BANNED FOR BEING

MODERATING ONLINE GAMES' PUBLIC SPHERE

ELENI MANIS
EVAN ENZER
DEREK SMITH
LETICIA MURILLO
ANNA SIPEK
SARAH ROTH
BRITTAN HELLER

MAY 19, 2023

STOP
SURVEILLANCE TECHNOLOGY OVERSIGHT PROJECT

Supported in part by a grant from the John D. and Catherine T. MacArthur Foundation

Executive Summary

- Online gamers aren't just playing games: they're joining lively public forums that reflect offline civil society with both its good and its evils.

- Game companies moderate content to promote civility in these forums, but they frequently make bad calls against historically marginalized gamers. These gamers get "banned for being" when they assert their identities and "banned for responding" when they reply to harassment by standing up for themselves.

- Games can begin to combat repeated bad calls against historically marginalized gamers—which amount to structural discrimination in online games—by offering robust notice and appeals systems.

- But to build truly inclusive public forums, games must also change the rules of moderation. Games must consider gamers' words in context to avoid penalizing historically marginalized gamers for asserting their identities or for responding to hate. They must carefully limit any rules against political and controversial speech to avoid forcing historically marginalized gamers off of gaming platforms.

This report includes graphic descriptions of online harassment.

I. Introduction

Online video games and gamer-centered platforms in the U.S. host lively public forums where people play, build communities, and engage in political activity. They are so much an extension of civil society that political campaigns have taken to organizing inside of them. But there's a persistent threat to these communities: widespread, severe harassment, including abuse targeted at historically marginalized gamers who reveal their identities. Online gaming communities can be so hostile to women (cis and trans), LGBTQ+ people, and BIPOC individuals that they leave online gaming altogether.

Game publishers have responded to harassment with content moderation, but moderation causes new problems for historically marginalized gamers. These gamers get "banned for being" when they assert their identities and run up against games' banned-word lists or rules forbidding political or controversial content. They get "banned for responding" when they reply to abuse by standing up for themselves. These bad calls amount to structural discrimination inside online games, reinforcing the perception that gaming is a straight, cis white man's world and pushing historically marginalized gamers off of games' public forums.

Games can begin to address bad calls against historically marginalized gamers by implementing robust notice and appeals systems. A blueprint for such systems exists in the Santa Clara principles,

a framework for social media moderation created by human rights advocates and endorsed by major social media platforms. But notice and appeals systems burden historically marginalized gamers with the task of justifying their innocent conduct and educating game moderators. We therefore suggest two changes to the rules of moderation themselves. Any enforcement of banned-word lists must take words' context into account and avoid flagging historically marginalized gamers for asserting their identities. Second, overbroad rules like "no politics" must be specified so that they cannot be used to force historically marginalized gamers off of game platforms or into hiding. These steps—better notice and appeals systems, and moderation rules built with historically marginalized gamers' input and interests in mind—will help game companies build the inclusive forums that many claim to support.

II. Gaming is Civil Society

Gaming is civil society.[1] People still find community in parks, cafes, and offline. But many gravitate to the internet—including an estimated 227 million American gamers.[2] Demographically, the gaming community resembles American society offline. Children meet friends on games after school[3] and adults find safe spaces to talk about shared interests.[4] Many gamers are men, but contrary to stereotype, at least 40% of women game regularly.[5] Gamers come from every educational background and are racially diverse: in the U.S., 41% of White individuals, 44% of Black individuals, and 48% of Latinx individuals game regularly.[6]

For Black, Latinx and other historically marginalized gamers, gaming offers extraordinary opportunities to build supportive communities. Historically marginalized gamers don't escape harassment by going online (*see* section III) but they have built welcoming groups for socializing and community organizing in games and gaming-adjacent platforms. Melanin Gamers, for example, organizes gaming tournaments for BIPOC people and promotes better representation of BIPOC people in games.[7] LGBTQ+ players host dozens of Minecraft servers dedicated to "non-toxic"

---

[1] Kat Schrier, *We the Gamers: How Games Teach Ethics and Civics*, 2021, Oxford University Press.

[2] This figure includes gamers playing multiplayer online games and other games. Entertainment Software Association, "2021 Essential Facts About the Video Game Industry" (Entertainment Software Association, 2021), https://www.theesa.com/wp-content/uploads/2021/08/2021-Essential-Facts-About-the-Video-Game-Industry-1.pdf.

[3] Brenda K. Wiederhold, "Kids Will Find a Way: The Benefits of Social Video Games," *Cyberpsychology, Behavior, and Social Networking* 24, no. 4 (April 2021): 213–14, https://doi.org/10.1089/cyber.2021.29211.editorial.

[4] Julia Kneer et al., "Same Gaming: An Exploration of Relationships Between Gender Traits, Sexual Orientation, Motivations, and Enjoyment of Playing Video Games," *Simulation & Gaming* 53, no. 5 (October 1, 2022): 423–45, https://doi.org/10.1177/10468781221113030. Evan Enzer, "Women Are a Massive Part of Nerd Culture…Society Still Thinks it's Only for Men," *Your Local Wizards* (blog), February 15, 2023, https://medium.com/@urlocalwizards/women-are-a-massive-part-of-nerd-culture-society-still-thinks-its-only-for-men-c802dd75d0a6.

[5] "Essential Facts About the Video Game Industry," Entertainment Software Association.

[6] Gamers who report gaming often or sometimes. Anna Brown, "Younger Men Play Video Games, but so Do a Diverse Group of Other Americans," *Pew Research Center* (blog), September 11, 2017, https://www.pewresearch.org/fact-tank/2017/09/11/younger-men-play-video-games-but-so-do-a-diverse-group-of-other-americans.

[7] "About: Melanin Gamers - News, Articles & Community," February 27, 2020, https://themelaningamers.com/about-us/.

gaming environments,[8] creating spaces where—as one gamer told S.T.O.P.—a person can realize that "maybe they are not the straight… guy they thought they were."[9] On Reddit and other game-related forums, and within games themselves, gamers can readily find places where Black gamers,[10] LGBTQ+ gamers,[11] women gamers,[12] Muslim gamers,[13] gamers with intersectional identities,[14] and other historically marginalized people congregate and organize for social change.

Gaming is such an authentic extension of civil society that even political campaigns have taken to organizing there. When the COVID-19 pandemic shut down offline gatherings, political campaigns were quick to realize that civil society was alive and well in games.[15] The Biden campaign launched "No Malarky Island," a virtual campaign headquarters in Animal Crossing, with 18 days left until the 2020 presidential election.[16] A candidate for prime minister in Canada followed suit.[17] Congresswoman Alexandria Ocasio Cortez streamed video on Twitch while gaming, reaching more viewers and potential voters than almost any prior video on the platform.[18] Game publishers themselves have sometimes balked at political and community organizing on their platforms (*see* section VI), but it is here: a lively, digital public sphere.

