INTERNET POLICY REVIEW

Volume 15 **Issue** 4

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Homogenising harm across realities: A comparative study of web 2.0 and XR community guidelines

Kyooeun Jang *University of Southern California*
Mike Ananny *University of Southern California*

DOI: https://doi.org/10.14763/2026.4.2133

Published: 26 November 2026
Received: 19 April 2025 **Accepted:** 1 March 2026

Funding: The authors did not receive any funding for this research.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.
Licence: This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 License (Germany) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. https://creativecommons.org/licenses/by/3.0/de/deed.en
Copyright remains with the author(s).

Citation: Jang, K., & Ananny, M. (2026). Homogenising harm across realities: A comparative study of web 2.0 and XR community guidelines. *Internet Policy Review*, *15*(4). https://doi.org/10.14763/2026.4.2133

Keywords: Platform governance, Extended reality, Content moderation, Institutional isomorphism, Trust

Abstract: This research examines how and why Extended Reality (XR) and Web 2.0 social media platforms, despite their technological differences, have remarkably standardised governance frameworks in defining and categorising online harm. Comparing the community guidelines of Meta's Horizon Worlds, VRChat, Roblox, and Zepeto against Meta's Community Standards, it identifies patterns of convergence and divergence. This analysis makes three key findings: 1) inter-organisational similarities between contemporary XR and Web 2.0 platforms exist alongside platform-specific innovations in addressing immersive harms; 2) historical echoes of early Web 2.0 guidelines appear in today's XR platforms, with adaptations for embodied interactions; and 3) intra-organisational similarities between Horizon Worlds and early Meta standards are complemented by newer approaches to immersive safety. This analysis highlights complexities of transplanting Web 2.0 governance models to immersive environments and suggests the need for governance approaches addressing unique characteristics of embodied digital interactions. By examining this transitional moment during which governance standards emerge, stabilise, and potentially constrain innovation, this study offers insights for navigating divergent regulatory frameworks such as the EU's Digital Services Act and US self-regulatory approaches as they confront embodied harms that fall outside frameworks built for text, image, and video content.

Introduction

In recent years, major technology companies have invested significantly in extended reality (XR) technologies, as demonstrated by Apple's Vision Pro launch, Meta's Quest 3S release, and rapid growth of XR platforms like Horizon Worlds and Rec Room, even as the commercial results have proven uneven. This technological expansion coincides with the maturation of Trust & Safety (T&S) frameworks across the platform industry, with Meta's comprehensive moderation framework becoming widely influential in shaping governance practices across platforms (Gillespie, 2018; Suzor, 2019).[1]

Despite this stabilisation of Meta's moderation framework as a dominant template, following Donald Trump's 2024 re-election, in January 2025 Meta announced significant policy changes (Kaplan, 2025). These included reduced focus on enforcing speech-related violations and elimination of fact-checking programmes, signalling a potential paradigm shift in platform governance approaches. As an industry leader whose practices often shape sector-wide frameworks, Meta's pivot raises questions about the future of platform governance, and calls into question how emerging XR technologies may inherit these evolving standards. These developments carry particular significance for European markets, where the Digital Services Act mandates transparency and accountability measures (European Union, 2022) that contrast sharply with US self-regulatory approaches. Yet the DSA was designed primarily with Web 2.0 social media platforms in mind, and researchers have noted its limited applicability to immersive environments (Owono & Morch, 2022; Lazerson, 2022). As XR platforms scale globally with their embodied, behavioural interactions, they increasingly navigate not only competing regulatory philosophies between the US and EU, but also the inadequacy of existing regulations designed for fundamentally different interaction modalities.

This inflection point highlights the need to better understand how governance frameworks change and, specifically, how they shift from Web 2.0 platforms and practices to newer immersive environments. Unlike traditional social media platforms that moderate discrete text and image content, XR platforms must govern embodied interactions including gestures, spatial proximity, and avatar-based expressions. The embodied nature of these interactions specifically means that governing harassment, consent, and user safety may require approaches that Web 2.0

1. Meta's Facebook/Instagram community standards have been extensively documented as influential templates for platform governance more broadly (Gillespie, 2018; Suzor, 2019). Newer platforms often adopt similar structures through direct emulation, hiring of former Meta employees, or responding to similar regulatory pressures that Meta has already navigated.

frameworks do not adequately address.

These challenges of governing immersive social spaces are not entirely new. Second Life, operating since 2003, pioneered approaches to moderating user behaviour, interactions, and creations in virtual environments (Boellstorff, 2008). Its early experiments offered potential community-driven templates for addressing challenges unique to immersive social spaces - templates that notably diverged from traditional web governance approaches. Yet contemporary XR platforms have largely drawn their governance frameworks not from this virtual world prehistory, but from Web 2.0 platforms, particularly Meta's content moderation approaches.

This study examines this tension across the community guidelines of four major XR platforms: Meta's Horizon Worlds, VRChat, Roblox, and Zepeto, comparing them against Meta's Community Standards. As the primary public expression of how platforms conceptualise harm, community guidelines are valuable proxies for understanding governance frameworks and their evolution during technological transitions (Gillespie, 2018).[2]

This study poses two primary research questions. These questions are interrelated, with the first examining the process of governance translation across platform types, and the second examining the institutional tensions that process produces. Building on the observation that contemporary XR platforms have largely adopted Web 2.0 governance templates rather than developing frameworks native to immersive environments, this study poses two primary research questions.

1. How do content governance mechanisms translate across platform types, particularly in migrations from Web 2.0 social networking platforms to social XR platforms? How are established frameworks reinterpreted and adapted for immersive environments?
2. How do platforms navigate between standardisation and platform-specific adaptations in their governance frameworks, and how do such navigations illustrate tensions among institutional templates and the novel challenges of governing immersive spaces?

For this study, Extended Reality (XR) refers to technologies that blend physical and digital environments, including virtual reality (VR), augmented reality (AR), and mixed reality (Morimoto et al., 2022). "Web 2.0" denotes the era of user-generated content and social networking platforms (roughly 2004-present) characterised by

2. This study focuses specifically on social XR platforms designed for socialisation and content creation rather than gaming, as these environments more closely parallel traditional social media in their governance challenges and user interaction patterns.

platforms like Facebook, Twitter, and YouTube that emphasise user participation and content creation (O'Reilly, 2005, Cormode & Krishnamurthy, 2008). By "immersive environments," this analysis refers to digital spaces where users experience embodied presence through avatars in synchronous, three-dimensional social contexts, regardless of whether they require VR headsets. The boundaries between these categories remain contested, particularly as technologies like Roblox and Zepeto enable immersive social experiences without requiring specialised XR hardware, as the platforms examined in this study illustrate.

This study focuses on content governance because of its lasting and often invisible power. As the middle layer of platform regulation that defines acceptable behaviour and guides enforcement decisions, it is situated between legal regulation (jurisdictional requirements) and content moderation (day-to-day enforcement) (Lazerson, 2022; Owono & Morch, 2022). Yet community guidelines occupy a hybrid legal status that resists easy categorisation. They are neither purely legal instruments nor merely aspirational codes of conduct. In some harm areas, platform governance intersects directly with legal obligations. Child sexual abuse material (CSAM), for instance, triggers mandatory reporting requirements to the National Center for Missing and Exploited Children (NCMEC) under U.S. law, while terrorism-related content implicates national security frameworks across jurisdictions. In other areas, community guidelines function as instruments of private risk management, addressing harms that are not illegal but carry reputational and commercial consequences for platforms, such as violent speech or content that degrades advertiser trust. These instruments operate as a form of private ordering that operates across multiple jurisdictions, functioning in practice as soft law whose force derives not from state enforcement but from contractual agreement and platform power (Harkai & Mezei, 2024). At the contractual level, users agree to these terms upon signing up and give platforms the right to moderate, making community guidelines enforceable through private means with real consequences for users (Mezei & Harkai, 2022).

In this study, community guidelines are best characterised as sociotechnical intermediary structures that span the full spectrum from legal compliance to platform-defined norms which translate both regulatory expectations and institutional values into operational governance. Through these governance frameworks, private platforms have become powerful architects of digital spaces (DeNardis & Hackl, 2015), making decisions that affect diverse stakeholders while raising concerns about the legitimacy and accountability of that power (Haggart & Keller, 2021).

