12 categories of online harms to assess how many each platform included in their community guidelines. The analysis found that only Facebook and Twitter have guidelines related to every category (see Figure 3).[55]

FIGURE 3: Existence of Platform Policies by Category

Policy category	Facebook	Gab	Instagram	LinkedIn	Pinterest	Reddit	Signal	Telegram	TikTok	Tumblr	Twitter	WhatsApp	YouTube
Authenticity and spam	✓	✓	✓	✓	✓	✓	✓	•	✓	✓	✓	✓	✓
Prohibited or restricted goods	✓	✓	✓	✓	✓	✓	•	✓	✓	✓	✓	•	✓
Trademark and copyright	✓	✓	✓	✓	✓	•	✓	✓	✓	✓	✓	✓	•
Harassment, threats, and discrimination	✓	✓	✓	✓	✓	✓	•	•	✓	✓	✓	✓	✓
Violent or sexual content	✓	✓	✓	✓	✓	✓	•	•	✓	✓	✓	✓	✓
Site use	✓	•	•	✓	✓	✓	✓	•	✓	✓	✓	✓	✓
Terrorist and criminal organizations	✓	•	✓	✓	✓	•	•	✓	✓	✓	✓	•	✓
Public health	✓	•	✓	•	✓	•	•	•	✓	✓	✓	•	✓
Child sexual abuse material	✓	•	•	•	✓	✓	•	•	✓	✓	✓	•	✓
Child safety	✓	✓	•	✓	•	•	•	•	✓	✓	✓	•	✓
Civic integrity	✓	•	•	✓	✓	•	•	•	•	✓	✓	•	✓
Deceased individuals	✓	•	•	•	•	•	•	•	•	•	✓	•	•

✓ **Policy** • **No Policy**

54 The other 12 include: Gab, Instagram, LinkedIn, Pinterest, Reddit, Signal, Telegram, TikTok, Tumblr, Twitter, WhatsApp, and YouTube.

55 The chart in Figure 3 was created by PCIO and visually adapted for this paper. Source: Jon Bateman, Natalie Thompson, and Victoria Smith, "How Social Media Platforms' Community Standards Address Influence Operations," Carnegie Endowment for International Peace, April 1, 2021, https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-

Of significance to transparency, the *Facebook Community Standards* provide a detailed explanation for each of the values used to create the community guidelines — authenticity, safety, privacy, and dignity — as well as a rationale for each policy level.[56] Additionally, the *Facebook Community Standards* also include a change log, or a history of edits made to the guidelines, further providing transparency into how the guidelines were developed. Collectively, these elements (accessibility, comprehensiveness, specificity, and transparency) make the *Facebook Community Standards* effective.

AMBIGUITY ABOUT WHETHER FACEBOOK COMMUNITY STANDARDS APPLY TO VR LEADS TO A SIGNIFICANT LOSS OF USER SAFETY

In clarifying that the *Facebook Community Standards* applied to all content and conduct in VR, Meta had established a baseline level of comprehensiveness and specificity for community guidelines in social VR, while the explained values, policy rationales, and change log of community guidelines set a strong baseline for transparency. Those expectations contributed to setting a strong standard for robust community guidelines across the nascent social VR industry.

With the update to the *Conduct in VR Policy,* that contribution is in doubt. In removing the clarification that the robust *Facebook Community Standards* apply to VR, Meta has regressed and diminished the baseline standards for platform moderation transparency and accountability across social VR. The former version of the *Conduct in VR Policy* offered a commitment from Meta to the public that it intends to include transparent and comprehensive community safety guidelines in VR. The current updated version of the *Conduct in VR Policy* puts that commitment in doubt, instead signaling a reversal in robust community guidelines.

---

standards-address-influence-operations-pub-84201. It is worth noting that since the *Facebook Community Standards* also apply to Instagram, Instagram's guidelines would have at least as many categories as the Facebook app does.

56 For example, the Policy Rationale for the policy category Adult Sexual Exploitation: "We recognize the importance of Facebook as a place to discuss and draw attention to sexual violence and exploitation. In an effort to create space for this conversation and promote a safe environment, we allow victims to share their experiences, but remove content that depicts, threatens or promotes sexual violence, sexual assault, or sexual exploitation. We also remove content that displays, advocates for or coordinates sexual acts with non-consenting parties to avoid facilitating non-consensual sexual acts. To protect victims and survivors, we remove images that depict incidents of sexual violence and intimate images shared without the consent of the person(s) pictured. As noted in the introduction, we also work with external safety experts to discuss and improve our policies and enforcement around online safety issues, and we may remove content when they provide information that content is linked to harmful activity. We have written about the technology we use to protect against intimate images and the research that has informed our work. We've also put together a guide to reporting and removing intimate images shared without your consent." Source: "Adult Sexual Exploitation," Meta Transparency Center, December 24, 2021, https://transparency.fb.com/policies/community-standards/sexual-exploitation-adults/.

If in fact Meta has reduced the coverage of the *Facebook Community Standards*, and those standards no longer apply to all Meta apps and technologies, it would be a significant reduction in transparency and commitment regarding what user content and behavior Meta moderates.

WITHOUT COMPREHENSIVE PLAYBOOKS, CONDUCT IN VR POLICY IS TOO BROAD TO BE EFFECTIVE

The *Facebook Community Standards* for content moderation on the Facebook platform are effective due in large part to their breadth and specificity. For example, the "Bullying and Harassment" policy in the *Facebook Community Standards* details over 60 playbooks, specific forms of bullying and harassment content that are prohibited on Facebook. This is in stark contrast to the *Conduct in VR Policy*, which gives only a few vague playbooks regarding what constitutes bullying and harassment conduct in social VR. For comparison, the *Facebook Community Standards* devotes over 1,700 words to detail what constitutes content-based bullying and harassment on the Facebook platform;[57] the *Conduct in VR Policy* provides less than 150 words to clarify what the platform considers conduct-based bullying and harassment. It is certainly not the case that there are fewer potential forms of bullying and harassment in conduct-interactions than content-interactions, nor is it the case that those harms are not already occurring in social VR, as discussed in Section 1 of this paper.

If detailed playbooks are not important elements of community guidelines, then why do the *Facebook Community Standards* include playbooks? If detailed playbooks are important elements of community guidelines, then aren't they necessary for users to understand what Meta considers to be conduct-based harms in social VR?

Indeed, specific playbooks delineated in the *Facebook Community Standards* have been hard fought for by the public and by civil society groups, precisely because their delineation is what makes such guidelines effective in ensuring user safety. For example, for years, Facebook pushed back on removing Holocaust denial posts from the platform. Finally, in October 2020, after years of public pressure, including the viral #StopHateForProfit campaign,[58] Facebook shifted its position. In a post, CEO Mark Zuckerberg attributed "data showing an increase in anti-Semitic violence" as evolving his point of view on "the tension between standing for free

57 See Appendix G for the full *Facebook Community Standards* Bullying and Harassment policy.

58 "One Year After Stop Hate for Profit: Platforms' Progress."

expression and the harm caused by minimizing or denying the horror of the Holocaust."[59] Facebook began banning Holocaust denial on the platform, finally including the position as a playbook-level rule among a number of other forms of hate speech in its *Facebook Community Standards*.[60] This example shows how the presence of just one small playbook within a much broader list of community guidelines can make a significant difference in establishing whether a platform's community guidelines are effective or not. As the case of Holocaust denial shows, if a rule is not listed clearly in community guidelines, it can be presumed that the platform is not taking meaningful action regarding that particular harm.

Unlike in the *Facebook Community Standards*, no such level of detail regarding how hate speech applies to content and conduct in VR is provided in the *Conduct in VR Policy*. In other words, no playbooks are provided regarding hate speech in VR. It may seem intuitive that Meta's policy for Holocaust denial posts in the *Facebook Community Standards* would mean that Holocaust denial shared through synchronous speech or other content in social VR is prohibited. However, without the playbook clearly delineated in the *Conduct in VR Policy,* there is inevitably ambiguity for users and opacity from the platform. Other forms of hate conduct may be even less intuitive to translate from the *Facebook Community Standards* as they are content-based policies. Is a Nazi salute in social VR prohibited? What about other hate gestures?[61] The *Conduct in VR Policy* only specifies sexual gestures as prohibited. Without specifically stating that hate gestures are also prohibited, there is ambiguity for the user. If even a small amount of ambiguity remains in what constitutes hate conduct, then the community guidelines in social VR are ineffective. Holocaust denial and Nazi salutes are just two examples of the general ambiguity remaining in what is allowed in Meta VR apps and technologies. This is a regression from the relative level of clarity that had been reached in Meta 2D apps, such as Facebook and Instagram. Without specifically detailing principles, policies, and playbooks, the *Conduct in VR Policy* remains far too nebulous to be useful in setting norms, enabling users to seek moderation, and providing transparency to users and the public for how content and conduct are moderated in social VR.

This ambiguity was also a problem in the former *Conduct in VR Policy*. Even though the former version stated that the extensive *Facebook Community Standards* applied to conduct in VR, it did not provide meaningful specificity or playbooks for *how* the *Facebook Community*

59 Mark Zuckerberg, "Today We're Updating Our Hate Speech Policy to Ban Holocaust Denial.," Post, *Facebook*, October 12, 2020, https://www.facebook.com/zuck/posts/10112455086578451.

60 See Appendix F

61 For example, Resolution Games, a gaming studio, explicitly prohibits offensive gestures. See "Player Code of Conduct," Resolution Games, April 16, 2021, https://www.resolutiongames.com/code-of-conduct.