III. Gaming Harassment

There is one particularly persistent threat to gaming communities: harassment. Games are *supposed* to be a place to relax and have fun with friends, but for many online gamers, the fun is marred by vile abuse. Over three quarters of adult players report being stalked, physically threatened, or otherwise severely harassed on games like Fortnite and World of Warcraft.[19] Over half of adult gamers report

[8] *See, for example*, "Minecraft LGBT Friendly Servers," Minecraft IP List, accessed April 14, 2023, https://www.minecraftiplist.com/server-tags/Lgbtfriendly. *See also* "Best Lgbt Minecraft Servers," *Minecraft Server List* (blog), accessed April 14, 2023, https://minecraft-servers-listing.com/category/lgbt/.
[9] Kofu, S.T.O.P. Interview with username Kofu, February 10, 2023.
[10] "About: Melanin Gamers - News, Articles & Community."
[11] "R/Gaymers," Reddit, n.d., https://www.reddit.com/r/gaymers/.
[12] "R/GirlGamers," Reddit, n.d., https://www.reddit.com/r/GirlGamers/.
[13] "R/Muslimgamers," Reddit, n.d., https://www.reddit.com/r/muslimgamers/.
[14] *See, for example*, Kishonna Gray-Denson, "Gaming Out Online: Black Lesbian Identity Development and Community Building in Xbox Live," *Journal of Lesbian Studies* 22 (November 22, 2017): 1–15, https://doi.org/10.1080/10894160.2018.1384293.
[15] Daisy Schofield, "Black Lives Matter Meets Animal Crossing: How Protesters Take Their Activism into Video Games," *The Guardian*, August 7, 2020, sec. Games, https://www.theguardian.com/games/2020/aug/07/black-lives-matter-meets-animal-crossing-how-protesters-take-their-activism-into-video-games.
[16] Makena Kelly, "The Official Biden HQ in Animal Crossing Has Poll Booths, Ice Cream, and No Malarkey," *The Verge*, October 16, 2020, https://www.theverge.com/2020/10/16/21519505/joe-biden-animal-crossing-new-horizons-biden-hq-campaign-election.
[17] Dakoda Trithara, "How Online Gaming Is a Growing Tool for Political Mobilization," *Policy Options*, December 13, 2021, https://policyoptions.irpp.org/magazines/december-2021/how-online-gaming-is-a-growing-tool-for-political-mobilization/.
[18] Allegra Frank, "AOC Met More than 400,000 Young Potential Voters on Twitch," *Vox*, October 22, 2020, https://www.vox.com/2020/10/22/21526625/aoc-twitch-stream-among-us-most-popular-twitch-streams-ever.
[19] Referring to massively multiplayer online games. "Hate Is No Game Hate and Harassment in Online Games 2022" (Anti-Defamation League, December 2022), https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2022.

being hounded over and over again.[20] This goes for young gamers, too: over two thirds of gamers aged 10-17 report being harassed, nearly 20% of them repeatedly.[21]

Worse yet, many gamers report being harassed for simply being who they are. Three quarters of women gamers (cis and trans) reported gender harassment in large 2021 and 2022 studies.[22] S.T.O.P.'s survey of online gamer forums revealed gamers reporting violently misogynistic hate (e.g., "A guy on Overwatch described in detail how he would torture rape and kill me."[23]) It is no surprise that nearly 60% of women gamers hide their gender online.[24] Other historically marginalized gamers face similar identity-based hate: 44% of Black gamers, 40% of Asian American gamers, 37% of LGBTQ+ gamers, and about 30% of Jewish, Muslim and Latino gamers have also been harassed for being who they are.[25] Hateful slurs are utterly commonplace in games.[26] Beyond that, actual calls to violence are common: in one regrettably typical example, gamers called on Bethesda Games for help after being chased in the game by self-proclaimed "gay eliminators" threatening to "lynch queers."[27] This hate follows gamers to every online platform where gamers gather: the video-streaming platform Twitch, the Discord chat app, the PC gaming marketplace Steam, YouTube, and more.[28] Gamers have even automated harassment for scale. In 2021, bots flooded the comments and messages sections of BIPOC and LGBTQ+ gamers with hate on Twitch.[29] The platform implemented account verification to stop these hate raids and sued two particularly active raiders, but white extremists took credit for a new wave of hate raids in 2022.[30]

As this suggests, gamer harassment is often ideologically motivated. White extremists frequent major online games, even games often marketed toward kids, like Roblox.[31] And extremists don't just play: they intimidate fellow gamers, spread hate, and recruit new members.[32] One in five adults and 15%

[20] 57% have experienced sustained harassment by some measures. "Online Gaming Forms of Harassment in the U.S. 2022," Statista, accessed March 17, 2023, https://www.statista.com/statistics/1133182/harassment-online-video-games/.
[21] "Hate Is No Game," Anti-Defamation League.
[22] "Hate Is No Game," Anti-Defamation League. *See also* Katherine Long, "A New Survey Confirms That Most Women Gamers Have Faced Discrimination," Paste Magazine, May 20, 2021, https://www.pastemagazine.com/games/a-new-survey-confirms-that-most-women-gamers-have.
[23] TheLadyPez, "Today's (Messaging) Ban from Xbox Because I Call It R*pe When a Player Texts about Having Three 14 Year-Old Wives," Reddit Post, *R/GirlGamers*, June 29, 2021, https://www.reddit.com/r/GirlGamers/comments/oa5vuf/todays_messaging_ban_from_xbox_because_i_call_it/.
[24] Long, "Most Women Faced Discrimination."
[25] "Hate Is No Game," Anti-Defamation League.
[26] James Cullen, "Does the F-Word Being a 'Gamer Word' Highlight Streaming's Homophobia Problem?," *NME* (blog), December 10, 2021, https://www.nme.com/features/gaming-features/does-the-f-word-being-a-gamer-word-highlight-streamings-homophobia-problem-3115023.
[27] AJpls [@AJpls_], "So @bethesda, How Do We Report People in @Fallout?" Tweet, *Twitter*, November 16, 2018, https://twitter.com/AJpls_/status/1063292165378502657.
[28] Grayson, "More than a Twitch Problem."
[29] Alexander Lee, "'Don't Let It Bother You, Just Continue Streaming': Confessions of a Twitch Streamer Who Received 'Hate Raids,'" *Digiday* (blog), March 8, 2022, https://digiday.com/marketing/dont-let-it-bother-you-just-continue-streaming-confessions-of-a-twitch-streamer-and-victim-of-online-hate-raids/. Grayson, "More than a Twitch Problem."
[30] Andy Chalk, "More Hate Raids Strike Twitch as White Supremacist Takes Credit," *PC Gamer*, March 14, 2022, https://www.pcgamer.com/more-hate-raids-strike-twitch-as-white-supremacist-takes-credit/.
[31] "Hate Is No Game," Anti-Defamation League.
[32] Anya Kamenetz, "Right-Wing Hate Groups Are Recruiting Video Gamers," *NPR*, November 5, 2018, sec. Health News, https://www.npr.org/2018/11/05/660642531/right-wing-hate-groups-are-recruiting-video-gamers.

of minors ages 10-17 report encountering white extremism when gaming.[33] Some game companies even appear to tolerate extremist content: white extremism has historically been common on Steam, the biggest online storefront and forum for PC games.[34] For years, Steam preserved and allowed users to continue "liking" the extremist hate speech of Brenton Tarrant, only removing his profile after he committed a mass shooting at a mosque in Christchurch, New Zealand.[35]

Gaming harassment doesn't stay online, either. One in six adult gamers polled in a large 2022 survey had their address, phone number, pictures or other personal information leaked online.[36] One in ten had false police reports filed against them.[37] This has sometimes led to police arriving at gamers' homes with guns drawn, a form of harassment called "swatting."[38] (Swatting involves searching out a person's home address and reporting a fake, severe public safety threat at that address, like a shooting or hostage situation.[39]) In one well-known incident, an online gaming dispute led to a fatal police shooting after one gamer swatted the other.[40] But even when the consequences are less extreme, it's no surprise that gamers report feeling depression, isolation, and fear due to being harassed, or that gamers often quit games where they're targeted for abuse.[41]

IV. Games' Responses and Their Shortfalls

Gaming companies know that they have a harassment problem and know that it's bad for business. Riot Games concluded that "toxic behavior" drove gamers away from the League of Legends game and hurt its image.[42] A top Meta executive, Andrew Bosworth, declared that harassment poses an "existential threat" to the company's virtual worlds and called for "almost Disney levels of safety."[43]

---

[33] "Exposure to White Supremacist Ideologies in Online Gaming Doubled in 2022, New ADL Survey Finds," Anti-Defamation League, December 7, 2022, https://www.adl.org/resources/press-release/exposure-white-supremacist-ideologies-online-gaming-doubled-2022-new-adl.

[34] "This Is Not a Game: How Steam Harbors Extremists" (Anti-Defamation League, April 2020), https://www.adl.org/resources/report/not-game-how-steam-harbors-extremists.

[35] "This Is Not a Game," Anti-Defamation League.