2. Theoretical framework

What forces drive the evolution of platform governance, and what are the implications of such change–for understanding platforms, governance, and institutions–as these frameworks emerge, spread, and stabilise? This question becomes particularly salient as online interactions move from text-based interactions to immersive environments, prompting an urgent need to appreciate how governance change reflects and impacts definitions and moderations of user behaviour. This study focuses through the theoretical lenses of neo-institutional sociology and science and technology studies (STS), situating governance framework evolutions in terms of how these fields characterise technological transition. Taken together, these models show how institutional forces drive standardisation across the socio-technical dynamics of platform governance.

2.1. What forces shape the evolution of platform governance?

To define what exactly is being governed, this analysis adopts the definition of “platform societies" as assemblages composed of technology companies creating data and commerce marketplaces, civil society groups and user communities shaping platform communication, and state actors developing legal and regulatory regimes in response to company practices and user expectations (Van Dijck, Poell, & De Waal, 2018; Poell, Nieborg, & Duffy, 2021).

As intertwined dynamics of private actors, economic interests, public values, and technological infrastructures, governance of these platform societies resists easy characterisation (Van Dijck, 2020). Governance unfolds through a mix of voluntary self-governance, externally imposed governance, and relational governance amongst the stakeholders that the platforms convene (Gorwa, 2019). What makes platforms especially challenging to govern is that their varied histories, aims, forms of power, and constantly shifting technological capabilities make them unlike past, more stable organisational forms (Napoli & Caplan, 2017; Gorwa, Binns, & Katzenbach, 2020). Their governance demands simultaneously accounting for and holding accountable a loose collection of users, companies, and states that all lack absolute power to shape platforms, while still keeping them recognisable as platforms (Gorwa, 2024). Especially as platforms change, simply describing their governance can be conceptually and empirically challenging.

This study addressed this challenge by examining how the field of platform governance persists and changes through official community guidelines, and in light of the particular technological innovation of XR. The analysis focuses on (1) the in-

ternal and external forces that social media organisations contend with as they devise platform governance regimes that make sense for them and their competitors; and (2) the shared pressures created by new technological forms like XR that challenge assumptions about platform communication being largely asynchronous and text-based, foregrounding the complexity and power of embodied, real-time communication. This research is particularly interested in how the *field* of platform governance reveals its stabilities and frictions as it reorients content moderation for new eras of communication.

2.1.1. Organisational pressures

Internal forces shaping platform governance operate through three forms of institutional isomorphism as described by neo-institutional sociology: coercive pressures of regulatory requirements, mimetic processes of imitating successful competitors, and normative influences of professional standardisation (DiMaggio & Powell, 1983). Organisations frequently implement governance frameworks primarily to demonstrate legitimacy to stakeholders, regardless of whether these frameworks directly address their specific operational contexts, a phenomenon Meyer and Rowan (1977) characterise as "ritual conformity." According to Zucker (1977), organisational practices become institutionalised when they become "objective" (repetition no longer requires conscious meaning-making), and "exterior" (perceived as elements of a shared, stable reality).

Such conformity is difficult to change. Hannan and Freeman (1984) describe structural inertia as the relative speed at which organisations respond to environmental changes, distinguishing core elements like organisational identity (with high inertia) from peripheral features that change more frequently. Content governance frameworks are often core identity elements that sustain platforms' legitimacy with users, regulators, and advertisers, becoming difficult to change once operationalised. Complicating this conformity is the fact that platforms are intertwined with algorithms that, like institutions, serve as opaque and authoritative systems (Napoli, 2014), and with data-drive practices that become institutionalised and core to platform functions (Caplan and boyd, 2018).

External forces shaping platform governance operate through what Ananny and Gillespie (2016) call "cycles of shock and exceptions." Public outcries following critical events often result in platforms taking temporary measures that mask genuine transformation. This pattern is evident in Meta's reactions to major crises: the Rohingya massacre (Warofka, 2018), the 2016 U.S. presidential elections (Allcott and Gentzkow, 2017), and the January 6th Attack on the U.S. Capitol (Rosen and Bickert,

2021) all prompted scenario-specific solutions rather than fundamental governance changes.

Platform governance is also deeply dependent on U.S. legal frameworks, with the protections provided by Section 230 of the 1996 Communications Decency Act shaping distinctive approaches to free speech and privacy (Klonick, 2018). These U.S.-centric foundations create tensions as platforms operate globally, implicitly or explicitly carrying forward these U.S.-centric approaches into internal ethical frameworks that dominate when drafting policies for non-U.S. locales (DeNardis & Hackl, 2015). This pattern sustains the dominance of U.S. approaches in ways that disproportionately affect marginalised groups in the Global South especially (Owono & Morch, 2022). As governance practices travel, they rely on complex and often invisible translation processes (Czarniawska & Joerges, 1996).

2.1.2. Technological affordances

The transition to XR introduces fundamentally new governance challenges. Blackwell et al. (2019) found that users experience harassment more acutely in immersive environments through synchronous audio and heightened embodiments that make proximity and simulated touching feel 'real'. These new interaction modes challenge fundamental assumptions about what constitutes 'content' versus 'speech.' In Web 2.0 environments, content typically refers to discrete units that could be moderated after posting. However, XR platforms must contend with ephemeral, embodied expressions that blur these distinctions. This study draws a distinction between content-based moderation, which governs discrete, archivable material, and conduct governance, which governs real-time behavioural interactions. In XR environments, these categories may blur, as a single avatar action can simultaneously function as content, speech, and conduct, which makes it unclear whether platforms should treat it as material to be removed, expression to be protected, or behaviour to be regulated. When an avatar performs a gesture or enters another user's personal space, is this content to be moderated or speech to be protected? These questions become particularly complex when considering that avatars themselves can be forms of expression (Nakamura, 2020).

The temporal dimension of XR interactions further complicates governance. Unlike asynchronous Web 2.0 content that persists and can be reviewed, XR interactions often occur in real-time without permanent records, making retrospective moderation challenging. Additionally, the spatial nature of virtual environments introduces new categories of potential harm (Jang et al., 2023), such as virtual stalking, unwanted proximity, or blocking access to virtual spaces that have no direct ana-

logues in traditional social media. These technological affordances also create new forms of context collapse (Marwick & boyd, 2011), where behaviours appropriate in one virtual setting may be harmful in another. The same avatar gesture might be playful in a gaming context but threatening in a professional virtual meeting, requiring governance that accounts for situational factors beyond content alone.

2.2. How does platform governance standardise?

2.2.1. Understanding governance through standards

Science and Technology Studies shows how technical standards become established and resistant to change. Pinch and Bijker (1984) argue that standardisation occurs when relevant social groups reach consensus on a technology's meaning, with standards becoming organised through classification systems that provide “spatio-temporal segmentation of the world." (Bowker and Star, 2000, p. 10) This segmentation ensures that objects are categorised through consistent operational principles, becoming stable standards that structure sociotechnical engagements.

Standards become authoritative and powerful through their invisibility, by becoming "taken for granted" techniques that are rarely questioned (Star & Lampland, 2009, p. 10). Their "irreversibility" stems from their "network externalities," or being embedded in *other* existing infrastructures that ensure their durability through relational stabilities (Star, 1991). This stability allows some systems to become boundary objects (Star & Griesemer, 1989) that convene actors and practices, mediate interpretations, and make complex relationships predictable. Community guidelines can be thought of in these terms – as standards that classify and organise, ensure stability and predictability, and bring together varied stakeholders while maintaining core principles. They facilitate collaboration among users, platforms, and policymakers by providing shared understanding of acceptable behaviour. This analysis posits that the "structural inertia" (Hannan & Freeman, 1984) that such guidelines provide compels platforms to maintain Web 2.0 approaches even when facing novel challenges like changing media technologies.

2.2.2. Standards in the era of XR

The tension between standardisation and innovation becomes particularly acute as XR platforms challenge traditional governance categories. Legal definitions of speech have evolved significantly over time, expanding beyond written messages to encompass visual symbols, physical gestures, and other forms of expression as new technologies and social contexts emerged (Petersen, 2022).

This evolution raises fundamental questions about the applicability of Web 2.0 governance standards to XR environments. Traditional content moderation frameworks assume clear distinctions between content (which can be removed), speech (which may be protected), and behaviour (which can be regulated). In XR environments, these categories blur: an avatar's appearance might simultaneously be content, speech, and behaviour. Similarly, spatial dynamics in virtual environments challenge traditional notions of harassment. Just as legal definitions of speech have expanded to include symbolic gestures (Petersen, 2022), platforms now face the challenge of reconceptualising their governance frameworks to account for embodied expressions online.

