Trust and Safety in Social XR: Mapping the Spatial Turn in Content Moderation

Dennis Redeker, Nikolas Pfannenschmidt, Manuel Baron Romero, Gabriel Durán, and Ana Sofia Villa Hernandez

Abstract. Social Extended Reality (XR) platforms introduce new challenges for content moderation. Unlike traditional social media, XR enables embodied, immersive interaction—intensifying the psychological and social impacts of online harms such as violence, sexual harassment, manipulation, and impersonation. Drawing on an analysis of platform policies and moderation practices, this paper examines how social XR platforms govern these risks. We find that legacy content moderation strategies, such as algorithmic content moderation, are insufficient for the novel characteristics of XR, where harmful material can consist of non-verbal, spatial, and highly engaging behavior. Comparing VRChat's structured policy framework with Horizon Worlds' (now Worlds') more fragmented approach, we highlight gaps in policy clarity, enforcement transparency, and user protection. The paper contributes to emerging debates on platform governance in immersive media, arguing that both state and platform actors should recalibrate their approaches to accountability, real-time moderation, and jurisdictional oversight. We argue that content moderation in XR is not merely a technical challenge—it is a socio-political dilemma requiring participatory, rights-respecting solutions rooted in human rights norms.

1 Introduction

Over the past few years, Extended Reality (XR) technologies have been heralded as the next paradigm shift in digital interaction (Tromp et al. 2022). As an umbrella term for immersive technologies, XR spans a spectrum from Augmented Reality (AR) and Augmented Virtuality (AV) to Virtual Reality (VR) (Grimm, Broll, Herold, Reiners, et al. 2019), with Mixed Reality (MR) encompassing the intermediate states between the real and the virtual world (Milgram and Kishino 1994) (see Figure 1). While early discourses as well as

doi:10.54501/jots.v3i2.290

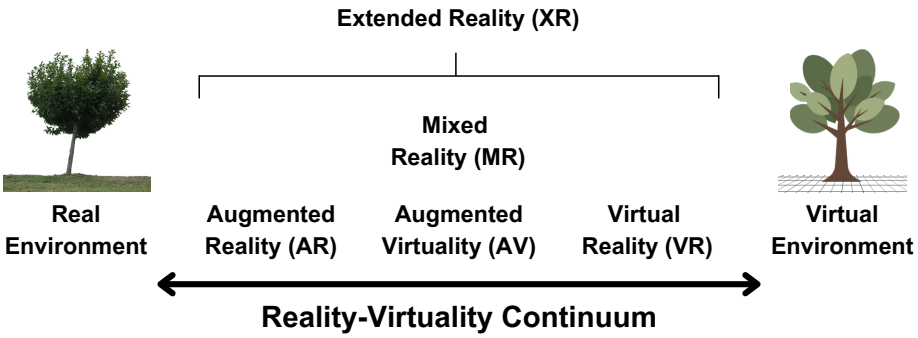

Figure 1: Reality-Virtuality Continuum—adapted from Milgram and Kishino 1994.[1]

investment decisions often leaned into uncritical techno-optimism— predicting societal transformation and digital emancipation—subsequent loss of attention has given rise to equally reductive skepticism of XR. Yet XR should be treated not merely as a trend but as a distinct medium with documented behavioral and perceptual effects, based on three decades of VR experimentation (Bailenson et al. 2025). Indeed, XR may well become one of the socio-technical infrastructures, amplifying the processes underlying 'deep mediatization' (Hepp 2019). Critical scholarship must therefore not be swayed by the upheavals of technological trend cycles and instead interrogate XR through empirically grounded lenses. This paper follows such an approach to examine the governance of online harms in social XR platforms, with a focus on content moderation.

Unlike conventional social media, social XR platforms such as VRChat and Horizon Worlds facilitate embodied, synchronous avatar interactions in shared virtual spaces. These environments feel convincingly real through deeply interwoven psychological states arising from the use of visualization technology (Witmer and Singer 1998). At the same time, this intensification of social experience also amplifies the potential for harm, as Zheng et al. (2023, 2) note: "[s]ocial VR safety risks are often unpredictable, highly personal, and real-time". Practices such as verbal abuse, non-consensual touch, impersonation, or manipulation are no longer disembodied acts on a screen but become enacted experiences with visceral effects that can carry offline world consequences (Zheng et al. 2023). Despite this, platform governance frameworks tend to remain grounded in moderation models developed for traditional asynchronous social media, which may be a mismatch in both form and function (Malik and Usman 2024), as novel avenues for abuse are not being adequately addressed (Trauthig and Woolley 2023). On the regulatory side, existing EU digital governance tools like the General Data Protection Regulation (GDPR) and the Digital Services Act (DSA) are not fully equipped for immersive XR experiences (Hine et al. 2024).

1. Photo credits: Emre Coskun via Unsplash

To critically explore how platforms respond to these novel risks, we conduct a comparative analysis of two prominent yet contrasting XR platforms: VRChat and Meta's Horizon Worlds. Horizon Worlds is the highly invested push into the metaverse by social media giant Meta Platforms, embedded within its broader platform governance architecture (Chow 2022). VRChat, by contrast, is a highly popular, standalone application by a medium-sized company, known for its subculturally active community and a permissive culture regarding copyright enforcement, despite official restrictions (Feldman 2018; VRChat, n.d.-a). Both cases offer distinct approaches to trust and safety — valuable for understanding emerging patterns in immersive content moderation. We center our analysis on four key challenges to user safety in social XR: violence, sexual harassment, manipulation, and impersonation. These harm types are both continuities from earlier social media and emerging in new ways under immersive conditions.

In the remainder of this paper, we outline the conceptual shift toward spatial platform governance (Section 2), map the four selected harm types in XR (Section 3), present our comparative analysis of platform policies (Section 4), examine moderation practices on both platforms (Section 5), and discuss the policy implications (Section 6).

2 The Spatial Turn in Platform Governance

XR is regarded by both policy-makers and industry stakeholders as a key enabling technology for digital transformation (European Parliamentary Research Service 2021; Antao and Likens 2024; Ivey 2024). Somewhat misleadingly, it has occasionally been depicted as a competitor to artificial intelligence (AI), particularly in light of recent shifts in public attention and investment priorities (Faughnder 2024), although the two are better understood as complementary. XR provides the spatial and interactive environments in which AI systems may operate and engage with users (Tromp et al. 2022). This synergy has recently begun to be recognized again, as awareness of both the strengths and limitations of AI grows, and a new generation of smart glasses with AI functions is being launched (Floemer 2025; Reilly 2025).

To date, XR has found applications in various fields, including healthcare, education, retail, and entertainment (El-Hajj 2024). In the latter, sharing technical similarities with VR games, social XR platforms have emerged as an immersive variant of social media. Unlike the flat-screen content of conventional platforms such as Facebook or TikTok, which is viewed on handheld devices or personal computers, the three-dimensional virtual environments of VRChat, Horizon Worlds, or Rec Room are typically accessed through VR head-mounted displays (HMDs) (McVeigh-Schultz, Kolesnichenko, and Isbister 2019). VR HMDs, such as the Meta Quest or Valve Index, cover the user's entire field of view, delivering visual input through integrated displays positioned behind binocular lenses, spatial audio via built-in stereo headphones (Grimm, Broll, Herold, Reiners, et al. 2019), and typically control input through wireless handheld controllers. A range of sensors,

cameras, and/or laser-based systems enable precise tracking of head, eye, and hand movements, allowing these to be accurately translated into the virtual environment that isolates the user from their surroundings. In addition to VR HMDs, there are dedicated AR headsets on the market that overlay digital information onto the perceived real world (Grimm, Broll, Herold, and Hummel 2019), such as the Microsoft HoloLens 2 and Magic Leap 2, as well as so-called MR HMDs, including the Apple Vision Pro and Meta Quest 3, which largely resemble VR headsets but feature a sophisticated pass-through function, blurring the distinctions further.

The resulting user experience is qualitatively distinct from that on other social media platforms. HMDs allow users to navigate vivid computer-generated worlds as first-person avatars, with physical movements mirrored in real time (Freeman and Acena 2022). Psychologically, this impression arises from the interplay of two perceptual states: presence, defined as the complete focus of thoughts and actions on virtuality (Witmer and Singer 1998) and sense of embodiment, understood as the identification of one's physical body with a virtual body (Kilteni, Groten, and Slater 2012). When these mechanisms converge, the depth of immersion can be such that users become so fully absorbed in the virtual environment that they do not consciously perceive their own movements in physical space. Experimental research demonstrates that individuals can be gradually guided to a predetermined real-world location through imperceptible manipulations of the virtual environment, remaining unaware that they have physically relocated—a phenomenon termed the "Human Joystick Attack". While such immersion enables rich interaction, it also carries heightened psychological and physiological implications that distinguish XR from other online platforms (Casey, Baggili, and Yarramreddy 2021).

Beyond its behavioral effects, this depth of immersion also transforms how users experience social interaction and self-representation within XR environments. Users report having "authentic social experiences" (Sabri et al. 2023, 3) and perceiving their physical body as "the immediate and sole interface between them and their avatar" (G. Freeman et al. 2020, 6). Further contributing to this integration is a strong identification with their avatar, which goes beyond that developed on non-immersive platforms. VR research has repeatedly shown that self-avatars can influence behavior, meaning that avatar choice and embodiment are not merely representational features but can shape how users act and respond in virtual environments (Bailenson et al. 2025). Accordingly, some users even playfully utilize this masquerade to explore alternative identities (G. Freeman et al. 2020). While the high degree of personal involvement enables deeply engaging interactions across cultural and geographic boundaries and holds potential for personal growth, the platforms must also account for the psychological impact of harmful experiences in the design of their content governance approach (Gray, Carter, and Egliston 2024).

From a broader perspective, social media platforms—being watched closely by pressure groups, advertisers and governments around the world—have been compelled to develop their own functional content governance regimes (Gillespie 2018; Roberts 2019; Celeste et al. 2023; Marchal et al. 2025). This is due to the fact that these companies do not fit neatly into the "host-editor-dichotomy" (Helberger, Pierson, and Poell 2018, 2) that has historically defined the distribution of institutional responsibility. They are neither classified as telecommunications companies, which are generally exempt from legal responsibility for transmitted content, nor as traditional media companies, which can be held legally accountable for their published content. Instead, as reflected in U.S. federal law, particularly Section 230 of the Communications Decency Act (CDA), they are regarded as intermediaries who are not liable for the misconduct of their users, provided they implement sufficient measures to address third-party complaints. Moreover, they enjoy relative immunity with respect to the enforcement of their chosen content governance approach (Suzor 2019).

In recent years, however, this laissez-faire model has come under increasing pressure from a wave of (supra)national laws, such as the German Netzwerkdurchsuchungsgesetz (NetzDG), the British Online Safety Act (OSA) or the European DSA (Heldt 2019; Kohl 2024). As such, two interrelated modes of governance can be observed: 'governance *of* platforms', where public institutions define the legal boundaries for content handling, and 'governance *by* platforms', where companies autonomously implement their governance mechanisms pursuing corporate interests (Gorwa 2019). More specifically, Poell, Nieborg, and Duffy (2022) identify three strategies of platform steering: regulation, curation and moderation. Regulation refers to the establishment of norms, guidelines, and policies that structure a platform's processes, while curation concerns the organization and prioritization of content and services. Moderation, in turn, entails enforcement of platform rules and policies through systematic monitoring and management of user activity. Lately, the increased use of down-ranking of specific types of content and "shadow banning" (Radsch 2021, 295) has blurred the distinction between the latter two.

On traditional social media, the asynchronous exchange of text, images and videos has gradually proven to be manageable through a combination of human oversight, community moderation and, increasingly, algorithmic moderation systems to cope with scale and reduce the psychological burden of review (Gorwa, Binns, and Katzenbach 2020; Sabri et al. 2023; Schulenberg et al. 2023). From a taxonomic perspective, post-based moderation can be categorized into ex ante (before publication) or ex post (after being published) (Grimmelmann 2015). In social XR, a similarly clear governance approach has yet to be found. In addition to the usual content types, these platforms are characterized by user encounters involving ephemeral speech, spatial behavior, gaze and proxemics. The nature of the exchanges therefore requires immediate and context-aware evaluation by human moderators, severely limiting scalability (Freeman and Acena 2022). Indeed, the strategies employed by corporate and community moderators remind of conventional police work (Ryan-Mosley 2023): vigilant roaming through populated

spaces, demonstrating presence, identifying trouble spots and deescalating conflicts. Accordingly, they wield special powers such as rapid teleportation across world areas—unlike regular users, who can only teleport next to friends—and the authority to remove users from spaces (Sabri et al. 2023; Gray, Carter, and Egliston 2024). Another notable shift is the growing reliance on user-driven content moderation, which transfers greater responsibility from platforms to individual users. This development has led to the implementation of further protective measures: shielded by a personal boundary bubble, users are expected to regulate interactions using tools such as blocking, vote-kicking, or reporting. However, during conflicts, this can place a significant burden on victims (Gray, Carter, and Egliston 2024). Appropriate automated moderation systems are not in sight, as they mostly seem to lack sufficient contextual sensitivity (Schulenberg et al. 2023). Still, our case example Meta Horizon Worlds demonstrates an early use of such in real-time interactions, employing systems that automatically suppress the audio of other users when they are disruptive or using foul language. All in all, since pre-screening content is not feasible in real-time communication, harm prevention gains renewed significance, as reflected in Gray, Carter, and Egliston (2024)'s proposed distinction between proactive and reactive moderation of ephemeral content. We believe that the scale of the difference between traditional social media services and social XR warrants to speak of a *spatial turn in content moderation* as a separate field in which moderation follows different necessities and provides different affordances.

3 Mapping Challenges to Trust and Safety on Social XR Platforms

There is a wide range of social, political and legal challenges related to XR technologies, including physical safety (Kenwright 2023), mental health (Taylor et al. 2024), privacy (Berrick and Spivack 2022) and data protection (Roesner, Kohno, and Molnar 2014). Given the focus on content moderation, in this paper we will examine the security concerns we consider most relevant for social XR: violence, sexual harassment, manipulation and impersonation. While platforms, users and scholars are familiar with these from traditional social media, the specific features of XR make them more complex, more severe and harder to tackle. In this mapping exercise, we explain how these four challenges to trust and safety can be understood in the context of social XR environments.