[36] "Hate Is No Game," Anti-Defamation League. "Online Harassment in the U.S.," Statista.

[37] "Hate Is No Game," Anti-Defamation League.

[38] "Hate Is No Game," Anti-Defamation League. "Online Harassment in the U.S.," Statista.

[39] Amelia Heidenreich, "Why Are There Live Streamers That Get Swatted?," *Quora*, accessed May 2, 2023, https://www.quora.com/Why-are-there-live-streamers-that-get-swatted. Kelsey Krahn, "What Is Swatting and How to Prevent It - 5 Tips to Protect Yourself," March 4, 2020, https://onerep.com/blog/how-to-reduce-your-risk-of-being-swatted-5-tips-that-will-get-you-off-the-hook.

[40] "Wichita Man Sentenced in 'Swatting' Case That Led to Death," KWCH, September 26, 2022, https://www.kwch.com/2022/09/26/wichita-man-sentenced-swatting-case-that-led-death/.

[41] "Hate Is No Game," Anti-Defamation League.

[42] Kenneth Shores et al., "The Identification of Deviance and Its Impact on Retention in a Multiplayer Game," 2014, 1356–65, https://doi.org/10.1145/2531602.2531724.

[43] Adi Robertson, "Meta CTO Thinks Bad Metaverse Moderation Could Pose an 'Existential Threat,'" *The Verge*, November 12, 2021, https://www.theverge.com/2021/11/12/22779006/meta-facebook-cto-andrew-bosworth-memo-metaverse-disney-safety-content-moderation-scale.

The gaming world notices when prominent players migrate to new games to protest toxic behavior in games and gaming studios.[44]

Game companies also do, by and large, take aim at harassment.[45] They publish codes of conduct and terms of service prohibiting toxic behavior. They collaborate with nonprofit organizations to improve their practices.[46] They publish win-loss statistics to show that teamwork, not trolling, is the most effective way to win games: Riot Games, for example, appealed to gamers' self-interest when it released statistics showing that "rage doesn't win games."[47] Games hire game designers to prevent bad behavior by "foster[ing] healthy social environments."[48] They reward players for prosocial behavior.[49]
But above all, game companies react to unacceptable behavior with sanctions. Game employees and automated tools comb over gamers' chats and other user-generated content to identify and penalize players for harassment. On game platforms, this content moderation falls into several key categories:

- **Muting**: the game blocks a gamer's voice or text chat for offenses such as using banned words.[50]
- **Shadowbanning**: the game mutes or isolates a player without telling them, boots a player from the game without an explanation,[51] or makes the game unusable for a player.[52]
- **Muting and Blocking (gamer-initiated):** the game allows a gamer to mute a fellow player temporarily or to block their communications entirely.[53]

---

[44] Matt Craig, "On YouTube and Twitch, the Activision Blizzard Harassment Lawsuit Leaves Creators Reeling," *Washington Post*, August 10, 2021, https://www.washingtonpost.com/video-games/2021/08/09/activision-blizzard-asmongold-content-creators/.
[45] In some states, moderation practices may be directed by law. *See*, for example, Karen Gullo, "Court's Decision Upholding Disastrous Texas Social Media Law Puts The State, Rather Than Internet Users, in Control of Everyone's Speech Online," *Electronic Frontier Foundation* (blog), October 6, 2022, https://www.eff.org/deeplinks/2022/10/courts-decision-upholding-disastrous-texas-social-media-law-puts-state-rather. Thomas Claburn, "Supreme Court Asked to Affirm Florida Content Moderation Law," The Register, September 23, 2022, https://www.theregister.com/2022/09/23/florida_supreme_court_moderation_law/. "California Enacts the California Age-Appropriate Design Code Act," *Hunton Andrews Kurth Privacy & Information Security Law Blog* (blog), September 16, 2022, https://www.huntonprivacyblog.com/2022/09/15/california-enacts-the-california-age-appropriate-design-code-act/.
[46] *See, for example,* Robert Lewington (Twitch) and The Fair Play Alliance Executive Steering Committee, "Being 'Targeted' about Content Moderation: Strategies for Consistent, Scalable and Effective Response to Disruption & Harm" (Fair Play Alliance, April 2021), https://fairplayalliance.org/wp-content/uploads/2022/06/FPA-Being-Targeted-about-Content-Moderation.pdf.
[47] Wesley Yin-Poole, "Riot Uses Stats to Prove League of Legends Ragers Lose More Often," *EuroGamer*, September 12, 2013, https://www.eurogamer.net/riot-uses-stats-to-prove-league-of-legends-ragers-lose-more-often.
[48] "Game Designer III - Behavior Evaluation and Moderation at Riot Games," GameJobs.co, accessed April 10, 2023, https://gamejobs.co/Systems-Designer-III-Player-Platform-Behavior-Evaluation-and-Moderation-at-Riot-Games.
[49] "Updated Honor Rewards," League of Legends, August 31, 2022 https://www.leagueoflegends.com/en-us/news/game-updates/updated-honor-rewards/.
[50] SeaFoamBeam, "Chat Moderation - What Happens When Players Use Words from the Blocked List," League of Legends Support, April 10, 2023, https://support-leagueoflegends.riotgames.com/hc/en-us/articles/10728400692243-Chat-Moderation-What-happens-when-players-use-words-from-the-Blocked-List.
[51] Joseph Yaden, "Warzone 2.0 Is Shadowbanning Players Who Get Too Many Kills," Digital Trends, January 9, 2023, https://www.digitaltrends.com/gaming/warzone-2-shadowban-players-for-earning-kills/.
[52] Brian Barrett, "Instead of Banning Cheaters, Pokémon Go Trolls Them," *Wired*, accessed May 29, 2017, https://www.wired.com/2017/05/pokemon-go-cheaters-shadowban/.
[53] See, for example, Skittle Sniper, "How to Mute and Block Players," League of Legends Support, April 10, 2023, https://support-leagueoflegends.riotgames.com/hc/en-us/articles/4405479481363-How-to-Mute-and-Block-Players.

- **Flagging (gamer-initiated)**: gamers report harassing players to a disciplinary committee for action.[54]
- **Temporary Suspensions and Permanent Bans:** the game temporarily or permanently removes offending players from a game.
- **Behavioral Ratings or Reports**: the game scores gamers using other players' feedback on their behavior.[55] High scores win privileges such as playing other high-ranked players.[56]

When content moderation goes right, it appears to benefit companies and players—though both sides express reservations. Proponents of moderation associate it with improved gamer retention, increased playing time,[57] an expanded user base, and an improved company image.[58] But content moderation is expensive, and it alienates players who go online to push the boundaries of acceptable speech (moderation's opponents call it "fascism," "communism," and worse[59]). These are two strong incentives for game makers to underinvest in the technology and manpower needed to rid platforms of harmful content.[60] Not all companies value historically marginalized gamers' business, either, especially in comparison to the significant cost of moderation tools. As Boston University Professor and Playmatics founder, Margaret Wallace, told S.T.O.P.:

> Third-party tools for content moderation are constantly evolving. They also provide varying degrees of efficacy. Once a game developer or publisher has committed time, money, and resources to use a specific community moderation solution, it can be difficult to migrate seamlessly from one offering to another should the game developer discover an alternative platform that better suits their needs. Additionally, because video games are a global business—not all key industry stakeholders will have the same mindset and approach to matters concerning LGBTQ+ players and those relating to diversity, equity, and inclusion (DEI) in general.[61]

For their part, gamers see companies' uncommitted approaches to moderation and accuse them of performing "moderation theater" to protect their bottom lines rather than actually ending harassment in games.[62] As one Twitch user put it, the platform expanded its banned words list and publicly congratulated itself while harassers switched to unbanned spellings of their favorite slurs.[63] And while moderation may be meant to protect historically marginalized gamers from abuse,