In light of this evolution from Web 2.0 platforms to XR environments, this study focuses on a twofold platform governance question: how do community guidelines translate across platform types as governance frameworks move from social media to immersive environments, and how do institutional forces, standardised templates, and platform-specific adaptations shape the tensions that translation produces?

3. Methods

This study uses thematic analysis to examine community guidelines across four social extended reality XR platforms and Meta's Facebook platform, following Braun and Clarke's (2006) six-step latent deductive approach. This analysis focuses on social XR platforms rather than gaming environments, as social platforms more closely parallel traditional social media in their governance challenges and user interaction patterns while introducing embodied interaction complexities that gaming platforms typically do not prioritise.

3.1. Sample selection and platform characteristics

This analysis centres on social XR platforms (SXPs): corporate-operated platforms designed for socialisation and creation, centrally moderated, enabling real-time virtual world creation. It examines platforms that share characteristics as the next generation of Second Life, where users design avatars and navigate virtual environments (Boellstorff, 2008).

The study focuses on four major XR platforms: Meta's Horizon Worlds, VRChat, Roblox, and Zepeto. For Web 2.0 comparison, this study considers Meta's Community Standards as a widely adopted governance template due to its comprehensive, transparent guidelines that have significantly influenced platform governance

norms (Lazerson, 2022).

TABLE 1: Platform selection and characteristics

PLATFORM NAME	PARENT COMPANY	ACCESS MEDIUM	AGE RESTRICTION	USER CONTENT RELATED GUIDELINES
HORIZON WORLDS	Meta (US)	PC, mobile[3], all Quest headsets	10 or older [4]	Meta Quest Code of Conduct for Virtual Experiences Meta Horizon Worlds Content Guidelines
VRCHAT	VRChat (US)	Windows, HTC Vive, HTC Vive Pro, Oculus Rift CV1, Oculus Rift S, Oculus Quest 2, Oculus Quest Link, Valve Index, WMR Headsets[5], Android (does not support iOS)[6]	13 or older[7]	VRChat Community Guidelines VRChat Creator Guidelines
ROBLOX	Roblox Corporation (US)	PC,[8]Xbox[9], Mobile: Apple iOS, Android, Amazon Fire OS[10] VR: Meta Quest 2, Meta Quest Pro, and Meta Quest 3[11]	No age restriction but users over 13 get expanded features[12]	Roblox Community Standards Roblox Restricted Content Policy
ZEPETO	Naver Z (South Korea)	PC and Mobile, no headset required[13]	13 or older, may restrict access based on the age of consent in each	Zepeto Community Guidelines Zepeto Studio Content Guidelines

3. https://www.meta.com/help/quest/140991407990979/
4. https://www.meta.com/blog/preteens-horizon-worlds-family-friendly-vr-mr/
5. https://help.vrchat.com/hc/en-us/articles/1500002378722-System-Requirements
6. https://help.vrchat.com/hc/en-us/articles/19616653355411-VRChat-Mobile-System-Requirements
7. https://hello.vrchat.com/community-guidelines#:~:text=In%20order%20to%20use%20VRChat,or%20guardian%20before%20using%20VRChat.
8. https://en.help.roblox.com/hc/en-us/articles/203312800-Computer-Hardware-Operating-System-Requirements
9. https://www.xbox.com/en-US/games/store/roblox/bq1tn1t79v9k
10. https://en.help.roblox.com/hc/en-us/articles/203625474-Roblox-Mobile-System-Requirements
11. https://en.help.roblox.com/hc/en-us/articles/17810433924628-Meta-Quest-FAQ
12. https://corp.roblox.com/newsroom/2024/11/major-updates-to-our-safety-systems-and-parental-controls

PLATFORM NAME	PARENT COMPANY	ACCESS MEDIUM	AGE RESTRICTION	USER CONTENT RELATED GUIDELINES
			country.[14]	

3.2. Platform selection rationale

These four platforms represent all major social XR platforms meeting criteria that enable analysis of governance framework standardisation and adaptation: (1) corporate-operated with centralised moderation (excluding decentralised platforms like Decentraland), (2) focused on social interaction and user-generated content creation rather than structured gameplay (excluding game-focused platforms like Fortnite), and (3) operating at substantial scale. These criteria identify platforms most comparable to Web 2.0 social media in governance challenges while introducing immersive environment complexities.

Within this universe of social XR platforms, the four selected represent important variations across multiple dimensions. First, they employ different technological approaches: Meta's Horizon Worlds was originally designed for VR headsets, VRChat operates primarily in VR but allows desktop access, while Roblox and Zepeto are accessible via mobile devices and computers without specialised hardware. This variation enables examination of how governance frameworks adapt across different levels of immersion and accessibility.

Second, these platforms vary in geographic origin (three US-based, one South Korean), enabling analysis of how governance templates circulate globally. Zepeto is particularly instructive for this analysis: despite its South Korean corporate ownership, the platform's trust and safety operations are based in the United States and staffed by professionals with prior experience at major US technology companies.[15]This demonstrates how US-centric governance frameworks disseminate through professional circulation across companies and borders, not only through mimetic isomorphism or regulatory pressure. This phenomenon where corporate ownership, trust and safety operations, and workforce span multiple jurisdictions reflects the jurisdictional complexities that challenge regulatory frameworks like

13. https://www.bluestacks.com/apps/entertainment/zepeto-on-pc.html
14. https://support.zepeto.me/hc/en-us/article s/360047786633-Terms-of-Use
15. Information about Zepeto's organisational structure based on author's ethnographic observations at Trust & Safety professional conferences (2023-2025) and informal conversations with platform governance practitioners.

the EU's Digital Services Act, which must determine accountability when platform governance functions are distributed across national boundaries.

Unlike the other three social XR platforms, Horizon Worlds shares a parent organisation with the analytical baseline. This makes it a useful case for examining whether governance transplantation operates differently within a single organisation than across organisational boundaries. Given their shared organisational origin, the closest alignment between Horizon Worlds and Meta's Community Standards would be a reasonable expectation, though this study treats that expectation as an empirical question rather than an assumption. Horizon Worlds was also selected on practical grounds, as the universe of social XR platforms meeting the study criteria is limited and Horizon Worlds remains among the most prominent examples currently operating at scale.

Finally, these platforms operate at substantial scale: Roblox reports over 111 million daily active users (Roblox, 2025), VRChat over 66,000 concurrent users (SteamDB, 2025), and Zepeto over 400 million registered users (Davies & Song, 2022). While they differ in target demographics (Roblox and Zepeto skew younger), technical affordances, corporate structures, and regulatory environments (Zepeto operates under South Korean laws criminalising online defamation, while US platforms operate under Section 230 protections), these differences enable examination of how governance frameworks both standardise across contexts and adapt to platform-specific needs.Horizon Worlds reported fewer than 200,000 monthly active users on VR as of 2022 (Horwitz et al., 2022), a figure Meta has not publicly updated since, though the platform's mobile app has since accumulated 45 million total downloads worldwide (Silberling, 2026).

This analysis encompasses both current and historical versions of platform guidelines. For Meta's historical documents, the study utilised the Platform Governance Archive managed by the University of Bremen, which hosts timestamped versions of platform community guidelines (Platform Governance Archive, n.d.). Documents were identified by searching the archive for Facebook's community guidelines and selecting versions that marked significant moments of policy change. For social XR platforms, the Internet Archive's Wayback Machine was used to access historical versions, analysing guideline evolution across different timeframes: Meta's XR guidelines (2022-2025), VRChat (2020-2025), Roblox (2016-2025), and Zepeto (2021-2025).

3.3. Data analysis and coding

Meta's Community Standards serve as the comparative benchmark for this analysis for several reasons. As one of the oldest continuously operating social media platforms, Meta has accumulated an extensively documented history of public-facing governance frameworks shaped by sustained regulatory and public scrutiny (Gillespie, 2018). Its Community Standards have become a widely referenced governance template in the platform industry, influencing how emerging platforms conceptualise and structure their own guidelines (Lazerson, 2022). In this sense, Meta's Community Standards function as a dominant institutional template against which other platforms' governance choices can be understood.

Platforms differing substantially from Facebook in technical affordances, user demographics, and interaction modalities risk inheriting guidelines designed for a different context, where they may function more as signals of institutional legitimacy and risk management than as effective tools for user protection. Meta's Community Standards are also not a static baseline, given that the January 2025 policy shift, which occurred within the analysis window, illustrates how the template itself is subject to change driven by factors beyond user safety considerations. These limitations are further addressed in the Discussion section.