3.1 Violence

To understand how violence is addressed on traditional social media, and by extension, on XR platforms, it is first necessary to distinguish its different forms. Galtung (1990) categorizes violence into three types: direct, structural, and cultural. These interconnected forms constitute a "vicious triangle" (302), within which they coexist, legitimize and reinforce one another. Direct violence is the most visible form, involving physical or psychological harm inflicted on individuals or groups. Structural violence refers to

the systematic ways in which social structures disadvantage or harm people. Cultural violence, though less visible, is equally harmful and describes cultural patterns that justify or normalize direct or structural violence (Galtung 1990). In the context of social media Mengü and Mengü (2015) define violence as the use of digital platforms to perpetrate or incite harm, aggression, or abuse against individuals or groups. This includes practices such as cyberbullying, stalking, hate speech, the dissemination of violent or graphic content, as well as certain forms of misinformation and propaganda. The relative anonymity and limited accountability often enjoyed by perpetrators further aggravate the problem, leaving victims feeling vulnerable and powerless. In some cases, their suffering may escalate into profound psychological distress, social withdrawal, and even instances of physical (self-)harm (Mengü and Mengü 2015). Arguably, the platform moderation practices of large social media platforms—if permissive to direct violence—may be considered structurally violent, enabling a culture of violence toward certain groups.

While issues of text-based aggression and the sharing of problematic content remain common on social XR platforms, the above-mentioned phenomena not only persist in their original form but reemerge transformed under the auspices of immersive technology. Voice-based harassment, enactments of physical violence through avatars, coordinated bullying and incitements to aggression, stalking, and intrusive avatar following are all part of an expanded catalogue of violations (Sabri et al. 2023). The potential for harm is also heightened. As users largely accept the illusion of being present in the virtual environment and perceive events around them as real (Slater 2009), they react as if these events were real (D. Freeman et al. 2017). This is especially true for threats arising in close proximity to the user, which elicit strong emotional responses comparable to those experienced in real-life confrontations (Rosén et al. 2019). Worse still, contemporary XR platforms import toxic social norms inter alia shaped by specific strands of ‘gamer culture’ (Gray, Carter, and Egliston 2024), thereby introducing new forms of cultural violence into the domain of social media.

3.2 Sexual Harassment

Sexual harassment has been a persistent and well-documented problem on traditional social media platforms and it “arguably reflects the systems of offline structural oppression to control women’s bodies and rights in today’s world” (Schulenberg et al. 2023, 1). It typically encompasses gender-based insults, unsolicited sexual attention, and coercive or degrading behavior, often targeting women, gender-diverse users and minors (Ramirez et al. 2023). These behaviors are frequently normalized or minimized within platform cultures, particularly those shaped by male-dominated or gaming-centric norms, contributing to an environment of exclusion and hostility (Gray, Carter, and Egliston 2024). Yet harmful harassment in these settings has generally remained disembodied — occurring via text, images, or voice.

In social XR environments, the experience of sexual harassment is transformed by the medium's core affordances: immersion, spatiality, and embodiment. Users often report a strong identification with their avatars, which are perceived not simply as representations but as extensions of the self (Freeman and Acena 2022; Kilteni, Groten, and Slater 2012). This intensified sense of bodily presence makes forms of harassment that were once symbolic or mediated, such as unwanted proximity, gaze, voice, or gestures, feel immediate and invasive. Interactions like simulated groping, sexually suggestive movements, or intrusive following are no longer abstract but are instead experienced as enacted violations of personal space. However, not only are established varieties of sexual harassment present in XR, new types have emerged, too. Ramirez et al. (2023, 9) identify four different "varieties of and vehicles for" sexual harassment that are unique to XR: (1) *UX harassment*— "when designers fail to provide fully inclusive embodiment opportunities for users that reflect problematic norms about acceptable and unacceptable bodies", (2) *embodiment harassment*—"when a user takes control of another user's XR body to alter its presentation in non-consensual and sexually explicit ways", (3) *pseudo-allyship* - "embodying oneself as a member of a marginalized group in order to undermine or harass group members", and (4) *deepfake harassment*—"use of deepfake technology to harass others, e.g., taking on the appearance of a person's abuser to traumatize or coerce".

Research has shown that users subjected to such behaviors in immersive environments often respond with physiological and emotional reactions similar to those triggered by real-world threats (Slater 2009; Rosén et al. 2019). The combination of spatial audio, haptic feedback, and real-time interaction heightens these effects, producing strong feelings of fear, shock, or helplessness (D. Freeman et al. 2017; Ramirez et al. 2023). Because these events unfold in real time and are often short-lived, the harm is compounded by the difficulty of registering or processing the violation as it occurs, reducing the "users' ability to successfully report abusive behavior" (Blackwell et al. 2019, 10). Avatar customization and embodiment introduce additional layers of vulnerability, as users' virtual bodies become direct targets for visual and spatial harassment (Ramirez et al. 2023). Overall, the immersive setting blurs the line between virtual and physical experiences, making such encounters more psychologically taxing than comparable incidents on screen-based platforms.

3.3 Manipulation

Manipulation in online spaces can be broadly defined as the covert use of information technologies to influence an individual's decision-making process by exploiting cognitive vulnerabilities (Susser, Roessler, and Nissenbaum 2019). It must be distinguished from persuasion, which involves the neutral presentation of arguments (such as product features or political pledges) that enable individuals to make autonomous and reflective decisions. Manipulation, in contrast, aims to induce individuals to act involuntarily or counter to their own best interests, potentially causing harm in the process (Mhaidli and Schaub 2021).

As XR platforms evolve into commercially viable and densely populated ecosystems, their attractiveness as targets for exploitation increases accordingly (Smaili and Rancourt-Raymond 2024). In anticipation of this development, Mhaidli and Schaub (2021) conducted a scenario-based study to explore five potential mechanisms for user manipulation. Their findings suggest that manipulative strategies may emerge through deceptive marketing experiences, the targeting of vulnerable populations, the distortion of users' perceived reality, the artificial induction of emotional states, and hyper-personalization techniques. Their analysis suggests that XR may expand the repertoire of manipulative practices by combining persuasive content with immersive sensory design and increasingly granular user profiling. Moreover, the persistent connectivity and multisensory character of XR systems enable the large-scale collection of user data, often beyond what is disclosed. Demographic information volunteered by users may be supplemented with biometric data, behavioral patterns, geolocation traces for the creation of highly detailed personal profiles. These profiles can be used to fine-tune persuasive interventions, which may range from targeted advertising to subtle behavioral nudging (El-Hajj 2024). By inferring emotional states, time windows of vulnerability can be detected to further increase users' susceptibility (Mhaidli and Schaub 2021). In the same way, a further line of concern relates to interface-level manipulation or UI attacks. Krauß et al. (2024) argue that XR-specific properties such as perception, spatiality, physical/virtual barriers, and device sensing can amplify deceptive design, enabling practices such as strategic content placement, spatial imbalance of options, and forced data capture. Complementing this, Cheng et al. (2024) show that XR systems can be vulnerable to interface-security failures such as clickjacking-like overlays, invisible interface elements, and synthetic-input attacks, which further blur the line between persuasion and deception. In this respect, manipulation in XR is not limited to what content says, but extends to how the interface itself structures attention, movement, and choice.

Compared to traditional screen-based social media, disinformation campaigns on XR may prove especially impactful because immersive environments can heighten realism, and the same properties that make VR persuasive may also amplify misinformation (Brown, Bailenson, and Hancock 2023; Han et al. 2022; Trauthig and Mimizuka 2022). This risk is heightened by the fact that manipulation in XR may also operate through asymmetries in perception between users. Research on transformed social interaction (TSI) has shown that collaborative virtual environments can present the same interaction differently to different users, which enables social cues and information flows to be modified in real time for strategic ends (Bailenson et al. 2004, 2005). In these contexts, media, human-like bots, and social cues should not be seen only as stand-alone pieces of content. They can also be built into the way a virtual environment, an avatar, or an interaction is presented to specific users. This helps explain why such manipulative elements may face less critical scrutiny in XR, since users encounter them within an immersive and tightly controlled sensory setting that can feel immediate, believable, and socially convincing (Falchuk, Loeb, and Neff 2018; Heller and Bar-Zeev 2021; Brown, Bailenson, and Hancock 2023).

3.4 Impersonation

Impersonation can generally be understood as the act of assuming another person's identity, typically with the intent to deceive or mislead others. On conventional social media, this commonly occurs through fake profiles, identity theft, or identity cloning. Fake profiles refer to accounts that do not correspond to an actual person. While not explicitly harmful, they are frequently abused for spamming and phishing real users. In cases of identity theft, adversaries gain unauthorized access to a genuine account in order to exploit the associated data for illicit activities. Identity cloning, on the other hand, involves creating a fabricated account with stolen data that imitates a real user, often for similarly deceptive or criminal purposes (Alharbi et al. 2022). Effective prevention of such practices greatly hinges on thorough authentication mechanisms during the registration process (Zhang and Gupta 2018), effective cryptographic techniques for secure communication and AI-based threat detection (El-Hajj 2024).

In the three-dimensional virtual environments of social XR platforms, the hurdles for malicious actors are higher. It is more difficult to recreate another user, because full-body replicas, voice, and movement patterns have to be emulated as well (Abraham et al. 2022). Nonetheless, when executed with a certain rigor, perpetrators may achieve a more convincing impersonation in a more intimate setting compared to social media (Lin and Latoschik 2022). As a result, it is easier to deceive and extract (corporate) information during interactions (Falchuk, Loeb, and Neff 2018), a tactic known as social engineering (Krombholz et al. 2015). With the further popularization of social XR platforms, these risks could turn even more far-reaching when professional roles are involved and assets such as cryptocurrencies or NFTs are transferred (Peukert et al. 2024). The protection of identity of users, beyond economic risks, appears as a central topic of debate on its own and "[p]reserving the integrity of individuals' identities and protecting them from impersonation or manipulation is crucial [and] not only a matter of privacy but also of personal security and trust in digital interactions" (Szita et al. 2025, 8).

These four types of harms—violence, sexual harassment, manipulation and impersonation, which represent challenges to trust and safety on social XR require platforms to put in place meaningful measures, both in terms of content and behavioral policies and internal moderation mechanisms.

4 The Policies of the Spatial Turn in Content Moderation

To assess the extent to which these four major harm types are addressed in current XR content governance approaches, we conducted a comparative analysis of platform policies and reconstructed moderation practices based on the material and additional official release notes as well as press coverage. Focusing on two prominent yet contrasting platforms with regard to their financial resources and business ethos, we compared the

more community-oriented, independent-style platform VRChat (VRChat Inc.) with the more centralized and strategically aligned Horizon Worlds (Meta Platforms Inc.) (Chow 2022; Feldman 2018; Meta, n.d.-ab; VRChat, n.d.-h). While doing so, we acknowledge that it is not only Western and mainstream platforms that are worthy of scholarly attention, a point well articulated by others (Gillespie et al. 2020; Hallinan et al. 2025). This comparative study on two cases can only be considered to be a further step of an expanding research agenda on trust and safety on social XR platforms.

4.1 Methods: Studying Platform Policies

Researchers of XR trust and safety at times conduct digital ethnography, directly engaging with safety measures in the virtual environments (Zheng et al. 2022). Yet other research avenues rely on user/usage data provided by platforms in order to investigate the mechanisms of content moderation. This practice has become increasingly difficult, however, as voluntary data sharing by platform companies decreased over the last few years (Coalition for Independent Technology Research 2025). In response, Article 40 of the DSA establishes a formal framework through which "vetted researchers" (European Union 2022, 25) may request access to data from Very Large Online Platforms (VLOPs). The 45 million monthly user threshold still significantly limits its applicability, as smaller platforms fall outside its reach. Since there is great value in better understanding the private ordering of trust and safety on these platforms, studying platform policies is a viable approach to better understand how the rules *ought to be*. After all, technical standards or platform rules can have effects similar to state regulation (Bygrave 2015; Celeste 2019; Suzor 2019). This "platform law" (Bygrave 2015) thus reveals to researchers something about the way platforms (and their owners) think about risks users encounter—including with regard to the four challenges in social XR outlined above. The framing of risks in these policies matters for effective engagement with them: systematic enforcement requires well-drafted rules—and for them to be publicly available builds trust.

Studying social media platform policies—such as legal terms of service documents and community guidelines—allows researchers to understand the 'private ordering' or attempts thereof of the digital environment (Celeste 2019; Quintais, De Gregorio, and Magalhães 2023). The scholarship on platform policies is blooming. Scholars study how platform content moderation policies conform with or relate to human rights (Celeste et al. 2023) or which normative values they express (Scharlach, Hallinan, and Shifman 2024; Hallinan et al. 2025). Depending on the kind of analysis, access to platform policies is key. While they are usually easily available from the platforms themselves, historical versions are harder to come by. Recently, the Platform Governance Archive (Katzenbach et al. 2023) has increased its scope to track changes in more platform policies over time. However, so far, social XR platforms are strikingly missing in the collection.

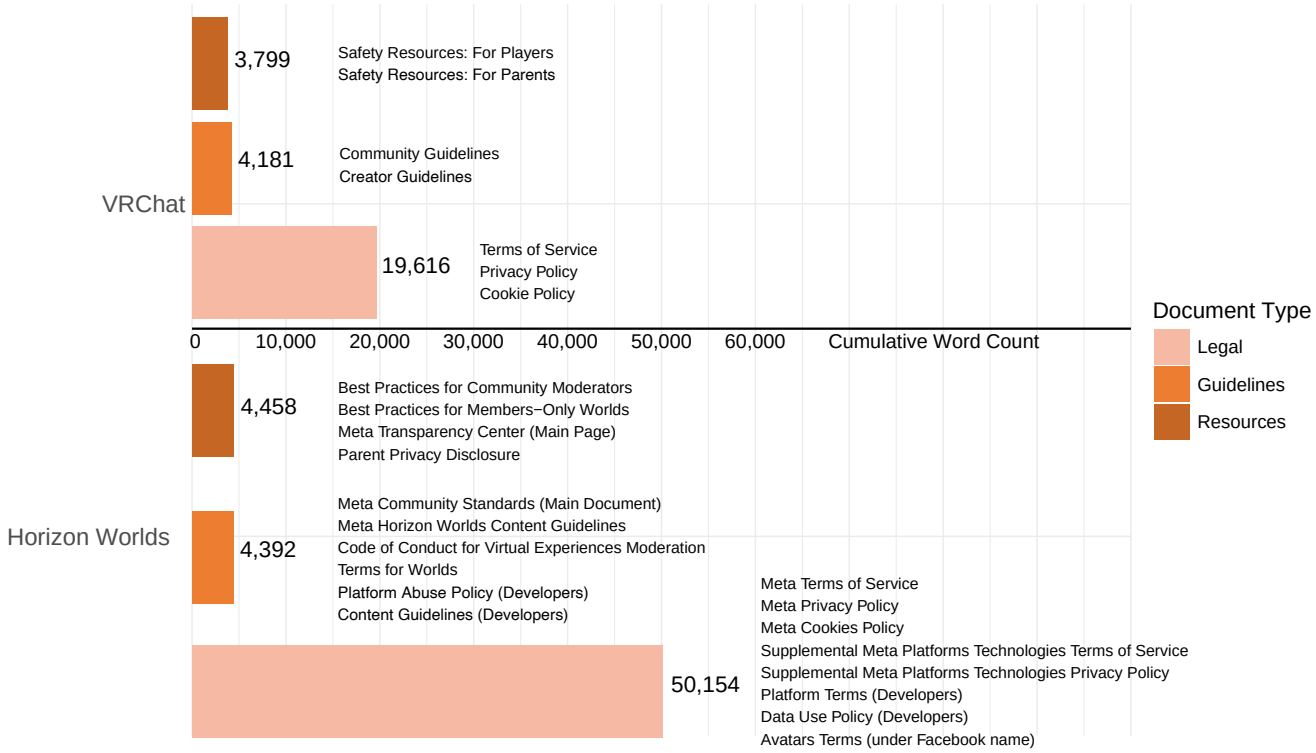

Figure 2: Policy documents included in analysis, by category and word count.