Standards, written for 2D content, would apply to conduct in VR.[62] If the *Facebook Community Standards* do still apply to VR, it remains unclear how they apply to the unique experience of VR. That said, the former *Conduct in VR Policy* did at least show a commitment from Meta that it intends to apply comprehensive community guidelines to social VR, even if their application required additional explanation.

In addition to the ambiguity surrounding what conduct is prohibited in VR, there is also ambiguity regarding how the community guidelines apply to forms of content unique to social VR, such as user creation of a virtual environment or object. In February 2022, reporters at Buzzfeed wrote about how they tested Meta's content moderation in social VR by building a private Horizon World, which they called "Qniverse," rife with audio and visual language that promoted overt election and vaccine misinformation, as well as support for QAnon.[63] After 36 hours passed and the Qniverse world remained on the platform, the BuzzFeed reporters submitted reports to Meta. Despite Qniverse's inclusion of language that clearly violated the *Facebook Community Standards*, such as "vaccines cause autism," Meta's response was that the world did not violate the *Content in VR Policy:* "Our trained safety specialist reviewed your report and determined that the content in the Qniverse doesn't violate our Content in VR Policy." Only after the Buzzfeed editors reached out directly to Meta's communications department was the Qniverse world removed.

The BuzzFeed case reveals that even some moderators at Meta are unclear about what content is prohibited in social VR. It also shows how they were unclear in how to interpret the former *Conduct in VR Policy*, which stated that the *Facebook Community Standards* apply to content and conduct in social VR. Even a blatant violation of the *Facebook Community Standards* policies regarding vaccine misinformation was found by Meta moderators to not be a violation of the *Conduct in VR Policy*.

THE CONDUCT IN VR POLICY IS UNCLEAR IN ITS DELINEATION OF PUBLIC VS. PRIVATE VR

The updated *Conduct in VR Policy* introduces a fundamental change to Meta's moderation policy: a delineation between public and private spaces. As the policy states:

62 For example, how does the policy around hate speech in the *Facebook Community Standards* apply to conduct in VR?

63 Emily Baker-White, "Meta Wouldn't Tell Us How It Enforces Its Rules In VR, So We Ran A Test To Find Out," BuzzFeed News, February 11, 2022, https://www.buzzfeednews.com/article/emilybakerwhite/meta-facebook-horizon-vr-content-rules-test.

> Developers and Meta may consider the context, including if certain behaviors took place in a public space, to determine what action is appropriate. Meta may also take action on our platform, such as suspending accounts.

What precisely constitutes a public space? For the user, this delineation is unclear. As a result, a user cannot know for certain when they can report behavior to be moderated. Similarly, a user who has been actioned by Meta is left without clear means to appeal the action because the policy is unclear.

Similarly, another portion of the *Conduct in VR Policy* discusses public and private spaces without providing clarity on their definitions:

> A lot depends on the context; speech and behavior that is acceptable in private or among friends may not be acceptable in more public spaces. But it's helpful for people to understand behaviors that don't align with our values in virtual spaces where you are around other groups.

What is a *more* public space? What does it mean to be around other groups—within earshot of other groups or perhaps when in direct interaction with other groups? Is a private space only a world that a user has created themself? Is a private space determined by the number of individuals in the space? If so, the number that delineates private from public needs to be clarified by Meta.

Additionally, does Meta's delineation between public and private mean that the community guidelines do not apply to a world that a user builds and invites only their friends to? Would the Buzzfeed Qniverse be acceptable under this delineation because individuals had to be invited by the world creator to enter it? If a user creates a private world devoted to extremist and racist ideology, is that no longer a violation of Meta's community guidelines under the updated *Conduct in VR Policy*?

How can a user be certain of whether they are in a private or public space? This ambiguity over what constitutes a public and private space is of severe consequence. It reduces transparency regarding where community guidelines apply, limiting users' ability to confidently report violations and appeal moderation.

THE CONDUCT IN VR POLICY LACKS TRANSPARENCY ELEMENTS THAT ARE IN THE FACEBOOK COMMUNITY STANDARDS

Whereas the *Facebook Community Standards* include three transparency elements — explanation of values, rationales for policies, and change log of the community guidelines — the *Conduct in VR Policy* does not. This leaves users, researchers, and the broader public with minimal insight into how Meta is deciding its guidelines regarding conduct, as well as how Meta has changed those guidelines.

It may be argued that the explanation of values and the rationales for policies in the *Facebook Community Standards* carry over to the *Conduct in VR Policy*. However, this would only be the case if the *Facebook Community Standards* apply to social VR, which remains ambiguous after the August update to the *Conduct in VR Policy*. Additionally, even if the *Facebook Community Standards* continue to apply in social VR, there should be some reimagination of explanations and rationales, considering the qualitative difference between user interactions in 2D Facebook and social VR.

SUMMARY OF FINDINGS

The review of Meta's community guidelines relevant to social VR finds four need-based opportunities for growth:

1. **Need for increased accessibility:** There are two, perhaps three, community guidelines that apply in Meta's social VR: *Horizon Policy*, *Conduct in VR Policy*, and *Facebook Community Standards*. There is an unclear relationship between the three community guidelines and it is unclear where each applies.
2. **Need for increased comprehensiveness:** The *Horizon Policy* and the *Conduct in VR Policy* do not thoroughly delineate principles or break them down by policies. As a result, the *Horizon Policy* and the *Conduct in VR Policy* do not address the breadth of online harms.
3. **Need for increased specificity:** a) The *Horizon Policy* and the *Conduct in VR Policy* do not provide playbooks, leaving ambiguity about what Meta considers violations; b) The *Conduct in VR Policy* does not provide enough clarity about how Meta distinguishes public from private VR spaces.
4. **Need for increased transparency:** The *Horizon Policy* and the *Conduct in VR Policy* lack a change log of edits, policy rationales, and explanation of values.

The analysis also finds that clarity is needed on whether the *Facebook Community Standards* continue to inform the *Conduct in VR Policy* and apply to social VR.

- August 2022 edits to the *Conduct in VR Policy* removed a statement that the *Facebook Community Standards* apply to VR.
- Applying the *Facebook Community Standards* to social VR set a positive expectation that the thoroughness of 2D community guidelines would be maintained and implemented in VR.
- If the *Facebook Community Standards* no longer apply to VR, there is increased urgency to make the *Conduct in VR Policy* and *Horizon Policy* more thorough and effective.

Recommendations

SOCIAL VR COMMUNITY GUIDELINES SHOULD BE ACCESSIBLE

To be effective, community guidelines must be easily found and understood by the end user. Meta has a need-based opportunity to clarify whether the *Facebook Community Standards* continue to apply in VR, and explain in clear language to which apps and technologies in VR the *Horizon Policy* and *Conduct in VR Policy* apply.

SOCIAL VR COMMUNITY GUIDELINES SHOULD BE SPECIFIC AND COMPREHENSIVE

Effective community guidelines do not merely contain broad statements barring harms on the platform. For example, simply prohibiting "hate," without clarifying what the platform considers "hate" or without providing playbooks for the policy, leaves too much ambiguity for the guidelines to be effective. Effective community guidelines include specific examples, such as how the *Facebook Community Standards* delineates more than 50 specific forms of hate speech content on the Facebook platform. (For the complete policy, see Appendix F: *Facebook Community Standards* Hate Speech Policy.)

The level of specificity of the *Facebook Community Standards* is what makes them effective; there is minimal ambiguity for the user about what constitutes hate or harassment on the Facebook platform. The *Facebook Community Standards* were also clearly written for the platform, which makes their specific policies relevant and effective for delineating acceptable user interactions. The *Facebook Community Standards* do have room for improvement, but they have consistently matured and improved in providing transparency and clarity for Facebook users.

The same cannot be said regarding users of Meta's social VR. Currently, Meta only provides broad ambiguous policies regarding what constitutes unacceptable user behavior in social VR. Additionally, as a result of updates to the *Conduct in VR Policy* and the *Horizon Policy*, there is ambiguity regarding whether the *Facebook Community Standards* apply to VR. Meta should clarify whether or not they do apply to VR.

If the *Facebook Community Standards* do apply to content and conduct in social VR, Meta should provide playbooks for each policy in the *Facebook Community Standards*, detailing how the policy specifically applies to the unique forms of conduct and content in immersive VR. As detailed, the experience and use of social VR is drastically different from that of the Facebook platform. Social VR is avatar-embodied and world-immersive, and user interactions are primarily through conduct (synchronous audio and avatar movement).

If Meta is in fact applying the *Facebook Community Standards* to social VR, then detailed specifics are needed regarding how each policy of the *Facebook Community Standards* can be interpreted for content and conduct unique to social VR. This means adding playbooks to the community guidelines for how each policy applies to synchronous audio, avatar gestures and movements, as well as to unique forms of content in social VR, e.g., environment or object creation. The BuzzFeed story illuminates that the application of the *Facebook Community Standards* to social VR requires specific translation to the unique forms of conduct and content in social VR.

If Meta is not in fact applying the *Facebook Community Standards* to social VR, then the company should create comprehensive community guidelines for social VR, complete with principles, policies, and playbooks, at least at the level of detail of the *Facebook Community Standards*.

This paper advocates for Meta's application of the *Facebook Community Standards* to social VR, with the addition of detailed specifics regarding how each policy of the *Facebook Community Standards* is to be interpreted for content and conduct unique to social VR. Since Meta intends that users will increasingly operate in the metaverse, switching seamlessly between apps (e.g., Facebook to Horizon and third-party apps), an overarching set of principles and policies should apply across Meta's apps and technologies. This standardization would contribute to the interoperability of apps and provide users with a shared level of content and conduct expectations across apps. However, as noted, Meta would need to develop comprehensive clarifying playbooks for how each policy applies to conduct in each environment.