The analysis developed a three-tiered classification system based on Meta's established framework: policy themes (broad headings such as "Safety," "Violence and Criminal Behaviour"), policy verticals (specific violations such as "Suicide and Self-Injury" or "Human Exploitation"), and policy clauses (detailed definitions and examples).

The study conducted a two-stage coding process to account for structural differences across platforms. For platforms with clear thematic organisation (Roblox and Zepeto), the analysis directly mapped policy verticals to Meta's Community Standards. For platforms using simpler structures (Meta's Social VR and VRChat), the research conducted line-by-line analysis, categorising each according to substantive content.

Using MAXQDA's qualitative analysis tools, the analysis developed a comprehensive codebook capturing structural (theme organisation, vertical categorisation), content (violation types, definitional approaches, examples provided), and linguistic elements (terminology, tone, phrasing patterns), enabling the identification of patterns across platforms with particular attention to how platforms adapted Web 2.0 terminology to describe XR-specific violations.

4. Findings

The study finds that platform governance homophily manifests not as a singular phenomenon, but in layered forms: across time (temporal homophily), across platforms (inter-organisational similarity), and within a single organisation (intra-organisational similarity). These patterns show how platforms draw from precedent, especially from Meta, while also adapting based on user demographics, platform functionality, and perceived risk.

4.1. Temporal homophily: governance evolution over time

4.1.1. The journey and impact of meta's community standards: a model for understanding platform evolution

Meta's Community guidelines evolved through four distinct phases reflecting the platform's growing influence and regulatory responsibilities (Table 1). This progression shows how platform governance complexity grows in proportion to scale and public scrutiny. The integration of value statements demonstrates how mature platforms articulate social responsibility beyond functionality, suggesting that emerging platforms may develop intricate frameworks earlier to anticipate similar pressures.

4.1.2. Patterns of policy evolution in XR platforms

Early-stage evolution: VRChat's mirror of early Meta (Facebook). VRChat's current guidelines resemble Meta's original Code of Conduct more than its present-day standards. Categories like "inappropriate content" appear under broad, moralistic language, i.e., "Do not engage in… illegal behavior" or "Treat people how they want to be treated." The platform frequently uses "any" to cast wide nets: "any form" of hate, or abuse of VRChat "in any way." This signals an early-phase regulatory approach based on user judgment and general behavioural expectations rather than specific enforcement logic.

Intermediate evolution: Zepeto, Roblox's similarities with MCS in 2015-2017. Both platforms mirror the period where Meta's guidelines emphasised values and user responsibility. Zepeto defines its policies around "safety, inclusion, diversity, dignity, and creativity," while Roblox seeks to ensure that "everyone feels safe and welcome." These reflect Meta's post-2015 shift toward normative framing. Roblox's evolution demonstrates the most faithful progression toward Meta's comprehensive model, accomplishing in five years what took Meta over a decade (Table 2).

However, examining historical versions reveals regression in XR-specific guidance.

Roblox's 2018 guidelines provided detailed examples of avatar-based violations in immersive spaces, but this attention to XR-specific harms is reduced in current guidelines, suggesting how standardisation toward Web 2.0 frameworks sacrifices context-specific guidance. This compressed timeline suggests accelerated learning, with newer platforms leveraging established governance templates while adopting value-driven approaches earlier than predecessors.

4.2. Inter-organisational similarity: Cross-platform parallels and convergences

4.2.1. Similarities in structure and types of harms covered

The community guidelines of Zepeto, Roblox, and VRChat show significant structural and content-level similarities to Meta's Community Standards. At the structural level, these platforms follow Meta's “cascading specificity” approach (Lazerson, 2022), a multi-tiered structure organising policies from broad overarching themes like "Safety" or "Violence and Criminal Behaviour" to specific policy verticals addressing distinct harm types, presented in a consistent format with policy rationales, subcategories, and examples.

This hierarchical approach enables consistent enforcement at scale by creating decision trees that guide moderators through increasingly specific categories (Newton, 2019). This pattern of cascading specificity appears consistently across these platforms, creating governance mechanisms that are simultaneously adaptable to platform-specific needs while maintaining consistency (Star & Griesemer, 1989).

Regarding harm classification, the platforms share 9 out of 24 policy verticals with Meta's Community Standards, including Dangerous Individuals and Organisations, Violence and Incitement, Suicide and Self-Injury, Bullying and Harassment, Hate Speech, Violent and Graphic Content, Adult Nudity, and Sexual Activity (Figure 1).

Roblox most comprehensively emulates Meta's structure. While its 2016 guidelines consisted of simple bullet points, policies since 2021 mirror Meta's framework with themed sections (Safety, Civility, Integrity) and detailed policy verticals. Zepeto follows a similar pattern with policy rationales and detailed dropdown examples. VRChat incorporates key structural elements through concise descriptions followed by bullet-pointed examples. Notably absent from all XR platforms is Meta's memorialisation policy, likely due to XR platforms allowing anonymous avatar-based interactions rather than real-world identity requirements.

The language defining these harms is strikingly similar, with platforms often incor-

porating elements from Meta's most severe (Tier 1) violations. For example, Zepeto's "Hateful Activities" policy prohibits users from posting "content, speech or behaviour that attacks, degrades, demeans, incites violence against, excludes, or dehumanizes people on the basis of the following protected characteristics," encompassing prohibited actions in Tier 1 of MCS's Hate Speech policy (Zepeto, 2025; Meta, 2024). Similarly, Roblox's "Threats, Bullying, and Harassment" vertical contains aspects mostly included in MCS' Tier 1 category, such as threats to expose personal information and threats of sexual violence (Roblox, 2025).

In contrast, social XR platforms provide relatively brief definitions compared to Meta's extensive Transparency Center portal. Notable differences exist around controversial content: VRChat restricts "inflammatory behaviour" including "controversial topics," Roblox prohibits discussion of political parties and geopolitical issues, and Zepeto bans "Political content and debated social issues" and territorial disputes, reflecting its South Korean origins and East/Southeast Asian user base with unresolved territorial disputes (Davies & Song, 2022).

Following Meta's public backlash after the 2021 'Facebook Files' expose (Wall Street Journal, 2021) and Frances Haugen's whistleblowing testimony, emerging platforms appear increasingly cautious about controversial content. Platforms like Roblox and Zepeto, which predominantly attract younger users - 54% of Roblox users are under 13 (Liao, 2022) and Zepeto has 70% female users between 13 and 21 (Davies and Song, 2022) - have prioritised child safety in their community guidelines, implementing strict restrictions such as prohibiting profanity and dating (Roblox, 2025) and preventing the creation of virtual lingerie (Zepeto, 2025-b).

4.2.2. Emergence of XR-specific policy categories

Social XR platforms are beginning to define harms unique to immersive contexts (Fig. 2). Common additions include policies on impersonating staff, audio conduct, personal space, and gating mature or public-facing environments, especially in VRChat, Horizon Worlds, and Zepeto. Some technical violations, like API tampering, have shifted from Terms of Service to community guidelines. Overall, coverage remains fragmented. VRChat, for example, distinguishes public vs. private spaces; Zepeto restricts political content and regulates virtual fashion assets. These inconsistencies suggest XR governance remains emergent, shaped by platform design rather than shared norms (Figure 2).

The core of this challenge is a tension between operational efficiency and specificity to the distinct risks and interaction patterns of immersive, embodied XR environments. While importing Web 2.0 models offers regulatory familiarity and en-

forcement scale, it often overlooks harms like spatial harassment, symbolic misuse of virtual objects, or consent violations involving avatars. The selective adaptation process shows how platforms consistently prohibit universally recognised risks while tightening control over political and controversial expression, signalling a governance logic prioritising liability management over user safety.

Zepeto's reporting mechanisms reveal these shortcomings. The platform's reporting options primarily address text and image-based content instead of interaction spaces like live audio, AR filters, and avatar behaviours. Users cannot report scenarios such as minors performing sexually suggestive choreography in AR photobooth features, or avatars configured in culturally insensitive ways. Such limitations expose a fundamental challenge of recognising and moderating complex interactions native to social XR platforms. The anonymity afforded by avatar-based interaction can intensify disinhibited behaviour (Suler, 2004), which further complicates frameworks for establishing user consent and interaction boundaries in spaces where behaviours appropriate in one virtual setting may be harmful in another (Blackwell et al., 2019).