---

[54] *See, for example*, Skittle Sniper, "Player Reporting Guide and FAQ."
[55] *See, for example*, "Conduct Summary," Dota 2 Wiki, accessed April 10, 2023, https://dota2.fandom.com/wiki/Conduct_Summary.
[56] "Conduct Summary," Dota 2 Wiki.
[57] "Gaming," Two Hat, accessed April 10, 2023, https://www.twohat.com/solutions/gaming/.
[58] Yi Liu, Pinar Yildirim, and Z. John Zhang, "Implications of Revenue Models and Technology for Content Moderation Strategies," SSRN Scholarly Paper (Rochester, NY, November 23, 2021), https://doi.org/10.2139/ssrn.3969938.
[59] Max Boot, "Elon Musk Is the Last Person Who Should Take over Twitter," *Washington Post*, April 15, 2022, sec. Opinion, https://www.washingtonpost.com/opinions/2022/04/14/elon-musk-should-be-last-person-twitter-hostile-takeover/.
[60] Liu, "Implications of Revenue Models. "Caught in a Vicious Cycle: Obstacles and Opportunities for Trust and Safety Teams in the Games Industry" (Anti-Defamation League, March 28, 2023), https://www.adl.org/resources/report/caught-vicious-cycle-obstacles-and-opportunities-trust-and-safety-teams-games.
[61] Margaret Wallace, S.T.O.P. Interview with Margaret Wallace, April 6, 2023.
[62] Ryan Hatch, "Can Gaming Save Itself from Its Content Moderation Problems?," Medium, January 1, 2021, https://uxdesign.cc/can-gaming-can-fix-its-content-moderation-problems-83dcfc8f7e83.
[63] Hatch, "Can Gaming Save Itself?"

centralizing the review of users' speech and other personal content is at odds with cybersecurity and privacy best practices.[64] It's not hard to imagine the danger that centralized data on historically marginalized gamers and their moderated content poses: gamers could be outed or ostracized, discriminated against by employers and others, and worse.

Industry-standard moderation (whether automated or human) don't just fall short of protecting historically marginalized gamers: it often punishes the very gamers it is supposed to protect. Our survey of gamers and their online conversations shows that historically marginalized gamers frequently feel that they are penalized for revealing their identities. According to one player, "I got banned for saying my nationality... I told him "I'm Paki" within 2 seconds I got banned… the ban system it's self is racist not the players" [this being a reference to Pakistani heritage that is often viewed as derogatory in the United Kingdom].[65] Black gamers in particular feel that they are penalized for asserting their Blackness through speech. As one gamer put it: "I'm a black man that lives in an urban community so I do use the N word but not in a man are that harms anyone… it's so deeply rooting in our culture… I don't think I deserve a permanent ban from Xbox."[66] Historically marginalized gamers also frequently feel that they are penalized for responding to harassment. As one gamer explained regarding the circumstances that got them banned from PlayStation: "I only responded to him calling me a gay boy saying he got killed by a gay boy."[67] We found that games make these bad calls over and over again (*see* Table 1).

[64] Jun Li, "Why Decentralization Is Essential for Protecting User Data and Privacy," *Forkast*, July 6, 2021, https://forkast.news/why-decentralization-protect-user-data-privacy/. Brooke Tanner and Samantha Lai, "Examining the Intersection of Data Privacy and Civil Rights," *Brookings* (blog), July 18, 2022, https://www.brookings.edu/blog/techtank/2022/07/18/examining-the-intersection-of-data-privacy-and-civil-rights/.

[65] This report retains the exact text written by gamers in quotations to preserve the original context. nessobeatz, "I Got Banned for Saying My Nationality...," Reddit Post, *R/Rainbow6*, July 14, 2018, https://www.reddit.com/r/Rainbow6/comments/8ysvtr/i_got_banned_for_saying_my_nationality/.

[66] FadeD InTruzion, "Banned for Inappropriate Language," January 13, 2022, https://answers.microsoft.com/en-us/xbox/forum/all/banned-for-inappropriate-language/7ad701ca-11f6-4309-ae42-510db8b1265e.

[67] "Banned of the Ps Network for Stating I Was Gay. Was Told to Wait It Out," Reddit Post, R/Playstation, June 23, 2022, https://www.reddit.com/r/playstation/comments/vj0qld/comment/idg4v74/?utm_source=share&utm_medium=web2x&context=3

TABLE 1: BAD CALLS AGAINST HISTORICALLY MARGINALIZED GAMERS

BANNED FOR BEING
My PSN is Gaymface. I'm a member of the LGBT community and have proudly been playing under this name for 5+ years. Suddenly today Ive been suspended for 7 days and my username changed to a Temp name because it "violates the community code of conduct." (PlayStation Network)[i]
I ran into a Lebron look alike. The guy wasn't playing well... I messaged him "You not bron [n word]…" Im a black man too and Im not trying to politic who can and cant use that word but i meant in no harm or ill manner. I received a permanent ban and it was my first offense. (Xbox Live)[ii]
A warning to any LGBT+ PlayStation players: Remove anything that even slightly hints that you are LGBT+ from your profiles... SIE [Sony Interactive Entertainment] moderation suspended me for a week after my profile's "About me" stated that I was gay, their reason being hate speech... my whole bio here for context: "Uber gay. That's about it for whatever I think I'm meant to put here lmao." (PlayStation Network)[iii]
I was permanently banned… they said it was because my Online ID was offensive… My online ID is/was: Kike_0615. Kike is short for my real name, Enrique, and in Mexico a lot of people call guys with that name "Kike". It is also pronounced "KEE-KAY". After I did some research it's apparently used as an offensive way of calling Jewish people… I'm really sorry if I offend anyone. (Playstation)[iv]
Got banned for saying (my n word) then i said you suck) im black and I didn't used as a racial slur... I would've understood if it was meant to be racist or something it wasn't offensive at all literally and my appeals did not get approved. (Xbox Live)[v]
Roblox banned me last week for saying "I'm Jewish"... they said it was hate speech. (Roblox)[vi]
Perma muted another girl for having trans rights in her user name. (Mount & Blade)[vii]
I got a seven-day ban off of Roblox for saying negro, but I'm black and I was joking with my friends. (Roblox)[viii]
I can not call my ship "The Sapphic Salmon"... Every game under Microsoft/Xbox Game Studio has this banned, because lesbians are nothing more than a porn category to them. (Sea of Thieves)[ix]
Hey, I recently noticed that the words "gay" and "homosexual" are censored and still don't understand why. (Overwatch)[x]
The user... claims she was banned from global chat for referring to herself as trans and gay. This user and others shared screenshots that appear to show Tabletop Simulator banning any user who says the phrase "I'm gay." (Tabletop Simulator)[xi]
BANNED FOR RESPONDING
Yes, counter speech is super important, but this is where mods usually remove and warn you instead of fascists.[xii]
A guy got me suspended from my playstation for a week… He said some extremely racist and antisemitic stuff on overwatch and then proceeded to message me on my ps account baiting me into paraphrasing what he said and reporting my message. (Overwatch) [xiii]
So I report them ofc, but say in chat "hey, let's not make fun of people for neurodivergence and mental illness"... I got a 1 day ban... They [appeals] sent back that line, listing it as discriminatory speech. That's upsetting to me as an also mentally ill person. (Roblox)[xiv]
He decided to start calling me racial slurs. The n word with the ER at the end... I did call him a "stupid *ss [removed]" and told him to leave me alone. I'm black and I didn't say anything about any group of people but he's allowed to say "[removed] are poor and can't make any money"… first my account was banned for 7 days… [then] banned for 60 days. (PlayStation Network) [xv]
I was banned from Roblox for a 7 day period for supposed "racism" and "harassment." When in reality... this man thought it was OK to tell me to go back to my plantation (my avatar is Lil Nas X), so ... I stood up for myself. (Roblox)[xvi]
I got a 24 hour ban for calling some guy a (won't repeat...but it is a BORDERLINE bad word) after he sent multiple sexist messages to me after dying to me in a video game. I blocked him, I reported him. Never received any sort of email that it was being looked at. (Xbox Live)[xvii]
I just got banned for 7 days for backing someone up when someone said they had looked gay. The other person was being homophobic towards them, but I get the ban for saying "how does he look gay?" (Roblox)[xviii]
I use a 'No Room for Racism' badge in Fifa, and when I refused to immediately concede a game after he scored first, I got a pile of 'White Lives Matter' type messages from some creep. I didn't respond with abuse, but reported him instead. Now, I found out that my account is suspended for 7 days!? ...there doesn't seem to be a way to appeal. (PlayStation Network)[xix]
I mean seriously this guy had (Mod Removed) in his name so I said to him you're a racist (Mod Removed) and MY ACCOUNT GOT BANNED?! (Xbox Live)[xx]
It's really fun also to be muted for an entire day because I say "toxic losers don't get opinions" to shut them down...Seriously, can we get actual humans checking appeals and stuff instead of just bots? (Mobile Legends)[xxi]
I got banned for defending people with autism. (Roblox)[xxii]
I got banned for telling someone to stop saying F*gg*t. (League of Legends][xxiii]
"I was banned for reporting about N-word usage on SoM [Season of Mastery] servers… [Blizzard said] "Calling for disciplinary action in the forums or discussing it in anyway is a violation of the code of conduct."(Blizzard Forums)[xxiv]