This call for specificity is aligned with expert recommendations regarding moderation policies across different types of platforms. The Oversight Board — an independent body with the ability to make certain content moderation decisions regarding Facebook and Instagram — has reiterated its call for Meta to provide users with transparency, stating in a recent LinkedIn post,

"tell people as clearly as possible how you make and enforce your decisions."[64] Similarly, the Anti-Defamation League (ADL), in a review of harassment in online gaming that is applicable to this analysis of social VR, wrote:[65]

> Strengthen policies and enforce terms of service. Many companies that created the games included in this survey have Codes of Conduct or Terms of Use that prohibit hate or harassment, but they are vague in the ways they describe protected groups and violative conduct. We recommend companies specify protected categories (including gender, gender identity, race/ethnicity, religion, sexual orientation, ability status) in their Terms of Use and explicitly prohibit doxing and swatting. In developing these Terms of Use, we advise games companies to consult with individuals and organizations representing groups that experience high rates of harassment.

Similarly, the Oasis Consortium — a trust and safety think tank — calls for clarity in its standards for user safety: "Set clear expectations for all users, starting with onboarding. A fine-print legal policy with a checkbox is not enough. Feature community guidelines where all users will see them, at key moments and repeatedly over time. Ensure that they are practical to enforce, and regularly updated."[66]

Brittan Heller, a leading expert on XR safety and moderation, discussed the need for companies to not simply transfer community guidelines from social media to social VR, saying: "When I advise AR and VR companies, the one thing I wish they remembered is that this is not social media. You cannot just take community standards or codes of conduct from a 2D content and conduct code and transfer it into a 3D environment, where you're going to need content, conduct, and environment."[67] Put in another way, simply transferring guidelines from 2D social media content to social VR conduct, without providing playbooks specific to social VR, could be comparable to applying rules about what a newspaper allows in its op-eds to rules about behavior acceptable at a theme park or a conference; the environments and modes of interaction are categorically different.

64 Oversight Board, "In Our Work, We Often See Users Left Guessing about Why Meta Removed Their Content," *LinkedIn*, accessed June 19, 2022, https://www.linkedin.com/posts/oversight-board-administration_what-is-the-impact-of-the-oversight-board-activity-6940284404220985344-w5iX/.

65 "Hate Is No Game: Harassment and Positive Social Experiences in Online Games 2021," ADL, accessed October 4, 2022, https://www.adl.org/hateisnogame.

66 Oasis Consortium, "User Safety Standards for Our Digital Future," 2021, https://www.oasisconsortium.com/usersafetystandards.

67 Atlantic Council, "What Happens When Toxic Online Behavior Enters the Metaverse?" Atlantic Council, June 6, 2022, https://www.atlanticcouncil.org/news/transcripts/what-happens-when-toxic-online-behavior-enters-the-metaverse/.

SOCIAL VR COMMUNITY GUIDELINES SHOULD MAINTAIN BASELINE LEVEL OF THOROUGHNESS OF 2D SOCIAL MEDIA COMMUNITY GUIDELINES

In order to be effective, social VR community guidelines need to be at least as detailed,[68] comprehensive, and platform-tailored as those now offered by Facebook. Traditional social media platforms have received a great deal of scrutiny from the public, and their community guidelines have evolved to their current state in part based on that scrutiny. As such, Meta should not compare their social VR community guidelines to other social VR platforms, which have not yet faced public scrutiny, and whose community guidelines may still be underdeveloped. The level of detail used in 2D social media community guidelines should be the barometer. Considering that conduct interactions may be even more nuanced and contextual than content interactions, there can be even greater ambiguity for users about what constitutes a violation of a broad policy in social VR. As such, detail and specificity are even more needed in the community guidelines of social VR.[69] The *Facebook Community Standards* goes out of its way to provide nuance regarding how various content-based harms are interpreted based on context;[70] similar elucidations are needed for conduct-based harms in social VR.

YouTube — a platform where context matters greatly due to the varying degree of voice, gesture, and visual content in a video — provides users with comprehensive community guidelines, complete with principles, policies, and playbooks. Additionally, due to the complex context of each case, YouTube clarifies numerous exceptions to its rules, and even provides video tutorials to help users understand the contextual application of each policy.[71] YouTube's community guidelines show that delineating specific playbooks for complex content and behavior contexts is possible and helps create a safer platform.

68 At least in their website form. In the in-VR app form, it is important to offer a shorter list that can reasonably be read by users, while noting that a more detailed list exists to be reviewed on a website.

69 Indeed, Microsoft's AltspaceVr community guidelines provide an example of how harassment is assessed: "If another community member expresses that something makes them uncomfortable, you must cease that behavior in their presence. Continued harassing behavior will result in an account suspension and subsequent determination of whether the account will be closed permanently."

Of course Microsoft still needs to provide specifics of harassment at the level of detail of Facebook's community guidelines, however it does show that a platform can use the community guidelines to clarify an ambiguity in conduct. It is, however, disturbing that Microsoft's clarification of what constitutes harassment is when someone asks for the behavior to stop. This puts the onus on victims, and enables perpetrators to get away with at least one harmful action before it is called out by the victim. Microsoft makes a similar unfortunate clarification of what constitutes "lewd or unwanted advances." See "Community Standards," Microsoft, January 20, 2022, https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/community/community-standards (accessed June 1, 2022).

70 See Appendix G. Meta provides numerous context caveats or clarifications in its Bullying and Harassment policy.

71 See Appendix H for an example of a YouTube policy, the *Harassment & cyberbullying policies.*

Upon releasing the first detailed *Facebook Community Standards* in 2018, Monika Bickert, then Vice President of Global Policy Management, told reporters, "You should, when you come to Facebook, understand where we draw these lines and what's OK and what's not OK."[72] Similarly, Bickert wrote:

> We decided to publish these internal guidelines for two reasons. First, the guidelines will help people understand where we draw the line on nuanced issues. Second, providing these details makes it easier for everyone, including experts in different fields, to give us feedback so that we can improve the guidelines – and the decisions we make – over time.[73]

Facebook made clear that the public release of detailed *Facebook Community Standards* was precisely to provide users with clarity on nuanced issues and to give transparency to users about where it draws lines. Users in social VR similarly deserve to understand where Meta draws the lines about "what's OK, and what's not OK." If interactions in social VR are more nuanced than in 2D social media, then Meta should provide more context and clarity for users in its community guidelines, not less, as the *Conduct in VR Policy* does. The greater the ambiguity, the greater the need for clarification. Additionally, as Bickert noted, the public release of detailed guidelines is crucial to enabling feedback from the public to improve the guidelines. Just as that transparent dialogue between the company and the public was needed for Facebook, so too it is needed for social VR.

Presumably, Meta has internal documentation that guides moderators of social VR in how to interpret nuanced context for specific user reports. In other words, if Meta has the ability to provide playbooks to its moderators regarding how to interpret the broad policies in the *Conduct in VR Policy* and how to assess nuanced user interactions in VR, then Meta has the ability to provide playbooks to its users. This follows the reasoning behind the 2018 public release of detailed *Facebook Community Standards*, which themselves were internal guidelines prior to being released to the public.[74] Likewise, Meta would have documentation for its moderators in how to delineate public and private spaces. Similarly, if the *Facebook Community Standards* continue to apply in VR, then presumably Meta has internal documentation that guides mod-

72 David Ingram, "Facebook Releases Long-Secret Rules on How It Polices the Service," Reuters, April 24, 2018, https://www.reuters.com/article/us-facebook-abuse/facebook-releases-long-secret-rules-on-how-it-polices-the-service-idUSKBN1HV0VR.

73 Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process," Meta, April 24, 2018, https://about.fb.com/news/2018/04/comprehensive-community-standards/.

74 Ibid.

erators in how to interpret the *Facebook Community Standards* for social VR. If so, then Meta should release a version of its internal guidelines for social VR to the public.

If Meta does not have internal guidelines for how to interpret the *Conduct in VR Policy* and *Facebook Community Standards* for social VR, that would underscore the urgent need to provide clarity about what user interactions are acceptable or unacceptable in social VR, both for moderators and for users.

SOCIAL VR PLATFORMS SHOULD PARTNER ACROSS SECTORS TO ASSESS HARMS AND PREPARE FOR FUTURE THREATS

Each social VR platform should assess the various forms of conduct and content that currently occur (and may occur) on their platform, including prevalent harms (bullying, hate speech, etc.) and other harms that may not be as prevalent but are very high risk (child safety, suicide, etc.). The platform's community guidelines, including principles, policies, and playbooks, should reflect those findings. Older social VR platforms with robust user reporting tools may have accrued significant internal data to identify many of the specific forms of harms that are already occurring. However, to ensure that no forms of harm are overlooked, all social VR platforms, both old and new, should supplement any internal data with external sources. Platforms should:

- Assess user reviews of platforms on VR app stores to identify forms of online harms occurring. Platforms should look both to reviews of their own platforms as well as those of similarly themed social VR platforms.
- Conduct surveys and interviews with users to identify forms of online harms occurring.
- Conduct surveys and interviews with users to assess whether there is clarity in the community guidelines.
- Consult with outside experts who study harms in social VR, such as XRSI[75] and XR Access.[76]
- Review research regarding harms in non-VR online gaming. As previously noted, non-VR online gaming platforms often utilize synchronous, conduct-based user interactions. Social VR platforms can consult that literature to identify content- and conduct-based harms that are likely to occur in social VR, albeit in a more immersive way.