For this paper, we collected platform policy documents from VRChat and Horizon Worlds in order to examine how these platforms frame and deal with the four identified challenges to trust and safety outlined above.[2] The relative length of these documents, for each platform and type of document, is shown in Figure 2. While document size is not a direct indicator for governance quality, the number and scale of policy documents nonetheless signal how each platform approaches the scope, structure, and accessibility of its ruleset. Qualitatively, understanding how different platforms perceive violence, sexual harassment, manipulation and impersonation tells us much about these problems and perhaps even more about how platform companies frame complex socio-technical constructs. Herein lies the strength of the comparative approach we take. We hope that in the future, an even more comprehensive analysis of social XR platform policies will be conducted to grasp a larger number of accounts of rules for XR. However, we also contend that studying how the rules *ought to be* on platforms— through studying policies—needs to be followed up by rigorous study of de facto moderation practices. Hence, in this paper, we also analyze publicly available documents and accounts of others in order to elucidate how the practices of content moderation relate to the analyzed policies. In doing so, we distill a number of challenges for content moderation on social XR. In the remainder of this section, we introduce the two platforms in question, identify relevant policy documents for each, and present a thematic analysis of these documents.

2. These data are available in Pfannenschmidt 2025.

4.2 VRChat

VRChat is a free social XR application released on the software and game distribution platform Steam (Valve) in February 2017 (SteamDB, n.d.). Since then, it has developed into a subcultural meeting point, featuring customized avatars, worlds, games, and "new social rituals and memes" (McVeigh-Schultz, Kolesnichenko, and Isbister 2019). As of March 2026, the platform counts approximately 45,000 average concurrent players on Steam alone (SteamDB, n.d.), with additional users across Google Play, the App Store, Pico and Meta Quest, Meta Rift and Viveport (VRChat, n.d.-h). The application is globally accessible by users aged 13+ under parental supervision, or without restrictions from the age of 18 via desktop PCs, Android/iOS mobile devices or a range of compatible HMDs from various manufacturers (VRChat, n.d.-j). The overall rating on the Steam Store is generally positive, with a satisfaction score of 79% out of about 250,000 ratings. In the ten most upvoted English-language reviews, criticism focuses on the high number of underage users, concerns about malicious actors, racist/harassing speech, and perceived deficiencies in moderation, as well as the subscriber-only third-party age verification system (VRChat, n.d.-i). The platform's policy documents are made readily accessible on its website, presented in a clear, hierarchically structured, and interlinked format. In this context, we identified and examined seven policies and guidelines as relevant to this mapping exercise (see Figure 2).

The platform's "Terms of Service" constitute the most comprehensive and binding collection of social norms, legal provisions, and disclaimers, forming the foundation for the "Community Guidelines" and "Creator Guidelines", which are both written in plain, direct language and supported by illustrative examples. As expected, it states that there is an unlimited license grant for all user-created content and that visiting worlds occurs at the user's own risk. While the "Terms of Service" are unlikely to be read by most users, together with the "Privacy Policy" and "Cookie Policy", all documents must be acknowledged and adhered to in order to access the application (VRChat 2024a, 2024b, 2024c, 2025a, 2025b).

The guidelines therefore serve as practical reference points for a broad audience. The "Community Guidelines" define expected behavior, including prohibitions against hate speech, sexual harassment, violence, misinformation, deception, and other unlawful conduct. They also prohibit the technical misuse of the platform, such as in-game cheating or the circumvention of moderation systems. A distinction is made between public and private spaces and groups. In private instances, "inflammatory behavior" is permitted, provided that all participants consent and it does not conflict with either the stated world rules or the authority of the instance owner. A comparable degree of autonomy applies to groups, which may establish their own rules as long as they remain "fair and [suitable] for the environment" (VRChat 2024a). Lastly, the guidelines describe the available self-moderation tools (VRChat 2024a).

The "Creator Guidelines" further specify the requirements for acceptable content for both content creators and regular users on the platform. These provisions cover language, behavior, avatars, worlds, iconography, groups, and all other content that users "upload, load, put into, put on, display in, make with, and/or make within VRChat" (VRChat, n.d.-a). Generally, all content must avoid discrimination, harassment, disruption, deception, or illegal activities, with no exceptions granted for "roleplay" or claims of "historical significance" (VRChat, n.d.-a). A notable feature is the requirement to use "Content Warning Labels", designed to ensure that avatars and worlds containing sexually suggestive, excessively violent, disturbing, or adult content can be automatically filtered (VRChat, n.d.-a).

Although there are minor redundancies and occasional overlaps across the policies, VRChat's policy framework is functional, coherent, and concise. The framework demonstrates an intention not only to inform but also to educate the player community and promote self-regulation of problematic behavior. This is supported by additional easily accessible material such as "Safety Resources: For Players" and "Safety Resources: For Parents", training videos, the official VRChat Wiki, and references to the policies of other platforms and online safety initiatives that users may consult (VRChat, n.d.-d, n.d.-e, n.d.-j). The effectiveness of the moderation system, as reflected in the user feedback, appears to be only partially influenced by these efforts.

4.3 Horizon Worlds

Meta's free social XR platform Horizon Worlds, recently rebranded as Worlds, has attracted wide media attention since its release due to its central role in the company's metaverse strategy, but has struggled to retain users and become profitable (Heath 2021; Murphy 2021; Mann 2025; Meta 2025a; Tassi 2025). This may have contributed to Meta's announcement in March 2026 to retire the platform's VR functionalities and transition it to mobile only. Yet, after user criticism, the company swiftly reversed the decision, upholding core VR features for the "foreseeable future" (Bosworth 2026) while disabling the creation of new worlds (Ashworth 2026; Meta 2026). Originally launched in late 2021 in the United States and Canada on Meta's proprietary digital store, the platform has since expanded to 25 countries across four continents and gradually lowered its age limit from adults to 13+ with parental supervision, now even permitting access to preteens aged 10+ in some countries (Meta 2021, 2024g, n.d.-y). Despite this expansion, official user numbers dropped from 300,000 monthly users in February 2022 to 200,000 in October 2022, after which Meta ceased publishing user statistics (Tassi 2022, 2025). Within these 25 countries, access is possible via desktop PCs, Android/iOS mobile devices and Meta-branded HMDs (Meta, n.d.-g, n.d.-ab). As of March 2026, its overall rating on the Meta store is mixed, averaging 3.4 out of 5 stars based on about 18,000 ratings. A recurring theme among the ten most upvoted reviews concerns frustration with the automatic installation of the application on Meta HMDs, the prevalence of underage users, the lack of engaging content, and repeated instances of toxic user behavior (Meta, n.d.-ab). Given

the large palette of Meta products, Horizon Worlds' policy framework consists of a vast set of documents applicable to different audiences, such as general users, developers and advertisers, spread across different domains and subdomains with inconsistent menu structures. Furthermore, products are often subject to multiple product classifications and overlapping policies. Horizon Worlds is simultaneously classified as a "Meta company product", as well as a "Meta Platforms Technologies Product" and "Meta Quest" software (Meta, n.d.-s). For the purpose of this study, eighteen policy, guideline and best practice documents affecting trust and safety parameters for end-users have been reviewed and analyzed (see Figure 2).

On the Facebook.com domain, the "Meta Terms of Service" form a general agreement describing the services offered, data usage practices, expected user conduct, and access restrictions across Meta's product portfolio. It includes instructions for reporting violations and references a long list of other additional policies. Although these obligations are initially framed in relation to "Facebook and our community" (Meta 2025d), the subsequent paragraphs make clear that they apply to all Meta products. In a similar way, the "Meta Platform Terms" set out requirements for developers related to data use, security, copyright, and third-party technologies (Meta, n.d.-d). Of particular interest are the "Meta Avatars Terms of Service", which cover usage rights and responsibilities related to avatars and digital clothing (Meta 2024a). Curiously, besides Meta claiming broad rights over the user-generated avatars, it requires users to acknowledge that "any Meta Avatar generated from your User Content is not intended to reflect your Persona" (Meta 2024a). Meta's "Privacy Policy" and "Cookie Policy" are presented in a "Privacy Center", designed for greater accessibility. It offers paragraph summaries, examples, modal dialogs, videos, and links to manage data settings. A summary table outlines how different data types are used (Meta 2023a, 2025c). Although the information is presented in a more interactive way, it is extensive, lacks intuitive structure, and is often only accessible by opening dialog windows.

The other policy and guideline documents are found on the Meta.com domain. Here, the "Code of Conduct for Virtual Experiences" is the main reference for Horizon Worlds users, creators, admins, and developers regarding acceptable behavior. To support the stated values of "voice, safety, authenticity, dignity, and privacy" (Meta, n.d.-c) a list of specific violations is named, covering platform abuse, deception, (sexual) harassment, spam, violence, copyright issues, and other criminal acts. These violations also appear in a far more extensive list in the Community Standards with separate policies, including a protection goal, called "Policy Rationale", as well as conditions and examples of what is and is not allowed. This section is also the only place where past versions of policies are stored in a chronicle—much like in the above-discussed comparative Platform Governance Archive, allowing users and researchers to trace changes over time (Meta, n.d.-d). Still, somewhat confusingly, it states that it applies to "Facebook, Instagram, Messenger and Threads" even though its integration into the "Terms of Service" renders it relevant for all Meta products, including Horizon Worlds (Meta 2025d, n.d.-d). Ultimately,

it remains unclear to what degree Horizon Worlds' content moderation regiment refers to this latter collection of more detailed policies.

On the more practical side, three resources from the Meta Quest Help Center, "Meta Horizon Worlds Content Guidelines", "Best Practices for Community Moderators", and "Members-Only Worlds Governance Best Practices", explain content moderation. The guidelines require creators to assign age ratings of 10+, 13+, or 18+ depending on content. For the 18+ category, the platform allows mature themes such as sexual content, violence, strong language, and depictions of recreational drug use (Meta, n.d.-q). The best practice material aims to support moderators, admins and creators of private worlds in managing their spaces "fair[ly] and respectful[ly]" (Meta, n.d.-a) in line with the "Code of Conduct" and their own set up rules. This includes maintaining clear communication and applying moderation tools progressively, from verbal warnings, to formal reports, and ultimately to the removal of a player (Meta, n.d.-c, n.d.-m).

In addition to policies for XR-specific use cases (Meta, n.d.-f) and associated privacy aspects (Meta, n.d.-e), Meta provides separate documents for other stakeholders. Developers are subject to rules on content restrictions (Meta, n.d.-c) data processing and system misuse (Meta, n.d.-b). Advertisers must follow ad policies (Meta, n.d.-i) and parents are given separate privacy terms (Meta 2025b). Along with setup guides for parental controls (Meta, n.d.-f) there is a Family Center offering educational resources and links to outside support pages (Meta, n.d.-o). It is worth noting that Meta has placed itself under the oversight of the multi-stakeholder "Oversight Board" and also operates a dedicated "Transparency Center", where a wide range of reports is compiled for various use cases. However, all these instruments deal solely with the matters of Facebook, Instagram, Whatsapp or Threads (Meta 2024f, n.d.-z).

In general, Meta's policy framework for Horizon Worlds tries to appear all-encompassing and provide interactive navigation, but oftentimes lacks a clear structure and overview. Although it may be suitable for legal compliance, it is less helpful for user orientation, since relevant resources are scattered across domains and the scope is not always clear. The decision to open the application for (pre)teens, and automatically install it on all Meta HMDs, yet continue to permit mature content equivalent to an "R-rating for movies" (Meta, n.d.-a) underscores the platform's risky approach of prioritizing user growth.

4.4 Focus: How Platform Policies Address the Four Sample Challenges

To bring a general idea on how Meta's Horizon Worlds and VRChat address violence, sexual harassment, manipulation and impersonation in their internal policies, Table 1 contrasts shared areas and differences in emphasis, coverage and enforcement logic across the two platforms.