4.3. Intra-organisational similarity: Meta's continued bifurcation of conduct and content

Meta's Horizon Worlds guidelines diverge from its broader Community Standards, both in form and framing. Unlike the comprehensive structure housed in the Transparency Center, Horizon's "Code of Conduct" appears separately within the Meta Store's Help Center and echoes older, more concise policies from 2012–2017. This document uses moral appeals and simple prohibitions: "Do not take advantage of…", "Be especially mindful…" (Meta, 2025-a), mirroring early Meta content that emphasised personal responsibility over nuanced policy. The language and structure lack the cascading specificity found in current Community Standards, instead combining multiple verticals into short, loosely defined categories.

Since the launch of its Quest 3S in late 2024, Meta has expanded XR safety documentation, but primarily through technical controls and product-level tools, not new policies. The *Code of Conduct for Virtual Experiences* (CCVE), for example, was relocated from the Help to Legal section in 2022, signalling a shift toward liability management. While it claims to address both content and conduct, its treatment of interpersonal harm is minimal, focusing instead on impersonation, scams, and other inauthentic behaviours already detailed in Meta's core guidelines.

This separation reflects a deeper bifurcation: content moderation remains cen-

tralised and elaborately codified in the Transparency Center, while conduct governance, especially for XR, is offloaded to product design (blocking, muting, spatial boundaries) not policy. Issues like harassment, hate speech, or symbolic aggression receive little attention in Horizon's specific guidelines, despite being central in Web 2.0 contexts. Overall, this dual system suggests a reluctance to build targeted XR governance. By treating speech and behaviour as distinct regulatory domains, Meta avoids confronting how embodied interactions blur that line. For instance, actions like "spatial encroachment" or symbolic object use may escape traditional content filters but still produce harm. The split between speech and action embedded in Meta's documentation architecture, risks leaving users without clear, cohesive protections in immersive settings.

5. Discussion

5.1. Hidden costs of pouring old wine into new bottles

Virtually embodied interactions prompt new questions about harassment, consent, and user safety in immersive spaces. The challenges of governing these spaces are not entirely new. Virtual worlds like Second Life, operating since 2003, and earlier text-based environments like LambdaMOO in the 1990s pioneered approaches to moderating embodied user behaviour and interactions (Boellstorff, 2008; Dibbell, 1993). Second Life developed governance frameworks that explicitly addressed embodied interactions, spatial dynamics, and user-created content in ways that anticipated many challenges facing contemporary XR platforms. For instance, Second Life's community standards distinguished between different types of virtual spaces through Maturity Ratings (General, Moderate, Adult), governed avatar-based expression and conduct, and addressed forms of harassment specific to virtual world environments (Linden Lab, 2023). These earlier virtual worlds demonstrated that governance frameworks could be designed specifically for immersive, avatar-based social interaction rather than adapted from text-based platforms.

However, contemporary XR platforms have largely bypassed these precedents, instead adopting governance templates from Web 2.0 social media platforms like Facebook and Instagram. This choice reflects both mimetic isomorphism (following perceived industry leaders) and the dominance of companies like Meta that bring their established Web 2.0 frameworks to new XR ventures. Yet bypassing these precedents comes with costs: Second Life's governance innovations, developed through two decades of managing virtual world communities, remain largely ignored in favour of frameworks designed for fundamentally different interaction modalities. The questions that LambdaMOO confronted about consent, harm, and

governance in virtual spaces in 1993 (Dibbell, 1993) remain unresolved in today's XR platforms, despite three decades of virtual world experience to draw from.

Following DiMaggio and Powell's institutional theory, platforms mimic dominant peers to gain legitimacy, even when mimicry misfits the medium. Meta's positioning of its own system as "industry standard" (Meta, 2024-d) intensifies this effect, encouraging emerging platforms to adopt technical fixes over governance innovation, even at the expense of addressing embodied harms.

These patterns reflect a tension that is productive rather than contradictory. Structural standardisation operates at the level of harm categories and policy architecture, where isomorphic pressures produce convergence across platforms. Operational fragmentation emerges at the level of implementation, shaped by platform-specific affordances and design choices. The two operate simultaneously because institutional forces govern what platforms include in their guidelines while technical factors govern how effectively those guidelines translate into protection for users in immersive environments.

5.2. Perpetuating existing power structures

The standardisation of governance frameworks across social XR platforms perpetuates the uncritical adoption of US-centric techno-legal frameworks. Meta's community guidelines are inextricably linked to U.S. First Amendment interpretations of free speech (Langvardt, 2017; Klonick, 2018). The exportation of these frameworks globally risks making US approaches to online safety a universal gold standard. While standardising definitions of harm can streamline organisational processes and help newer companies achieve legitimacy, this can inhibit localised approaches with region-specific nuances.

European regulations recognise categories of harm with different emphases than US frameworks, including stronger protections for children's data rights under the GDPR and the Digital Services Act's requirements for risk assessments and transparency. XR platforms thus face competing pressures: adopting Meta's governance templates provides legitimacy and operational efficiency within US regulatory contexts, but these frameworks require substantial adaptation to satisfy European requirements. This challenge intensifies for immersive technologies, as European data protection laws designed for Web 2.0 must now address new forms of personal data in XR environments, including biometric information, spatial tracking, and behavioural patterns that have no clear precedent in existing regulations.

This is particularly evident in how hate speech policies define protected cate-

gories. The platforms examined in this study largely adhere to classifications derived from the 1964 U.S. Civil Rights Act that specifies race, colour, religion, sex, and national origin as protected categories (U.S. Department of Justice, Civil Rights Division, n.d.). However, as platforms expand globally, such categories are insufficient to cover hate speech targeting people from the same country but different regions (i.e. "Skatefish" to degrade people from Jeolla province in South Korea (Son, 2015)) or people from different countries simplistically collapsed into one linguistic group (i.e. "Sudaca," used to degrade Hispanic South Americans (Garcia-Guirao, 2021)).

Furthermore, platform governance remains vulnerable to geopolitical pressures. Meta's January 2025 strategic shift shows that its content moderation practices are susceptible to U.S. political climates (Chan et al., 2025; Lunden, 2025), shaped by electoral considerations over consistent, user-centreed principles. By transplanting these US-centric frameworks globally, platforms solidify a governance model that legitimises tech corporations' role in shaping online behaviour worldwide, perpetuating concerns about lack of public oversight and transparency in content moderation (Gillespie, 2018; Owono & Morch, 2022).

5.3. The burden of self-governance in immersive spaces

These patterns of homogenisation are not coincidental. Institutional pressures actively lock in Web 2.0 governance frameworks even as XR environments demand something fundamentally different. Platforms adopt Meta's Community Standards not because those standards work well for immersive spaces, but because adoption signals legitimacy (DiMaggio & Powell, 1983). Once embedded, these frameworks become difficult to change as they are entrenched in core organisational structures, enforcement systems, and cross-platform expectations that resist easy modification (Hannan & Freeman, 1984).

Community guidelines in the XR era continue to function as boundary objects (Star & Griesemer, 1989), as these documents can mean different things to different stakeholders without requiring agreement on how governance should actually proceed. Platforms use them to signal responsibility to regulators and advertisers; researchers and civil society use them as accountability benchmarks; users treat them as behavioural expectations. This flexibility is precisely what makes community guidelines durable, but it also allows platforms to claim they are governing immersive spaces without their guidelines actually addressing embodied, real-time, or spatially dependent harms.

The result is an outsourcing of governance responsibility to individual users, reflecting the concept of responsibilisation within neoliberal political rationality (Brown, 2017). The imbalance documented across platforms, where safety tools are extensively documented but behavioural policies remain thin, illustrates this dynamic. Providing users with controls while maintaining ambiguous conduct policies creates the appearance of empowerment while shifting governance burdens downward. In XR environments, this disproportionately affects users who are least equipped to manage risk in real time, including those with limited technical literacy, newcomers, and younger users.

Several XR-specific governance challenges remain undertheorised and warrant further research, such as the role of virtual world creators in moderating user-generated environments; questions of avatar identity and liability as platforms host both human-controlled and AI-generated entities (Cheong, 2022); the identification of embodied harm including toxic re-embodiment across demographic groups (Nakamura, 2020); and the governance implications of haptic technology, which raises questions of consent and privacy that existing frameworks cannot yet address (Jewitt et al., 2020; Jang et al., 2023). Addressing these challenges will require governance frameworks that are developed alongside emerging technologies rather than retrofitted from existing ones.