75 Website: XRSI.org
76 Website: xraccess.org

- Consult with civil-society organizations and other stakeholders that have expertise in harms faced by particular marginalized communities, and investigate how those harms may be expressed in social VR.[77] For example, in 2020, Facebook added a number of policies to its community standards, including specific forms of antisemitic hate speech, after consulting with outside experts.[78] In the same period, Facebook also added a number of playbooks regarding other forms of hatred and bigotry based on how each uniquely manifests on Facebook (see Appendix B). Similarly, social VR platforms should consult with outside experts to design community guidelines that recognize how various forms of hate manifest in social VR, before exclusionary norms are established by bad actors.
- Good policies and policy implementation include effective remedy and appeal processes for aggrieved parties. Remedy is an important part of the UN Guiding Principles on Business and Human Rights.[79] Platforms should make it clear to users how they can find remedy, including how to appeal a disciplinary action taken by the platform.
- In a crisis, the ability to take action for violations of community guidelines at scale is significant, as is the ability to adjust policies.[80] Social VR platforms should have a response plan in place for future crises that may arise on their burgeoning platforms.
- Platforms should continue to make efforts to introduce the community guidelines to users in creative ways throughout their experience on the platform in order to increase user proficiency of the guidelines and make the guidelines more effective at rule-setting. The immersive characteristic of VR expands the creative opportunities to teach users about the community guidelines. If a platform chooses to delineate between public and private spaces in regards to the application of community guidelines, then the platform must clarify:

    1. What exactly constitutes a public space;[81]
    2. What exactly constitutes a private space;
    3. What guidelines specifically apply in a public space; and
    4. What guidelines specifically apply in a private space.

77 This aligns with ADL's aforementioned recommendation to gaming platforms, "In developing these Terms of Use, we advise games companies to consult with individuals and organizations representing groups that experience high rates of harassment." Source: "Hate Is No Game."

78 Monika Bickert, "Removing Holocaust Denial Content," Meta, October 12, 2020, https://about.fb.com/news/2020/10/removing-holocaust-denial-content/.

79 "Guiding Principles on Business and Human Rights" (United Nations Human Rights, 2008), https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

80 Brian Fishman, "Live-Streaming of Terrorist and Violent Extremist Content: Moderation and Crisis Response" (Panel, TAT-GIFCT webinar, June 23, 2022).

81 Platforms may consider referring to pages 23–26 in *Ethical Approaches to Closed Messaging Research: Considerations in Democratic Contexts* for a parallel discussion regarding how to define public vs private conversations in closed messaging applications.

SOCIAL VR COMMUNITY GUIDELINES SHOULD BE TRANSPARENT, WITH A CHANGE LOG, AN EXPLANATION OF VALUES, AND POLICY RATIONALES

As discussed in Section 2, Meta's *Conduct in VR Policy* lacks explanation of values, policy rationales, and edit histories — three key transparency elements that are included in the *Facebook Community Standards*. Meta should add an easily accessible tab to the *Horizon Policy* and the *Conduct in VR Policy* so that the public can find past versions of the community guidelines. In writing community guidelines, all social VR platforms should incorporate the values that stand behind their guidelines, the rationales behind each policy, as well as the history and corresponding dates of how and when the guidelines have been edited.

FUTURE RESEARCH

Outside researchers and policymakers should collaborate with social VR platforms, and serve as independent reviewers, in efforts to develop comprehensive community guidelines on social VR platforms.

Researchers could:

- Identify forms of conduct and content that are unique to social VR and should be included in community guidelines. For example:
  - Assess user reviews of platforms on VR app stores to identify forms of online harms occurring.
  - Conduct surveys and interviews with VR users to identify forms of online harms occurring.
  - Conduct surveys and interviews with VR users to assess whether there is clarity in the community guidelines.
- Create a scale or rubric for examining the incorporation of VR-centric policies and play-books into community guidelines.[82]
- Conduct regular surveys regarding hate and harassment in social VR, assessing forms of harms and frequently targeted victims. Consistent surveying could also provide important insight into which platforms are reducing experiences of hate and harassment compared with prior surveys.

82 Researchers may consider drawing inspiration for the rubric from the Election Integrity Partnership report, *The Long Fuse*, which includes an evaluation of election related platform policies (ch. 6). Carnegie Endowment for International Peace's *How Social Media Platforms' Community Standards Address Influence Operations* may also provide ideas for how to assess a policy's comprehensiveness.

- Conduct regular surveys that assess positive, inclusive, and safe user experiences in social VR.[83] This can provide important insight into opportunities that can be tested and adopted by other VR platforms.
- Consider how alternative forms of moderation, such as decentralized volunteer moderation, are currently or may be applied in social VR. What are the pros and cons of decentralized vs. centralized moderation in social VR?
- Consider how the use of avatars and synchronous voice in social VR could introduce new manifestations of unconscious bias in moderation enforcement decisions.
- Users in social VR will increasingly interact with digital beings, i.e., interactive, AI-generated characters. Research shows that humans quickly develop emotional connections with digital beings.[84] This suggests that harms caused by interacting with a digital being may be experienced similarly to harms from a human-embodied avatar. Some digital beings may look just like human-embodied avatars. How do community guidelines and safety practices in social VR need to account for AI-generated digital beings? Do AI-generated digital beings need to be coded to follow community guidelines? Should digital beings be required to have identifiers that differentiate them from human-embodied avatars?[85]
- Considering how diverse social VR platforms and applications may be from each other (e.g., gaming, education, work), as well as expectations for user interoperability, can there be a baseline expectation for some community guidelines that apply across all social VR platforms?

Outside researchers, civil society organizations, and policymakers have an important opportunity to apply pressure on social VR platforms to be transparent in their community safety policies, products, and operations. It is precisely now, at this early stage of social VR adoption and development of the metaverse, where clear, transparent norms must be established.

83 This can follow methods similar to those used in ADL's survey of positive and negative user experiences in online gaming.

84 Linda Ricci, "AI and Digital Beings: Exploring the Intersection of Artificial Intelligence and XR" (Augmented World Expo, Santa Clara, California, June 2, 2022).

85 In the European Union, for example, the proposed EU AI Act seeks to codify previous guidelines that call for AI systems to identify themselves as non-human.

References

Meta Transparency Center. "Adult Sexual Exploitation," December 24, 2021. https://transparency.fb.com/policies/community-standards/sexual-exploitation-adults/.

Allie, Rey. "Developing and Operationalizing Policy." Panel, Safety Matters Summit, May 19, 2022.

IDC: "AR/VR Headset Shipments Grew Dramatically in 2021, Thanks Largely to Meta's Strong Quest 2 Volumes, with Growth Forecast to Continue, According to IDC," March 21, 2022. https://www.idc.com/getdoc.jsp?containerId=prUS48969722.

Atlantic Council. "What Happens When Toxic Online Behavior Enters the Metaverse?" Atlantic Council, June 6, 2022. https://www.atlanticcouncil.org/news/transcripts/what-happens-when-toxic-online-behavior-enters-the-metaverse/.

Baker-White, Emily. "Meta Wouldn't Tell Us How It Enforces Its Rules In VR, So We Ran A Test To Find Out." BuzzFeed News, February 11, 2022. https://www.buzzfeednews.com/article/emilybakerwhite/meta-facebook-horizon-vr-content-rules-test.

Basu, Tanya. "The Metaverse Has a Groping Problem Already." MIT Technology Review, December 16, 2021. https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/.

Bateman, Jon, Natalie Thompson, and Victoria Smith. "How Social Media Platforms' Community Standards Address Influence Operations." Carnegie Endowment for International Peace, April 1, 2021. https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201.

Belamire, Jordan. "My First Virtual Reality Groping." *Athena Talks* (blog), October 22, 2016. https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee.

Bickert, Monika. "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process." Meta, April 24, 2018. https://about.fb.com/news/2018/04/comprehensive-community-standards/.

Bickert, Monika. "Removing Holocaust Denial Content." Meta, October 12, 2020. https://about.fb.com/news/2020/10/removing-holocaust-denial-content/.

Blackwell, Lindsay, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. "Harassment in Social Virtual Reality: Challenges for Platform Governance." *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 100:1-100:25. https://doi.org/10.1145/3359202.

Bosworth, Andrew, and Nick Clegg. "Building the Metaverse Responsibly." Meta, September 27, 2021. https://about.fb.com/news/2021/09/building-the-metaverse-responsibly/.

Microsoft. "Community Standards," January 20, 2022. https://learn.microsoft.com/en-us/windows/mixed-reality/altspace-vr/community/community-standards.

Meta Quest. "Conduct in VR Policy," n.d. Accessed March 8, 2022.

Cortese, Michelle, and Jessica Outlaw. "Social and Multi-User Spaces in VR: Trolling, Harassment, and Online Safety." *The IEEE Global Initiative on Ethics of Extended Reality (XR) Report*, December 15, 2021, 1–17.

PCMag. "Definition of Social VR." Accessed June 17, 2022. https://www.pcmag.com/encyclopedia/term/social-vr.

Feltham, Jamie. "Rec Room Passes 3 Million Monthly Active VR Users." UploadVR, April 14, 2022. https://uploadvr.com/rec-room-3-million-vr-users/.

Fishman, Brian. "Live-Streaming of Terrorist and Violent Extremist Content: Moderation and Crisis Response." Panel, TAT-GIFCT webinar, June 23, 2022.