V. Notice and Appeals: Santa Clara Principles for Games

Repeated bad calls silence and alienate historically marginalized gamers. It may not be game companies' intention to create a hostile environment for historically marginalized gamers. But that isn't clear, especially because moderation decisions are often announced without any justification:

- "There was no explanation on why I was banned." (after responding to racism)[68]
- "I woke up today and found out my account was banned for 60 days out of nowhere." (after an initial week-long ban for responding to racist hate)[69]
- "I recently noticed that the words "gay" and "homosexual" are censored and still don't understand why."[70]

In other cases, game publishers may provide an explanation for their decisions—but their misguided reasons leave gamers feeling excluded and aggrieved, and often without an obvious path to right perceived wrongs:

- "The word gay is not a slur, and was not used as an insult but as a personal adjective. No policies were broken…. I did approach support but they just brushed me off and gave copy/paste answers, unfortunately." (re: ban for user bio saying "Uber gay")[71]
- "I just got a one day ban for saying I'm autistic, which I am…I tried to appeal but was told that "autistic" is a slur that constitutes hate speech. So yeah, I guess they want us to shut up and keep it to ourselves. While it's only day and not a big deal, it still feel like crap to be marginalized by the moderators. It is clear what was happening in the context of the conversation."[72]

The Santa Clara principles for content moderation—crafted by human rights advocates for social media platforms, readily adapted for games—provide helpful guidance here. Endorsed (and imperfectly implemented) by platforms including Apple, Meta, Google, Reddit, Twitter, and GitHub,[73] the principles address the concerns of gamers who suspect bad calls by providing reasons for moderation to moderated individuals and the ability to appeal suspected bad calls.

---

[68] Dialogue Amongst Friends, "Banned for What," July 9, 2022, https://answers.microsoft.com/en-us/windows/forum/all/banned-for-what/d5eda784-1c39-4619-9ea8-70bd487c3b21.

[69] "How Long Has PlayStation Been Supporting Racism (Banned after I Was Called Racial Slurs)," January 11, 2022 https://us.community.sony.com/s/question/0D54O00007GwWP8SAN/how-long-has-playstation-been-supporting-racism-banned-after-i-was-called-racial-slurs?language=en_US.

[70] "Why Are These Words Censored? - General Discussion," Overwatch Forums, August 5, 2019, https://us.forums.blizzard.com/en/overwatch/en/overwatch/t/why-are-these-words-censored/381985.

[71] justpostingbugsifind, "A Warning to Any LGBT+ PlayStation Players," Reddit Post, *R/Playstation*, October 27, 2021, www.reddit.com/r/playstation/comments/qgwo2d/a_warning_to_any_lgbt_playstation_players/.

[72] Shadocat42, "I Got Banned for Defending People with Autism. Since When Was Any of This against Roblox Rules? If Roblox Was Boycotted or Sued for Their Moderation System I Say That They Deserve It 100%. [Putting Full Story in Comments]," Reddit Post, *R/ROBLOXBans*, August 26, 2022, https://www.reddit.com/r/ROBLOXBans/comments/wyk3bu/comment/in3ziq1/?utm_source=share&utm_medium=web2x&context=3.

[73] Access Now et al., "Santa Clara Principles on Transparency and Accountability in Content Moderation," Santa Clara Principles, accessed May 18, 2023, https://santaclaraprinciples.org/.

The notice principle requires companies to be transparent about how and why they moderate speakers. Applied to video games, notice would require that companies give gamers adequate notice following content moderation by:

- Informing a gamer of the specific behavior that resulted in moderation;
- Indicating who or what flagged the behavior (while preserving gamers' privacy)[74];
- Specifying which clause in the game's terms of service was violated;
- Providing gamers who flag others' behavior with a log of the outcomes of their reports.[75]

Notice begins to build a fairer system. But it's not enough—as Emma Llansó, director of the Free Expression Project at the Center for Democracy and Technology, points out, bad actors will always find a way to abuse systems, and companies need a way to remedy that abuse.[76] To do justice by individual gamers, moderation programs must provide an easy way for gamers to appeal seemingly bad calls. The Santa Clara appeal principle—modified slightly for games—requires that games provide moderated gamers with:

- Easy initiation of an appeal;
- A timeline of the appeal process once it begins;
- Human review by persons who are culturally competent;
- The opportunity for the user to provide additional information; and
- A final decision with sufficient reasoning for the final determination.[77]

Appeals processes help human and automated moderators continually learn cultural competence from their gamers. Benign words become slurs and are in turn rehabilitated by the people targeted with those slurs.[78] Symbols that continue to stand for hate in some contexts get repurposed for different ends in others.[79] Unless game companies commit to learning from gamers' appeals, they are bound to miss when they moderate.

In fairness, gaming moderation poses challenges that social media moderation does not. Social media platforms like Facebook moderate easier-to-scan text and images as well as harder-to-moderate videos and live interactions.[80] Game companies, by contrast, struggle to keep up with an enormous volume of constant, live voice chat using tools that are slow and inaccurate.[81] Virtual reality games poses the same problem of keeping up and the additional and possibly insurmountable

[74] In the gaming context, it may be easy to identify a reporting player if the reported and reporting gamers are participating in a small group, such as a raid or player *vs.* player event. This risk of identification increases the risk of retaliation against the player who made the report. It may be best in many cases to not provide enough information to identify the reporting player unless the reporting player indicates that they are comfortable with the risk of identification.
[75] "Santa Clara Principles."
[76] Emma Llansó, S.T.O.P. Interview with Emma Llansó, April 20, 2023.
[77] "Santa Clara Principles."
[78] Brittan Heller, "Is This Frog a Hate Symbol or Not?," *The New York Times*, December 25, 2019, sec. Opinion, https://www.nytimes.com/2019/12/24/opinion/pepe-frog-hate-speech.html.
[79] Heller, "Hate Symbol or Not?"
[80] Christianna Silva, "How Is Facebook Going to Moderate Notoriously Difficult Live Audio?," Mashable, April 27, 2021, https://mashable.com/article/facebook-audio-moderation.
[81] Otto Söderlund, "Why Games Need Better Voice Chat Moderation," Speechly, October 24, 2022, https://speechly.com/blog/why-games-need-better-voice-chat-moderation.

challenge of interpreting individuals' behavior correctly.[82] And while game companies may turn to automated and AI-based tools to deal with the sheer volume of content that gamers generate, this may simply scale up games' biased moderation practices: as Brittan Heller, human rights lawyer and technologist, told S.T.O.P., "AI may increase gaming companies' capacity to provide content moderation. However, since preexisting biases in data sets can reemerge in unexpected ways, game makers and worldbuilders need to anticipate challenges. It's essential to understand the way moderation systems have harmed users in the past – and may continue to do so in the future, if we are not mindful."[83]

If anything, games' increased likelihood of making bad calls means that notice and appeals systems are as or more important for gaming than for social media. But not a single major gaming company appears to have endorsed or even publicly discussed the Santa Clara principles.[84] Until now, no one even appears to have asked them to do so.