6. Conclusion

This study shows how the homogenisation of content governance frameworks across Web 2.0 and social XR platforms challenges platform governance in three ways. First, the wholesale adoption of Web 2.0 frameworks creates fundamental misalignments between governance mechanisms and immersive interactions. While Web 2.0 guidelines were designed for discrete, archivable content, XR platforms must contend with ephemeral, embodied, and spatially dependent interactions that resist traditional content moderation approaches. The regression observed in platforms like Roblox–moving from XR-specific guidelines to more generic Web 2.0 frameworks–suggests that standardisation pressures may actually reduce platforms' capacity to address immersive harms.

Second, this analysis reveals a concerning bifurcation between content and conduct in governance approaches. Rather than developing comprehensive frameworks that address the unique nature of immersive interactions, platforms increasingly rely on technical solutions and user controls to manage behavioural issues while maintaining traditional content moderation frameworks. This separation becomes problematic in XR environments where distinctions between content and

conduct become murky, as seen in how avatar appearances, gestures, and spatial positioning can simultaneously be both speech and behaviour.

Third, current governance approaches increasingly shift responsibility for safety onto users themselves. In XR contexts, this manifests as an emphasis on user tools and controls over direct platform moderation, disproportionately burdening vulnerable users, particularly those with limited technical literacy, newcomers to immersive environments, and younger users who may struggle to deploy complex protective tools in real time. What appears as user empowerment through safety tools is a transfer of responsibility that obscures the power relations that structure risky and harmful environments.

Ultimately, the significance of these findings extends beyond current XR platforms to raise important questions about the future of platform governance more widely. As new technologies emerge, will platforms continue to pour old governance frameworks into new bottles? Or can new approaches be developed that better account for novel forms of social interaction? Breaking free from Web 2.0 paradigms demands more than technical innovation; it requires fundamentally reimagining how harm manifests across emerging technological contexts.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Ananny, M., & Gillespie, T. (2016, September 22–23). *Public platforms: Beyond the cycle of shocks and exceptions* [Conference paper]. IPP2016: The Platform Society, Oxford Internet Institute, University of Oxford, Oxford, United Kingdom. https://blogs.oii.ox.ac.uk/ipp-conference/sites/ipp/files/documents/anannyGillespie-publicPlatforms-oii-submittedSept8.pdf

Blackwell, L., Ellison, N., Elliott-Deflo, N., & Schwartz, R. (2019). Harassment in social virtual reality: Challenges for platform governance. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–25. https://doi.org/10.1145/3359202

Boellstorff, T. (2008). *Coming of age in Second Life: An anthropologist explores the virtually human*. Princeton University Press. https://doi.org/10.2307/j.ctvc77h1s

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. The MIT Press. https://doi.org/10.7551/mitpress/6352.001.0001

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, W. (2017). *Undoing the demos: Neoliberalism's stealth revolution*. Zone Books.

Caplan, R., & boyd, d. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, *5*(1). https://doi.org/10.1177/2053951718757253

Chan, K., Ortutay, B., & Riccardi, N. (2025, January 8). Meta eliminates fact-checking in latest bow to Trump. *AP News*. https://apnews.com/article/meta-facts-trump-musk-community-notes-413b8495939a058ff2d25fd23f2e0f43

Cheong, B. C. (2022). Avatars in the metaverse: Potential legal issues and remedies. *International Cybersecurity Law Review*, *3*, 467–494. https://doi.org/10.1365/s43439-022-00056-9

Cormode, G., & Krishnamurthy, B. (2008). Key differences between web 1.0 and web 2.0. *First Monday*, *13*(6). https://doi.org/10.5210/fm.v13i6.2125

Czarniawska, B., & Joerges, B. (1996). Travels of ideas. In B. Czarniawska & G. Sevón (Eds), *Translating organizational change* (pp. 13–47). De Gruyter. https://doi.org/10.1515/9783110879735.13

Davies, C., & Song, J. (2022, September 27). Asia's largest metaverse platform Zepeto ramps up global expansion. *Financial Times*. https://www.ft.com/content/14c88e84-f3c8-485e-a9df-31ead34e48f0

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, *39*(9), 761–770.

Dibbell, J. (1993). A rape in cyberspace. *The Village Voice*. https://www.villagevoice.com/a-rape-in-cyberspace/

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, *48*(2), 147–160. https://doi.org/10.2307/2095101

European Union. (2022). *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)* (No. 2022/2065; pp. 1–102). European Union. https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

Facebook. (n.d.). *Community standards*. Facebook. Retrieved https://transparency.fb.com/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards

Facebook. (2007). *Content code of conduct*. Facebook. https://github.com/PlatformGovernanceArchive/pga-corpus/blob/main/Versions/PDF/Facebook/Community%20Guidelines/20070830_Facebook_CommunityGuidelines.pdf

Facebook. (2011). *Facebook community standards*. Facebook. https://github.com/PlatformGovernanceArchive/pga-corpus/blob/main/Versions/PDF/Facebook/Community%20Guidelines/20110127_Facebook_CommunityGuidelines.pdf

Facebook. (2012). *Facebook community standards*. Facebook. https://github.com/PlatformGovernanceArchive/pga-corpus/blob/main/Versions/PDF/Facebook/Community%20Guidelines/20121210_Facebook_CommunityGuidelines.pdf

Facebook. (2015, April 9). *Facebook community standards*. Facebook. https://github.com/PlatformGovernanceArchive/pga-corpus/blob/main/Versions/PDF/Facebook/Community%20Guidelines/20150409_Facebook_CommunityGuidelines.pdf

Facebook. (2018). *Understanding the community standards enforcement report*. Facebook. https://abou

t.fb.com/ltam/wp-content/uploads/sites/14/2018/05/understanding_the_community_standards_enforcement_report.pdf

Facebook. (2023). *Hate speech*. Facebook. https://transparency.fb.com/policies/community-standards/hate-speech/

García-Guirao, P. (2021). The language impact of COVID-19 in the Spanish online press: A critical discourse analysis of hate speech. *European Research Studies Journal*, *24*(4, pt 2), 756–770. https://doi.org/10.35808/ersj/2769

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press. https://doi.org/10.12987/9780300235029

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R. (2024). Introduction. In *The politics of platform regulation: How governments shape online content moderation* (pp. 1–10). Oxford University Press. https://doi.org/10.1093/oso/9780197692851.003.0001

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Haggart, B., & Keller, C. I. (2021). Democratic legitimacy in global platform governance. *Telecommunications Policy*, *45*(6). https://doi.org/10.1016/j.telpol.2021.102152

Hannan, M. T., & Freeman, J. (1984). Structural inertia and organizational change. *American Sociological Review*, *49*(2), 149–164. https://doi.org/10.2307/2095567

Harkai, I., & Mezei, P. (2024). Recalibrating territoriality in platform end-user license agreements. In D. Popović & R. Kulms (Eds), *Repositioning platforms in digital market law* (pp. 55–81). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-69678-7_3

Horwitz, J., Rodriguez, S., & Bobrowsky, M. (2022, October 15). Company documents show Meta's flagship metaverse falling short. *The Wall Street Journal*. https://www.wsj.com/articles/meta-metaverse-horizon-worlds-zuckerberg-facebook-internal-documents-11665778961

Jang, K., Kollannur, S., Zhu, X., & Culbertson, H. (2023). When virtual hugs become sexual harassment: Conceptualizing acceptable forms of mediated touch in extended reality platforms. *2023 IEEE World Haptics Conference (WHC)*. IEEE World Haptics Conference. https://2023.worldhaptics.org/wp-content/uploads/2023/06/1162-doc.pdf

Jewitt, C., Price, S., Leder Mackley, K., Yiannoutsou, N., & Atkinson, D. (2020). *Interdisciplinary insights for digital touch communication*. Springer International Publishing. https://doi.org/10.1007/978-3-030-24564-1

Kaplan, J. (2025, January 7). *More speech and fewer mistakes*. Facebook. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE guide no. 131. *Medical Teacher*, *42*(8), 846–854. https://doi.org/10.1080/0142159x.2020.1755030

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*(6), 1598–1670.