YouTube Help. "Harassment & Cyberbullying Policies," n.d. https://support.google.com/youtube/answer/2802268?hl=en.

ADL. "Hate Is No Game: Harassment and Positive Social Experiences in Online Games 2021." Accessed October 4, 2022. https://www.adl.org/hateisnogame.

Heaney, David. "Meta Vision Of The Metaverse Shows Futuristic Headset Design." UploadVR, June 16, 2022. https://uploadvr.com/meta-concept-future-headset-design/.

Heath, Alex. "Meta's Social VR Platform Horizon Worlds Hits 300,000 Users." The Verge, February 17, 2022. https://www.theverge.com/2022/2/17/22939297/meta-social-vr-platform-horizon-300000-users.

Meta. "Horizon Mature Worlds Policy," July 2022. https://store.facebook.com/help/quest/articles/horizon/create-in-horizon-worlds/restrictions-to-worlds-in-horizon.

Meta Quest. "Horizon Worlds Prohibited Content Policy," n.d. Accessed March 9, 2022.

Ingram, David. "Facebook Releases Long-Secret Rules on How It Polices the Service." Reuters, April 24, 2018. https://www.reuters.com/article/us-facebook-abuse/facebook-releases-long-secret-rules-on-how-it-polices-the-service-idUSKBN1HV0VR.

Meta. "Introducing Meta Accounts and Meta Horizon Profiles for VR," July 7, 2022. https://about.fb.com/news/2022/07/meta-accounts-and-horizon-profiles-for-vr/.

Isaac, Mike. "Meta Spent $10 Billion on the Metaverse in 2021, Dragging down Profit." The New York Times, February 2, 2022. https://www.nytimes.com/2022/02/02/technology/meta-facebook-earnings-metaverse.html.

Jaloza, Lisa Brown. "Passing the Visual Turing Test: The inside Story of Our Quest for Visual Realism in VR." Tech at Meta, June 21, 2022. https://tech.fb.com/ar-vr/2022/06/passing-the-visual-turing-test-the-inside-story-of-our-quest-for-visual-realism-in-vr/.

Kat Lo [@lolkat]. "I Jumped on VR Chat Recently for Research and within 1 Minute of Entering the First Public Room I Witnessed: 1. a Dozen People Chanting the n Word and Laughing for 5 Minutes 2. Several Men Crowding and Making Sexual Comments to a Nervous Girl Who Said She Was 15 in Her Profile." Tweet. *Twitter*, May 31, 2022. https://twitter.com/lolkat/status/1531608673117667328.

Lawler, Richard. "Meta's VR Roadmap Reportedly Plans Four New Headsets for Release through 2024." The Verge, May 2, 2022. https://www.theverge.com/2022/5/2/23053888/meta-virtual-reality-headset-cambria-quest-vr-mr.

Oasis Consortium. "User Safety Standards for Our Digital Future," 2021. https://www.oasisconsortium.com/usersafetystandards.

Stop Hate for Profit. "One Year After Stop Hate for Profit: Platforms' Progress," June 16, 2021. https://www.stophateforprofit.org/platforms-progress-year-later.

ADL. "Online Hate and Harassment: The American Experience 2021," May 3, 2022. https://www.adl.org/online-hate-2021.

Outlaw, Jessica. "Harassment in Social VR: Stories from Survey Respondents." *Medium* (blog), May 8, 2018. https://jessica-outlaw.medium.com/harassment-in-social-vr-stories-from-survey-respondents-59c9cde7ac02.

Oversight Board. "In Our Work, We Often See Users Left Guessing about Why Meta Removed Their Content." *LinkedIn*. Accessed June 19, 2022. https://www.linkedin.com/posts/oversight-board-administration_what-is-the-impact-of-the-oversight-board-activity-6940284404220985344-w5iX/.

Patel, Nina Jane. "Fiction vs. Non-Fiction." *Medium* (blog), December 21, 2021. https://ninajanepatel.medium.com/fiction-vs-non-fiction-d824c6edfbe2.

Meta. "Promoting Safety with Policy, Product and Operations," November 15, 2018. https://about.fb.com/news/2018/11/inside-feed-community-integrity-keeping-people-safe/.

Ravenscraft, Eric. "What Is the Metaverse, Exactly?" Wired, April 25, 2022. https://www.wired.com/story/what-is-the-metaverse/.

Resolution Games. "Player Code of Conduct," April 16, 2021. https://www.resolutiongames.com/code-of-conduct.

Reach3. "Reach3 Insights' New Research Reveals 59% of Women Surveyed Use a Non-Gendered/Male Identity to Avoid Harassment While Gaming," May 19, 2021. https://www.reach3insights.com/women-gaming-study.

Ricci, Linda. "AI and Digital Beings: Exploring the Intersection of Artificial Intelligence and XR." Presented at the Augmented World Expo, Santa Clara, California, June 2, 2022.

Soon, Weilun. "A Researcher's Avatar Was Sexually Assaulted on a Metaverse Platform Owned by Meta, Making Her the Latest Victim of Sexual Abuse on Meta's Platforms, Watchdog Says." Business Insider, May 29, 2022. https://www.businessinsider.com/researcher-claims-her-avatar-was-raped-on-metas-metaverse-platform-2022-5.

United Nations Human Rights. "Guiding Principles on Business and Human Rights." United Nations Human Rights, 2008. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

Meta. "Upcoming Conduct in VR Policy," 2022. https://store.facebook.com/help/quest/articles/accounts/privacy-information-and-settings/conduct-in-vr-policy-updated.

"Virtual Harassment: The Social Experience of 600+ Regular Virtual Reality Users." Slidedeck. The Extended Mind, 2018. https://drive.google.com/file/d/1afFQJN6QAwmeZdGcRj9R4ohVr00ZNO4a/view.

Educause. "XR (Extended Reality) Community Group." Accessed October 4, 2022. https://www.educause.edu/community/xr-extended-reality-community-group.

Zuckerberg, Mark. "Founder's Letter, 2021." Meta, October 28, 2021. https://about.fb.com/news/2021/10/founders-letter/.

Zuckerberg, Mark. "Today We're Updating Our Hate Speech Policy to Ban Holocaust Denial." Post. *Facebook*, October 12, 2020. https://www.facebook.com/zuck/posts/10112455086578451.

Zuylen-Wood, Simon van. "'Men Are Scum': Inside Facebook's War on Hate Speech." Vanity Fair, February 26, 2019. https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech.

Appendix A

Horizon Policy (July 2022 Version)[86]

HORIZON MATURE WORLDS POLICY

We want Horizon to be a safe and welcoming environment for everyone. When worlds are meant for mature audiences, Meta requires that the world is marked as mature. The following worlds will be allowed in Horizon but must be marked as Mature.

- Content that is sexually suggestive; for example, near nudity, depictions of people in implied or suggestive positions, or an environment focused on activities that are overly suggestive.
- Worlds that are dedicated to or have a core focus on the promotion of marijuana, alcohol, tobacco, or age-regulated activities (including gambling).
- Intense or excessively violent fictional content, including blood and gore, that could shock or disgust users.

Please refer to the Prohibited Worlds Policy below, and Conduct in VR Policy and for a reminder of what is not allowed even with a 18+ tag.

HORIZON PROHIBITED WORLDS POLICY

We want Horizon to be a fun and welcoming environment for our diverse community where people feel safe and respected. That's why we prohibit certain content in worlds that users might find offensive. In addition to what's not allowed under the Conduct in VR Policy, we do not allow the following,

- Content that is sexually explicit or provocative, including nudity, depictions of people in explicit positions, or content or worlds that are sexually provocative or implied.
- Content that depicts or promotes the use of illegal drugs or abuse of prescription drugs.
- Content that promotes criminal or dangerous activity.
- Content that depicts real life intense violence, including blood and gore.

Content that attempts to buy, sell or trade real life regulated goods such as firearms, blades, alcohol and tobacco.

86 "Horizon Mature Worlds Policy," Meta, July 2022, https://store.facebook.com/help/quest/articles/horizon/create-in-horizon-worlds/restrictions-to-worlds-in-horizon. Accessed July 31, 2022.

Appendix B

Horizon Policy (June 2022 Version)[87]

HORIZON WORLDS PROHIBITED CONTENT POLICY

We want Horizon to be a fun and welcoming environment for our diverse community where people feel safe and respected. That's why we only allow content appropriate for all our users. This means that we do not allow the following, in addition to what's not allowed under the Conduct in VR Policy and the Community Standards.

- Content that is adult or sexual in nature; for example, nudity, depictions of people in explicit or suggestive positions, or activities that are suggestive or sexually provocative.
- Content that depicts regulated goods or activities, including prescription medication, marijuana, alcohol, tobacco, firearms, and gambling.
- Intense or violent content, including blood and gore.

87 "Horizon Worlds Prohibited Content Policy," Meta Quest, n.d., accessed March 9, 2022.

Appendix C

Conduct in VR Policy (August 2022 Version)[88]

Our mission is to give people the power to build community and bring the world closer together. Through virtual experiences, people around the world have new ways to work, play, and connect. As you enjoy these experiences, make sure your conduct, which includes content you create or share, is respectful and appropriate for diverse audiences.

We want to empower people to create spaces that inspire and give people **voice**. We are actively committed within spaces, public and private, to values such as **safety** (e.g., no credible threats of violence), **authenticity** (e.g., no scams or impersonating other people), **dignity** (e.g., no bullying), and **privacy** (e.g., no doxing).