VI. A Word of Caution about Over-Policing Gamers

Game publishers can begin to address bad calls against marginalized gamers by implementing the Santa Clara principles, which would revolutionize the game industry's fairness and cultural competence. But appeals systems still burden historically marginalized gamers with justifying their innocent conduct and educating game companies. To be truly inclusive, games need to do more—or rather, they need to do less, and exercise restraint when moderating.

In this vein, we suggest two changes to games' rules of moderation: (1) any keyword bans must take context into account and avoid flagging historically marginalized gamers for asserting their identities or responding to abuse, and (2) overbroad rules like "no politics" must be specified and limited so that they, too, can't be used to force historically marginalized gamers off platforms or into hiding.

Context Matters

Any keyword bans must take context into account and avoid flagging historically marginalized gamers for asserting their identities or for responding to hate. Game companies say that they welcome all gamers, but many persist in banning historically marginalized gamers' "key words." A leaked copy of Minecraft Bedrock's 2023 banned words list includes words reappropriated by some members of the LGBTQ+ community.[85] It also includes several variations on the "n word," a measure that penalizes Black gamers for asserting their Blackness while failing to protect them from abusive gamers using alternate spellings of the word.[86] Many censured words take meaning from

[82] Aaron Mak, "Metaverse Moderation: I Was a Virtual Reality Bouncer," *Slate*, May 9, 2022, https://slate.com/technology/2022/05/metaverse-content-moderation-virtual-reality-bouncers.html.
[83] Brittan Heller, S.T.O.P. Interview with Brittan Heller, April 21, 2023.
[84] S.T.O.P. checked each of the companies in this list. "The Top 20 Gaming Companies," Yahoo Life, March 8, 2023, https://www.yahoo.com/lifestyle/top-20-gaming-companies-210000307.html.
[85] Minigamerguy123, "I Found the Full List of Blacklisted Words on Bedrock Edition," Reddit Post, *R/PhoenixSC*, July 14, 2022, www.reddit.com/r/PhoenixSC/comments/vz7fd1/i_found_the_full_list_of_blacklisted_words_on/.
[86] Minigamerguy123, "Blacklisted Words on Bedrock."

their context, but game publishers routinely refuse to take context into consideration, citing scaling issues as a reason for contextless moderation approaches.[87] Some companies refuse outright, as a point of principle. PlayStation, for example, refuses to allow historically marginalized gamers to reclaim and defang slurs: "[t]here are no exceptions for using hateful language or slurs in any form on PSN, even if the context is lighthearted, non-serious, or used as reappropriation."[88] Other game companies leave principles aside, but use blunt automated processes to flag words regardless of context.[89]

The effect is the same: historically marginalized gamers get "banned for being" and "banned for responding" (*see* Table 1). (As one BIPOC gamer put it, "if I was at the table with Xbox whenever they decided this, I coulda told them black people are gonna get punished."[90]) When historically marginalized gamers get banned for innocent speech, they hear "you are not welcome" and conclude, as one autistic gamer put it, that "they want us to shut up."[91] Blake Goodman, an LGBTQ+ gamer and gaming scholar, agrees: "I don't feel welcome in online games. These days, you are either getting biased AI keyword monitoring or getting nothing. But banned keywords are not the best way to deal with the harassment problem."[92]

In conversation with S.T.O.P., game scholars and game designers emphasized the need to consider context when allowing or moderating gamers' speech. "There has to be a contextual understanding in content moderation," says Kat Schrier, Director of the Games and Emerging Media Program at Marist College.[93] "What is the culture and what is the context?"[94] said Mitu Khandaker, founder of Glow up Games. "We let players curse in our game, to a limit. It's a game about rap, and profanity is an important part of expression in hip-hop."[95]

Still, the gaming industry has balked at the supposed difficulty of taking context seriously, given the challenge of resolving rare edge cases.[96] But leaving perfection aside, moderation programs could implement available measures to make many fewer bad calls against historically marginalized gamers.

---

[87] Mike Masnick, "Masnick's Impossibility Theorem: Content Moderation At Scale Is Impossible To Do Well," Techdirt, November 20, 2019, https://www.techdirt.com/2019/11/20/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well/. Furthermore, while these tools recognize protected categories like race or sex, they are unable to grapple with the nuance of quasi- or nonprotected categories like age. This leads to instances where categories like "Black children" (race-age) are provided less protection from harassment than "white men" (race-sex). *See* Ángel Díaz and Laura Hecht-Felella, "Double Standards in Social Media Content Moderation" (Brennan Center for Justice), August 21, 2021, https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation.

[88] "PlayStation Policy against Hate Speech on PlayStation Network (US)," accessed April 14, 2023, https://www.playstation.com/en-us/support/account/playstation-hate-speech-policy/.

[89] Jonas Strandell, "5 Content Moderation Methods You Should Understand," *Besedo* (blog), March 15, 2016, https://besedo.com/knowledge-hub/blog/5-moderation-methods-you-should-understand/.

[90] Emmanuel Ocbazghi, "Gaming While Black: How Racist Trolls Are Still Dominating Video Games," *Business Insider*, April 16, 2018, https://www.businessinsider.com/online-gaming-racist-xbox-live-playstation-twitch-microsoft-black-gamers-2018-4.

[91] Shadocat42, "I'm Glad You Got It …," Reddit Comment, *R/ROBLOXBans*, September 4, 2022, www.reddit.com/r/ROBLOXBans/comments/wyk3bu/i_got_banned_for_defending_people_with_autism/in3ziq1/.

[92] Blake Goodman, S.T.O.P. Interview with Blake Goodman, April 6, 2023.

[93] Kat Schrier, S.TO.P. Interview with Kat Schrier, March 16, 2023. Kat Schrier is the Author of *We the Gamers: How Games Teach Ethics and Civics*, 2021, Oxford University Press.

[94] Schrier, S.TO.P. Interview with Kat Schrier.

[95] Mitu Khandaker, S.TO.P. Interview with Mitu Khandaker, March 13, 2023.

[96] Masnick, "Masnick's Impossibility Theorem."

Automated tools struggle to understand context,[97] but games can trade automated tools that ignore context for ones that at least try to understand and charitably interpret historically marginalized gamers.[98] Game moderators can ask gamers to greenlight or redlight conversation partners' speech before making moderating calls. They can offer gamers a chance to be heard *before* penalties are applied, instead of post-moderation. They can periodically meet with historically marginalized gamers to make and revise moderated word lists,[99] reducing the burden on those gamers to later justify themselves in adversarial appeals. Likewise, they could permit players to create their own communities within the game, set their own moderation policies for those environments, and give them the tools to enforce their rules.[100] These measures don't preclude the need for appeals to address bad calls. But an appeals-centered system puts the burden on blameless historically marginalized gamers to defend themselves—and as one BIPOC gamer put it, "I don't feel I should have to."[101]

Banning Politics is a Political Statement

Overbroad rules such as "no politics," "no offensive content," and "no ruining others' experiences" must be specified so that they cannot be used to force historically marginalized gamers, whose identities are inherently politicized, off of platforms or into hiding. Telling marginalized gamers to "leave politics at the door"—or even, effectively, "check your identity at the door"—reinforces the notion that straight white cisgender men are the intended audience for video games in the U.S. Gaming is supposed to be an escape from everyday reality, but admonishments to avoid politics and to avoid offending anyone put historically marginalized gamers on notice: watch out, revealing your identity or voicing concerns may get you banned. Consider, for example:

- We encourage showing off what makes you unique and awesome, but it's not cool to post something that keeps others from having positive experiences…. [Don't] use Xbox as a platform for discussing politics.[102]
- The spirit of inclusion lives in our values, which are key to sustaining a vibrant and welcoming community… [but] Mojang Studios affirms the Xbox Community Standards [banning not-positive content and politics].[103] (Minecraft)

[97] Evan Enzer, "TikTok Shadowbanned My Cat Videos," *In Less Than 30* (blog), April 2, 2023, https://medium.com/@InLessThan30/tiktok-shadowbanned-my-cat-videos-d9cec9030d6. Karen Hao, "AI Still Doesn't Have the Common Sense to Understand Human Language," *MIT Technology Review*, January 31, 2020, https://www.technologyreview.com/2020/01/31/304844/ai-common-sense-reads-human-language-ai2/. Charles Simon, "Your AI Doesn't Really Understand You," *Forbes*, April 12, 2022, https://www.forbes.com/sites/forbestechcouncil/2022/04/12/your-ai-doesnt-really-understand-you/.