Table 1: How Platform Policies Address the Four Challenges

Harm Type	Platform	Policies
Violence	Both	• Broad prohibitions on direct violence • Rather limited addressing of structural and cultural violence
	VRChat	• Focus on doxxing and weight-based bullying • Zero tolerance for critical harms, such as threats, terrorism, child sexual exploitation and abuse material (CSEAM) • Protection of minors through content warnings
	Horizon Worlds	• Focus on outing and swatting • Sub-policies covering diverse cases of violence on the platform • Protection of minors via age gated world ratings
Sexual Harassment	Both	• General bans on harassment, abuse, stalking, threats, and illegal sexual conduct • Restrictions on sexually explicit content in public spaces
	VRChat	• Public/private spatial separation for sexual or intimate content • Reliance on private consent for allowing sexual conducts • Zero-tolerance banning of CSEA/CSAM and non-consensual intimate imagery
	Horizon Worlds	• Explicit bans on sexual assault, sexual exploitation, unwanted sexual contact (in diverse sub-policies) • Explicit bans on lewd/obscene/vulgar avatars • Prohibition of sexually explicit or provocative public worlds/events • Focus on sexual exploitation of adults and minors, solicitation, explicit language
Manipulation	Both	• Prohibitions on scams, misleading, deceptive, or fraudulent conduct • Limits around data use and advertising personalization
	VRChat	• Explicit bans on malware, hacking, phishing, and social engineering • Prohibitions on attempts to mislead, trick, fool, or confuse users • Explicit bans on misinformation/false information • Use of personal data for own and third-party promotional/advertising purposes
	Horizon Worlds	• Explicit bans on hacking, scamming and doxing • Personalized advertising through platform AI recommendation systems • Statement that user data are not sold or directly shown to third-party advertisers
Impersonation	Both	• General bans on impersonation, identity theft, and use of fake accounts
	VRChat	• Explicit bans on impersonating authority roles • Explicit bans on impersonating other users or communities, including age misrepresentations for fraudulent activities • Link between impersonation, platform integrity and account-security harms such as unauthorized access
	Horizon Worlds	• Advisory to label parody/role-play clearly to avoid confusion

4.4.1 Violence

As argued at the outset, the experience of violence in virtual environments can have a lasting psychological impact on users and therefore requires an effective policy response. Generally, both platforms cover direct violence and its depictions quite well, though with markedly different tones. VRChat's approach is notably authoritative in character. As the first rule of its community guidelines, VRChat prohibits any conduct that may "hurt or harm other people, [...] upset, [...] cause them trouble, or disturb them" (VRChat 2024a). This is supplemented with a zero-tolerance policy for critical harm, imposing "immediate, permanent, and irreversible" (VRChat 2024a) user bans. Here, powerful vocabulary is coupled with intentional vagueness, in that "offenses encompass, but are not limited to, the distribution of child sexual abuse material (CSAM), child sexual exploitation and abuse (CSEA), non-consensual intimate imagery, credible threats of violence posing real-world physical harm, and terrorist content"(VRChat 2024a). This is complemented by provisions forbidding "[h]armful, hateful, or illegal activity [...], even if everyone is informed or consents" (VRChat 2024a), effectively closing the loophole of consensual harm. Tackling distinct digital forms of direct violence, VRChat specifically focuses on "doxxing" and "weight-based bullying" (VRChat 2024a). Its policies against structural violence follow the same prescriptive tone, as VRChat states that it "do[es] not tolerate intolerance" (VRChat 2024a), exemplified in the protection of minors through content warnings (VRChat, n.d.-a). The more abstract form of cultural violence is tangentially covered through the prohibition of "organization, promotion, or support of violent extremism, including the glorification of violent events, perpetrators of said events or acts, or similar behaviors" (VRChat 2024a).

Meta Horizon Worlds, by contrast, adopts a more descriptive, rather than moralizing rhetorical approach. Its community guidelines refer to various forms of violence, including "[b]ullying, harassing, stalking or hateful behavior" and "[p]romoting or coordinating acts of physical harm, such as sexual or physical assault, or suicide or self-harm" (Meta, n.d.-c). Meta's "Community Standards" divide related issues further into sub-policies covering violence, hate, incitement, self-injury, and crime (Meta, n.d.-d), featuring an almost comprehensive collection of diverse cases. This structuring may demonstrate the platform's alertness and assists moderators in case-specific decisions, but does little to improve user orientation. Among distinct digital forms of direct violence, Horizon Worlds specifically addresses "outing" and "swatting" (Meta, n.d.-d). On structural violence, Meta retains its descriptive style, forbidding "targeting a person or group of people on the basis of their protected characteristics" (Meta, n.d.-d). Also notable in this context is the protection of minors through age-gated world ratings (Meta, n.d.-q). Cultural violence is, similar to VRChat, addressed marginally through the prohibition of "[a]dvocating, engaging in or promoting violence" (Meta, n.d.-c).

4.4.2 Sexual Harassment

Given the long history of gender-based harassment in virtual environments, both platforms are expected to show particular diligence in mitigating such harms. However, VRChat's guidelines do not explicitly mention sexual harassment, although it is addressed indirectly. In addition to the fundamental "no hurt or harm" imperative, they prohibit "sexist, racist, [and] hateful" conduct (VRChat 2024a). Users may not "harass, abuse, stalk, threaten, defame, or [...] use the platform for any illegal purpose" (VRChat 2025a). To protect users from disturbing encounters, VRChat maintains a clear distinction between public and private instances. Only in the latter are "sensitive, intimate, or provocative" (VRChat, n.d.-a) avatars, content or conduct permissible among "consenting adults" (VRChat 2024a), supported by content warnings. In public spaces, users are advised to behave as one would in "a shopping center, or a family-friendly beach" (VRChat, n.d.-a), a practical orientation. As noted above, the zero-tolerance policy applies to child sexual exploitation and non-consensual intimate imagery, equally relevant in the context of sexual harassment (VRChat 2024a).

Meta's Horizon Worlds addresses sexual harassment mainly within the broader context of physical assaults and bullying. The "Code of Conduct for Virtual Experiences" prohibits illegal or harmful acts, including sexual or physical assault, exploitation, and harassment (Meta, n.d.-c). The "Community Guidelines" reinforce this in several issue-specific guidelines about sexual violence, prohibiting "content that depicts, threatens or promotes sexual violence, sexual assault or sexual exploitation" (Meta, n.d.-d) and "[u]nwanted contact that is: [r]epeated, [or] [s]exually harassing", "[a]ttacks based on [...] experience of sexual assault, sexual exploitation, sexual harassment" and "sexualized commentary" (Meta, n.d.-d). Special protection applies to individuals based on protected characteristics, such as "sexual orientation, sex, [and] gender identity"(Meta, n.d.-d). More restrictive than VRChat regarding content types, Meta prohibits the creation of "obscene,[...] vulgar, [and] lewd" (Meta 2024a) avatars. Likewise, public worlds and events may not be "sexually explicit or provocative, including nudity, depictions of people in explicit positions" (Meta, n.d.-q). Overall, Meta's policy framework provides broader protection against sexual harassment, while VRChat offers clearer guidance and stricter boundaries for consensual interactions.

4.4.3 Manipulation

Immersion brings a range of entry points for manipulative endeavours and has therefore a special place in both platforms' policy documents. In VRChat, users are prohibited from engaging in "activities intended to cause damage or gain unauthorized access to another user's account, network, or system" (VRChat 2024a). This includes distributing malware, engaging in phishing, using other hacking or social engineering techniques or any attempt to "mislead, trick, fool, or confuse people" (VRChat 2024a). Furthermore, "false information or misinformation" is strongly prohibited (VRChat 2024a). Regarding data use, personal information may be utilized for "developing and providing [own

and third-party] promotional and advertising materials" (VRChat 2024c). Ultimately, corporate own cookies and third-party cookies are said to be employed deliberately (VRChat 2024b). Several of these practices are also broadly prohibited in the platform's Community Guidelines. Users may not employ the service for any unlawful purpose or in any manner that infringes the rights of other users, or other third parties. The platform reserves wide discretionary powers to enforce these rules: it may suspend or terminate a user's account at any time, "for any reason or no reason" and with or without prior notice (VRChat 2025a).

Meta Horizon Worlds on the other hand prohibits "unlawful, misleading, [...] fraudulent" conduct on their platform (Meta, n.d.-g). Its Code of Conduct mentions this by banning any behavior "designed to deceive other users", including any activity engaging in fraud or scams, and any form of unauthorized access or doxing (Meta, n.d.-c). These rules directly target manipulative and deceptive practices within immersive environments. While the platform uses its own advertising service to deliver AI-personalised ads, it states that it does not sell or disclose personal user data to third parties (Meta 2023c). In addition, creators, developers, and admins bear primary responsibility for enforcing the Code of Conduct within their own virtual spaces, with Meta intervening only when persistent or severe violations occur (Meta, n.d.-c).

4.4.4 Impersonation

Both platforms recognize the issue of impersonation and have included specific sections in their community standards to address it. As stated, VRChat has a clear stance on prohibiting the impersonation of authority roles, and generally forbids impersonating other users and misrepresenting communities with the intent to mislead or confuse others. As mentioned above, VRChat interdicts activities intended to harm or gain unauthorized access to another user's account, network, or system (VRChat 2024a). The Terms of Service in VRChat specifically refer to a prohibited conduct category that bans "any fraudulent activity" (VRChat 2025a), which includes impersonating any person or entity, misrepresenting one's affiliation or identity, accessing another user's account without authorization, or falsifying one's age or date of birth.

Horizon Worlds, while not explicitly detailing impersonation in the same way, deals with this concern in their Code of Conduct, emphasizing the importance of authenticity and safety. Although the platform's policies may not explicitly enumerate every form of impersonation, they strongly discourage any behavior that involves pretending to be someone else, committing identity theft, or using fake accounts. Such actions are deemed inconsistent with the platform's values of authenticity, safety, and privacy. To promote transparency, Meta advises users to clearly indicate when they are engaging in parody or role-playing to prevent misunderstandings (Meta, n.d.-c).

5 The Practices of the Spatial Turn in Content Moderation

While examining platform policies provides essential insights into how content moderation should function in theory, we argue this must be complemented by investigation of actual moderation practices. Yet, this proves difficult as neither platform gives access to moderation performance indicators. To nonetheless analyze empirical material, we first conducted a press review based on two keyword-based Google News searches for major English-language news outlets, collecting the most relevant reports published before 1 August 2025, followed by further issue-focused Google News searches[3]. We drew primarily on established online newspapers, supplemented by coverage from technology outlets. Journalistic accounts are particularly suited to this purpose, as they are, in contrast to user reviews and forum entries, subject to editorial standards and platforms are able to contest false portrayals. Still, there are limitations to this approach. Google News is an algorithm-based news aggregator useful for detecting prominent online news articles, but not providing a comprehensive sample. Furthermore, as effective moderation tends to remain invisible, incidents without media resonance are excluded from our sample so that the cases covered are not representative. Accordingly, we make no claims about incident frequency or objective performance. What this approach does allow is the identification of platform-specific patterns across high-profile cases, which carry documentary value in terms of their development and the platform responses they provoked. Thus, in a second step, the resulting content governance regime is described on the basis of platform documents, including release notes, best practice documents, guidelines, and user resources, providing a detailed account of the actors, instruments, and procedures involved for each platform. Through this dual analytical approach—examining both what platforms claim to do and what they actually do in terms of creating structures and affordances, or direct intervention—we identify several critical challenges facing content moderation in social XR.

5.1 VRChat

As one of the earliest social XR platforms, VRChat has received polarized media attention, with much of the criticism concerning its ineffective content moderation approach. The platform gained widespread attention between late 2017 and early 2018, following a rapid surge in user numbers that attracted a handful of first-person reports. While early articles highlighted the platform's potential for "creativity", "friendship" (Feldman 2018), and the "[appealing] chaos" (Gault 2018), the "[virtually endless] possibilities" of user-generated content (Fagan 2018), they also pointed out to its regulatory shortcomings, describing it as a "new virtual Wild West" (Fagan 2018). During this period of rapid growth, VRChat was operated by a small development team and lacked essential moderation

3. The Google News searches were conducted using the queries "Horizon Worlds content moderation" and "VRChat content moderation", reviewing results on all displayed search pages and supplementing these with targeted Google News searches on the specific issues violence, sexual harassment, impersonation, manipulation, child safety, as well as the Horizon Worlds Buzzfeed experiment and the Ugandan Knuckles incident on VRChat

infrastructure (VRChat 2018). As a result, users from marginalized groups were subjected to repeated harassment, trolling, sexual misconduct, and hate speech (Fagan 2018; Feldman 2018; Gault 2018). A prominent culmination of these issues was the rise of the "Ugandan Knuckles" meme phenomenon, which became closely associated with VRChat's early public image. Users adopted distorted avatars of the character Knuckles from the video game "Sonic the Hedgehog" and used voice clips from the Ugandan action film "Who Killed Captain Alex?" (Feldman 2018). This led to large groups of users mimicking African accents, clicking their tongues, spitting at female avatars, and making remarks about Ebola. This form of group harassment was widely condemned as both racist and sexist (Alexander 2018a; Feldman 2018; Gault 2018). Although VRChat's developers publicly stated that "hatred and other negative behavior that negatively affect the VRChat community do not have a place" on the platform (VRChat 2017), they did not implement restrictions on the use of the avatars associated with the meme (Alexander 2018a). In contrast, Roblox, a gaming platform where the meme also appeared, chose to ban it (Alexander 2018b). An additional case of gender-based harassment was made public during the same period, when VR designer Katie Chironis released a two-minute video showing her being harassed while speaking in a crowded public VRChat world (Gault 2018). As a reaction to this series of events the developers published an open letter to their community, in which they excused for the insufficient moderation mechanisms of their early stage business and educated about self-moderation tools like muting, blocking, and reporting and introduced a new feature that allowed users to automatically mute all players not being part of their friends list (Gault 2018; VRChat 2018).

During the pandemic, user numbers continued to increase as everyday life turned back to normality (Faber 2022). At that time, two field studies in VRChat were picked up by major news sources. In the first, a researcher from the Center for Countering Digital Hate spent over 11 hours on the platform and observed more than 100 incidents of problematic behavior, including reports of harassment and instances of shared sexual content with users who identified as minors. The researcher submitted his findings to the platform but received no response (Frenkel and Browning 2021). In the second, a BBC News journalist created a fake account posing as a 13-year-old girl. In publicly accessible areas resembling red-light districts, she encountered "grooming, sexual material, racist insults and a rape threat" (Crawford and Smith 2022). Another report, at the end of the metaverse hype, described how a comic artist was repeatedly harassed by a known user, despite having blocked him. Hidden from her view, the user began seeking out her friends, joining their conversations, while all she could see was her friends talking to themselves. This case of deliberate evasion of protection mechanisms caused the victim to develop broader anxieties about online interactions in general (Metz 2022).

VRChat employs a content moderation approach characterized by being anticipatory and personalized. Three different groups of actors exercise varying moderation actions within the platform (VRChat, n.d.-f). Regular users have access to self-moderation functions, while owners, moderators, or administrators of groups or instances possess

more extensive powers. At the platform level, VRChat's Trust and Safety Team enforces decisions concerning user accounts and access (VRChat, n.d.-f). These actions are embedded within the underlying moderation infrastructure, which unites several security functions—including trust scoring, avatar visibility, and user interactability—all managed through an adjustable security dashboard with quick settings known as "shield levels" (VRChat, n.d.-f). There are eight visible trust ranks, ranging from "friends" to "nuisance", that determine core functionalities, including the ability to upload and report content, as well as a user's default appearance. This is necessary in VRChat, as voice, avatar skins, particles and animations may be overwhelming or affect system performance for some users. For example, users classified as nuisances are generally muted and shown as standard robots, while friends are always audible and fully featured (VRChat, n.d.-g). Furthermore, content warning labels assigned by creators allow for age-sensitive display of violent, sexually suggestive or otherwise adult properties (VRChat 2023). This is complemented by an age verification function currently limited to paying VRC+ subscribers to confirm their adult status, display such a marker in their profile, and unlock access to otherwise age-restricted 18+ instances. The verification process is outsourced to the third-party provider Persona, requiring users to scan a government-issued ID via a camera-enabled device (VRChat, n.d.-b). Additionally, although vaguely described, automated systems are said to be in use to detect harmful or hateful content. Because the platform lacks a distinct parent account function, VRChat instead refers users to the parental control tools provided by the platforms that host its app (VRChat, n.d.-d).