Consistent with these values, developers, including Meta, may take action, such as limiting functionality or restricting features in their respective apps and products. Developers and Meta may consider the context, including if certain behaviors took place in a public space, to determine what action is appropriate. Meta may also take action on our platform, such as suspending accounts. As experiences and innovations change we'll continue to update this guidance to help people understand their evolving responsibilities to the ecosystem and each other.

A lot depends on the context; speech and behavior that is acceptable in private or among friends may not be acceptable in more public spaces. But it's helpful for people to understand behaviors that don't align with our values in virtual spaces where you are around other groups.
Examples:

- Don't do or promote anything that's **abusive**, **violent**, or **illegal**, such as:
    - Bullying, harassment, or stalking
    - Sexualizing, exploiting, or abusing minors
    - Engaging in hate speech or promoting hateful ideologies or hate groups
    - Encouraging or promoting physical world violence or people with a violent mission. We do not allow individuals or groups that are engaged in or advocate for terrorism, violence, crime, or hate

88 "Upcoming Conduct in VR Policy," Meta, 2022, https://store.facebook.com/help/quest/articles/accounts/privacy-information-and-settings/conduct-in-vr-policy-updated. Accessed July 31, 2022.

    - Human exploitation, trafficking, or smuggling
    - Promoting suicide or self-harm
    - Engaging in, encouraging, or threatening of any form of non-consensual sexual activity, including sharing intimate images of others without consent
    - Sharing personal information, such as social security numbers, credit card numbers, account login information, or residential information, about yourself or others that could lead to financial or other harms
    - Violating intellectual property rights, such as improperly using copyrighted or trademarked materials
    - Selling or exchanging real-world restricted goods, such as firearms, weapons, or pharmaceutical and non-medical drugs
- Don't engage in, solicit, create, or share:
    - Pornographic or sexually explicit content or behavior
    - Excessively violent content or behavior
- Don't do or promote anything that is designed to deceive other users, Meta, or developers, or that otherwise abuse our products or services. For example, do not:
    - Pretend to be another person or entity, steal someone's identity, or create or use fake accounts. If role playing or parodying, ensure it's clear to others
    - Attempt to gather sensitive information, compromise user accounts, or engage in unauthorized access, such as by creating, hosting, or sharing malware
    - Engage in fraud, scams, or other deceptive activities
    - Spam users, such as repeated offers of commercial services, goods or requests
    - Create or use a Meta account if you're under the age of 13

Appendix D

Conduct in VR Policy (July 2022 Version)[89]

CONDUCT IN VR POLICY

With Oculus, we're creating new ways for people to defy distance and connect with each other and the world around them. Through virtual reality, we can radically redefine the way people work, play and connect. This is a new environment for many people and it's important to have clear guidelines for respectful behavior. We want everyone to feel safe while they enjoy an immersive virtual experience. The Community Standards outline what is and is not allowed on Facebook apps and technologies, and apply to both content and conduct in VR. You can read the full Community Standards here. To help you better understand how the Community Standards apply to conduct in a virtual space, we highlighted areas of our policy and how they apply to conduct in VR. Oculus users come from many different backgrounds, so make sure that your conduct(as well as any content created or shared)is appropriate for a diverse audience and does not violate the Community Standards. Do not:

- Harass or bully other users through conduct, including:
  - Stalking or repeatedly following others against their wishes
  - Cornering, blocking normal movement, physically intimidating or invading personal space without consent
  - Encouraging intimidation or bullying of others, including threats to SWAT, hack, dox, or DDOS.

- Conduct yourself in an offensive or abusive way, including:
  - Touching someone in a sexual way or making sexual gestures
  - Sexualizing minors in any way. In cases of sexual exploitation of children, we report content to the National Center of Missing and Exploited Children.
  - Supporting or representing hateful ideologies or groups by using symbols or attacking people on the basis of race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.

- Impersonate a Facebook employee, partner, representative, or other real person or encourage other users to do so.

89 "Conduct in VR Policy," Meta Quest, n.d., accessed March 8, 2022.

If we find that you have violated the Community Standards or this Conduct in VR Policy across Facebook or Oculus products, we may take action on your account, including temporarily restricting or suspending your account. For repeated or egregious offenses, we may permanently disable your account.

Appendix E

Facebook Community Standards

Principles and Policies[90]

VIOLENCE AND CRIMINAL BEHAVIOR

- Violence and Incitement
- Dangerous Individuals and Organizations
- Coordinating Harm and Promoting Crime
- Restricted Goods and Services
- Fraud and Deception

SAFETY

- Suicide and Self-Injury
- Child Sexual Exploitation, Abuse and Nudity
- Adult Sexual Exploitation
- Bullying and Harassment
- Human Exploitation
- Privacy Violations

OBJECTIONABLE CONTENT

- Hate Speech
- Violent and Graphic Content
- Adult Nudity and Sexual Activity
- Sexual Solicitation

INTEGRITY AND AUTHENTICITY

- Account Integrity and Authentic Identity
- Spam
- Cybersecurity
- Inauthentic Behavior
- Misinformation
- Memorialization

RESPECTING INTELLECTUAL PROPERTY

- Intellectual Property

CONTENT-RELATED REQUESTS AND DECISIONS

- User Requests
- Additional Protection of Minors

90 "Facebook Community Standards," Meta Transparency Center, accessed April 5, 2022, https://transparency.fb.com/policies/community-standards/.

Appendix F

Facebook Community Standards

Hate Speech Policy[91]

Policy Rationale

We believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why we don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, and in some cases may promote offline violence.

We define hate speech as a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We also prohibit the use of harmful stereotypes, which we define as dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics like occupation, when they're referenced along with a protected characteristic. Sometimes, based on local nuance, we consider certain words or phrases as code words for PC groups.

We recognize that people sometimes share content that includes someone else's hate speech to condemn it or raise awareness. In other cases, speech that might otherwise violate our standards can be used self-referentially or in an empowering way. Our policies are designed to allow room for these types of speech, but we require people to clearly indicate their intent. If the intention is unclear, we may remove content.

91 "Hate Speech," Meta Transparency Center, n.d., https://transparency.fb.com/policies/community-standards/hate-speech/.

Learn more about our approach to hate speech.

Do not post:

Tier 1

Content targeting a person or group of people (including all groups except those who are considered non-protected groups described as having carried out violent crimes or sexual offenses or representing less than half of a group) on the basis of their aforementioned protected characteristic(s) or immigration status with:

- Violent speech or support in written or visual form

- Dehumanizing speech or imagery in the form of comparisons, generalizations, or unqualified behavioral statements (in written or visual form) to or about:
  - Insects.
  - Animals that are culturally perceived as intellectually or physically inferior.
  - Filth, bacteria, disease and feces.
  - Sexual predator.
  - Subhumanity.
  - Violent and sexual criminals
  - Other criminals (including but not limited to "thieves," "bank robbers," or saying "All [protected characteristic or quasi-protected characteristic] are 'criminals'").
  - Statements denying existence.

- Mocking the concept, events or victims of hate crimes even if no real person is depicted in an image.

- Designated dehumanizing comparisons, generalizations, or behavioral statements (in written or visual form) that include:
  - Black people and apes or ape-like creatures.
  - Black people and farm equipment.
  - Caricatures of Black people in the form of blackface.
  - Jewish people and rats.
  - Jewish people running the world or controlling major institutions such as media networks, the economy or the government.
  - Denying or distorting information about the Holocaust.
  - Muslim people and pigs.

* Muslim person and sexual relations with goats or pigs.
* Mexican people and worm-like creatures.
* Women as household objects or referring to women as property or "objects".
* Transgender or non-binary people referred to as "it".
* Dalits, scheduled caste or 'lower caste' people as menial laborers.

Tier 2

Content targeting a person or group of people on the basis of their protected characteristic(s) with:

- Generalizations that state inferiority (in written or visual form) in the following ways:
  * Physical deficiencies are defined as those about:
    - Hygiene, including but not limited to: filthy, dirty, smelly.
    - Physical appearance, including but not limited to: ugly, hideous.

  * Mental deficiencies are defined as those about:
    - Intellectual capacity, including but not limited to: dumb, stupid, idiots.
    - Education, including but not limited to: illiterate, uneducated.
    - Mental health, including but not limited to: mentally ill, retarded, crazy, insane.

  * Moral deficiencies are defined as those about:
    - Character traits culturally perceived as negative, including but not limited to: coward, liar, arrogant, ignorant.
    - Derogatory terms related to sexual activity, including but not limited to: whore, slut, perverts.

- Other statements of inferiority, which we define as:
    - Expressions about being less than adequate, including but not limited to: worthless, useless.
    - Expressions about being better/worse than another protected characteristic, including but not limited to: "I believe that males are superior to females."
    - Expressions about deviating from the norm, including but not limited to: freaks, abnormal.

- Expressions of contempt (in written or visual form), which we define as:
  * Self-admission to intolerance on the basis of a protected characteristics, including but not limited to: homophobic, islamophobic, racist.

    - Expressions that a protected characteristic shouldn't exist.
    - Expressions of hate, including but not limited to: despise, hate.

- Expressions of dismissal, including but not limited to: don´t respect, don't like, don´t care for

- Expressions of disgust (in written or visual form), which we define as:
    - Expressions that suggest the target causes sickness, including but not limited to: vomit, throw up.
    - Expressions of repulsion or distaste, including but not limited to: vile, disgusting, yuck.

- Cursing, except certain gender-based cursing in a romantic break-up context, defined as:
    - Referring to the target as genitalia or anus, including but not limited to: cunt, dick, asshole.
    - Profane terms or phrases with the intent to insult, including but not limited to: fuck, bitch, motherfucker.
    - Terms or phrases calling for engagement in sexual activity, or contact with genitalia, anus, feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit.