Langvardt, K. (2017). Regulating online content moderation. *Georgetown Law Journal*, *106*, 1353–1388. https://doi.org/10.2139/ssrn.3024739

Lazerson, R. (2022). *A secure and equitable metaverse: Designing effective community guidelines for social VR* (CLTC White Paper Series) [White paper]. Center for Long-Term Cybersecurity, University of California, Berkeley. https://cltc.berkeley.edu/wp-content/uploads/2022/11/Secure_Equitable_Metaverse.pdf

Liao, S. (2022, September 22). Roblox wants to advertise to gamers ages 13 and up in the metaverse. *Washington Post*. https://www.washingtonpost.com/video-games/2022/09/09/roblox-ads-metaverse/

Linden Lab. (2023). *Community standards*. Linden Lab. https://www.lindenlab.com/legal/community-standards

Lunden, I. (2025, January 7). Meta drops fact-checking, loosens its content moderation rules. *TechCrunch*. https://techcrunch.com/2025/01/07/meta-drops-fact-checking-and-loosens-its-content-moderation-rules/

Marwick, A. E., and boyd, d. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

Meta. (2023). *Manage social features*. Meta. https://www.meta.com/help/quest/articles/accounts/parental-supervision/manage-social-features/

Meta. (2024a). Celebrating 10 years of Reality Labs. *Meta Newsroom*. https://about.fb.com/news/2024/04/celebrating-10-years-of-reality-labs/

Meta. (2024b). *Hate speech*. Meta. https://transparency.meta.com/policies/community-standards/hateful-conduct/

Meta. (2024c). *Learn about parental supervision for a teen in Meta Horizon Worlds*. Meta. https://www.meta.com/help/quest/1737463343292580/

Meta. (2024d). *Learn about the personal boundary setting in Meta Horizon Worlds*. Meta. https://www.meta.com/help/quest/758869488544227/

Meta. (2024e). *Moderators and community guides in Meta Horizon Worlds*. Meta. https://www.meta.com/help/quest/2259751227488308/

Meta. (2025a). *Code of conduct for virtual experiences*. Meta. https://www.meta.com/help/quest/articles/accounts/privacy-information-and-settings/code-of-conduct-for-virtual-experiences/

Meta. (2025b). *Meta Horizon Worlds content guidelines*. Meta. https://www.meta.com/help/quest/481214418021533/

Mezei, P., & Harkai, I. (2022). End-user flexibilities in digital copyright law–an empirical analysis of end-user license agreements. *Interactive Entertainment Law Review*, *5*(1), 2–21. https://doi.org/10.4337/ielr.2022.0003

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83(2), 340–363. https://doi.org/10.1086/226550

Morimoto, T., Kobayashi, T., Hirata, H., Otani, K., Sugimoto, M., Tsukamoto, M., Yoshihara, T., Ueno, M., & Mawatari, M. (2022). XR (extended reality: Virtual reality, augmented reality, mixed reality) technology in spine medicine: Status quo and quo vadis. *Journal of Clinical Medicine*, 11(2), Article 470. https://doi.org/10.3390/jcm11020470

Nakamura, L. (2020). Feeling good about feeling bad: Virtuous virtual reality and the automation of racial empathy. *Journal of Visual Culture*, *19*(1), 47–64. https://doi.org/10.1177/1470412920906259

Napoli, P. M. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication Theory*, *24*(3), 340–360. https://doi.org/10.1111/comt.12039

Napoli, P. M., & Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday*, *22*(5). http://firstmonday.org/ojs/index.php/fm/article/view/7051

Newton, C. (2019, February 25). The trauma floor. *The Verge*. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

O'Reilly, T. (2005). *What is web 2.0: Design patterns and business models for the next generation of software*. https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Owono, J., & Morch, L. (2022). *Content governance in the metaverse*. Content Policy and Society Lab. https://www.skeyesmedia.org/documents/bo_filemanager/Content-Governance-in-the-Metaverse.pdf

Petersen, J. (2022). *How machines came to speak: Media technologies and freedom of speech*. Duke University Press. https://doi.org/10.1215/9781478021827

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399–441.

Platform Governance Archive. (n.d.). *Platform Governance Archive*. Retrieved https://platformgovernancearchive.uni-bremen.de/

Poell, T., Nieborg, D. B., & Duffy, B. E. (2021). *Platforms and cultural production*. John Wiley & Sons.

Roblox. (n.d.). *Roblox safety center*. Roblox. Retrieved https://about.roblox.com/safety

Roblox. (2018). *Roblox community rules*. Roblox. https://web.archive.org/web/20190126061245/https://en.help.roblox.com/hc/en-us/articles/203313410

Roblox. (2019, September 8). *Roblox rules of conduct*. Roblox. https://web.archive.org/web/20160908173545/https://en.help.roblox.com/hc/en-us/articles/203313410

Roblox. (2020, December 4). *Roblox community rules*. Roblox. https://web.archive.org/web/20160908173545/https://en.help.roblox.com/hc/en-us/articles/203313410

Roblox. (2021, November 23). *Roblox community standards*. Roblox. https://web.archive.org/web/20250218070016/https://en.help.roblox.com/hc/en-us/articles/203313410-Roblox-Community-Standards

Roblox. (2023, December 6). *Roblox community standards*. Roblox. https://web.archive.org/web/20241206231344/https://en.help.roblox.com/hc/en-us/articles/203313410-Roblox-Community-Standards

Roblox. (2024, December 6). *Roblox community standards*. Roblox. https://web.archive.org/web/20160908173545/https://en.help.roblox.com/hc/en-us/articles/203313410

Roblox. (2025a). *Roblox community standards*. Roblox. https://en.help.roblox.com/hc/en-us/articles/203313410-Roblox-Community-Standards

Roblox. (2025b, September 4). *Annual Roblox economic impact report: How Roblox and creators are driving economic opportunities*. Roblox. https://corp.roblox.com/newsroom/2025/09/roblox-annual-economic-impact-report

Rosen, G., & Bickert, M. (2021, January 7). *Our response to the violence in Washington*. Facebook. https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/

Silberling, A. (2026, March 19). Meta decides not to shut down Horizon Worlds on VR after all. *TechCrunch*. https://techcrunch.com/2026/03/19/meta-decides-not-to-shut-down-horizon-worlds-on-vr-after-all/

Son, S. (2015, February 24). ?? ????? '??'…??·???? ?? ?? [Hong-eo' on the Google search screen. *Yonhap News*. https://www.yna.co.kr/view/AKR20150224074600054

Star, S. L. (1991). Power, technology, and the phenomenology of conventions: On being allergic to onions. In J. Law (Ed.), *A sociology of monsters: Essays on power, technology and domination* (pp. 26–56). Routledge.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, translations, and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, *19*(3), 387–420. https://doi.org/10.1177/030631289019003001

Star, S. L., & Lampland, M. (2009). Reckoning with standards. In M. Lampland & S. L. Star (Eds), *Standards and their stories: How quantifying, classifying, and formalizing practices shape everyday life* (pp. 3–24). Cornell University Press.

SteamDB. (2025, January 1). VRChat peaks at 66 k players. *Valve Corporation*. https://store.steampowered.com/news/group/4777282/view/7285229827084255391

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Suzor, N. P. (2019). *Lawless: The secret rules that govern our digital lives*. Cambridge University Press. https://doi.org/10.1017/9781108666428

The Wall Street Journal. (2021, October 25). *The Facebook files*. The Wall Street Journal. https://www.wsj.com/articles/the-facebook-files-11631713039

U.S. Department of Justice, C. R. Division. (n.d.). *Title VII of the Civil Rights Act of 1964*. Retrieved https://www.justice.gov/crt/laws-we-enforce

Van Dijck, J. (2020). Governing digital societies: Private platforms, public values. *Computer Law & Security Review*, *36*, 105377. https://doi.org/10.1016/j.clsr.2019.105377

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

VRChat. (2025a). *Community guidelines*. VRChat. https://hello.vrchat.com/community-guidelines

VRChat. (2025b). *Creator guidelines*. VRChat. https://hello.vrchat.com/creator-guidelines

Warofka, A. (2018). Taking action against hate speech in Myanmar. *Facebook Newsroom*. https://about.fb.com/news/2018/11/myanmar-hria/

Zepeto. (2024). *Contents guidelines*. Zepeto. https://docs.zepeto.me/studio/docs/contents_guidelines

Zepeto. (2025). *Community guidelines*. Zepeto. https://support.zepeto.me/hc/en-us/articles/900005448466-Community-Guidelines

Zucker, L. G. (1977). The role of institutionalization in cultural persistence. *American Sociological Review*, *42*(5), 726–743. https://doi.org/10.2307/2094862