[98] *See*, for example, "Contex AI - Content Moderation through A.I.," accessed May 16, 2023, https://www.contex.ai. While S.T.O.P. does not endorse the performance of any specific context considering AI tools, Cotex AI is one company attempting to provide a contextual solution to automated content moderation.

[99] Brittan Heller, "Hate Symbol or Not?"

[100] Emma Llansó, S.T.O.P. Interview with Emma Llansó. *See*, for example, Reddit provides community moderators with backend tools to moderate their community forums. "Moderation Tools - Overview," Reddit Mods, accessed May 16, 2023, https://mods.reddithelp.com/hc/en-us/articles/360008425592-Moderation-Tools-overview.

[101] Dialogue Amongst Friends, "Banned for What."

[102] "Xbox Community Standards," Xbox.com, accessed April 14, 2023, https://www.xbox.com/en-US/legal/community-standards.

[103] "Community Standards," Minecraft.net, April 13, 2023, https://www.minecraft.net/en-us/community-standards.

- PSN is for everyone, regardless of race, gender identity, sexual orientation… Do not create, upload, stream, or share any[thing] that is… offensive [to anyone].[104] (PlayStation)
- To promote our civil environment, Roblox prohibits… Real-World Sensitive Events… Political Content.[105]

Historically marginalized gamers get harassed for being who they are, but these rules discourage them from expressing concerns about identity-based hate or from asserting their identities in positive ways. To Black gamers, "no politics" means "you'll endure racist hate when gaming, but don't talk about Black Lives Matter." On Minecraft, "no politics" means that gamers can't talk about abortion, even though they routinely endure misogyny and gender-based harassment while gaming (the word "abortion" is banned in Minecraft Bedrock Edition).[106] One wouldn't expect people who deal with systemic, regular discrimination (in life and in gaming) to check their politics at the door and "just have fun"—but that's exactly what games seem to demand.

Rules against politics, offense and controversy make some gamers' intolerance historically marginalized gamers' problem. One World of Warcraft moderator shut down a guild of gay gamers so that it wouldn't attract homophobes who would spoil everyone's good time.[107] Its maker, Blizzard Games, backtracked under legal pressure,[108] but the game's standing policy of disallowing "language that could be offensive" and of requiring names to "be appropriate and inoffensive"[109] maintains the same line: historically marginalized gamers, don't get too comfortable.

"No politics" and similar rules also deputize gamers to harass their diverse peers in the name of community rules. As historically marginalized gamers reported:

- "We protested [transphobic behavior] by putting trans rights in our name too, they [community moderators] went on this big tirade how it's unfair to them to force politics on them."[110]
- "I had a post listing the accomplishments of black people after a person said that they didn't create anything. This post was removed because of guidelines… I get flagged all the time because people don't like the truth."[111]

Moderators (and moderation programs) must differentiate between legitimate user reports and simply racist, sexist, or intolerant ones. Requiring gamers who flag other gamers to submit detailed

[104] "PlayStation Network Code of Conduct (US)," accessed April 14, 2023, https://www.playstation.com/en-us/support/account/community-code-of-conduct.
[105] "Roblox Community Standards," Roblox Support, accessed February 10, 2023, https://en.help.roblox.com/hc/en-us/articles/203313410-Roblox-Community-Standards.
[106] Minigamerguy123, "Blacklisted Words on Bedrock."
[107] Elizabeth Winterhalter, "Venn Diagram of LGBTQ+ and Gaming Communities Goes Here," JSTOR Daily, June 2, 2021, https://daily.jstor.org/venn-diagram-of-lgbtq-and-gaming-communities-goes-here/ .
[108] Winterhalter, "Venn Diagram Goes Here."
[109] "Blizzard's In-Game Code of Conduct," Blizzard Support, accessed February 10, 2023, https://us.battle.net/support/en/article/42673.
[110] flanneluwu, "Devs Siding with Nazis," Reddit Post, *R/GirlGamers*, March 1, 2023, https://www.reddit.com/r/GirlGamers/comments/11erz7b/devs_siding_with_nazis/.
[111] Dialogue Amongst Friends, "Banned for What."

and specific reports may cut down on false flags and will certainly help human and automated moderators make better, more efficient determinations about flagged behavior.[112]

VII. Conclusion

Online games are vibrant sites of civil society. They also have a public safety problem—one that mainstream games address with policing-type "solutions." Even the best-intentioned content moderation systems surveil, police, adjudicate and punish gamers: as one guide to targeted moderation put it, "we can think of this as akin to the offline world of police work, where 'chasing down leads' is often only as effective as the quality of the leads provided…we can tailor the method of submitting a 'police report' such that the evidence required to judge the offence is built in."[113]

It's clear that these "policing" solutions have failed to establish an environment of public safety in games: hate and harassment are rampant. It's also clear that games' methods continue to harm historically marginalized gamers for simply being online and responding to seemingly inevitable identity-related abuse.

Game designers and gaming scholars told S.T.O.P. that the solution to games' public safety problem isn't more policing-centric moderation and isn't rocket science—but it does go against longstanding practices in the industry. Making games safe and welcoming for historically marginalized gamers should begin with hiring diverse teams of game developers to build games that "feel like they are for everyone."[114] Non-diverse teams have produced too many stereotyped and racist representations of diverse characters: BIPOC men cast as criminals and drug dealers; a Black woman whose superpower developers codenamed "Feminist Whore."[115] Diverse groups of game developers can do better: Khandaker told S.T.O.P. that Glow up Games' diverse team of developers draw from their own experiences to "build a space and set a context [that's authentic], influencing who comes to the game in the first place."[116] Hateful gamers may visit online worlds built by and for historically marginalized gamers, but they won't feel at home.

Major game studios are still overwhelmingly homogeneous,[117] but to their credit, they are experimenting with non-policing measures to discourage abuse.[118] In-game rewards for sustained good behavior are effective.[119] Announcing community rules works, too: when online communities

---

[112] Lewington, "Being 'Targeted.'"
[113] Lewington, "Being 'Targeted.'"
[114] Khandaker, S.T.O.P. Interview with Mitu Khandaker.
[115] Kishonna L. Gray, *Intersectional Tech: Black Users in Digital Gaming* (Baton Rouge: Louisiana State University Press, 2020).
[116] Our Team," *Glow Up Games* (blog), accessed April 12, 2023, http://glowup.games/our-studio/.
[117] Amber Burton and Biz Carzon, "Gaming Industry Diversity: How Activision Blizzard, EA, Unity and Riot Games Compare," *Protocol*, July 9, 2022, https://www.protocol.com/workplace/gaming-industry-diversity-reports. Jamal Michel, "Black Game Developers: Diversity Push Is Lots of Talk, Little Progress," *Washington Post*, January 31, 2022, sec. Perspective, https://www.washingtonpost.com/video-games/2022/01/31/black-game-developers-diversity-push-is-lots-talk-little-progress/.
[118] Brendan Maher, "Can a Video Game Company Tame Toxic Behaviour?," *Nature* 531, no. 7596 (March 1, 2016): 568–71, https://doi.org/10.1038/531568a.
[119] Rewards are statistically shown to incentivize player behavior. "Game Reward Systems," *Learning Theories* (blog), January 15, 2016, https://learning-theories.com/game-reward-systems.html.

state their rules against hate and harassment during discussions, more people feel welcome to participate.[120] If major game studios were to play to win, they'd move beyond moderation-heavy systems that center after-the-fact punishment to models that prioritize prosocial behavior, and the experiences of historically marginalized gamers, from the start.