Tier 3

Content targeting a person or group of people on the basis of their protected characteristic(s) with any of the following:

- Segregation in the form of calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting segregation.

- Exclusion in the form of calls for action, statements of intent, aspirational or conditional statements, or statements advocating or supporting, defined as
    - Explicit exclusion, which means things like expelling certain groups or saying they are not allowed.
    - Political exclusion, which means denying the right to right to political participation.
    - Economic exclusion, which means denying access to economic entitlements and limiting participation in the labour market.
    - Social exclusion, which means things like denying access to spaces (physical and online)and social services, except for gender-based exclusion in health and positive support Groups.

Content that describes or negatively targets people with slurs, where slurs are defined as words that are inherently offensive and used as insulting labels for the above characteristics.

For the following Community Standards, we require additional information and/or context to enforce:

Do not post:

- Content explicitly providing or offering to provide products or services that aim to change people's sexual orientation or gender identity.
- Content attacking concepts, institutions, ideas, practices, or beliefs associated with protected characteristics, which are likely to contribute to imminent physical harm, intimidation or discrimination against the people associated with that protected characteristic. Facebook looks at a range of signs to determine whether there is a threat of harm in the content. These include but are not limited to: content that could incite imminent violence or intimidation; whether there is a period of heightened tension such as an election or ongoing conflict; and whether there is a recent history of violence against the targeted protected group. In some cases, we may also consider whether the speaker is a public figure or occupies a position of authority.
- Content targeting a person or group of people on the basis of their protected characteristic(s) with claims that they have or spread the novel coronavirus, are responsible for the existence of the novel coronavirus, are deliberately spreading the novel coronavirus, or mocking them for having or experiencing the novel coronavirus.

In certain cases, we will allow content that may otherwise violate the Community Standards when it is determined that the content is satirical. Content will only be allowed if the violating elements of the content are being satirized or attributed to something or someone else in order to mock or criticize them.

Appendix G

Facebook Community Standards

Bullying and Harassment Policy[92]

Policy Rationale

Bullying and harassment happen in many places and come in many different forms from making threats and releasing personally identifiable information to sending threatening messages and making unwanted malicious contact. We do not tolerate this kind of behavior because it prevents people from feeling safe and respected on Facebook.

We distinguish between public figures and private individuals because we want to allow discussion, which often includes critical commentary of people who are featured in the news or who have a large public audience. For public figures, we remove attacks that are severe as well as certain attacks where the public figure is directly tagged in the post or comment.

For private individuals, our protection goes further: We remove content that's meant to degrade or shame, including, for example, claims about someone's sexual personal activity. We recognize that bullying and harassment can have more of an emotional impact on minors, which is why our policies provide heightened protection for users between the ages of 13 and 18.

Context and intent matter, and we allow people to post and share if it is clear that something was shared in order to condemn or draw attention to bullying and harassment. In certain instances, we require self-reporting because it helps us understand that the person targeted feels bullied or harassed. In addition to reporting such behavior and content, we encourage people to use tools available on Facebook to help protect against it.

We also have a Bullying Prevention Hub, which is a resource for teens, parents, and educators seeking support for issues related to bullying and other conflicts. It offers step-by-step guidance, including

92 "Bullying and Harassment," Meta Transparency Center, n.d., https://transparency.fb.com/policies/community-standards/bullying-harassment/.

information on how to start important conversations about bullying. Learn more about what we are doing to protect people from bullying and harassment here.

Note: This policy does not apply to individuals who are part of designated organizations under the Dangerous Organizations and Individuals policy or individuals who died prior to 1900.

Do not:

Tier 1: Target anyone maliciously by:

- Repeatedly contacting someone in a manner that is:
  - Unwanted, or
  - Sexually harassing, or
  - Directed at a large number of individuals with no prior solicitation.

- Attacking someone based on their status as a victim of sexual assault, sexual exploitation, sexual harassment, or domestic abuse.

- Calling for self-injury or suicide of a specific person, or group of people.

- Attacking someone through derogatory terms related to sexual activity (for example: whore, slut).

- Posting content about a violent tragedy, or victims of violent tragedies that include claims that a violent tragedy did not occur.

- Posting content about victims or survivors of violent tragedies or terrorist attacks by name or by image, with claims that they are:
  - Acting/pretending to be a victim of an event.
  - Otherwise paid or employed to mislead people about their role in the event.

- Threatening to release an individual's private phone number, residential address or email address.

- Making statements of intent to engage in a sexual activity or advocating for them to engage in a sexual activity.

- Making severe sexualized commentary

- Sharing derogatory sexualized photoshopped imagery or drawings

- Calling for, or making statements of intent to engage in, bullying and/or harassment.

- Posting content that further degrades or expresses disgust toward individuals who are depicted in the process of, or right after, menstruating, urinating, vomiting, or defecating

- Creating Pages or Groups that are dedicated to attacking individual(s) by:
  * Calling for death, or to contract or develop a medical condition.
  * Making statements of intent of advocating to engage in sexual activity.
  * Making claims that the individual has or may have a sexually transmitted disease.
  * Sexualizing another adult.

- Sending messages that contain the following attacks when aimed at an individual or group of individuals in the thread:
  * Attacks referenced in Tier 1, 2 and 4 of this policy.
  * Targeted cursing.
  * Calls for death, serious disease, disability, epidemic disease or physical harm.

Tier 2: Target private individuals, limited scope public figures (for example, individuals whose primary fame is limited to their activism, journalism, or those who become famous through involuntary means) or public figures who are minors with:

- Calls for death, or to contract or develop a medical condition.

- Female-gendered cursing terms when used in a derogatory way.

- Claims about sexual activity or sexually transmitted diseases except in the context of criminal allegations against adults about non-consensual sexual touching.

- Pages or Groups created to attack through:
  * Targeted cursing.
  * Negative physical descriptions.
  * Claims about religious identity or blasphemy.
  * Expressions of contempt or disgust.
  * Female-gendered cursing terms when used in a derogatory way.

Tier 3: Target public figures by purposefully exposing them to:

- For adults and minors:
  - Calls for death, or to contract or develop a medical condition.
  - Claims about sexually transmitted disease
  - Female-gendered cursing terms when used in a derogatory way.
  - Content that praises, celebrates or mocks their death or medical condition.
  - Attacks through negative physical descriptions.

- For minors:
  - Comparisons to animals or insects that are culturally perceived as intellectually or physically inferior or to an inanimate object ("cow," "monkey" "potato").
  - Content manipulated to highlight, circle or otherwise negatively draw attention to specific physical characteristics (nose, ear and so on).

Tier 4: Target private individuals or limited scope public figures (for example, individuals whose primary fame is limited to their activism, journalism, or those who become famous through involuntary means) with:

- Comparisons to animals or insects that are culturally perceived as intellectually or physically inferior or to an inanimate object ("cow," "monkey" "potato").
- Content manipulated to highlight, circle or otherwise negatively draw attention to specific physical characteristics (nose, ear and so on).
- Attacks through negative physical descriptions.
- Content that ranks individuals on physical appearance or personality.
- Content sexualizing another adult.
- Content that further degrades individuals who are depicted being physically bullied except in self-defense and fight-sport contexts.
- Content that praises, celebrates, or mocks their death or serious physical injury.
- In addition to the above, attacks through Pages or Groups:
  - Negative character or ability claims.
  - First-person voice bullying only if the object targets more than one private individual.

Tier 5: Target private adults (who must self-report) or any private minors or involuntary minor public figures with:

- Targeted cursing.
- Claims about romantic involvement, sexual orientation or gender identity.
- Coordination, advocacy or promotion of exclusion.
- Negative character or ability claims, except in the context of criminal allegations and business reviews against adults. We allow criminal allegations so that people can draw attention to personal experiences or offline events. In cases in which criminal allegations pose off-line harm to the named individual, however, we may remove them.
- Expressions of contempt or disgust, except in the context of criminal allegations against adults.

Tier 6: Target private individuals who are minors with:

- Allegations about criminal or illegal behavior.
- Videos of physical bullying shared in a non-condemning context.

Tier 7: Target private individuals (who must self-report) with:

- First-person voice bullying.
- Unwanted manipulated imagery.
- Comparison to other public, fictional or private individuals on the basis of physical appearance
- Claims about religious identity or blasphemy.
- Comparisons to animals or insects that are not culturally perceived as intellectually or physically inferior ("tiger," "lion").
- Neutral or positive physical descriptions.
- Non-negative character or ability claims.
- Any bullying or harassment violation, when shared in an endearing context.
- Attacks through derogatory terms related to a lack of sexual activity.

We add a cover to this content so people can choose whether to see it:

Videos of physical bullying against minors shared in a condemning context
For the following Community Standards, we require additional information and/or context to enforce:

Do not:

- Post content that targets private individuals through unwanted Pages, Groups and Events. We remove this content when it is reported by the victim or an authorized representative of the victim.

- Create accounts to contact someone who has blocked you.

- Post attacks that use derogatory terms related to female gendered cursing. We remove this content when the victim or an authorized representative of the victim informs us of the content, even if the victim has not reported it directly.

- Post content that would otherwise require the victim to report the content or an indicator that the poster is directly targeting the victim (for example: the victim is tagged in the post or comment). We will remove this content if we have confirmation from the victim or an authorized representative of the victim that the content is unwanted.

- Post content praising, celebrating or mocking anyone's death. We also remove content targeting a deceased individual that we would normally require the victim to report.