There are six essential moderation actions available to common users, such as muting (a user's voice or chatbox), blocking (users or groups), reporting (users, groups, instances, worlds or communication), initiating a vote kick (which may result in a one-hour ban from the instance), activating safe mode (a menu option or controller shortcut that disables all user features except those of friends) and adjusting the previously mentioned security dashboard. Instance and group instance moderators, administrators and owners are given four additional actions. They are capable of forcing a user's microphone off, issuing a warning, and kicking or banning a user (VRChat, n.d.-f). In-game reporting is said to be useful for flagging malicious content, but "effective reporting" (VRChat, n.d.-c) requires submitting a moderation report on the platform's website, ideally with a written description and supporting media evidence. Reporters are notified when a case has been processed, though the consequences for the accused user are not disclosed (VRChat, n.d.-c). Violations against VRChat's zero-tolerance policy result in instant and irrevocable bans. Such sanctions may also apply to conduct occurring outside the platform, especially unsolicited contact, harassment and harming minors (VRChat 2024a). The platform reserves the right to monitor user activity, including audio and video, either automatically or in specific cases, and may "block, filter, mute, remove or disable access to any [u]ser [c]ontent" (VRChat 2025a). Appeals can be submitted through the platform's website, and users are notified of the outcome (VRChat 2024a).

5.2 Horizon Worlds

From the very outset, Horizon Worlds has been criticized in the media for its insufficient content moderation. Even before its official launch, a female beta tester reported that her avatar had been groped in the world's lobby and that "there were other people there who supported this behavior" (Heath 2021). Two other prominent cases of attacks also involved hostile groups. A female psychologist was "verbally and sexually harassed [by] three to four male avatars" (Patel 2021) in the live event area Horizon Venues, while a female researcher from a non-profit experienced a "non-consensual [...] sexual act" as others "watched and passed around a vodka bottle" (Smith 2022). Although women were disproportionately targeted, a male journalist also reported being virtually groped by a male avatar while setting off to a virtual concert with his friend (Rifkind 2021). Another common theme in critical news reports on Horizon Worlds concerned underage players. Despite the platform's original age restrictions, children apparently found ways around them. Embodying adult-looking avatars, a New York Times journalist encountered obvious underage users in all 24 separate sessions across all hours of the day (Hill 2022). This was due to the lack of an age verification process upon entry into the virtual world, as access was permanently granted "once it's linked to an adult's Facebook account" (Oremus 2022). News outlets also highlighted that while "young users [were] inappropriately harassing other people" in Horizon Worlds (Nix 2023), they simultaneously represented potential targets for "sexual predators" (Oremus 2022).

Another incident that gained significant attention in both news media (Maiberg 2022; McCarthy 2022) and academic literature (Hine 2023; Trauthig and Woolley 2023; Gray, Carter, and Egliston 2024) involved an investigative experiment. After Meta rejected their list of questions regarding content enforcement, journalists from Buzzfeed decided to test the platform's policies themselves. They created the private world "Qniverse" and decorated it with "misinformation slogans" and other media content typically violating Facebook's "Community Standards" (Baker-White 2022). The outcome was revealing. Not only did Meta fail to respond within 48 hours despite the world being reported three times, but a "trained safety specialist [...] determined that the content in the Qniverse [did not] violate [Meta's] Content in VR Policy" (Meta, quoted in Baker-White 2022). Only after the journalists contacted "Meta's comms department, a channel not available to ordinary people" (Baker-White 2022), was the world removed without further clarification regarding its compliance with Meta's content policies.

Meta Horizon's moderation approach can best be described as reactive and standardized. Like VRChat, it operates through three tiers of moderating actors. First, common users may protect themselves by hiding, reporting or distancing from malicious content while seeking help or allies (Meta, n.d.-x). Second, world creators, their assigned admins (Meta, n.d.-m), and voluntary community moderators (Meta 2024e), possess additional intervention options. Third, Meta-employed "community guides" and "trained safety specialists" (Meta, n.d.-j) enjoy the highest moderation authority through the platform-wide and potentially permanent enforcement of measures and sanctions. Notably,

parents are equipped with different tools and capabilities to supervise their child (Meta, n.d.-v). As of date there are three kinds of accessible spaces. These include the shareable "personal space" (Meta 2022) in which the player starts, invitation-based "members-only worlds" (Meta, n.d.-n) and searchable public 10+/13+/18+ age-rated worlds (Meta, n.d.-r), the latter of which may include "intense violence, sexual content, strong language and/or regulated goods like tobacco or alcohol" (Meta, n.d.-n). Although machine learning is explicitly listed as a technology used to enforce policies (Meta, n.d.-h), Meta was for a long time " 'exploring how best to use AI [in Horizon Worlds]' " and admitting it was " 'not built yet' " (Murphy 2021). But recently, evidence of automated content moderation can be found in the "mute-assist" function that can automatically "detect profanity and potentially offensive words, shouting and other loud noises" (Meta, n.d.-t).

The repertoire of moderation tools and capabilities varies between actors, depending on their place in the hierarchy. Common users are given ten actions to curate their VR experience, such as muting oneself/others, auto-muting loud/offensive users ("mute-assist"), blocking, reporting others/worlds, initiating and answering a vote kick poll, removing a user from own "personal space", activating a one-meter protection bubble ("personal boundary"), obscuring voices ("voice mode"), retreating to a safe zone ("pausing") and adjusting the voice/world chat (Meta, n.d.-b, n.d.-e, n.d.-k, n.d.-l, n.d.-t, n.d.-w, n.d.-aa). World-creators and their admins and community moderators are additionally authorized to mute users (for 15 minutes), issue warnings via pop-up windows, and remove problematic users from their respective domain (Meta, n.d.-a). In order to assess the problematic situations a two-minute "rolling buffer stored locally" (Bosworth 2021) on the HMD is transmitted to the responsible moderators. The content of the data might consist of "audio, video and other interactions" and be "stored on [Meta's] servers" for up to two years and be used to "train models to better combat harmful behavior" (Meta, n.d.-u). Meta also reserves the right to store recordings for an even longer time period in order "to comply with applicable law" (Meta, n.d.-u).

If necessary, sanctions are imposed by the so-called trained safety specialists. For a long time, these human review teams routinely observed problematic situations invisibly in real-time. However, in June 2024, their tasks were reduced to "reviewing reports and taking appropriate actions" (Meta 2024d). Their inflicted consequences follow a "strike" system (Meta, n.d.-p) and range from the offender's microphone being muted, to the suspension of access to some of the software's core functions or Horizons Worlds entirely (Meta, n.d.-u). Ultimately, the disabling and deletion of the Meta account is at stake (Meta 2024b). When any of these actions are enforced, the cause and associated policy is stated via a notification and an email (Meta 2023b). Answering back a self-initiated dismissed report, routinely leads to a second review. The option to appeal a sanction, implemented relatively late in 2024, requests Meta to reevaluate "suspected violations and restrictions placed on [a] profile" (Meta 2024b) or a world (Meta 2024c). It can be started off through a link in a notification or warning mail and is sometimes limited by an appeals deadline (Meta, n.d.-h, n.d.-p).

Both platforms, although having developed sophisticated processes and tools over time, struggle with achieving effective, large-scale content moderation. Drawing from a distinct audience, due to market positioning and corporate size, they face surprisingly comparable threats and societal hardships. Yet, VRChat's anticipatory and personalized moderation approach tends to be more risk-averse and user-empowering, possibly resulting in a stronger self-regulating community. On the other hand, Horizon Worlds has introduced intelligent moderation tools, including a protection bubble and an automated muting function, while its in situ policing with human review teams does not appear to have been a good cultural or economic fit. As its user numbers are dwindling, in contrast to VRChat, the decision to expand its customer base to preteens in a weakly moderated environment containing adult content raises strong concerns about corporate responsibility.

6 Discussion and Conclusion

The comparative analysis of Horizon Worlds and VRChat demonstrates that these platforms have developed evolving but still limited content governance policies and practices to ensure trust and safety in immersive social media environments. Each has assembled policy frameworks that seek to regulate embodied interaction and mitigate risks such as violence, sexual harassment, manipulation, and impersonation, yet the accessibility, scope, coherence, and enforceability of these regimes remain uneven. While VRChat's community-centric approach results in relatively coherent policies (which also only have to apply to their core platform), Horizon Worlds' fragmented and overlapping regulatory structure renders the applicable standards rather difficult to comprehend. Meta's governance architecture—spread across several domains and subject to shifting product classifications—obscures what rules actually apply to the XR environment. In neither case do policy measures fully match the immersive affordances of the medium. These findings highlight a broader structural problem: legacy content moderation models, designed for text- or image-based media, struggle to adapt to the spatial and behavioral complexity of social XR.

The challenges to develop content moderation for social XR arise out of the processes of deep mediatization, in which technologies and media not only structure communication but lived experience. The nature of XR, embodiment, and the immersion into the environment amplify and make more critical the opportunities and risks of mediated interaction, with the impact of increasing the stakes of governance. Consequently, we already see the impact of the *spatial turn*. Governance is no longer primarily concerned with discrete pieces of content but with situated, embodied interactions and user experience unfolding spatially and in real time. This inadequacy of existing XR content governance systems suggests that the spatial turn in content moderation has not yet been followed by a normative turn in moderation practice. The transition from screen-based to embodied environments requires not only new technical tools but also new institutional capacities for context-aware judgment, and likely new shared understanding

of what is acceptable expression. However, as industry attention and investment shift away from XR toward artificial intelligence, it could be that an *XR winter* may entail a *XR content-moderation winter as well*. Reduced funding for safety teams and research, as part of the overall shift of attention and values associated with the current U.S. federal administration, may lead to stagnation precisely when immersive harms are becoming more complex, and when they are better understood.

A particularly illustrative difference between the two platforms lies in their division of public and private spaces. VRChat's content warning label system enables users to classify avatars and worlds containing sexually suggestive, violent, or disturbing content so that they can be filtered automatically. This explicit delineation between public and private instances institutionalizes a form of spatial governance that mirrors offline expectations of contextual integrity (Nissenbaum 2004). Horizon Worlds, by contrast, requires creators to assign age ratings of 10+, 13+, or 18+ depending on the maturity of the content. In 18+ spaces, Meta explicitly permits mature themes, ranging from sexual content and violence to strong language and depictions of recreational drug use (Meta, n.d.-q). Both mechanisms reflect an effort to manage exposure rather than to regulate behavior. Yet in Horizon Worlds the boundaries between these spaces remain porous, and the rationale for age classification is neither transparent nor publicly auditable.

Meta's approach further illustrates the challenge of determining which governance regime even applies. The company's general "Terms of Service" and "Community Standards" are written with Facebook and Instagram in mind, leaving it unclear whether XR-specific content— produced and experienced in real-time, embodied form—is subject to the same oversight mechanisms. Crucially, the Meta Oversight Board, established as an external accountability body for the company's social networks, has no jurisdiction over Horizon Worlds. Combined with the limited deployment of AI and machine-learning tools for real-time moderation—despite public references to such systems—this absence of institutional oversight undermines user trust in the platform's safety commitments. Nick Clegg, until recently Meta's President of Global Affairs, articulated the company's reluctance to embrace real-time speech moderation by analogy: "We wouldn't hold a bar manager responsible for real-time speech moderation in their bar, as if they should stand over your table, listen intently to your conversation, and silence you if they hear things they don't like" (Clegg 2022). This analogy risks downplaying the heightened responsibility of platforms that design, own, and algorithmically mediate the very environments in which such interactions occur. Unlike a bar manager, platforms control the architecture, affordances, and enforcement systems of their virtual yet embodied spaces and therefore should not negate accountability for harms that are structurally enabled by those designs. At the regulatory level, the applicability of existing frameworks such as the EU's DSA to immersive environments remains partly unsettled. As Hine et al. (2024) observe, the DSA—alongside related instruments including the NIS2 Directive, the Directive on Liability for Defective Products, the GDPR, the Digital Markets Act (DMA), the AI Act, and

the Terrorism Regulation —constitutes the principal legal architecture governing platform responsibility for online safety and privacy within the European Union.

While the DSA has applied to all intermediary services since February 2024, its most stringent transparency, auditing, and systemic-risk obligations are reserved for Very Large Online Platforms (VLOPs), defined as those reaching more than 45 million monthly active users in the EU. Platforms that do not meet this threshold are nonetheless subject to significant procedural obligations where they qualify as "online platforms" within the meaning of Article 3(i) DSA when they store and disseminate user-generated content to the public. In particular, online platforms must implement accessible and user-friendly notice-and-action mechanisms enabling users to report illegal content (Art. 16), provide a detailed statement of reasons when imposing content moderation measures (Art. 17), and establish internal complaint-handling systems (Art. 20), as well as engage with certified out-of-court dispute settlement (ODS) bodies (Art. 21). Statements of reasons must, at a minimum, specify the restriction imposed, the facts and circumstances underlying the decision, the legal or contractual basis relied upon, whether automated means were used, and the available avenues of redress. ODS bodies operate as independent and impartial entities across the EU, offering users an additional layer of review for moderation decisions (Husovec 2024).

Unlike VLOPs, however, there is no formal designation procedure for ordinary online platforms by the European Commission. The classification follows directly from the functional definition contained in the Regulation. This raises interpretative questions in the context of immersive environments, as the DSA was primarily designed with traditional, screen-based platforms in mind. Although the definition of "online platform" is technologically neutral and may encompass XR, the application of DSA mechanisms to spatial, real-time environments has not yet been judicially tested, even though some would classify them as online platforms (Kerikmäe, Hamuľák, and Mesarčík 2025).

The central challenge, therefore, lies not in the absence of regulation, but in adapting procedural safeguards originally conceived for conventional social media to the behavioral and immersive dynamics of XR ecosystems. This raises the broader question of who regulates social XR. Ongoing policy processes within the European Union and the Council of Europe are beginning to address XR's implications for human rights, democracy, and the rule of law (Council of Europe and IEEE Standards Association 2024). These efforts reflect the emerging field of "digital constitutionalism", which seeks to translate constitutional principles —such as freedom of expression, due process, and privacy—into the governance of digital infrastructures (Padovani and Santaniello 2018; Celeste 2022; Celeste et al. 2023). Extending these principles to immersive environments will require institutional innovation capable of reconciling the real-time, borderless nature of XR with the procedural guarantees of constitutional law. Doing so may require seeking out the input from a variety of stakeholders, including civil society groups and academics (Celeste et al. 2023; Palladino, Redeker, and Celeste 2025).

At the same time, the constitutional challenge is not merely institutional but architectural. XR moderation operates within a data environment fundamentally different from that of traditional social media. Immersive platforms rely on continuous body and movement tracking to render interaction. VR research identifies body tracking as central to the medium's distinctiveness, while also noting that such tracking may render users uniquely identifiable through behavioral data (Bailenson et al. 2025). The very architecture of VR intensifies the legal stakes of moderation practices, as these may rely on highly behavioral data generated through continuous tracking. Such data falls within the scope of EU data protection law. Consequently, XR moderation mechanisms implicate core GDPR principles, including data minimisation, purpose limitation, and accountability. In this, like in a variety of other attributes as well as its roots, XR may resemble more closely virtual games and gaming than social media platforms.

Ultimately, the comparative evidence presented here suggests that neither Horizon Worlds nor VRChat yet offer a sufficiently mature framework to guarantee user trust and safety in social XR. Both have taken important steps toward formalizing behavioral norms and providing user-level protection tools, but they fall short of the transparency, consistency, and oversight necessary for accountable governance. As the spatial turn reconfigures digital interaction into embodied, co-present experience, the governance of these environments must evolve accordingly —and do so normatively. Whether through the enforcement of the DSA, the development of industry-wide standards, or the establishment of new multi-stakeholder oversight mechanisms, the next phase of social-XR governance will determine whether immersive technologies can mature without reproducing the failures of earlier social-media ecosystems. We argue that it is important that platforms fully commit to their systems and engage in a normative turn with regard to XR content governance, which may go beyond the implementation of a few new security features.

It should be noted that social XR content governance did not emerge and will not develop in a vacuum. The influences from traditional social media content governance are as important as the influences from online gaming environments, including instance-based moderation, avatar-centered interaction, and community-driven enforcement, which bear striking similarities to social XR. There is some evidence that, rather than community sizes similar to traditional social media, social XR and online gaming share smaller, more tightly knit communities as a basis for interaction (Tsutsui et al. 2025). In turn, the intensified embodiment and the realism of XR amplify the dynamics of game-like interactions into socially consequential experiences. Even if XR platforms "get it right" and develop norms, policies and moderation systems suitable to keep in check these risks to trust and safety of its users, there might still be societal risks. Floridi (2022) reminds us in that context that even if all goes well otherwise, XR platforms may be the new frontier for the digital divide, with many people unable to experience spatial social connection the way others can.

References

Abraham, Melvin, Pejman Saeghe, Mark McGill, and Mohamed Khamis. 2022. "Implications of XR on Privacy, Security and Behaviour: Insights from Experts." In *Nordic Human-Computer Interaction Conference,* 1–12. https://doi.org/10.1145/3546155.3546691.

Alexander, Julia. 2018a. "Ugandan Knuckles Is Overtaking VRChat," January. https://www.polygon.com/2018/1/8/16863932/ugandan-knuckles-meme-vrchat.

———. 2018b. "Understanding Ugandan Knuckles in a Post-Pepe the Frog World," February. https://www.polygon.com/2018/2/2/16951684/ugandan-knuckles-pepe-frog-meme-vrchat.

Alharbi, Ahmed, Hai Dong, Xun Yi, Zahir Tari, and Ibrahim Khalil. 2022. "Social Media Identity Deception Detection: A Survey." *ACM Computing Surveys* 54 (3): 1–35. https://doi.org/10.1145/3446372.

Antao, Rohit, and Scott Likens. 2024. "The Essential Eight Technologies: What You Need to Know," February. https://www.pwc.com.au/digitalpulse/the-essential-eight-technologies-what-you-need-to-know.html.

Ashworth, Boone. 2026. "Meta Will Keep Horizon Worlds Alive in VR 'for the Foreseeable Future'," March. https://www.wired.com/story/meta-will-keep-horizon-worlds-alive-in-vr-for-the-foreseeable-future/.

Bailenson, Jeremy N., Andrew C. Beall, Jack Loomis, Jim Blascovich, and Matthew Turk. 2004. "Transformed Social Interaction: Decoupling Representation from Behavior and Form in Collaborative Virtual Environments." *Presence: Teleoperators and Virtual Environments* 13 (4): 428–41. https://doi.org/10.1162/1054746041944803.

———. 2005. "Transformed Social Interaction, Augmented Gaze, and Social Influence in Immersive Virtual Environments." *Human Communication Research* 31 (4): 511–37. https://doi.org/10.1111/j.1468-2958.2005.tb00881.x.

Bailenson, Jeremy N., Cyan DeVeaux, Eugy Han, David M. Markowitz, Monique Santoso, and Portia Wang. 2025. "Five Canonical Findings from 30 Years of Psychological Experimentation in Virtual Reality." *Nature Human Behaviour* 9 (7): 1328–38. https://doi.org/10.1038/s41562-025-02216-3.

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

Berrick, Daniel, and Jameson Spivack. 2022. "Understanding Extended Reality Technology & Data Flows: Privacy and Data Protection Risks and Mitigation Strategies," November. https://fpf.org/blog/understanding-extended-reality-technology-data-flows-privacy-and-data-protection-risks-and-mitigation-strategies/.

Blackwell, Lindsay, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 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.

Bosworth, Andrew. 2021. "Keeping People Safe in VR and Beyond," November. https://www.oculus.com/blog/keeping-people-safe-in-vr-and-beyond/.

———. 2026. "ICYMI This from My AMA Yesterday (@boztank)," March. https://x.com/boztank/status/2034747828724277433.

Brown, James, Jeremy Bailenson, and Jeffrey Hancock. 2023. "Misinformation in Virtual Reality." *Journal of Online Trust and Safety* 1 (5). https://doi.org/10.54501/jots.v1i5.120.

Bygrave, Lee A. 2015. *Internet Governance by Contract.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199687343.001.0001.

Casey, Peter, Ibrahim Baggili, and Ananya Yarramreddy. 2021. "Immersive Virtual Reality Attacks and the Human Joystick." *IEEE Transactions on Dependable and Secure Computing* 18 (2): 550–62. https://doi.org/10.1109/TDSC.2019.2907942.

Celeste, Edoardo. 2019. "Terms of Service and Bills of Rights: New Mechanisms of Constitutionalisation in the Social Media Environment?" *International Review of Law, Computers & Technology* 33 (2): 122–38. https://doi.org/10.1080/13600869.2018.1475898.

———. 2022. *Digital Constitutionalism: The Role of Internet Bills of Rights.* 1st ed. London: Routledge. https://doi.org/10.4324/9781003256908.

Celeste, Edoardo, Nicola Palladino, Dennis Redeker, and Kinfe Yilma. 2023. *The Content Governance Dilemma: Digital Constitutionalism, Social Media and the Search for a Global Standard.* Springer International Publishing. https://doi.org/10.1007/978-3-031-32924-1.

Cheng, Kaiming, Arkaprabha Bhattacharya, Michelle Lin, Jaewook Lee, Aroosh Kumar, Jeffery F. Tian, Tadayoshi Kohno, and Franziska Roesner. 2024. "When the User Is Inside the User Interface: An Empirical Study of UI Security Properties in Augmented Reality." In *Proceedings of the 33rd USENIX Conference on Security Symposium.* https://dl.acm.org/doi/10.5555/3698900.3699052.

Chow, Andrew R. 2022. "A Year Ago, Facebook Pivoted to the Metaverse. Was It Worth It?," October. https://time.com/6225617/facebook-metaverse-anniversary-vr/.

Clegg, Nick. 2022. "Making the Metaverse: What It Is, How It Will Be Built, and Why It Matters," May. https://nickclegg.medium.com/making-the-metaverse-what-it-is-how-it-will-be-built-and-why-it-matters-3710f7570b04.

Coalition for Independent Technology Research. 2025. *The State of Independent Technology Research 2025: Power in Numbers.* Technical report. https://independenttechresearch.org/wp-content/uploads/2025/08/The-State-of-Independent-Technology-Research-Power-in-Numbers.pdf.

Council of Europe and IEEE Standards Association. 2024. *The Metaverse and Its Impact on Human Rights, the Rule of Law and Democracy.* Council of Europe. https://rm.coe.int/the-metaverse-and-its-impact-on-human-rights-the-rule-of-law-and-democ/1680b178b0.

Crawford, Angus, and Tony Smith. 2022. "Metaverse App Allows Kids into Virtual Strip Clubs," February. https://www.bbc.com/news/technology-60415317.

European Parliamentary Research Service. 2021. *Key Enabling Technologies for Europe's Technological Sovereignty.* https://www.europarl.europa.eu/RegData/etudes/STUD/2021/697184/EPRS_STU(2021)697184_EN.pdf.

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 (Digital Services Act)." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R2065.

Faber, Tom. 2022. "Finding Community, and Freedom, on the Virtual Dance Floor," December. https://www.nytimes.com/2022/12/27/arts/music/vrchat-virtual-reality-clubbing.html.

Fagan, Kaylee. 2018. "A Large Number of People Have Come Out Saying VRChat Has Saved Their Lives—Here's What It's Like to Experience the Online Meeting Place of the 21st Century," March. https://www.businessinsider.com/vrchat-explained-2018-2.

Falchuk, Ben, Shoshana Loeb, and Ralph Neff. 2018. "The Social Metaverse: Battle for Privacy." *IEEE Technology and Society Magazine* 37 (2): 52–61. https://doi.org/10.1109/MTS.2018.2826060.

Faughnder, Ryan. 2024. "AI vs. the Metaverse: How Artificial Intelligence Might Change the Future of the Internet," July. https://www.latimes.com/entertainment-arts/business/newsletter/2024-07-09/theories-of-the-future-of-the-internet-revisited-the-wide-shot.

Feldman, Brian. 2018. "What Is VRChat and Who Is Ugandan Knuckles?," January. https://nymag.com/intelligencer/2018/01/what-is-vrchat-and-who-is-ugandan-knuckles.html.

Floemer, Andreas. 2025. "Android XR: Google zeigt seine KI-Brille mit Display," May. https://www.heise.de/news/Android-XR-Google-ist-bereit-fuer-smarte-Brillen-mit-und-ohne-Bildschirm-10390304.html.

Floridi, Luciano. 2022. "Metaverse: A Matter of Experience." *Philosophy & Technology* 35. https://doi.org/10.1007/s13347-022-00568-6.

Freeman, Daniel, Sarah Reeve, A. Robinson, Anke Ehlers, David Clark, Bernhard Spanlang, and Mel Slater. 2017. "Virtual Reality in the Assessment, Understanding, and Treatment of Mental Health Disorders." *Psychological Medicine* 47 (14): 2393–400. https://doi.org/10.1017/S003329171700040X.

Freeman, Guo, and Dane Acena. 2022. ""Acting Out" Queer Identity: The Embodied Visibility in Social Virtual Reality." *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2): 1–32. https://doi.org/10.1145/3555153.

Freeman, Guo, Samaneh Zamanifard, Divine Maloney, and Alexandra Adkins. 2020. "My Body, My Avatar: How People Perceive Their Avatars in Social Virtual Reality." In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–8. https://doi.org/10.1145/3334480.3382923.

Frenkel, Sheera, and Kellen Browning. 2021. "The Metaverse's Dark Side: Here Come Harassment and Assaults," December. https://www.nytimes.com/2021/12/30/technology/metaverse-harassment-assaults.html.

Galtung, Johan. 1990. "Cultural Violence." *Journal of Peace Research* 27 (3): 291–305. https://doi.org/10.1177/0022343390027003005.

Gault, Matthew. 2018. "VR's Hit Social App Is a Dank Meme-Soaked Chat Room," January. https://www.vice.com/en/article/vrchat-review/.

Gillespie, Tarleton. 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, Tarleton, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich, and Sarah Myers West. 2020. "Expanding the Debate about Content Moderation: Scholarly Research Agendas for the Coming Policy Debates." *Internet Policy Review* 9 (4). https://doi.org/10.14763/2020.4.1512.

Gorwa, Robert. 2019. "What Is Platform Governance?" *Information, Communication & Society* 22 (6): 854–71. https://doi.org/10.1080/1369118X.2019.1573914.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 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.

Gray, Joanne E., Marcus Carter, and Ben Egliston. 2024. *Governing Social Virtual Reality: Preparing for the Content, Conduct and Design Challenges of Immersive Social Media.* Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-61831-4.

Grimm, Paul, Wolfgang Broll, Rigo Herold, and Johannes Hummel. 2019. "VR/AR-Eingabegeräte und Tracking" [in de]. In *Virtual und Augmented Reality (VR/AR),* edited by Ralf Dörner, Wolfgang Broll, Paul Grimm, and Bernhard Jung, 117–62. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-58861-1_4.

Grimm, Paul, Wolfgang Broll, Rigo Herold, Dirk Reiners, and Carolina Cruz-Neira. 2019. "VR/AR-Ausgabegeräte" [in de]. In *Virtual und Augmented Reality (VR/AR),* edited by Ralf Dörner, Wolfgang Broll, Paul Grimm, and Bernhard Jung, 163–217. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-58861-1_5.

Grimmelmann, James. 2015. "The Virtues of Moderation." *Yale Journal of Law & Technology* 17:42–109. https://doi.org/10.31228/osf.io/qwxf5.

El-Hajj, Mohammed. 2024. "Cybersecurity and Privacy Challenges in Extended Reality: Threats, Solutions, and Risk Mitigation Strategies." *Virtual Worlds* 4 (1). https://doi.org/10.3390/virtualworlds4010001.

Hallinan, Blake, C. J. Reynolds, Rebecca Scharlach, Dana Theiler, Noa Niv, Omer Rothenstein, Isabell Knief, and Yehonatan Kuperberg. 2025. "Priorities and Exclusions within Trust and Safety Industry Standards." *New Media & Society,* https://doi.org/10.1177/14614448251357225.