[120] J. Nathan Matias, "Preventing Harassment and Increasing Group Participation through Social Norms in 2,190 Online Science Discussions," *Proceedings of the National Academy of Sciences* 116, no. 20 (May 14, 2019): 9785–89, https://doi.org/10.1073/pnas.1813486116.

Table 1 endnotes:

[i] Gaymface, "LGBT Username Suspended for 'Violating Code of Conduct,'" Reddit Post, *R/Playstation*, September 22, 2021, www.reddit.com/r/playstation/comments/psy88z/lgbt_username_suspended_for_violating_code_of/.

[ii] Miles3210, "Banned for Inappropriate Language," January 15, 2022, https://answers.microsoft.com/en-us/xbox/forum/all/banned-for-inappropriate-language/7ad701ca-11f6-4309-ae42-510db8b1265e.

[iii] justpostingbugsifind, "A Warning to Any LGBT+ PlayStation Players," Reddit Post, *R/Playstation*, October 27, 2021, www.reddit.com/r/playstation/comments/qgwo2d/a_warning_to_any_lgbt_playstation_players/.

[iv] Kikep0912, "Sony Banned Me for a Common Nickname in My Culture.," Reddit Post, *R/PS4*, November 22, 2018, www.reddit.com/r/PS4/comments/9zaddy/sony_banned_me_for_a_common_nickname_in_my_culture/.

[v] ..... :, "XBOX Enforcement TEAM," August 11, 2022, https://answers.microsoft.com/en-us/xbox/forum/all/xbox-enforcement-team/8c965d46-cc02-4d8c-a4fc-55598fe146fd.

[vi] blackpinkbmw, "Roblox Bans You for …," Reddit Comment, *R/ROBLOXBans*, August 27, 2022, www.reddit.com/r/ROBLOXBans/comments/wyk3bu/i_got_banned_for_defending_people_with_autism/ilzaery/.

[vii] flanneluwu, "Devs Siding with Nazis," Reddit Post, *R/GirlGamers*, March 1, 2023, https://www.reddit.com/r/GirlGamers/comments/11erz7b/devs_siding_with_nazis/ .

[viii] "I Got a Seven-Day Ban off of Roblox for Saying Negro, but I'm Black and I Was Joking with My Friends. Should I Appeal It or Wait It Out?," *Quora*, accessed April 14, 2023, https://www.quora.com/I-got-a-seven-day-ban-off-of-Roblox-for-saying-negro-but-Im-black-and-I-was-joking-with-my-friends-Should-I-appeal-it-or-wait-it-out.

[ix] lady_haybear, "'Sapphic' and 'Sappho' Are Banned Words in Sea of Thieve's Ship Naming System.," Reddit Post, *R/LesbianGamers*, August 23, 2022, www.reddit.com/r/LesbianGamers/comments/wvrbd0/sapphic_and_sappho_are_banned_words_in_sea_of/.

[x] WarZan, "Why Are These Words Censored? - General Discussion," *Overwatch Forums*, August 5, 2019, https://us.forums.blizzard.com/en/overwatch/en/overwatch/t/why-are-these-words-censored/381985.

[xi] Charlie Hall, "Tabletop Simulator Removes Global Chat, Developers Say Moderation 'Has Failed' Its Customers," *Polygon* (blog), January 12, 2022, https://www.polygon.com/tabletop-games/22879963/tabletop-simulator-moderation-homophobic-transphobic-global-chat.

[xii] flanneluwu, "Yes, Counter Speech …," Reddit Comment, *R/GirlGamers*, February 15, 2022, www.reddit.com/r/GirlGamers/comments/ssu13b/im_so_tired_of_right_wing_extremism_in_gaming/hx3n83a/.

[xiii] MistyWinchester, "A Guy Got Me Suspended from My Playstation for a Week," Reddit Post, *R/GirlGamers*, October 20, 2021, www.reddit.com/r/GirlGamers/comments/qby5wm/a_guy_got_me_suspended_from_my_playstation_for_a/.

[xiv] Lilly Murphy, "Why Did Roblox Ban Me for 7 Days Because of Saying the Word 'Gay'?," *Quora*, accessed February 17, 2023, https://www.quora.com/Why-did-Roblox-ban-me-for-7-days-because-of-saying-the-word-gay.

[xv] "How Long Has PlayStation Been Supporting Racism (Banned after I Was Called Racial Slurs)," accessed February 22, 2023, https://us.community.sony.com/s/question/0D54O00007GwWP8SAN/how-long-has-playstation-been-supporting-racism-banned-after-i-was-called-racial-slurs?language=en_US.

[xvi] Alexander Protan, "Why Did Roblox Ban Me for 7 Days Because of Saying the Word 'Gay'?," *Quora*, accessed April 14, 2023, https://www.quora.com/Why-did-Roblox-ban-me-for-7-days-because-of-saying-the-word-gay.

[xvii] DoeeKat, "I Reacted to an Inappropriate Sexist Message on Xbox and Got Banned," August 5, 2019, https://answers.microsoft.com/en-us/xbox/forum/all/i-reacted-to-an-inappropriate-sexist-message-on/1793fd38-d5b4-40c3-878c-36311cab42d9.

[xviii] Niko [@LawhornNiko], "@notedicista @Roblox_RTC @Crashstyler747_ @railworks2rblx I Know I'm a Little Late, but I Just Got Banned for 7 Days for Backing Someone up When Someone Said They Had Looked Gay. The Other Person Was Being Homophobic towards Them, but I Get the Ban for Saying 'How Does He Look Gay?,'" Tweet, *Twitter*, December 28, 2021, https://twitter.com/LawhornNiko/status/1475729421726269446.

[xix] Normanhouse, "Banned for Calling out a Racist," Reddit Post, *R/Playstation*, April 6, 2021, www.reddit.com/r/playstation/comments/mkz8xl/banned_for_calling_out_a_racist/.
[xx] kasemcmulin, "I Got Com Banned for Calling a Racist a Racist," February 12, 2022, https://answers.microsoft.com/en-us/xbox/forum/all/i-got-com-banned-for-calling-a-racist-a-racist/e5880acd-e0fc-4300-b64b-ee4748277a9a.
[xxi] CoreBear-was-taken, "So since the in Game Moderation Doesn't Want to Help...," Reddit Post, *R/MobileLegendsGame*, November 27, 2022, www.reddit.com/r/MobileLegendsGame/comments/z5ylrp/so_since_the_in_game_moderation_doesnt_want_to/
[xxii] Storycxde, "I Got Banned for Defending People with Autism. Since When Was Any of This against Roblox Rules? If Roblox Was Boycotted or Sued for Their Moderation System I Say That They Deserve It 100%. [Putting Full Story in Comments]," Reddit Post, *R/ROBLOXBans*, August 26, 2022, www.reddit.com/r/ROBLOXBans/comments/wyk3bu/i_got_banned_for_defending_people_with_autism/.
[xxiii] Xijinpingthirstpost, "riot games sucks" Reddit Post, *R/GirlGamers*, February 3, 2022, https://www.reddit.com/r/GirlGamers/comments/sjoe4l/comment/hvhhpxr/?utm_source=reddit&utm_medium=web2x&context=3.
[xxiv] Hathos_, "Update: I Was Banned for Reporting about N-Word Usage on SoM Servers," Reddit Post, *R/Classicwow*, October 24, 2021, www.reddit.com/r/classicwow/comments/qemhgw/update_i_was_banned_for_reporting_about_nword/.

Surveillance Technology Oversight Project, Inc.

40 Rector Street
9th Floor
New York, NY 10006

www.STOPSpying.org

STOP
SURVEILLANCE TECHNOLOGY OVERSIGHT PROJECT