- Post content calling for or stating an intent to engage in behavior that would qualify as bullying and harassment under our policies. We will remove this content when we have confirmation from the victim or an authorized representative of the victim that the content is unwanted.

- Post content sexualizing a public figure. We will remove this content when we have confirmation from the victim or an authorized representative of the victim that the content is unwanted.

- Repeatedly contact someone to sexually harass them. We will remove this content when we have confirmation from the victim or an authorized representative of the victim that the content is unwanted.

- Engage in mass harassment against individuals that targets them based on their decision to take or not take the COVID-19 vaccine with:
  * Statements of mental or moral inferiority based on their decision, or
  * Statements that advocate for or allege a negative outcome as a result of their decision, except for widely proven and/or accepted COVID-19 symptoms or vaccine side effects.

- Remove directed mass harassment, when:
  - Targeting, via any surface, 'individuals at heightened risk of offline harm', defined as:
    - Human rights defenders
    - Minors
    - Victims of violent events/tragedies
    - Opposition figures in at-risk countries during election periods
    - Government dissidents who have been targeted based on their dissident status
    - Ethnic and religious minorities in conflict zones
    - Member of a designated and recognizable at-risk group
  - Targeting any individual via personal surfaces, such as inbox or profiles, with:
    - Content that violates the bullying and harassment policies for private individuals or,
    - Objectionable content that is based on a protected characteristic

- Disable accounts engaged in mass harassment as part of either
  - State or state-affiliated networks targeting any individual via any surface.
  - Adversarial networks targeting any individual via any surface with:
    - Content that violates the bullying and harassment policies for private individuals or,
    - Content that targets them based on a protected characteristic, or,
    - Content or behavior otherwise deemed to be objectionable in local context

Appendix H

YouTube Harassment & Cyberbullying Policies[93]

Content that threatens individuals is not allowed on YouTube. We also don't allow content that targets an individual with prolonged or malicious insults based on intrinsic attributes. These attributes include their protected group status or physical traits.

If you find content that violates this policy, report it. Instructions for reporting violations of our Community Guidelines are available here. If you've found multiple videos or comments that you would like to report, you can report the channel. For tips and best practices to stay safe, keep your account secure, and protect your privacy, check out this Help Center article.

If specific threats are made against you and you feel unsafe, report it directly to your local law enforcement agency.

What this policy means for you

If you're posting content

Don't post content on YouTube if it fits any of the descriptions noted below.

> Content that features prolonged name calling or malicious insults (such as racial slurs) based on someone's intrinsic attributes. These attributes include their protected group status, physical attributes, or their status as a survivor of sexual assault, non-consensual intimate imagery distribution, domestic abuse, child abuse and more.
>
> Content uploaded with the intent to shame, deceive or insult a minor. A minor is defined as an individual under the legal age of majority. This usually means anyone younger than 18 years old, but the age of a minor might vary by geography.

Other types of content that violate this policy

93 "Harassment & Cyberbullying Policies," YouTube Help, n.d., https://support.google.com/youtube/answer/2802268?hl=en.

Revealing someone's personally identifiable information (PII), such as their home address, email addresses, sign-in credentials, phone numbers, passport number, medical records, or bank account information.

> *Note: This doesn't include posting widely available public information. Public information can include an official's office phone number or the phone number of a business.*

Content that incites others to harass or threaten individuals on or off YouTube.

Content that encourages abusive fan behavior such as doxxing, dogpiling, brigading or off-platform targeting.

Content that targets an identifiable individual as part of a harmful conspiracy theory where the conspiracy theory has been linked to direct threats or violent acts.

Content making implicit or explicit threats of physical harm or destruction of property against identifiable individuals.

> *Note: "Implicit threats" include threats that don't express a specific time, place or means, but may feature weapon brandishing, simulated violence and more.*

Content posted by vigilantes restraining or assaulting an identifiable individual.

Content reveling in or mocking the death or serious injury of an identifiable individual.

Content that depicts creators simulating acts of serious violence against others (executions, torture, maimings, beatings and more).

Content featuring non-consensual sex acts, unwanted sexualization or anything that graphically sexualizes or degrades an individual.

Content that displays or shows how to distribute non-consensual sexual imagery.

This policy applies to videos, video descriptions, comments, live streams, and any other YouTube product or feature. Keep in mind that this isn't a complete list. Please note these policies also apply to external links in your content. This can include clickable URLs, verbally directing users to other sites in video, as well as other forms.

Exceptions

If the primary purpose is educational, documentary, scientific, or artistic in nature, we may allow content that includes harassment. These exceptions are not a pass to harass someone. Some examples include:

Debates related to high-profile officials or leaders: Content featuring debates or discussions of topical issues concerning individuals who have positions of power, like high-profile government officials or CEOs of major multinational corporations.

Scripted performances: Insults made in the context of an artistic medium such as scripted satire, stand up comedy, or music (such as a diss track). Note: This exception is not a pass to harass someone and claim "I was joking."

Harassment education or awareness: Content that features actual or simulated harassment for documentary purposes or with willing participants (such as actors) to combat cyberbullying or raise awareness.

Note: 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 person.

Monetization and other penalties

In some rare cases, we may remove content or issue other penalties when a creator:

Repeatedly encourages abusive audience behavior.

Repeatedly targets, insults and abuses an identifiable individual based on their intrinsic attributes across several uploads.

Exposes an individual to risks of physical harm based on the local social or political context.

Creates content that harms the YouTube community by persistently inciting hostility between creators for personal financial gain.

Examples

Here are some examples of content that's not allowed on YouTube:

Repeatedly showing pictures of someone and then making statements like "Look at this creature's teeth, they're so disgusting!", with similar commentary targeting intrinsic attributes throughout the video.

Targeting an individual based on their membership in a protected group, such as by saying: "Look at this filthy [slur targeting a protected group], I wish they'd just get hit by a truck."

Targeting an individual and making claims they are involved in human trafficking in the context of a harmful conspiracy theory where the conspiracy is linked to direct threats or violent acts.

Using an extreme insult to dehumanize an individual based on their intrinsic attributes. For example: "Look at this dog of a woman! She's not even a human being — she must be some sort of mutant or animal!"

Depicting an identifiable individual being murdered, seriously injured, or engaged in a graphic sexual act without their consent.

Accounts dedicated entirely to focusing on maliciously insulting an identifiable individual.

More Examples

Targeting an individual based on their intrinsic attributes to wish for their death or serious injury, for example "I wish someone would just bring a hammer down on that [Member of a Protected Group's] face."

Threatening someone's physical safety. This includes implied threats like "when I see you next, things will end badly for you." It also includes explicit threats like "when I see you on Saturday I'm going to punch you in the face." Threatening or implying violence by saying things such as, "You better watch out" while brandishing a weapon is also an example.

Posting an individual's nonpublic personal identifying information like a phone number, home address, or email to direct abusive attention or traffic toward them. For example: "I got a hold of their phone number, keep on calling and leaving messages until they pick up!"

"Raiding" or directing malicious abuse to identifiable individuals through in-game voice chat or messages during a stream.

Directing users toward a YouTuber's comment section for malicious abuse. For example: "everyone needs to go over to this individual's channel right now and just go crazy, let them know how much we want them to die."

Linking to off platform sites that host or feature non-consensual intimate imagery.

Requesting that other users get in touch to share non-consensual intimate imagery.

"Swatting" or other prank calls to emergency or crisis response services, or encouraging viewers to act in this or any other harassing behavior.

Stalking or attempting to blackmail users.

Zooming in or prolonged focused or emphasis on the breasts, buttocks or genital area of an identifiable individual for the purposes of degrading, objectifying, or sexualizing.

Video game content which has been developed or modified ("modded") to promote violence or hatred against an individual with the attributes noted above.

Remember these are just some examples, and don't post content if you think it might violate this policy.

What happens if content violates this policy

If your content violates this policy, we'll remove the content and send you an email to let you know. If this is your first time violating our Community Guidelines, you'll likely get a warning with no penalty to your channel. If it's not, we may issue a strike against your channel. If you get 3 strikes within 90 days, your channel will be terminated. You can learn more about our strikes system here.

We may terminate your channel or account for repeated violations of the Community Guidelines or Terms of Service. We may also terminate your channel or account after a single case of severe abuse, or when the channel is dedicated to a policy violation. You can learn more about channel or account terminations here.

Acknowledgments

It is with deep gratitude that the author thanks his advisor at the Center for Long-Term Cybersecurity, Jessica Newman, as well as the following individuals for their role in supporting this work, providing expert knowledge, pivotal resources, or essential revisions: Chuck Kapelke, Sarah Ryan, Richmond Wong, Jeeyun Baik, Jordan Famularo, Juliana Friend, Gregg Muragishi, Kavya Pearlman, and Matthew Soethe. The judgments and conclusions of this paper are solely those of the author, and are not necessarily endorsed by individuals consulted during the project. Special thanks to Nicole Hayward for her expert design and formatting of this paper.

About the Author

Rafi Lazerson examines security and human rights issues related to emerging technologies. He holds an MPA from the Goldman School of Public Policy at UC Berkeley, a BA in political science from Brooklyn College, and Rabbinical Ordination. As an MPA he conducted a graduate internship with the U.S. Department of State and was an Alternative Digital Futures Researcher with the Center for Long-Term Cybersecurity. He is a Research Affiliate with the Center for Security in Politics, UC Berkeley. Previously, he worked as Assistant Regional Director at the ADL and as CMO of an online marketplace.


CLTC
Center for Long-Term
Cybersecurity
UC Berkeley