Han, Eugy, Mark R. Miller, Nilam Ram, Kristine L. Nowak, and Jeremy N. Bailenson. 2022. "Understanding Group Behavior in Virtual Reality: A Large-Scale, Longitudinal Study in the Metaverse." https://papers.ssrn.com/abstract=4110154.

Heath, Alex. 2021. "Meta Opens Up Access to Its VR Social Platform Horizon Worlds," December. https://www.theverge.com/2021/12/9/22825139/meta-horizon-worlds-access-open-metaverse.

Helberger, Natali, Jo Pierson, and Thomas Poell. 2018. "Governing Online Platforms: From Contested to Cooperative Responsibility." *The Information Society* 34 (1): 1–14. https://doi.org/10.1080/01972243.2017.1391913.

Heldt, Amélie. 2019. "Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports." *Internet Policy Review* 8 (2). https://doi.org/10.14763/2019.2.1398.

Heller, Brittan, and Avi Bar-Zeev. 2021. "The Problems with Immersive Advertising: In AR/VR, Nobody Knows You Are an Ad." *Journal of Online Trust and Safety* 1 (1). https://doi.org/10.54501/jots.v1i1.21.

Hepp, Andreas. 2019. *Deep Mediatization.* 1st ed. Routledge. https://doi.org/10.4324/9781351064903.

Hill, Kashmir. 2022. "This Is Life in the Metaverse," October. https://www.nytimes.com/2022/10/07/technology/metaverse-facebook-horizon-worlds.html.

Hine, Emmie. 2023. "Content Moderation in the Metaverse Could Be a New Frontier to Attack Freedom of Expression." *Philosophy & Technology* 36 (3). https://doi.org/10.1007/s13347-023-00645-4.

Hine, Emmie, Isadora Neroni Rezende, Huw Roberts, David Wong, Mariarosaria Taddeo, and Luciano Floridi. 2024. "Safety and Privacy in Immersive Extended Reality: An Analysis and Policy Recommendations." *Digital Society* 3. https://doi.org/10.1007/s44206-024-00114-1.

Husovec, Martin. 2024. *Principles of the Digital Services Act.* 1st ed. Oxford University Press. https://doi.org/10.1093/law-ocl/9780192882455.001.0001.

Ivey, Rebecca. 2024. "Is XR the Unsung Hero of the Digital Revolution?," August. https://www.weforum.org/stories/2024/08/why-xr-is-key-to-unlocking-the-next-digital-revolution/.

Katzenbach, Christian, Adrian Kopps, João C. Magalhães, Dennis Redeker, Tom Sühr, and Larissa Wunderlich. 2023. "The Platform Governance Archive V1—A Longitudinal Dataset to Study the Governance of Communication and Interactions by Platforms and the Historical Evolution of Platform Policies [Data Paper]," September. Accessed April 6, 2026. https://doi.org/10.31235/osf.io/5vcfz.

Kenwright, Benjamin. 2023. "Impact of XR on Mental Health: Are We Playing with Fire?" *arXiv,* https://doi.org/10.48550/arXiv.2304.01648.

Kerikmäe, Tanel, Ondrej Hamuľák, and Matúš Mesarčík. 2025. "Disinformation Tackling in the Metaverse and the Digital Services Act." *Cogent Social Sciences* 11 (1). https://doi.org/10.1080/23311886.2025.2485386.

Kilteni, Konstantina, Raphaela Groten, and Mel Slater. 2012. "The Sense of Embodiment in Virtual Reality." *Presence: Teleoperators and Virtual Environments* 21 (4): 373–87. https://doi.org/10.1162/PRES_a_00124.

Kohl, Uta. 2024. "Toxic Recommender Algorithms: Immunities, Liabilities and the Regulated Self-Regulation of the Digital Services Act and the Online Safety Act." *Journal of Media Law* 16 (2): 301–35. https://doi.org/10.1080/17577632.2024.2408912.

Krauß, Veronika, Pejman Saeghe, Alexander Boden, Mohamed Khamis, Mark McGill, Jan Gugenheimer, and Michael Nebeling. 2024. "What Makes XR Dark? Examining Emerging Dark Patterns in Augmented and Virtual Reality Through Expert Co-Design." *ACM Transactions on Computer-Human Interaction* 31 (3): 1–39. https://doi.org/10.1145/3660340.

Krombholz, Katharina, Heidelinde Hobel, Markus Huber, and Edgar Weippl. 2015. "Advanced Social Engineering Attacks." *Journal of Information Security and Applications* 22:113–22. https://doi.org/10.1016/j.jisa.2014.09.005.

Lin, Jinghuai, and Marc Erich Latoschik. 2022. "Digital Body, Identity and Privacy in Social Virtual Reality: A Systematic Review." *Frontiers in Virtual Reality* 3. https://doi.org/10.3389/frvir.2022.974652.

Maiberg, Emanuel. 2022. “Being a Facebook Metaverse ’Community Guide’ Seems Like a Nightmare Job,” February. https://www.vice.com/en/article/being-a-facebook-metaverse-community-guide-seems-like-a-nightmare-job.

Malik, Tammana, and Noah Usman. 2024. “Physical Solutions in Virtual Spaces: Challenges to Content Moderation in XR,” May. https://www.techpolicy.press/physical-solutions-in-virtual-spaces-challenges-to-content-moderation-in-xr/.

Mann, Jyoti. 2025. “Meta’s CTO Said the Metaverse Could Be a ’Legendary Misadventure’ If the Company Doesn’t Boost Sales, Leaked Memo Shows,” February. https://www.businessinsider.com/meta-cto-metaverse-reality-labs-legendary-misadventure-memo-2025-2.

Marchal, Nahema, Emma Hoes, K. Jonathan Klüser, Felix Hamborg, Meysam Alizadeh, Mael Kubli, and Christian Katzenbach. 2025. “How Negative Media Coverage Impacts Platform Governance: Evidence from Facebook, Twitter, and YouTube.” *Political Communication* 42 (2): 215–33. https://doi.org/10.1080/10584609.2024.2377992.

McCarthy, Dan. 2022. “A New Challenge for Meta: How to Moderate the Metaverse,” February. https://www.emergingtechbrew.com/stories/2022/02/23/a-new-challenge-for-meta-how-to-moderate-the-metaverse.

McVeigh-Schultz, Joshua, Anya Kolesnichenko, and Katherine Isbister. 2019. “Shaping Pro-Social Interaction in VR: An Emerging Design Framework.” In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–12. https://doi.org/10.1145/3290605.3300794.

Mengü, Murat, and Seda Mengü. 2015. “Violence and Social Media.” *Athens Journal of Mass Media and Communications* 1 (3): 211–28. https://doi.org/10.30958/ajmmc.1-3-4.

Meta. 2021. “Horizon Worlds Opens to Those 18+ in the US and Canada,” December. Accessed July 30, 2025. https://www.meta.com/blog/horizon-worlds-opens-to-those-18-in-the-us-and-canada/.

———. 2022. “Meta Horizon Worlds Continues Rolling Out Personal Space to More People,” September. Accessed July 30, 2025. https://www.oculus.com/blog/horizon-worlds-personal-space/.

———. 2023a. “Cookies Policy,” December. Accessed July 30, 2025. https://www.facebook.com/privacy/policies/cookies?locale=en_US.

———. 2023b. “Meta Horizon Worlds V91 Release Notes,” January. Accessed July 30, 2025. https://www.oculus.com/blog/meta-horizon-worlds-v91-release-notes/.

———. 2023c. “Our Approach to Explaining Ranking,” December. Accessed November 1, 2025. https://transparency.meta.com/features/explaining-ranking/.

———. 2024a. “Meta Avatars Terms of Service,” October. Accessed July 30, 2025. https://www.facebook.com/legal/avatars_terms?locale=en_US.

———. 2024b. “Meta Horizon Worlds V149 Release Notes,” February. Accessed July 30, 2025. https://www.meta.com/blog/meta-horizon-worlds-v149-release-notes/.

———. 2024c. “Meta Horizon Worlds V150 Release Notes,” February. Accessed July 30, 2025. https://www.meta.com/blog/quest/meta-horizon-worlds-v150-release-notes.

———. 2024d. “Meta Horizon Worlds V167 Release Notes,” June. Accessed July 30, 2025. https://www.meta.com/de-de/blog/quest/meta-horizon-worlds-v167-release-notes/.

———. 2024e. “Meta Horizon Worlds V191 Release Notes,” December. Accessed July 30, 2025. https://www.meta.com/de-de/blog/quest/meta-horizon-worlds-v191-release-notes/.

———. 2024f. “Oversight Board Cases,” February. Accessed July 30, 2025. https://transparency.meta.com/de-de/oversight/oversight-board-cases.

———. 2024g. “Welcoming Preteens to Meta Horizon Worlds,” November. Accessed July 30, 2025. https://www.meta.com/blog/preteens-horizon-worlds-family-friendly-vr-mr/.

———. 2025a. “Content Guidelines,” June. Accessed August 3, 2025. https://developers.meta.com/horizon/policy/content-guidelines.

———. 2025b. “Developer Data Use Policy,” May. Accessed July 30, 2025. https://developers.meta.com/horizon/policy/data-use?locale=en_US.

———. 2025c. “Developer Policy Overview,” June. Accessed July 30, 2025. https://developers.meta.com/horizon/policy/policy-overview?locale=en_US.

———. 2025d. “Terms of Service,” January. Accessed July 30, 2025. https://www.facebook.com/terms.

———. 2026. “Updates to Your Meta Quest Experience in 2026,” March. Accessed April 5, 2026. https://communityforums.atmeta.com/blog/AnnouncementsBlog/updates-to-your-meta-quest-experience-in-2026/1369435.

———. n.d.-a. “Best Practices for Community Moderators in Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/761587246153030.

———. n.d.-b. “Block or Unblock Someone in Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/block-or-unblock-horizon/.

———. n.d.-c. “Code of Conduct for Virtual Experiences.” Accessed July 30, 2025. https://www.meta.com/legal/quest/code-of-conduct-for-virtual-experiences.

Meta. n.d.-d. “Community Standards.” Accessed July 30, 2025. https://transparency.meta.com/policies/community-standards.

———. n.d.-e. “Frequently Asked Questions About Worlds on Meta Horizon.” Accessed July 30, 2025. https://www.meta.com/help/quest/796650364312013/.

———. n.d.-f. “Get Parents Started with Supervision for Meta Horizon.” Accessed July 30, 2025. https://www.meta.com/help/quest/529890378674396.

———. n.d.-g. “Horizon Worlds.” Accessed July 30, 2025. https://horizon.meta.com.

———. n.d.-h. “How We Restrict the Use of Meta Platforms Technologies Products and Handle Complaints.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/2182719088748080/.

———. n.d.-i. “Introduction to the Advertising Standards.” Accessed July 30, 2025. https://transparency.meta.com/policies/ad-standards.

———. n.d.-j. “Learn about Moderators and Community Guides in Worlds.” Accessed July 30, 2025. https://www.meta.com/help/quest/2259751227488308/.

———. n.d.-k. “Learn about Personal Space in Worlds.” Accessed July 30, 2025. https://www.meta.com/help/quest/1917313465137559/.

———. n.d.-l. “Learn about the Personal Boundary Setting in Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/personal-boundary-horizon-worlds/.

———. n.d.-m. “Members-Only Worlds Governance Best Practices in Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/1735973196872947.

———. n.d.-n. “Members-Only Worlds in Meta Horizon Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/articles/horizon/explore-horizon-worlds/members-only-worlds/.

———. n.d.-o. “Meta Family Center.” Accessed July 30, 2025. https://familycenter.meta.com/?locale=en_US.

———. n.d.-p. “Meta Horizon Profile Restrictions and Suspensions on Meta Quest.” Accessed July 30, 2025. https://www.meta.com/help/quest/489834268719871/.

———. n.d.-q. “Meta Horizon Worlds Content Guidelines.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/481214418021533/.

———. n.d.-r. “Meta Horizon Worlds V180 Release Notes.” Accessed July 30, 2025. https://www.meta.com/de-de/blog/meta-horizon-worlds-v180-release-notes/.

———. n.d.-s. “Meta Platforms Technologies Products Definition.” Accessed July 30, 2025. https://www.meta.com/legal/meta-platforms-technologies-products/.

———. n.d.-t. “Mute and Unmute Your Microphone in Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/mute-mic-horizon/.

———. n.d.-u. “Notice of Recording to Improve Your Experience in Worlds.” Accessed July 30, 2025. https://www.meta.com/legal/quest/monitoring-recording-safety-horizon/.

———. n.d.-v. “Parental Supervision Settings.” Accessed July 30, 2025. https://www.meta.com/help/quest/304866315041200/.

———. n.d.-w. “Report Someone in Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/report-someone-horizon-worlds/.

———. n.d.-x. “Safety and Privacy in Meta Horizon Worlds.” Accessed July 30, 2025. https://www.meta.com/help/quest/1737463343292580.

———. n.d.-y. “Supported Countries for Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/759652432136500.

———. n.d.-z. “Transparency Reports.” Accessed July 30, 2025. https://transparency.meta.com/reports.

———. n.d.-aa. “Use Voice Channel in Meta Horizon Worlds.” Accessed July 30, 2025. https://www.meta.com/en-gb/help/quest/articles/horizon/safety-and-privacy-in-horizon-worlds/use-voice-channel-horizon-worlds.

———. n.d.-ab. “Worlds.” Accessed July 27, 2025. https://www.meta.com/en-gb/experiences/worlds/2532035600194083/.

Metz, Rachel. 2022. “Harassment Is a Problem in VR, and It’s Likely to Get Worse,” May. https://edition.cnn.com/2022/05/05/tech/virtual-reality-harassment.

Mhaidli, Abraham Hani, and Florian Schaub. 2021. “Identifying Manipulative Advertising Techniques in XR Through Scenario Construction.” In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* 1–18. https://doi.org/10.1145/3411764.3445253.

Milgram, Paul, and Fumio Kishino. 1994. “A Taxonomy of Mixed Reality Visual Displays.” *IEICE Transactions on Information and Systems* 77 (12): 1321–29. https://www.alice.id.tue.nl/references/milgram-kishino-1994.pdf.

Murphy, Hannah. 2021. “How Will Facebook Keep Its Metaverse Safe for Users?,” November. https://www.ft.com/content/d72145b7-5e44-446a-819c-51d67c5471cf.

Nissenbaum, Helen. 2004. “Privacy as Contextual Integrity.” *Washington Law Review* 79:119–57. https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/.

Nix, Naomi. 2023. "Meta Doesn't Want to Police the Metaverse. Kids Are Paying the Price," March. https://www.washingtonpost.com/technology/2023/03/08/metaverse-horizon-worlds-kids-harassment/.

Oremus, Will. 2022. "Kids Are Flocking to Facebook's 'Metaverse.' Experts Worry Predators Will Follow," February. https://www.washingtonpost.com/technology/2022/02/07/facebook-metaverse-horizon-worlds-kids-safety.

Padovani, Claudia, and Mauro Santaniello. 2018. "Digital Constitutionalism: Fundamental Rights and Power Limitation in the Internet Eco-System." *International Communication Gazette* 80 (4): 295–301. https://doi.org/10.1177/1748048518757114.

Palladino, Nicola, Dennis Redeker, and Edoardo Celeste. 2025. "Civil Society's Role in Constitutionalising Global Content Governance." *Internet Policy Review,* https://doi.org/10.14763/2025.1.1830.

Patel, Nina Jane. 2021. "Fiction vs. Non-Fiction," December. https://ninajanepatel.medium.com/fiction-vs-non-fiction-d824c6edfbe2.

Peukert, Christian, Hamed Qahri-Saremi, Ulrike Schultze, Jason B. Thatcher, Christy M. K. Cheung, Adeline Frenzel-Piasentin, Maike Greve, Christian Matt, Manuel Trenz, and Ofir Turel. 2024. "Metaverse: A Real Change or Just Another Research Area?" *Electronic Markets* 34 (1). https://doi.org/10.1007/s12525-024-00711-5.

Pfannenschmidt, Nikolas. 2025. "Spatial Turn in Content Moderation Archive." Accessed November 6, 2025. https://github.com/nikolaspfannenschmidt/spatial_turn_in_content_moderation_archive.

Poell, Thomas, David B. Nieborg, and Brooke Erin Duffy. 2022. *Platforms and Cultural Production.* Cambridge: Polity Press.

Quintais, João Pedro, Giovanni De Gregorio, and João C. Magalhães. 2023. "How Platforms Govern Users' Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications." *Computer Law & Security Review* 48. https://doi.org/10.1016/j.clsr.2023.105792.

Radsch, Courtney C. 2021. "Shadowban / Shadow Banning." In *IGF Glossary of Platform Law and Policy Terms,* 295–96. FGV Direito Rio. https://direitorio.fgv.br/sites/default/files/2022-08/2b03170e3f20446e0a6035f7494321cb.pdf.

Ramirez, Erick J., Shelby Jennett, Jocelyn Tan, Sydney Campbell, and Raghav Gupta. 2023. "XR Embodiment and the Changing Nature of Sexual Harassment." *Societies* 13 (2). https://doi.org/10.3390/soc13020036.

Reilly, Liam. 2025. "Tech Promised Virtual Reality Would Revolutionize Entertainment. That Moment Might Finally Be Closer Than We Think," July. https://edition.cnn.com/2025/07/12/tech/virtual-reality-entertainment-apple-meta-google-disney.

Rifkind, Hugo. 2021. "'Everything Facebook Did Badly Could Be Much Worse in the Metaverse.'," December. https://www.thetimes.co.uk/article/everything-facebook-did-badly-could-be-much-worse-in-the-metaverse-glrttr7rt.

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v.

Roesner, Franziska, Tadayoshi Kohno, and David Molnar. 2014. "Security and Privacy for Augmented Reality Systems." *Communications of the ACM* 57 (4): 88–96. https://doi.org/10.1145/2580723.2580730.

Rosén, Jörgen, Granit Kastrati, Aksel Reppling, Klas Bergkvist, and Fredrik Åhs. 2019. "The Effect of Immersive Virtual Reality on Proximal and Conditioned Threat." *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-53971-z.

Ryan-Mosley, Tate. 2023. "How an Undercover Content Moderator Polices the Metaverse: 'We Can Be the First Line of Defense.'," April. https://www.technologyreview.com/2023/04/28/1072393/undercover-content-moderator-polices-the-metaverse/.

Sabri, Nazanin, Bella Chen, Annabelle Teoh, Steven P. Dow, Kristen Vaccaro, and Mai Elsherief. 2023. "Challenges of Moderating Social Virtual Reality." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3544548.3581329.

Scharlach, Rebecca, Blake Hallinan, and Limor Shifman. 2024. "Governing Principles: Articulating Values in Social Media Platform Policies." *New Media & Society* 26 (11): 6658–77. https://doi.org/10.1177/14614448231156580.

Schulenberg, Kelsea, Lingyuan Li, Guo Freeman, Samaneh Zamanifard, and Nathan J. McNeese. 2023. "Towards Leveraging AI-Based Moderation to Address Emergent Harassment in Social Virtual Reality." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3544548.3581090.

Slater, Mel. 2009. "Place Illusion and Plausibility Can Lead to Realistic Behaviour in Immersive Virtual Environments." *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1535): 3549–57. https://doi.org/10.1098/rstb.2009.0138.

Smaili, Nadia, and Audrey de Rancourt-Raymond. 2024. "Metaverse: Welcome to the New Fraud Marketplace." *Journal of Financial Crime* 31 (1): 188–200. https://doi.org/10.1108/JFC-06-2022-0124.

Smith, Adam. 2022. "Woman Says She Was Virtually 'Raped' in the Metaverse While Others 'Passed Around a Bottle of Vodka.'," May. https://www.independent.co.uk/tech/rape-metaverse-woman-oculus-face-book-b2090491.html.

SteamDB. n.d. "VRChat Player Count Chart." Accessed March 12, 2026. https://steamdb.info/app/438100/charts/#max.

Susser, Daniel, Beate Roessler, and Helen F. Nissenbaum. 2019. “Online Manipulation: Hidden Influences in a Digital World.” *Georgetown Law Technology Review* 4 (1). https://doi.org/10.2139/ssrn.3306006.

Suzor, Nicolas P. 2019. *Lawless: The Secret Rules That Govern Our Digital Lives.* 1st ed. Cambridge University Press. https://doi.org/10.1017/9781108666428.

Szita, Kata, Lauren Buck, Nicola Palladino, Qian Xiao, Pat Treusch, Dalila Burin, Jennifer O’Meara, and Vincent Wade. 2025. “Considerations on User Identity within Metaverse Environments.” *Open Research Europe* 5. https://doi.org/10.12688/openreseurope.20411.1.

Tassi, Paul. 2022. “Meta’s Horizon Worlds Has Somehow Lost 100,000 Players in Eight Months,” October. https://www.forbes.com/sites/paultassi/2022/10/17/metas-horizon-worlds-has-somehow-lost-100000-players-in-eight-months/.

———. 2025. “Meta Still Believes in Horizon Worlds, Its Metaverse, but Likely Not for Long,” February. https://www.forbes.com/sites/paultassi/2025/02/05/meta-still-believes-in-horizon-worlds-its-metaverse-but-likely-not-for-long/.

Taylor, Alexandra, M. Claudia tom Dieck, Timothy Jung, Justin Cho, and Ohbyung Kwon. 2024. “XR and Mental Wellbeing: State of the Art and Future Research Directions for the Metaverse.” *Frontiers in Psychology* 15. https://doi.org/10.3389/fpsyg.2024.1360260.

Trauthig, Inga, and Kayo Mimizuka. 2022. “WhatsApp, Misinformation, and Latino Political Discourse in the U.S.” https://www.techpolicy.press/whatsapp-misinformation-and-latino-political-discourse-in-the-u-s/.

Trauthig, Inga, and Samuel Woolley. 2023. “Addressing Hateful and Misleading Content in the Metaverse.” *Journal of Online Trust and Safety* 1 (5). https://doi.org/10.54501/jots.v1i5.109.

Tromp, Jolanda G., Gabriel Zachmann, Jerome Perret, and Beatrice Palacco. 2022. “Future Directions for XR 2021-2030: International Delphi Consensus Study.” In *Roadmapping Extended Reality,* 1st ed., edited by Mariano Alcañiz, Marco Sacco, and Jolanda G. Tromp, 1–34. Wiley. https://doi.org/10.1002/9781119865810.ch1.

Tsutsui, Hiroto, Takefumi Hiraki, Yuichi Hiroi, and Shoichi Hasegawa. 2025. “Community Analysis of Social Virtual Reality Based on Large-Scale Log Data of a Commercial Metaverse Platform.” In *2025 IEEE International Symposium on Emerging Metaverse (ISEMV),* 95–103. https://doi.org/10.1109/ISEMV67326.2025.00026.

VRChat. 2017. “Important Notice Regarding Hatred and Other Behavior in VRChat,” December. Accessed August 1, 2025. https://www.reddit.com/r/VRchat/comments/7lc9za/important_notice_regarding_hatred_and_other/.

———. 2018. “An Open Letter to Our Community,” January. Accessed November 6, 2025. https://medium.com/@vrchat/an-open-letter-to-our-community-1b7aa5d9026f.

———. 2023. “Content Gating,” September. Accessed October 25, 2025. https://hello.vrchat.com/blog/content-gating.

———. 2024a. “Community Guidelines,” August. Accessed July 24, 2025. https://hello.vrchat.com/community-guidelines.

———. 2024b. “Cookie Policy,” August. Accessed July 24, 2025. https://hello.vrchat.com/cookie-policy.

———. 2024c. “Privacy Policy,” November. Accessed July 24, 2025. https://hello.vrchat.com/privacy.

———. 2025a. “Creator Guidelines,” April. Accessed July 24, 2025. https://hello.vrchat.com/creator-guidelines.

———. 2025b. “Terms of Service,” May. Accessed July 24, 2025. https://hello.vrchat.com/legal.

———. n.d.-a. “About.” Accessed July 30, 2025. https://www.linkedin.com/company/vrchat/about.

———. n.d.-b. “Age Verification.” Accessed March 13, 2026. https://wiki.vrchat.com/wiki/Age_Verification.

———. n.d.-c. “Reporting.” Accessed October 25, 2025. https://wiki.vrchat.com/wiki/Reporting.

———. n.d.-d. “Safety Resources for Parents.” Accessed October 25, 2025. https://help.vrchat.com/hc/en-us/articles/33301610887443-Safety-Resources-For-Parents.

———. n.d.-e. “Safety Resources for Players.” Accessed October 25, 2025. https://help.vrchat.com/hc/en-us/articles/33302819755539-Safety-Resources-For-Players.

———. n.d.-f. “Trust and Safety.” Accessed October 25, 2025. https://wiki.vrchat.com/wiki/Trust_and_Safety.

———. n.d.-g. “Trust Rank.” Accessed October 25, 2025. https://wiki.vrchat.com/wiki/Trust_Rank.

———. n.d.-h. “VRChat.” Accessed July 27, 2025. https://hello.vrchat.com.

———. n.d.-j. “VRChat.” Accessed July 24, 2025. https://wiki.vrchat.com/wiki/VRChat.

———. n.d.-i. “VRChat.” Accessed July 24, 2025. https://store.steampowered.com/app/438100/VRChat/.

Witmer, Bob G., and Michael J. Singer. 1998. “Measuring Presence in Virtual Environments: A Presence Questionnaire.” *Presence: Teleoperators and Virtual Environments* 7 (3): 225–40. https://doi.org/10.1162/105474698565686.

Zhang, Zhiyong, and Brij B. Gupta. 2018. “Social Media Security and Trustworthiness: Overview and New Direction.” *Future Generation Computer Systems* 86:914–25. https://doi.org/10.1016/j.future.2016.10.007.

Zheng, Qingxiao, Tue Ngoc Do, Lingqing Wang, and Yun Huang. 2022. “Facing the Illusion and Reality of Safety in Social VR.” *arXiv,* https://doi.org/10.48550/arXiv.2204.07121.

Zheng, Qingxiao, Shengyang Xu, Lingqing Wang, Yiliu Tang, Rohan C. Salvi, Guo Freeman, and Yun Huang. 2023. “Understanding Safety Risks and Safety Design in Social VR Environments.” *Proceedings of the ACM on Human-Computer Interaction* 7 (CSCW1): 1–37. https://doi.org/10.1145/3579630.

Authors

Dennis Redeker is a postdoctoral researcher at the Centre for Media, Communication and Information Research (ZeMKI) at the University of Bremen, Germany. His research focuses on analyses of global digital governance, including the governance of the Internet, platforms and artificial intelligence from a political science/international studies perspective. He is the principal investigator of "Fostering Digital Pan-Africanism in AI Governance through Evidence and Action" (AI PAN-AFRICANISM). Email: redeker@uni-bremen.de

Nikolas Pfannenschmidt is a student research assistant at ZeMKI at the University of Bremen, Germany. He is currently pursuing a bachelor's degree in Political Science and Media and Communication Studies, focusing on digital communication, media and the public, and political communication.

Manuel Baron Romero is a lawyer and a graduate of the European Master in Law, Data and Artificial Intelligence (EMILDAI).

Gabriel Durán is a legal practitioner specialising in EU digital regulation and platform governance. He serves as Head of Adjudication at ADROIT, an out-of-court dispute settlement body under Article 21 of the Digital Services Act (DSA). He is also a graduate of EMILDAI.

Ana Sofia Villa Hernandez is a data privacy lawyer and a graduate of EMILDAI.

Acknowledgements

An earlier version of this article was presented at the PlatGovNet 2025 conference and benefited from valuable feedback from the session chair and conference participants. We are grateful to the students of the undergraduate-level course "Global Governance of Extended Reality"—Sergino Apenuvon, Emma Brandewiede, Sofía Estrella, Simona Falkaite, Mala Jürgens, Lois Kiwala, Julie-Obehi Okojie, Anissa Saâdaoui, Elena Sánchez, Maja Wahl and Christophorus Wicaksono—whose research contributions and discussions helped shape the focus and orientation of the article. The manuscript also benefited from feedback received when it was presented at the colloquium of the Platform Governance, Media and Technology Lab at the Centre for Media, Communication and Information Research (ZeMKI), University of Bremen. Maitê Alegre Gonzalez supported the editing of the manuscript and Sebastian Kuhnke helped with visualizations. Finally, we thank the reviewers of the International Communication Association's Communication Law & Policy Division (ICA CLP), as well as the journal's editors and anonymous reviewers, for their generous and constructive feedback, which helped improve the manuscript.

Data availability statement

The archived platform policy data is available for download in Pfannenschmidt 2025.

Funding statement

Not applicable.

Ethical standards

Research is based on publicly available documents and reporting; no specific ethical issues are being reported.

Keywords

Extended Reality (XR); content moderation; platform governance; immersive harms; trust and safety; spatial turn