Tables and figures

TABLE 1: Meta's community guidelines evolution through four phases

PHASE	GUIDELINE NAME	KEY FEATURES	EMPHASIS
2007	Code of Conduct	"Inappropriate Content" umbrella; emphasis on IP protection	Legalistic tone, minimal categorisation
2011	Community Standards	Introduction of "Community Standards"; Emergence of thematic groupings by title (i.e.: "Graphic Violence," "Bullying & Harassment")	Recognition that Facebook is a community "where millions of people connect with each other"
2015	Community Standards	Introduction of value-oriented organisation (i.e: "Safety," "Voice")	Integration of social responsibility, addressing "harm" through principles
2018–2024	Community Standards	Introduction of policy rationales, examples, transparency reports	Detailed enforcement tiers, appeals mechanisms

TABLE 2: Roblox's policy evolution timeline (2016–2024)

YEAR	POLICY LABEL	STRUCTURE	LANGUAGE (HATE SPEECH POLICY)	EXAMPLE PROHIBITIONS
2016	*Rules of Conduct*	Simple bullet list; minimal formatting; no categories	Directive tone focused on respectful behaviour and "appropriate content." Included both broad moral appeals and some direct restrictions	Explicitly banned "Nazi symbols and imagery"; no formal "hate speech" category
2020	*Community Rules*	Still a flat list; grouped under high-level themes	Slightly more structured; concise rule statements like "Discriminatory speech, symbols, or actions are not tolerated."	Named symbols: "Swastikas," "Confederate flags," "KKK imagery," "racist slogans"
2021	*Community Standards*	Thematic organisation with 4 sections (Safety, Civility, Fairness, Security); examples provided	More narrative framing; rationale-based. Broader phrasing such as "symbols and flags associated with discrimination" introduced	Includes examples like "derogatory memes," "avatar clothing associated with hate groups"
2024	*Community Standards*	Streamlined themes; policy verticals more abstract and principle-oriented	Abstract and proactive; includes off-platform behaviour, broader ideological framing, and protections based on caste, gender identity, etc.	"Content that condones extremist ideologies," "glorifies hate groups," "organized crime"

No.	Policy Theme	Policy Vertical	Horizon Worlds Code of Conduct Mature and Prohibited Worlds	VRChat Community Guidelines Creator Guidelines	Roblox Community Guidelines Restricted Content Polic	Zepeto Community Guidelines Studio Guidelines
1	VIOLENCE AND CRIMINAL BEHAVIOR	Restricted Goods and Services	☑	☐	☑	☑
		Coordinating Harm and Promoting Crime	☑	☑	☐	☑
		Dangerous Individuals and Organizations	☑	☑	☑	☑
		Violence and Incitement	☑	☑	☑	☑
		Fraud and Deception	☑	☐	☑	☑
2	SAFETY	Suicide and Self-Injury	☑	☑	☑	☑
		Child Sexual Exploitation, Abuse and Nudity	☑	☐	☑	☑
		Adult Sexual Exploitation	☑	☐	☑	☑
		Bullying and Harassment	☑	☑	☑	☑
		Human Exploitation	☑	☐	☐	☑
		Privacy Violations	☐	☑	☑	☑
3	OBJECTIONABLE CONTENT	Hate Speech	☑	☑	☑	☑
		Violent and Graphic Content	☑	☑	☑	☑
		Adult Nudity and Sexual Activity	☑	☑	☑	☑
		Sexual Solicitation	☐	☐	☐	☑
4	INTEGRITY AND AUTHENTICITY	Account Integrity and Authentic Identity	☑	☑	☑	☑
		Spam	☑	☐	☑	☑
		Cybersecurity	☑	☑	☑	☑
		Inauthentic Behavior	☐	☑	☑	☑
		Misinformation	☐	☑	☐	☐
		Memorialization	☐	☐	☐	☐
5	RESPECTING INTELLECTUAL PROPERTY	Intellectual Property	☑	☐	☑	☑

FIGURE 1: Matrix showing overlap between XR platform policies and Meta's Community Standards policy verticals. *Note: This comparison is against the version that existed through 2024, before Meta's significant restructuring of content moderation approaches in January 2025.*

New Policy Areas	Horizon Worlds Code of Conduct Mature and Prohibited Worlds	VRChat Community Guidelines Creator Guidelines	Roblox Community Guidelines Restricted Content Policy	Zepeto Community Guidelines Studio Guidelines
Distinguishes & regulates both on-platform and off-platform behavior	☐	☑	☑	☑
Differentiates private and public spaces	☐	☑	☐	☐
Impersonation of platform personnel	☐	☑	☐	☑
Provides separate age gated policies for mature worlds	☑	☐	☑	☐
Prohibits use of platform for medical reasons	☐	☑	☐	☐
Defines personal space	☐	☑	☐	☐
Audio-related policies	☐	☑	☑	☐
NFT integration	☐	☑	☐	☐

FIGURE 2: Matrix of new types of policy verticals not covered in Meta's Community Standards

Appendix

Community Guidelines consulted

Note: Meta's Community Standards are listed through December 2018, the point at which the core harm categories and policy structure that inform the current framework largely stabilised. Post-2018 versions were consulted selectively for specific policy areas cited in the analysis, including the Hate Speech and Harassment policies, and are referenced directly in the text. For Meta's historical documents, versions were accessed through the Platform Governance Archive managed by the University of Bremen. For XR platform guidelines, versions were accessed through platform websites and the Internet Archive's Wayback Machine.

Meta's Community Standards (Facebook)

DATE	DOCUMENT
August 2007	Community Guidelines
January 2011	Community Guidelines
March 2012	Community Guidelines
May 2012	Community Guidelines
August 2012	Community Guidelines
November 2012	Community Guidelines
December 2012	Community Guidelines
November 2013	Community Guidelines
March 2014	Community Guidelines
July 2014	Community Guidelines
March 2015	Community Guidelines

DATE	DOCUMENT
April 2015	Community Guidelines
June 2016	Community Guidelines
July 2017	Community Guidelines
April 2018	Community Guidelines
May 2018	Community Guidelines
June 2018	Community Guidelines
July 2018	Community Guidelines
August 2018	Community Guidelines
October 2018	Community Guidelines
November 2018	Community Guidelines
December 2018	Community Guidelines

Meta Horizon Worlds

DATE	DOCUMENT
May 2022	Horizon Worlds Prohibited Content Policy
November 2022	Meta Horizon Worlds Policy
November 2022	Code of Conduct for Virtual Experiences
November 2022	Personal Space in Meta Horizon Worlds
December 2022	Code of Conduct for Virtual Experiences
May 2023	Meta Horizon Worlds Policy
August 2023	Meta Horizon Worlds Policy
September 2023	Code of Conduct for Virtual Experiences
December 2024	Meta Horizon Worlds Content Guidelines
January 2025	Meta Horizon Worlds Policy
January 2025	Code of Conduct for Virtual Experiences
January 2025	Personal Space in Meta Horizon Worlds

Roblox

DATE	DOCUMENT
September 2016	ROBLOX Rules of Conduct
November 2018	Roblox Community Rules

DATE	DOCUMENT
September 2019	Roblox Community Rules
December 2020	Roblox Community Rules
December 2021	Roblox Community Standards
May 2022	Roblox Community Standards
November 2022	Roblox Community Standards
December 2022	Roblox Community Standards
June 2023	Roblox 17+ Policy Standards
August 2023	Roblox Community Standards
August 2023	Roblox 17+ Policy Standards
October 2023	Roblox Community Standards
December 2023	Roblox Community Standards
October 2024	Roblox Community Standards
December 2024	Restricted Content Policy
January 2025	Roblox Community Standards
January 2025	Restricted Content Policy

VRChat

DATE	DOCUMENT
December 2020	Community Guidelines
December 2021	Community Guidelines
December 2022	Community Guidelines
September 2023	Creator Guidelines
December 2023	Community Guidelines
December 2024	Community Guidelines
December 2024	Creator Guidelines
January 2025	Community Guidelines
January 2025	Creator Guidelines

Zepeto

DATE	DOCUMENT
May 2021	Community Guidelines

DATE	DOCUMENT
November 2021	Community Guidelines
May 2022	Community Guidelines
December 2022	Community Guidelines
September 2023	Content Guidelines (Studio Guide)
December 2023	Community Guidelines
September 2024	Community Guidelines
September 2024	Content Guidelines (Studio Guide)
January 2025	Content Guidelines (Studio Guide)
February 2025	Community Guidelines

Data availability

The community guidelines analysed in this study are publicly available through the respective platform websites and archived versions accessed through the Internet Archive's Wayback Machine and the Platform Governance Archive (University of Bremen).

Published by

Alexander von Humboldt Institute for Internet and Society

CAIS Research for the Digital Age

in cooperation with

CREATe

cnrs centre internet et société

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies