ITIF | INFORMATION TECHNOLOGY & INNOVATION FOUNDATION

Content Moderation in Multi-User Immersive Experiences: AR/VR and the Future of Online Speech

DANIEL CASTRO | FEBRUARY 2022

Multi-user immersive experiences (MUIEs)—three-dimensional, digitally rendered environments where multiple users can interact with other people and virtual objects in real time—present new content-moderation challenges. Policymakers should work with those developing MUIEs to balance user safety, privacy, and free expression.

KEY TAKEAWAYS

- Augmented and virtual reality (AR/VR) technologies allow individuals to communicate in immersive spaces, thereby creating new content moderation challenges such as restricting offensive gestures or virtual signs at a private property.
- As multi-user immersive experiences (MUIEs) become more popular, not only for entertainment, but also in educational and professional contexts, more work will be necessary to create safe and welcoming environments.
- Policymakers should take steps while MUIEs are still in their nascent stages, such as by establishing channels for platforms and law enforcement to work together to identify and respond to dangerous or illicit content in MUIEs.
- They should also strengthen protections against potential harms, including nonconsensual pornography and other defamatory content, fraud, and threats to child safety, and create a working group to develop guidance on intellectual property issues.
- Finally, policymakers should ensure that intermediary liability protections extend to providers of MUIEs, and that any changes to intermediary liability law consider the potential impacts on AR/VR communications platforms and their users.

CONTENTS

Introduction 2
Defining Features of MUIEs 4
Components of MUIEs 4
Forms of MUIEs 6
Communications Mediated by MUIEs 7
The Current Landscape for Content Moderation and Online Speech 8
The Regulatory Landscape for User-Generated Content 9
How Communications Platforms Define and Identify Unacceptable Content 11
Content Moderation Practices on 2D Platforms 11
User Activities and Content Moderation in Multi-User Immersive Experiences (MUIEs) 12
Elements of MUIEs 13
Content Moderation Practices and Challenges in Immersive Experiences 18
Considerations for Online Speech in an Immersive Future 24
Recommendations 26
Recommendations to Mitigate Potential Harms From Malicious Actors 26
Recommendations to Empower Providers to Develop Innovative Solutions 28
Conclusion 30
Endnotes 31

INTRODUCTION

For over a century, innovations in communication technologies have often been accompanied by concerns about the quality and integrity of the conversations they facilitate. For example, in 1844 the telegraph introduced the possibility of near-instantaneous communication across distances, but Samuel Morse, recognizing the growing political interest in his newly invented technology, instructed his assistant in Washington DC to "be especially careful not to give a partisan character to any information you transmit" during the upcoming elections.[1] Radio and television enabled mass communication with the public, but some feared "hypnotized audiences falling under the sway of irrational forces like fascism, communism, or even a corrupt and bankrupt capitalism."[2] Further technological innovations, from the telephone to videoconferencing, instant messaging, and social media, engendered new concerns about their risks and dangers and questions about the role and responsibilities of the intermediaries that enable these technologies.

Augmented and virtual reality (AR/VR) may be the next major phase in this evolution of communication technologies. In addition to offering immersive experiences for individual users, such as single-player games or 360-degree videos, many companies are building AR/VR technologies that will allow users to interact with others within immersive spaces. These **multi-user immersive experiences (MUIEs)**—three-dimensional, digitally rendered environments where multiple users can interact with other people and virtual objects in real-time—can transform the way people connect and share information. As personal tools, they can enhance social experiences, encourage creative expression, and create new opportunities for knowledge exchange for a broad, global user base. As enterprise technologies, MUIEs also can enable real-time remote collaboration, enhance training and knowledge retention, and overcome barriers of physical space to convene participants from around the world for meetings and events.

In some cases, MUIEs may be unmoderated channels for private communications between individuals—similar to conversations over a telephone or private meetings in a conference room—in others, MUEIs will be moderated by the platform, the operator of a particular virtual space (such as an employer), or by other members of the community. MUIEs present unique moderation scenarios such as:

- A group of users "spamming" a virtual space with verbal harassment or hate speech;
- A user's avatar intentionally violating another user's sense of personal space in a virtual environment, including through inappropriate actions that simulate sexual harassment or assault;
- Someone creating pornographic content that uses another person's likeness without their consent, such as by creating an avatar that has their face and body shape or using sensitive video footage; or
- An individual (or group) using AR to add misinformation or defamatory content to physical spaces, such as road signs, homes, or storefronts.

Other social platforms, including multiplayer online games, audio-only social media channels, videoconferencing and live streaming services, and photo-sharing apps, are already raising similar challenges and offer insights on how MUEIs might respond to hate speech, harassment,

and other unwelcome communications.[3] MUIEs will need to be accompanied by best practices for moderating this new medium to address challenges, including:

- Developing moderation approaches for not just images, videos, and words, but also novel forms of digital content such as environments, virtual nonverbal communication by avatars, and three-dimensional virtual objects;
- Creating technical systems that can effectively identify harmful content and differentiate it from surrounding digital elements (for example, a virtual offensive sign placed on a virtual table);
- Establishing terms of use and other guidelines that allow for multiple uses of a single platform (such as social gatherings, multiplayer games, and professional collaboration) as well as varying means of access (such as fully immersive headsets or mobile phones); and
- Balancing safety measures (such as monitoring conversations or limiting gestures) with user privacy and free expression in immersive virtual interactions.

Further complicating these considerations is the wide range of communication that MUIEs can facilitate, from private conversations with people users know to group interactions with strangers. MUIEs allow users to communicate with people they have an existing relationship with off the platform (such as family, friends, and coworkers), as well as strangers where the platform facilitates the connection. MUEIs allow users to communicate in different formats, including one-on-one (such as a private conversation), one-to-many (such as publicly sharing content on a social platform), and many-to-many (such as in a group conversation or a multi-player game).

To create safe and welcoming environments, many MUIEs will require moderation tools designed to prevent misuse. Instances of users "assaulting" others' embodied avatars, engaging in offensive and disruptive activities, or putting others at greater risk of harm in the "real world" have already arisen on some of the most popular social MUIEs.[4] As user bases continue to expand, platforms should assume that bad-faith actors will continue to attempt to behave in ways that could cause emotional distress and real-world harm and take preemptive actions to address potential instances of misuse. It is important to be prepared for these challenges: "Zoom-bombing" arose during the COVID-19 pandemic because many organizations quickly began using videoconferencing services without taking measures to address potential threats. At the same time, MUIEs offer opportunities for community building, information-sharing, and meaningful political speech and advocacy, often among individuals who would not be able to meet or organize in the real world due to physical distance.[5] Over-moderation in response to potential abuses could stifle these activities.

With MUIEs, third-party platforms will increasingly mediate channels for exercising speech rights. Without proper consideration for these shifting parameters of speech in immersive spaces, content moderation approaches—and the policies that enable or restrict them—could have a chilling effect on individual expression or allow harmful speech to proliferate.

Effective platform self-regulation will be critical, but policymakers should work with industry leaders to mitigate the greatest potential harms from immersive content. At the same time, they should ensure platforms have the necessary tools and knowledge to implement content moderation approaches that protect users from harm while enabling online speech. To this end, policymakers should take the following actions while MUIEs are still in their nascent stages:

- Establish channels for platforms and law enforcement to work together to identify and respond to dangerous or illicit content in MUIEs;
- Strengthen protections against real-world harms that could arise from activities in MUIEs, including non-consensual pornography and other defamatory content, fraud, and threats to child safety;
- Ensure intermediary liability protections extend to providers of MUIEs, and that any changes to intermediary liability law consider the potential impacts on AR/VR communications platforms;
- Create a working group to develop guidance on intellectual property and copyright protections to promote innovation, fair compensation, and creative expression in immersive experiences; and
- Establish voluntary guidelines for identifying, responding to, and reporting on harmful content and content moderation activities in MUIEs.

This report reviews the policies and practices in content moderation today that will inform online speech protections in new communications technologies. It explores the ways in which MUIEs are similar to existing communications platforms, identifies where they differ, and discusses the unique challenges that arise from this distinction. It then discusses the broader implication of these considerations for online speech and free expression in an immersive future and offers recommendations to policymakers to balance concerns about potential harms with measures to protect users' speech rights in immersive spaces.

DEFINING FEATURES OF MUIES

MUIEs can be partially virtual (i.e., AR) or fully immersive (i.e., VR) environments. AR experiences merge real-world surroundings with virtual elements. In contrast, users in VR experiences interact within a fully digitally rendered environment, typically through embodied avatars (virtual representations of themselves that can mirror their physical movements). Importantly, users may participate in MUIEs with various devices, including mobile devices, computers, heads-up displays (HUDs) such as glasses or visors, or head-mounted displays (HMDs) that fully obscure their physical surroundings. MUIEs can serve many purposes, including as professional collaboration tools and learning environments; as venues for social gatherings and events; and as platforms for multiplayer games. Many can serve multiple, overlapping (and sometimes conflicting) purposes. These immersive features and wide range of use cases differentiate MUIEs from other communications platforms and raise considerations for content moderation specific to this medium.

Components of MUIEs

Like the devices and underlying technologies that enable them, the individual components of MUIEs are not necessarily unique.[6] Video conferencing (e.g., Zoom or Microsoft Teams) and audio chat (e.g., Discord or Clubhouse) platforms, multiplayer games (e.g., Fortnite or World of Warcraft), 2D virtual worlds (e.g., Second Life or Roblox), and even conventional social media platforms (e.g., Facebook or Tik Tok) offer similar elements and raise similar concerns. However, MUIEs require specific considerations for content moderation and online speech because they combine these elements in ways that other multi-user communications platforms do not.

Table 1: Comparing components of MUIEs with other digital platforms

	Ephemeral	Interactive	Individual	Mediated	Embodied	Physical
Video conferencing	Yes	No	No	No	No	No
Voice chat	Yes	No	No	No	No	No
Multiplayer games	Yes	Yes	No	Yes	Yes	No
2D virtual worlds	Yes	Yes	Yes	Yes	Yes	No
Social media	No	No	Yes	No	No	No
Multi-User VR	Yes	Yes	Yes	Yes	Yes	Yes
Multi-User AR	Yes	Yes	Yes	Yes	Yes	Yes

- **Real-time, ephemeral communication:** Because MUIEs are designed to mirror real-world interactions, much of the content they will moderate will happen in real-time and without lasting records. This raises challenges similar to those faced by audio and video chat platforms such as Discord, Zoom, or Clubhouse.
- **Interactive virtual content:** Digital content in MUIEs is not static; it includes immersive spaces and three-dimensional, interactive objects. In VR, these objects exist in purely virtual space, and in AR they may also augment or interact with physical surroundings. This raises challenges similar to those found in 2D virtual environments such as Roblox or Second Life.
- **Individualized experiences:** MUIEs are driven by user choice and active engagement, rather than passive observation. Each user will have a unique perception of a virtual experience and can choose where to go, whom to engage with, and what objects to interact with. This combines the considerations raised by recommendation algorithms on 2D platforms with those found in interactive multiplayer games.
- **Mediated conduct:** User behavior is just as important in MUIEs as their digital interactive elements. MUIE platforms have the ability to restrict user behavior, such as by stopping avatars from making certain motions. This raises content policy decisions similar to those in both 2D interactive environments and multiplayer games.
- **Embodied avatars:** In most VR MUIEs, and some AR MUIEs, users will interact with one another in the form of 3D avatars that mirror the real-world actions of individual users. This allows for more lifelike experiences, but exacerbates challenges for user safety and conduct monitoring similar to those raised by 2D platforms that include avatars and first-person perspective.
- **Physical experiences:** Perhaps the most clearly defining feature of MUIEs is their ability to merge physical and virtual experiences. VR MUIEs do this through embodied avatars, which give users the sensation that they are physically experiencing what their avatar sees and feels. Meanwhile, AR MUIEs create the illusion of virtual objects positioned within physical space. These features are largely unique to MUIEs.

Forms of MUIEs

MUIEs cover three general types of services. First, there are social experiences, or the immersive equivalent of conventional social networking platforms. Also referred to as social VR or AR, these services allow users to interact with each other one-on-one or in groups of varying sizes, and to share experiences with virtual objects or environments.[7] For example, a group of friends could meet in a virtual living room—or in AR, as virtual avatars in each other's physical space—or multiple users could share AR artwork in a public park.[8] Second, there are enterprise collaboration platforms, which are more targeted to professional users. For example, a team of office workers could hold a meeting in VR, allowing them to communicate "face-to-face" from different locations, or a group of designers or engineers could remotely share and interact with 3D models using AR.[9]

Finally, multiplayer gaming and entertainment platforms are perhaps the most well-known types of consumer MUIEs. These include games similar to 2D massive multiplayer online games (MMOs) as well as multi-user entertainment and events similar to livestreamed content on 2D platforms.[10] For example, users could compete against other players in a virtual battle royale arena (either fully virtual in VR or overlayed on a physical space in AR), or stand in front of the stage at a virtual concert.[11]

Table 2: Different forms of multi-user immersive experiences (MUIEs)

MUIE Type	Definition	2D Examples	Immersive Examples
Social experiences	A service that allows multiple users to communicate through one-on-one conversations or in widely accessible for a, with rules established and maintained by a platform provider	Social networking platforms	Socializing in a virtual living room; sharing virtual artwork in a public park
Enterprise collaboration	A service built to bring individuals together to communicate and share information for professional purposes, with some guidelines set by a platform provider but with most rules and enforcement actions managed by an organization administrator	Internal communication and videoconferencing services	Holding meetings in virtual spaces; collaborating remotely with 3D models and AR
Multiplayer games & entertainment	A service that brings together multiple individuals for a specific entertainment purpose, such as playing a game or viewing an event	PC or console-based MMOs, livestreams and virtual events	Playing immersive battle royale games; attending a virtual concert

These services may also overlap. For example, a social platform could also serve as a tool for professional networking or enterprise events. Similarly, multiplayer gaming and entertainment services could offer social features, such as non-combat spaces in a game or the ability to meet and interact with other attendees at a virtual concert or live event.

Communications Mediated by MUIEs

Because they serve such a wide range of purposes, MUIEs can facilitate multiple forms of interpersonal interactions. As with the components of MUIEs, the types of interaction are not unique to this medium. However, immersive experiences are meant to mirror "real world" interactions—meaning they often facilitate all of these interactions simultaneously across a single platform.

Interactions in MUIEs can be defined by familiarity as well as scale. Familiarity considers whether and how individual users may already know one another. Users who are known to one another and/or have a relationship off-platform (familiars) may require different content moderation and safety tools than users who were not previously known to one another, and whose relationship was facilitated by the platform (strangers). Familiars might include friends, family, colleagues, or new acquaintances. Strangers, on the other hand, could be introduced to one another via a platform that allows for public interaction or includes matching features, such as pairing up players in a game.

Table 3: Examples of interactions facilitated by MUIEs and 2D communications platforms

	One-to-One	One-to-Many	Many-to-Many
Familiars	• *One-on-one conversations in MUIEs** • Private video calls • Direct messaging*	• *Inviting known users to virtual environments in MUIEs* • *Sharing virtual content in MUIEs** • Photo- and video-sharing applications* • Conference call or videoconference	• *Large-group gatherings with known people in MUIEs* • *Collaborative projects with known people in MUIEs* • Group chats on messaging services • MMOs*
Strangers	• *One-on-one conversations in MUIEs** • Dating apps • Randomized chats • Direct messaging*	• *Inviting all users to virtual environments in MUIEs* • *Sharing virtual content in MUIEs** • Photo-sharing applications* • Livestreams and webinars	• *Large-group gatherings with strangers in MUIEs* • *Collaborative projects open to unknown people in MUIEs* • Wikis • MMOs*

**Most platforms will allow users to configure whether these communication channels are open to strangers*

Scale considers the breadth of users that distribute and receive immersive content. In one-to-one interactions, as one user shares content and one user receives it. This could include private conversations or privately sharing a virtual object, much like direct messages or private chats on 2D. In one-to-many interactions, multiple users receive content shared by a single individual, such as a virtual object placed in a public space. This is similar to most user-generated content on existing social media platforms. Finally, MUIEs can facilitate many-to-many interactions, in which multiple users both contribute and receive content. This is comparable to group chats, chat rooms, collaborative platforms such as wikis, and MMOs.

THE CURRENT LANDSCAPE FOR CONTENT MODERATION AND ONLINE SPEECH

Online platforms continually innovate to make digital communications safer, more engaging, and easier to use, encourage creative expression, and meet their growing user bases' evolving needs and expectations.[12] MUIEs can draw from the tools and best practices that existing platforms have developed when shaping their own content-moderation approaches. The expansion of the Internet and Internet-enabled devices has allowed individuals to exercise their right to freedom of expression and opinion globally. At the same time, the open and user-driven nature of Internet communications also creates opportunities for malicious misuse. Bad actors can exploit online platforms for their own gain. As people increasingly rely on these platforms to interact with and understand the world around them, these malicious users' online activities can translate into real-world harms. Platforms therefore develop content moderation approaches to mitigate the potential for harm by preventing bad actors from abusing their services and minimizing the impacts of those who do.

Real-World Impacts of Online Content

After over two decades of evolution in consumer Internet services, it has become clear that user-generated content reflects the best and the worst of the "offline" world. The proliferation of digital communications services demonstrates that, overall, there is value to the types of content that appear on these platforms. But although a mostly open Internet can positively impact people's lives, there should still be limits online to prevent less desirable, or outright illegal, content from harming individuals.[13]

Online communications platforms have expanded opportunities to exercise free speech and creative expression in many ways. They have created virtual spaces to build new communities across physical distance, often offering systems of support that would not otherwise be possible. User-generated content have also enriched civic engagement, advocacy, and political speech and amplified voices from marginalized communities. And digital communications platforms continually create new channels for creative expression and innovation. This expanded access to information, community, and opportunity can positively impact users' daily lives.

Unfortunately, not all user-generated content is positive. At a certain point, content can cross the line from welcome forms of free expression to unwanted or harmful activity. Where to draw that line depends largely on laws, as well as the norms and expectations of individual platforms and their users. But certain forms of content are widely viewed as unacceptable, including mis- and disinformation, harassment, and hate speech. This kind of content can have a wide-reaching

impact in the real world, including social and political consequences, emotional harms, and, in some cases, targeted violence.

The most egregious instances of malicious activity online violate laws and norms of acceptable behavior and can cause lasting real-world harm. For example, non-consensual pornography (NCP, aka "revenge porn") can inflict emotional, economic, and in some cases even physical harm on victims.[14] Other exploitative content, such as child sexual abuse material (CSAM), can also proliferate on online platforms if sufficient safeguards are not in place. And violent content, or content threatening or inciting violence, can cause emotional harm to targets even if it does not translate into action.

Online platforms can significantly impact how individuals express themselves and experience the world. While they must follow local laws in the United States and other countries where they operate, they are ultimately private arbiters of speech and the policies governing them can play an outsized role in shaping user speech and safety.[15]

The Regulatory Landscape for User-Generated Content

Content moderation practices represent a combination of regulatory compliance measures, industry best practices, and policies or mechanisms that address the unique needs and expectations of individual platforms' users. Although a significant portion of content moderation and user conduct policies and practices are self-regulatory, platforms develop these within the framework of laws and regulations that apply to both the service providers and individual users.

Relevant Laws and Regulations in the U.S. Context

In the United States, Section 230 of the Communications Decency Act protects online communications platforms, from social media to crowdsourced review websites, from liability for the content that users share on their platform.[16] Many experts and advocates credit this law for allowing an open Internet to flourish as platforms had the freedom to experiment with different moderation approaches without fear of legal repercussions for good-faith efforts to mediate the interactions on their services. Because the law applies broadly to interactive computer services, it will also apply to future communications platforms—including immersive experiences.

However, this allowance does not mean current or future digital communications platforms are wholly exempt from legal compliance. First, there are exceptions to the intermediary liability protections under Section 230. Since its introduction in 1996, courts have determined several exceptions to the liability shield Section 230 provides, namely when platforms directly induce or develop (rather than unknowingly facilitate) illegal content; if they fail to remove content that constitutes a breach of contract, or fail to warn users of known illegal activity; and if they selectively repost such content (thereby acting as a publisher) or fail to act in good faith.[17] Further, under the Stop Enabling Sex Traffickers and Fight Online Sex Trafficking Acts (SESTA/FOSTA), service providers also lose Section 230 protections if they appear to host content promoting prostitution.[18] Platforms are also responsible for removing content that violates intellectual property rights. Under the Digital Millennium Copyright Act (DMCA), service providers are expected to remove copyright-infringing content when notified by copyright owners.[19]

Due to intermediary liability protections and safe harbor provisions, other laws that address online content primarily target harmful or illegal activities by individual users. Activities prohibited in "real life" are still against the law if they take place online. These include credible threats of violence, fraud and identity theft, and trafficking. Further, the most extreme instances of online abuse can be prosecuted under related laws, such as those that address defamation, harassment, stalking, or extortion.[20] A handful of state-level laws further dictate acceptable content and conduct online. For example, most states and the District of Columbia have implemented "revenge porn" laws that hold perpetrators (rather than platforms) accountable for disseminating non-consensual pornography.[21] A few states have recently introduced laws meant to address new threats from synthetic media, including "deepfake porn" and targeted mis- and disinformation.[22]

Policy Gaps

Although the U.S. regulatory landscape addresses many of the potential harms from online content and conduct, there are still notable policy gaps that, to date, have fallen on platforms and other industry actors to fill. Policymakers are already trying to catch up to rapid changes in communications technologies and two-dimensional media. If left unaddressed, these gaps will only widen as new, immersive mediums are more widely adopted. Most of these shortfalls come from inadequate protections against malicious activity and platform abuse. For example, there is no federal law or guidance on identifying and removing harmful synthetic media. Policymakers should ensure that the regulatory landscape provides sufficient guidance for platforms, necessary guardrails to protect against malicious misuse, and adequate protections against or legal recourse for victims of online abuses.

Global Considerations

Although this report focuses on U.S. companies and regulatory considerations, it is important to note that content policy decisions (and the laws that dictate them) will impact a global user base. Outside the United States, most democratic countries have intermediary liability laws in place with similar goals to balance protecting free speech and mitigating risks.[23] However, not all countries in which platforms operate will have similar commitments to free expression. In many cases, platforms may find themselves responsible for making decisions about these trade-offs in precarious environments.[24] Further, content policies that promote online speech in one context may facilitate malicious misuse in another. For example, community guidelines that allow users to post anonymously can offer valuable protections for dissident voices online but also conceal the identity of harassers. Or, on a broader scale, policies that turn to government sources to determine parameters for acceptable content might help platforms successfully identify and remove dangerous organizations or misinformation in countries with strong protections for speech rights but inadvertently limit online speech in more precarious human rights environments. Thus, platforms should develop moderation and monitoring approaches that are adaptable enough to balance privacy, safety, and free expression across a global user base.[25]

Platforms often have to comply with laws and regulations in many countries, which may put more restrictions on their content moderation practices than those in their host countries. For example, Germany's *Netzwerkdurchsetzungsgesetz* (NetzDG) requires platforms to remove unlawful content within 24 hours of notice.[26] Other regulations, such as strict anti-defamation laws, may place additional responsibilities on companies operating across a wide range of regulatory environments with varying parameters of legally permissible speech. In some cases, governments

can use vaguely worded laws prohibiting certain types of content such as defamation or extremist propaganda to justify demands to remove dissenting or other speech they deem unfavorable.[27]

How Communications Platforms Define and Identify Unacceptable Content

Platforms typically set additional parameters for acceptable content beyond what may be limited by law based on the nature of their service and the expectations of their users. Even for illegal content, platforms generally choose how to monitor for and respond to this kind of content. There is a broad consensus on what platforms should moderate and how to do so for some of the most egregious forms of illegal, harmful, or otherwise unwanted content. For example, platforms and regulators can agree that content such as CSAM and violent terrorist content should be expeditiously removed from online platforms, and laws such as the DMCA and federal criminal law provide clear definitions to identify and respond to unlawful content.

Other forms of content are generally viewed as unacceptable but may be more difficult to define or identify accurately. For example, it can be difficult to identify whether intimate media was posted with consent, set the parameters for defamatory content or hate speech, or determine whether violent imagery is dangerous or informative. Similarly, there is no clear definition of mis- or disinformation or guidelines for when to remove false or misleading information. The responsibility largely falls on platforms to determine the parameters for this content and develop policies to respond to it. As social media and other digital communications platforms have evolved, they have developed increasingly complex content policies and practices to respond to harmful content and protect the free speech of users.

Content Moderation Practices on 2D Platforms

Most 2D platforms, including social media platforms, multiplayer games, and even enterprise collaboration tools, have developed robust content moderation approaches that take into consideration the needs and expectations of their user base, advertisers, and content creators; relevant laws and regulations; and the trade-offs that may exist between promoting speech and protecting their users from harm both on- and offline. Based on these considerations, platforms will adopt content moderation models that usually include some combination of community guidelines, user reporting, and proactive moderation from both human moderators and machine-learning tools.[28]

There are many reasons for taking action on content. In some instances, these actions are a direct response to legal takedown demands for unlawful content or copyright infringement, while in others, they may be due to content that violates a platform's terms of service or community guidelines but is not expressly illegal. In the former case, the necessary action is usually straightforward, but in the latter, it is largely up to the individual platform or service to decide how to respond. Further, content that may present some risk of harm (or may not be desirable to most users), but that does not directly violate terms of service, could benefit from different enforcement actions than content that is in direct violation of relevant laws or platform rules.[29]

Content moderation has evolved significantly over the past several decades as the need for more nuanced approaches has become increasingly evident. Platforms that host user-generated content rely on a set of tools to identify potentially harmful content and take necessary actions. First, they utilize both automated systems and human moderators to detect and respond to unacceptable content. Automated systems can be a valuable tool as they can process higher

volumes of content more efficiently—but not necessarily more effectively—than human moderators. Automated systems are particularly useful for preemptively identifying and removing or flagging content that has well-defined parameters, such as CSAM or copyright infringement. However, automated systems are susceptible to over- or under-moderation of more context-dependent content, such as extremist content, hate speech, harassment, or other violent or offensive content. [30] For these decisions, human moderators can make more accurate calls on whether content violates internal policies.

User controls can also be a part of a platform's overall content moderation approach. Users may find some content that does not reach the threshold of rules violations offensive or harmful, and tools such as blocking other users or hiding unwanted content can protect individual users without limiting others' speech. Further, due to the volume of content, most platforms rely on user reporting to identify potentially rules-violating content.

If content does violate platform rules (but not content-related laws), platforms can take actions that will impact user speech and safety in different ways. Rather than simply removing the content or user from the platform, either permanently or for a specific period, they may choose to add "friction" to slow the speed at which other users share it. Platforms may also use this type of response to give more time for fact-checkers to review potential mis- and disinformation.[31] They may also choose to include more contextual information, such as fact-check labels or violent content warnings. Or they may demonetize the content or impose other financial restrictions on the account to take away any financial incentive users might have to post inflammatory content that violates community guidelines in a bid to attract viewers. The consequences for violating a platform's content rules, especially in professional or educational situations, may be offline, such as someone getting fired from their job or suspended from school. These approaches form a valuable foundation for content moderation, but no system is perfect. The sheer volume of information flowing across these platforms makes it virtually impossible to always catch even the most clearly harmful content, and this effort is made even more difficult by the ambiguity around other forms of content where the line between free expression and harmful speech can be difficult to draw. For example, moderators require additional context to decide whether a user is airing legitimate grievances or engaging in hate speech or incitement to violence. And in some instances, content that is acceptable in one context may violate terms in another, such as hate groups using coded language or even individuals using words or phrases that have different meanings across cultures. Further, new forms of content, such as real-time audio and video communications and synthetic media, present new challenges that existing industry standards may not be sufficient to address. The immersive, multimodal nature of MUIEs will likely exacerbate these challenges.

USER ACTIVITIES AND CONTENT MODERATION IN MULTI-USER IMMERSIVE EXPERIENCES (MUIES)

Content moderation is a complex effort across platforms, and MUIEs are no exception. Many of the considerations, challenges, and components from 2D platforms are also necessary components of MUIE content policies. However, immersive experiences contain unique elements that present specific challenges for developing and enforcing content policies. First, there is a greater variety in types of content that MUIE platforms mediate, including not just digital media, but also virtual environments and real-time verbal and non-verbal communication. Second,

content policies will vary based on the intended (as well as potential and unintended) use cases for platforms, including professional collaboration, casual socializing, and gaming and entertainment. Third, MUIE users may access experiences with varying levels of immersion, from fully immersive head-mounted displays to two-dimensional computer screens—which means platforms must develop policies that address varying perceptions of the same experience. And finally, MUIEs can use a combination of direct moderation approaches, community standards, and user controls to shape user experiences—but the immersive nature of these interactions raises unique challenges in balancing user safety with privacy and free speech objectives.

Table 4: Elements of MUIEs that can impact content policies

Element	Components	Moderation Challenges
Content Type	Environments, objects, actions	Ephemeral content and nonverbal communication; adding friction to three-dimensional, real-time, interactive content; technical challenges in differentiating harmful and non-harmful elements; challenges identifying and removing unlawful content
Use Context	Professional, social, entertainment	Establishing norms of behavior through content policy and user controls; policies for multiple use contexts on the same platform
Immersion Level and Means of Access	Fully immersive (HMD), partially immersive (HUD or mobile device), 2D (personal computer or mobile device)	Different perceptions of the same experience based on level of immersion; definitions of unacceptable content will vary by level of immersion; need to implement adaptive, contextually relevant moderation approaches
Moderation Approaches	Moderated, community-based, delegated	Need to balance safety with privacy and free expression

Elements of MUIEs

MUIEs share a number of traits with their two-dimensional counterparts. However, their ability to merge physical space with virtual elements, or mirror physical actions in virtual spaces, distinguishes them from other digital communications platforms. In order to understand how these technologies can shape online speech, it is important to recognize the content that comprises these experiences, the different contexts in which they are used, how users access virtual spaces, and the different ways that platforms monitor and moderate user activities.

Types of Content

Unlike more traditional digital media platforms, content in multi-user experiences includes not just the media that users produce, but also their actions, real-time interactions and activities, and the spaces in which these take place.

First, there are ***environments***, or the digitally rendered, three-dimensional spaces that users navigate in an immersive experience. These may be fully virtual (i.e., the entire space is digitally constructed in VR) or partially virtual (i.e., digital elements augment or alter physical surroundings). Environments may be built by a platform provider, users, or a combination of the two. For example, many social VR platforms allow users to alter templates or build environments from scratch. Environments are an important differentiating feature for MUIEs that quite literally set the stage for all interactions and activities within an immersive experience.

Second, ***objects*** are digitally rendered elements that users can manipulate or interact with. Objects include two-dimensional digital media (such as images, text, or video) within a virtual space as well as digitally rendered, three-dimensional elements placed within a virtual environment or overlayed on physical surroundings. Like environments, objects may include platform-provided elements as well as user-created content. In some instances, users may purchase or trade these virtual assets. Although this type of content is likely the most analogous to traditional understandings of user-generated content, it presents unique challenges for both automated and supervised content moderation.

Finally, ***actions*** broadly comprise the ways in which users conduct and express themselves in immersive experiences. In fully virtual environments, this includes the avatar a user selects. Depending on what the platform allows, this could range from a photorealistic replica of their physical appearance to a completely different animated character. It also includes how individuals interact with others through these avatars, including gestures, actions, and proximity. Real-time interactions, including audio conversations, also fall within this category. Unlike environments or objects, actions are solely user-driven, making this an important area for robust content moderation approaches that allow users to interact with one another in safe and engaging ways.

Use Context

MUIEs can also be defined by their use context, i.e., their intended purpose. Although social applications and multiplayer games are perhaps the most well-known example of MUIEs, immersive technologies can enrich collaboration and communication in a variety of settings. For example, MUIEs have significant potential in professional settings. Fully immersive meeting spaces present a more interactive and realistic alternative to in-person meetings for remote teams. And AR and mixed reality (MR)—an extension of AR technologies that not only displays virtual objects over physical space, but also allows users to manipulate or interact with those objects using physical motions such as hand gestures—can allow individuals to share, interact with, and even manipulate digital objects in real-time, whether or not they are in the same physical space. Similar solutions can also be valuable to distance learning, research, and other educational uses. In this context, the users represent a very specific group of individuals interacting for a specific purpose, who may also interact with each other outside of the MUIE setting. Although such enterprise and education-focused applications are still in their early stages, this segment is expected to grow rapidly in the coming years.[32]

In contrast, MUIEs created for entertainment, such as multiplayer games or interactive media, can host a broad set of users who might only interact within the context of that experience. However, like professional or education contexts, users will interact with one another for a specific purpose (i.e., within the structure and objectives of the entertainment experience, such

as a game or performance). Further, in addition to specific games or entertainment experiences, MUIEs can also serve as multi-purpose social platforms.[33] Although these may have a general theme or objective, they are most analogous to the two-dimensional social platforms that carry the bulk of digital communications today. Here, users may represent a broad range of interests and objectives, as well as varying levels of familiarity with one another.

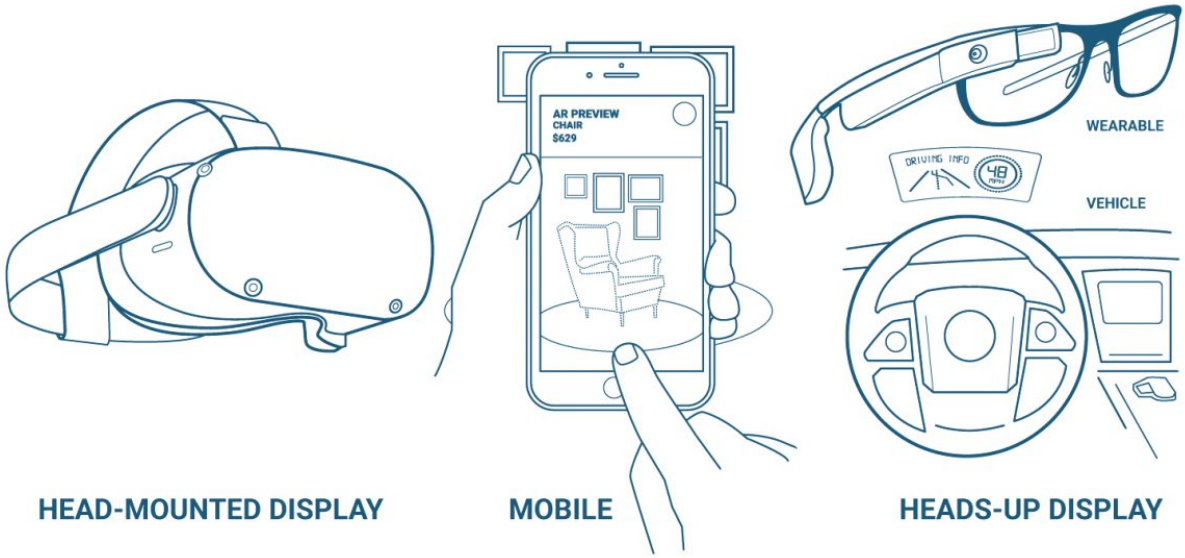
In many instances, MUIEs will serve multiple purposes. For example, multiplayer games can also involve socializing beyond the necessities of play, and a platform that offers social experiences could also serve as a professional meeting space. Because of this, content moderation practices will have to be flexible enough to adapt to different use cases, while still providing sufficient safeguards against harmful or unwanted content across a given platform.

Level of Immersion

The term "immersive experience" can apply to a wide range of technologies. Users may access MUIEs using computers or mobile devices, heads-up displays that allow them to maintain situational awareness, or head-mounted displays that fully submerge them in a virtual environment. The means of access—and subsequent level of immersion—can impact how a user engages with and responds to a virtual environment. For example, in fully immersive VR environments in which users embody avatars and access a virtual environment through an HMD, they may perceive the experiences of virtual versions of themselves as their own, and conversely, the characteristics of their avatar can impact their perceptions of their physical selves.[34] This sense of being "really there" is a unique and valuable attribute of immersive experiences, but also means that both positive and negative interactions can have greater impacts on individuals' emotional wellbeing. [35] At this level of immersion, every aspect of the experience is digitally rendered—and therefore, also mediated. However, users may also experience fully virtual environments through two-dimensional screens, such as a personal computer or mobile device. They may encounter other users who are using similar devices as well as those using fully immersive HMDs. In this case, every aspect of their experience is similarly rendered, but they do not have the same first-person, embodied experience as they would when using an HMD.[36]

Meanwhile, partially immersive AR experiences allow users to view and interact with digital elements, including other users, in their physical space using HUDs or mobile devices. Although this experience may be more "hands-on" than two-dimensional media, users still have an awareness of their real-world surroundings, removing the "really there" element of immersion. Here, only the digital elements are directly rendered (i.e., mediated) by third-party providers—but they are not independent from physical surroundings. For example, a digital overlay could be used to alter someone's appearance, or add objects to a space. Therefore, this collision of virtual and physical elements is also an element of content policy for these experiences, as platforms must set parameters for both which physical spaces are acceptable to interact with (e.g., through geofencing or other technical tools) as well as parameters for acceptable augmentation of those spaces (e.g., monitoring for defamatory or otherwise offensive additions to buildings, open spaces, or other people).

Figure 1: Examples of levels of immersion for MUIEs



Moderation Approaches

Drawing from existing content moderation as well as real-world mediation approaches, MUIEs implement content monitoring and moderation models that vary based on use context, platform objectives, and user expectations. Most employ a mixture of moderation approaches in order to enforce their community guidelines and other policies.

Currently, many popular MUIE platforms employ a community-based model, in which users create and moderate spaces according to both platform-wide guidelines and rules that they establish specifically for that space.[37] Here, enforcement is driven by user reporting and decision-making by community moderators and, when necessary, platform staff. This model is similar to community-driven social discussion forums such as Reddit, Facebook Groups, or Discord.[38]

MUIE platforms may also elect to include more direct monitoring in their content moderation approaches. Conceptually, this is most analogous to social media platforms that rely on a more centralized content moderation approach based on user reporting, automated decisions, and human moderators. However, implementation is unique to MUIEs, as some may elect to have moderators "physically" present within the virtual environment, or otherwise monitor or review real-time interactions in addition to visual content.

In some use contexts, such active moderation may not be necessary. For example, workplace or education-oriented platforms may rely on employers or instructors to develop and enforce policies that align with other policies unique to each organization. This delegated moderation approach is most similar to enterprise communications platforms, such as Zoom, Slack, or Microsoft Teams. Here, while providers may include guidelines for appropriate use in their user agreements, they will primarily rely on user and administrator controls to meet the unique needs of the implementing organization.

Examples of MUIE Elements in Existing Platforms

Although MUIEs are still in their early stages, there are a number of platforms currently available for both personal and professional use. Below are examples of some of the most well-known platforms on the market and the elements that make up their content policy approaches.

Social and Entertainment Platforms

AltspaceVR, owned by Microsoft, is an immersive social platform focused on multi-user events.[39] Individuals and organizations can host events using platform-provided or user-generated spaces, or create their own custom spaces. Event hosts and participants interact through avatars which users create from a selection of options within the application. AltspaceVR serves a variety of use contexts, from casual gatherings to events and conferences. While it is primarily built for fully immersive VR, AltspaceVR is also accessible through Windows Mixed Reality headsets and desktop computers.[40] The platform's community standards establishes parameters for acceptable content pertaining to harassment, bullying, unwanted advances, disclosing personal information, impersonating employees, sharing inappropriate content, and respecting personal space; however, the parameters for inappropriate content only apply to publicly available spaces.[41] The platform enforces these rules through a combination of user reporting and "concierges," or company representatives present within these spaces. Users can mute, block, or kick out other users, and control their experience with options such as activating a "space bubble," or virtual barrier that prevents other avatars from coming too close.[42]

Rec Room is an immersive platform focused on multi-user games.[43] The content in Rec Room is largely user-generated, including spaces and activities, objects, and outfits and accessories for avatars. Although the purpose of Rec Room—playing games with others—is fairly straightforward, the variety of activities means that acceptable conduct may vary slightly depending on the context of each unique space. Players can access Rec Room from multiple devices with varying levels of immersion, including HMDs, computers, game consoles, and smartphones. The platform has a basic code of conduct as well as safety guidelines for players to follow, and relies primarily on volunteer community moderators and user reporting for content monitoring.[44] The platform offers user safety controls including an adjustable personal space bubble and the ability to mute, block, or report other users. Players can also create invitation-only private spaces that allow them to interact only with users they have personally selected. Rec Room also offers "junior accounts" for users 12 and under.[45]

VRChat is a social VR platform driven by user-generated content, including virtual environments and activities, virtual objects, and fully customizable avatars.[46] The platform is used almost entirely for socializing and entertainment, but like other immersive platforms with user-generated spaces, the use context and acceptable conduct will vary between different spaces. VRChat users agree to baseline community guidelines and terms of service, but different experiences within the platform may impose additional rules.[47] Although the platform is geared toward fully immersed VR users, it is also accessible via a desktop application.

Professional Platforms

Horizon Workrooms is the professional collaboration application in Meta's "Horizon" suite of VR applications, which also includes Venues for virtual events and Worlds for social experiences and

activities. Unlike social and entertainment platforms, the environments and most spaces in Workrooms are built by the platform.[48] Users may add content to the spaces, such as by drawing on a whiteboard or sharing a screen, but content moderation focuses primarily on user actions. Users in VR interact through avatars, while those on computers join in the format of a more traditional video call. Workrooms is built explicitly for professional collaboration, and content policies are specific to this context. For example, the first level of moderation comes from the team leader or system administrator, who has the ability to remove users. However, teams and individual users agree to follow the Facebook Community Standards or Conduct in VR Policy, and users can report violations of these policies directly to the platform.[49]

Spatial is an immersive communication and collaboration platform that allows users to meet virtually in either pre-built spaces or custom environments.[50] Users interact through three-dimensional avatars generated from 2D images, or through video chat.[51] They can add notes in real time also "bring in" digital media including documents, web pages, and virtual 3D models. Although focused on enterprise use, there are not restrictions on the kinds of gatherings that can be hosted in Spatial; use contexts include education, professional collaboration, and events and presentations, and individuals can access these gatherings from VR or AR headsets, AR-enabled mobile devices, or desktop computers.[52] Perhaps because of this versatility, Spatial utilizes a variety of content moderation tools including human moderators, technical controls, and user reporting to enforce relatively robust community guidelines that specify what is considered acceptable conduct and content to upload and share on the platform.[53] Room hosts also have the ability to mute or remove disruptive users at their own discretion.[54]

Content Moderation Practices and Challenges in Immersive Experiences

As with more established two-dimensional platforms, user interactions and activities in MUIEs can have lasting real-world impact. The underlying concerns are similar: securing users' speech rights while mitigating the potential for emotional, economic, and even physical harm as a result of online activities. The best of 2D multi-user platforms can also be seen in MUIEs: they can connect individuals across distances and offer otherwise inaccessible experiences to a small but growing community of users.[55] But the reverse is also true: hate speech, harassment, and other potentially harmful interactions can take place in these virtual environments.[56] Because of this, some of the best practices that these 2D platforms have developed can serve as a blueprint for content policy in multi-user AR/VR. However, while some challenges (and their solutions) may be similar and replicable, others will not translate directly into immersive experiences.

It is worth noting that the different elements of MUIEs will often interact and overlap. For example, user-generated content and conduct will vary between use contexts (e.g., using different avatars for social and professional experiences). But they also present distinct considerations when developing appropriate content moderation practices.

Content

As discussed, user-generated content in MUIEs encapsulates a much broader set of elements than their two-dimensional counterparts. The different types of content—and the ways in which they depend, build on, and interact with one another—require a more holistic approach to mediating these virtual spaces. User-generated content in the form of environments, objects, and

actions presents unique considerations for content moderation by platforms as well as speech from users.

First, because the full experience is digitally rendered, it is more difficult to establish clear definitions of what constitutes user-generated content and establish clear parameters for acceptable conduct. Monitoring for offensive language or imagery is one thing, but fully immersive platforms also have to consider gestures and nonverbal actions that users may present through avatars. Indeed, nonverbal cues are a necessary component of realistic, engaging multi-user immersive experiences—but they are also more difficult to identify or moderate.[57] Platforms moderating nonverbal actions have to distinguish rude gestures, coded hand signals, or violent actions from other nonverbal gestures—for example, distinguishing a violent slap or punch from a high-five or fist bump. Further, the three-dimensional and often interactive nature of immersive content distinguishes it from traditional digital media. Not only can users observe this content from multiple vantage points, they can also manipulate or interact with it as they would physical objects. And the wide range of user interactions with one another and the digital elements in their environment are often synchronous and ephemeral, meaning much of this immersive content is generated in real-time and is not easily retained for future review.

These distinctions are exacerbated by the different types of content. Environments, objects, and actions each serve distinct purposes in MUIEs, so platforms cannot approach all immersive content in the same way. User-built environments will likely have different guidelines for acceptable use than user-created objects, since environments serve as spaces for user interaction while objects may be a component of these interactions. Further, objects themselves may be acceptable content, but the ways in which users interact with them may break platform rules. One oft-cited example of this is the "Ugandan Knuckles" incident, in which some users on the social VR platform VRChat used a 3D model of a cartoon character to harass other users by physically blocking space and making offensive and racist comments.[58] Put simply, there are both more avenues for user speech and new vectors for malicious misuse in MUIEs than in other platforms for multi-user interaction.

Because of this, immersive content introduces new technical, behavioral, and legal compliance considerations for platforms. At the most basic technical level, it is challenging to distinguish individual digitally rendered objects. For example, technical tools alone might not be able to identify an offensive poster on a virtual wall, weapons or other violent objects on a virtual table, or inappropriate or even copyright-infringing imagery on an avatar's virtual clothing. Because of this, it can be difficult to identify and flag unacceptable content for removal through automated systems—and there is also the possibility that objects or environments that fall within a platform's rules for acceptable content could be mistakenly removed.

Immersive content will also require different understandings of user behavior and behavior-based moderation approaches. As this report has discussed, many two-dimensional multi-user platforms have found some success balancing speech and safety by introducing measures that add "friction" to posts (i.e., reduce engagement without removing the post). For example, a platform might add fact-check labels to potentially misleading posts, or put a content warning on violent imagery. However, these approaches might not have the same effect in three-dimensional spaces. User engagement with MUIE content is meant to feel like interactions with real-world objects and individuals: they can view immersive content from multiple angles, interact with it,

and in some instances even manipulate it in real-time. It seems unlikely that a two-dimensional content label would significantly discourage engagement with this type of content. Some experts have suggested more hands-on approaches, such as bystander intervention training, but these are largely voluntary and could be difficult to scale.[59] Further, because most immersive experiences are driven by real-time interaction, a significant portion of the user-generated content in MUIEs will require new and highly adaptive monitoring and moderation approaches that address the challenges of ephemeral content—something existing platforms are already struggling to accomplish.[60]

Finally, immersive content could raise new legal and compliance questions. As this report has discussed, there are certain types of content and communications that platforms are legally obligated to remove or otherwise address, and MUIE providers must also comply with these requirements. Many of the content challenges discussed above will also apply to identifying and removing illegal material in MUIEs. For example, it is more difficult to stop real-time, verbal conversations than it is to remove written messages or posts, or to distinguish illicit objects from an "acceptable" environment. There are also definitional concerns when it comes to areas such as intellectual property law, particularly as MUIEs become more hyper-realistic. For example, immersive experiences may contain unauthorized virtual replicas of patented works, or serve as a channel to distribute unlicensed digital media. Companies may be faced with "virtual knockoffs" of their works, and some are already taking steps to maintain ownership over key elements such as logos, slogans, and designs in immersive spaces.[61] Further, location-based AR experiences blur the lines of physical property and virtual additions—that is, real-world objects or locations may remain physically untouched but significantly altered virtually for users within an AR experience. Conversely, the virtual media that AR devices may superimpose on physical space (including public spaces or private homes) could include copyrighted material or reveal protected trade secrets.[62] While intellectual property protection on websites, distribution platforms, and in video games or other interactive media may inform approaches for MUIEs, the immersive and lifelike nature of these platforms may raise new questions about acceptable use, definitions of parody, and First Amendment protections for creative expression.[63]

Context

In the "real world," different social settings have distinct norms of behavior—and the same is true of different virtual spaces. The definitions of acceptable content and conduct will vary between different use contexts for MUIEs. For example, users will behave differently in a multiplayer gaming environment than in a professional meeting or collaborative space. What is acceptable or even expected in one context could be inappropriate or even dangerous in another.[64] Therefore, understanding use context, and communicating these expectations to individual users, is critical to enable individual expression while also ensuring user safety.

The idea that individuals will alter their behavior in different virtual environments is not unique to MUIEs. In today's digital world, a person may present a different version of themselves on any number of platforms. For example, they may post casual photos, use profane language, or provide extensive details about their personal lives on the social media platforms they use to connect with friends and family while behaving in a much more professional manner on workplace collaboration platforms or career-oriented websites. MUIEs will exacerbate this effect by combining digital media with real-time user actions and simulated face-to-face interactions.

Because immersive experiences are meant to mirror or enhance the "real world," norms of behavior are foundational to these spaces as the underlying social fabric of virtual worlds. Research into virtual environments has found that these expectations often form organically, driven by user engagement and self-policing.[65] However, as MUIEs become more widely used in different contexts, providers will have to develop a more concrete set of rules for user behavior that take into consideration factors such as experience objectives, user demographics, users' familiarity with one another, and types of user-generated content that can exist in an experience.

Use context is also rarely clear-cut, and platforms often have to adapt their services—and corresponding content policies and guidelines—to meet evolving use cases. This evolution can enrich virtual experiences by expanding the ways in which users can engage with one another and their virtual environments, but it can also complicate how platforms approach content moderation or limitations on user activities. For example, the popular online multiplayer game Fortnite introduced a "Party Royale" mode, a non-combat multiplayer environment where participants can socialize and attend events.[66] But when the entertainment-focused platform hosted a panel discussion about race in America, some attendees used an in-game feature to throw virtual tomatoes at the Black panelists on screen as well as at other players.[67] In the context of the battle royale game or most activities in "Party Royale" mode, this would fall within the parameters of acceptable conduct. But in the context of a serious discussion on a sensitive topic, these features inadvertently enabled harassment and other inappropriate behavior. These challenges will inevitably arise in more immersive multiplayer and multi-user experiences as well, particularly as both their user bases and scope of potential use expand.

Regardless of the intended or actual use context, platforms also have to consider users' likely familiarity with one another when developing moderation approaches. In a private one-on-one conversation between two individuals who are friends, colleagues, or otherwise known to one another, privacy needs will likely outweigh safety concerns. Here, content policies should focus on blocking mechanisms or other user-controlled tools. But a one-on-one conversation between strangers (such as unsolicited messages or game features that match random players) presents greater risk of abuse, as users could involuntarily find themselves in an experience with someone engaging in lewd, harassing, or otherwise inappropriate behavior. This may require more active moderation approaches, such as automated tools to detect inappropriate content or more robust user reporting features. Similarly, a private group or gathering (such as an internal company meeting) may turn to community or administrator moderation tools such as muting, blocking, or booting disruptive members rather than direct monitoring, while an open multi-user environment (such as a social experience open to anyone) may necessitate more active moderation from platforms themselves.

In between these two extremes, individual users may share information with multiple known users, such as sending a message or sharing a virtual object to a group of friends, or to multiple unknown users, such as adding items to a virtual environment or broadcasting a message on a public channel. Here, moderation approaches should seek to balance the privacy and speech rights of the user sharing the content, as well as the safety of users receiving it. For example, in a one-to-many interaction among familiar users, those on the receiving end have elected to participate in an interaction, while in the same kind of interaction among strangers, it is possible that users will involuntarily receive communications that are offensive or otherwise unwanted.

Because it is possible that most MUIEs could host any of these combinations, platforms should develop content policies that can adapt to the different concerns they present.

Table 5: How priorities and enforcement mechanisms may vary on different communications platforms

	One-to-One	One-to-Many	Many-to-Many
Familiars	Policies emphasize privacy; rely on user-controlled tools	Policies balance privacy and safety; rely on user-controlled tools	Policies emphasize privacy; rely on community or administrator moderation
Strangers	Policies emphasize safety; rely on user reporting and automated tools	Policies balance privacy and safety; rely on user reporting, automated tools, and active monitoring	Policies emphasize safety; rely on user reporting and active monitoring

MUIE providers must not only develop contextually relevant content moderation and monitoring approaches, but also help new users understand the norms and expectations of a given space. Entering an immersive experience for the first time can be jarring—it is more difficult to simply observe and get a sense of how these environments operate than it would be on a two-dimensional multi-user platform.[68] This means that even well-intentioned users might violate community guidelines or unwritten norms of a virtual space when they first enter. Further, dynamics that exist in real-world social environments also translate into virtual space. For example, one study found that women who enter social VR experiences take actions similar to those they would in a real-world scenario, such as avoiding large groups of male-presenting avatars.[69] Because interactions in MUIEs can feel more high-stakes than less immersive alternatives, it is necessary to find ways to safely and effectively introduce new users to virtual environments. For social and entertainment MUIEs, this responsibility will largely fall on platforms themselves. However, other use contexts, such as workplace or education settings, may also require similar efforts from the organizations providing them. Understanding these dynamics and developing moderation approaches to both enforce the rules and encourage appropriate behavior is crucial for MUIEs.

Immersion

The level of immersion in MUIEs establishes how users perceive the virtual elements of an experience as well as other participants. In AR, they will understand these elements as additions to their physical space; in fully immersive VR, they will experience a virtual environment as its own independent space. This presents important considerations for content moderation at different levels of immersion.

In AR, multi-user experiences encompass either shared perception of virtual objects in the same physical space (such as virtual artwork in a park) or real-time, shared perception of virtual objects by users in different physical spaces (such as remotely sharing 3D models). Thus, in AR, content moderation approaches have to consider not only the digital content that users generate, but also how that content augments real-world locations, objects, or even people. This is particularly important when AR elements can be "anchored" in physical space—that is, when multiple users can view and even interact with the same object in the same space. Here, the

virtual content can have direct real-world impacts.[70] For example, virtual signposts could misdirect users and put them in dangerous situations (such as criminals luring victims to areas where it would be easier to rob them).[71] And users could place derogatory or defamatory messages on a private home or business—or even "pin" content on a person—without the individual or owner even knowing they were there. In addition, how users interact with the digital elements of AR can also present content policy challenges. For example, a multi-user experience that allows participants to create virtual objects and anchor them in physical space would have to find ways to encourage creative expression while discouraging defacing or destroying others' virtual works. [72]

Meanwhile, in addition to moderating virtual elements themselves, VR content policies should be primarily concerned with how people interact with virtual objects and other avatars. Although they can transcend the laws of physics, fully immersive environments should be governed primarily as physical spaces because users will often experience them as they would the "real world." For example, users should not be permitted to engage in actions that would be inappropriate in physical space within virtual experiences. This includes relatively low-impact actions such as standing too close to another person as well as malicious or violent activity such as stalking, harassment, or assault.[73] And as with AR content, users should also be discouraged from defacing or destroying virtual objects or environments. This necessitates a complex system of permissions and content monitoring that encourages users to engage and interact with one another while also reducing instances of platform abuse.

Means of access will also impact experience, as well as the ways in which platforms can approach content. If users access a virtual space from different kinds of devices, they will have different perceptions of their environments as well as actions. For example, someone accessing a VR space from a headset may have a more acute perception of personal space, while someone accessing the same environment from a computer could be less sensitive to these boundaries. Similarly, someone viewing AR content through a mobile device will be more anchored in physical space than someone using a HUD. Some actions, for example standing close to someone else's avatar or seeing an object thrown at their screen, will therefore feel less "lifelike," and therefore present less risk of psychological harm. Thus, even in the absence of malicious misuse, this makes it more challenging to develop and enforce norms of behavior when individual users experience varying levels of immersion.

Because levels of immersion will have a significant impact on user experience as well as the parameters of acceptable use, MUIE providers must be prepared to implement adaptive, contextually relevant content moderation approaches that consider the complex set of actions and interactions that form the experience. Further, user controls should anticipate the different means by which users might access these environments and provide them with tools to shape the experience to their personal safety and comfort preferences. For example, many multi-user VR environments allow users to set invisible perimeters around their avatars to prevent others from invading their personal space.

Approach

All of these considerations must come together to form perhaps the most challenging aspect of immersive content platforms: moderation approaches. Private communications channels can rely largely on user-driven moderation (e.g., the ability to leave an interaction or block certain users

from contacting them), but public and semi-private platforms require a much more comprehensive approach. These platforms need to establish approaches that mitigate harms while still allowing for meaningful interactions in experiences. Community-based moderation relies heavily on established norms within user communities, which may be difficult to maintain organically as platforms scale. Further, the few industry standards and best practices that do exist developed largely based on after-the-fact user feedback, rather than more preemptive considerations.[74]

On one hand, MUIEs can sometimes offer preemptive measures to restrict user conduct and monitor for harmful content. For example, avatar movements can be automatically restricted, such as by limiting the hand gestures that are displayed or preventing avatars from standing too close to one another.[75] They also allow for more participatory moderation, such as the ability to capture real-time conversations for future review or place moderators directly in an immersive experience, either as avatars themselves or invisible observers. On the other hand, these same controls can overly limit user expression and also raise notable privacy concerns. Critics of extensive monitoring in immersive spaces have noted that people will likely behave differently and limit what they do and say if they know they are constantly being watched, or potentially being watched—and while this might achieve the objective of discouraging harmful conduct, it will almost certainly have a more widespread chilling effect on user speech. [76]

Platforms need to find the right balance between established rules, active monitoring, user reporting, and individual user controls. Most platforms have baseline requirements for conduct that users agree to when signing up for their service. For example, all users must abide by Facebook's community guidelines as well as a Conduct in VR Policy when using an Oculus headset, regardless of whether they are using one of the company's software applications.[77] These guidelines are valuable for users who may be new to these experiences and need help understanding the standard norms of behavior that are expected of them. But unless they are truly enforceable by either hardware providers or individual platforms, they might not deter users who would intentionally violate these rules. Therefore, some level of moderation and content monitoring must exist in order to properly identify and respond to rules-violating behavior.

One way to strike the right balance between safety and privacy is to implement user controls that allow individuals to shape experiences to meet their needs and expectations. For example, immersive experiences might let users choose how wide of a perimeter—if any—they would like to have around their avatar, set filters that would prevent them from engaging with certain types of objects or environments, or mute or block other users in addition to reporting inappropriate behavior. This allows platforms to develop moderation and monitoring approaches that fill in these gaps and prevents them from making wide-ranging moderation decisions for a diverse set of users.

CONSIDERATIONS FOR ONLINE SPEECH IN AN IMMERSIVE FUTURE

As this report demonstrates, MUIEs raise new considerations for user safety and free speech protections. Because immersive spaces can feel just as tangible as physical space, users may expect to exercise speech rights in immersive spaces as they would in the real world or in certain small- and large-group communication platforms, such as Zoom meetings, which may include some administrative moderation controls but generally do not feature moderation actions from platforms themselves. But because these are mediated, virtual spaces, there can be more

restrictions on user activity as platforms attempt to shape these experiences to meet the needs and expectations of their users. Indeed, the governance role of immersive platforms—especially fully immersive experiences—is perhaps even more critical than their 2D counterparts, because they mediate every aspect of user interactions from where they can go to whether and how they can interact with others. For large-group social or entertainment experiences, they must not only make direct content moderation decisions (such as removing or flagging content or suspending users), but also decide how community moderators and individual users will be able to shape their own experiences and the experiences of others through tools such as muting, blocking, kicking out, and reporting other users.[78] And while platforms may not engage in direct, active moderation of private discussions, they still shape the contexts in which these interactions take place. For example, a platform that offers both large-group gatherings and private conversations may apply certain limitations across the board (e.g., restricting where virtual objects can be placed in AR or preventing avatars from making violent motions in VR). Further, platforms that are primarily intended for professional use make decisions about the kinds of controls that will be made available to system administrators (e.g., muting users) as well as individual users. This responsibility will only grow as these platforms continue to scale up and expand to a wide variety of contexts.

Beyond individual or small-group interactions, organized activity will also look different in virtual spaces than it would on 2D platforms. For example, on two-dimensional platforms there might be a page or group dedicated to a cause or political position, or a "viral" post that raises awareness among new groups of people. But in MUIEs, this could take the form of large real-time gatherings or mass coordinated efforts to alter physical space with digital imagery. For example, a multi-user VR platform might find users organizing a protest within one of its environments—but unlike a "real world" protest, the platform provider would have the ability to limit access to that environment, restrict what users could do in that environment, or even decide whether to permit counter-protestors in the same space. Meanwhile, an AR platform might find the equivalent of a viral social media post displayed in public spaces, on buildings, or even attached to individuals—and moderators would have to determine whether this falls within the rules for acceptable use, or constitutes virtually defacing these spaces and should be removed.

These unique capabilities—and subsequent challenges—raise important considerations for both platforms and the policymakers shaping the regulatory landscape in which they must operate. First, as MUIEs gain more widespread adoption, it will be important to consider how network effects impact the governance role of providers. Today's MUIEs host only a fraction of the world's digital communications, and users are spread across a number of different platforms. If one platform does not meet an individual user's needs or expectations, they have several options for alternatives. But this structure is not guaranteed in the future—just as social media user bases condensed around platforms where they found existing social ties, it is possible that only a handful of MUIEs will dominate the market, with the leading platform (or platforms) mediating a significant share of person-to-person and large-group communications. This would mean that their content moderation decisions could have outsized impacts on individuals' abilities to share and receive information as well as exercise their speech rights in digital spaces. While this may not present significant concerns if the dominant platform(s) and the countries in which they operate share values of free expression and rule of law, such wide-reaching control over how individuals interact could stifle online speech when these values are not present, such as in countries with overbearing censorship laws.

Second, the wide range of use contexts raises questions about who is responsible for monitoring and regulating user conduct in different types of immersive experiences. For example, to what extent should an employer using a virtual meeting platform be responsible for handling misconduct within that experience? This has implications not only for online safety and content moderation, but also employee privacy and workplace safety. Although platforms may be able to provide guidelines or user controls, they are unlikely to provide direct moderation services. Therefore, it would become the employer's responsibility to balance privacy and safety within these environments. Similarly, in social or general-use MUIEs, policymakers should consider the extent to which law enforcement should be permitted to engage in virtual spaces and provide guidance for platforms.

Finally, MUIE providers face challenges similar to those of their 2D counterparts when developing policies and practices that will apply to a global user base. Not only will they have to determine how existing laws governing online speech apply to immersive content; they also have to adapt these experiences to multiple cultural contexts. Actions or objects that are offensive in one part of the world might be acceptable elsewhere. The challenges around real-time content will only become greater as more languages come into play, and in avatar-based immersive experiences, varying cultural norms around nonverbal communications as well as dress codes could further complicate content moderation and monitoring approaches. Over-moderation of these nuances can limit online speech and free expression—including in parts of the world where there are already precarious environments for free expression.

RECOMMENDATIONS

Immersive experiences have the potential to bring people together across distances to socialize, collaborate, and share knowledge in an environment that fosters free expression. However, as this report has discussed, the feeling of being "really there" and the ability to alter physical space raises unique challenges for content moderation and online speech. As with 2D platforms, the onus should fall primarily on immersive platforms themselves to develop and implement the necessary safety measures to meet the needs and expectations of their users. However, users should still be held accountable for their speech and conduct. To that end, policymakers should work with platforms to establish the necessary tools to protect and enforce the legal rights of individuals and businesses. Policy efforts should focus on two key areas: first, ensuring the legal and regulatory landscape sufficiently protects both users and non-users from the most egregious potential harms from platform abuse; and second, establishing necessary guidelines to help platforms navigate novel challenges for content policy and online speech.

Recommendations to Mitigate Potential Harms From Malicious Actors

Policy approaches for immersive experiences should prioritize identifying and mitigating potential for actual, real-world harm to individuals—both those who are using the platforms, or those who may be impacted by platform abuse. Although the medium may be different, the greatest potential harms from MUIEs are similar to those from other multi-user platforms, including social media and other digital communication channels. As these platforms become more widely used, they will become increasingly valuable targets for bad actors. In order to mitigate the potential for malicious misuse of these platforms, policymakers should ensure that laws that address violent, defamatory, or otherwise illegal online content sufficiently cover the potential harms from immersive experiences. Further, law enforcement bodies should start working with platforms now

to develop clear definitions of unlawful content or conduct and establish channels to improve coordination on efforts to remove the most dangerous forms of content.

Strengthen Protections Against Potential Real-World Harms

One content policy challenge that MUIEs will almost certainly inherit from their two-dimensional counterparts is the potential for virtual actions to lead to real-world harms. Policymakers should ensure that there are legal measures in place to protect individuals from physical, emotional, and reputational harms that could arise from malicious misuse of digital communications platforms, regardless of the format.

There is a serious risk to personal autonomy with a technology that makes it relatively simple to impersonate another individual (e.g., using their likeness to create an avatar) or falsify recorded media (e.g., manipulating virtual environments or modifying physical surroundings without a user's knowledge). Platforms will almost certainly have terms of use that prohibit this kind of activity and may also implement technical controls to better detect and respond to violations of these terms. But because of the content moderation challenges discussed in this report, such as context-based moderation and limits on technical tools, some instances of abuse might still make it through.

Policymakers should ensure that there are sufficient protections in place for victims of fraud or identity theft who may suffer reputational, emotional, or economic harms. Further, laws protecting victims of harmful online material, such as CSAM and non-consensual pornography, should address the ways in which this content might be created and distributed on immersive platforms. For example, VR pornography is already available; it would be entirely possible to not only share sensitive or explicit imagery without consent, but also to use non-consensual images or videos to create immersive pornographic experiences.[79]

Create Channels for Platforms and Law Enforcement to Work Together to Tackle Highly Dangerous Content

As with two-dimensional platforms, whether certain content is acceptable will vary among MUIEs based on their intended use, user base, and available user controls. However, there is a role for policymakers to support platforms in identifying and responding to the most egregious instances of harmful content and abuse. This includes CSAM, illicit activity and transactions, malicious mis- and disinformation, organized violence, and terrorist content.

With input from stakeholders in civil society and the private sector, federal law enforcement and intelligence agencies, such as the FBI, DHS, and ODNI, should provide guidance to help immersive platforms create definitions for this kind of activity across content types (i.e., environments, objects, and actions). These definitions would enable MUIE providers to better identify this kind of content on their platforms and help law enforcement understand how threats might appear in this new medium. This would also create a valuable foundation for future efforts to combat online threats in immersive experiences.

Further, in order to coordinate effective responses to this kind of content, these agencies should establish a forum that brings together immersive platforms—as well as their 2D counterparts—and government actors to share information about dangerous or illicit activity and other potential threats to real-world safety. There is already precedence for this kind of collaboration, such as the coordinated effort to track disinformation campaigns in the lead-up to the 2020 election.[80]

This kind of formal coordination will allow platforms to more effectively identify and refer credible threats of violence, organized disinformation, and other unlawful activity to the appropriate enforcement agencies. In addition, this can enhance information-sharing across jurisdictions at the state and federal level, which will be increasingly important as immersive experiences reduce barriers of physical distance.

It is important to recognize that this kind of content will likely constitute a small fraction of content or users. Therefore, such an initiative should mitigate online threats on MUIEs while ensuring the speech rights of the majority of users who are not engaging in dangerous or unlawful activities. There should be clear guidelines regarding the information that platforms disclose: platforms should not be required to share any private information beyond their legal obligations, and the scope of additional information that agencies request should be limited and disclosure fully voluntary.

Recommendations to Empower Providers to Develop Innovative Solutions

In order to preserve free speech, policymakers should clearly and narrowly define specific forms of illegal content. Beyond that, they should ensure the regulatory environment for MUIEs allows developers to take an iterative approach to content moderation as their user bases and use cases expand. Policymakers can enable platforms to develop policies and practices that encourage online speech while also protecting user safety by supporting laws that encourage good-faith content moderation and providing guidance to address the unique considerations immersive experiences raise for both creative expression and user safety.

Develop Voluntary Guidelines for Addressing Harmful Content in Immersive Experiences

Protecting users and enabling speech in immersive experiences will require significant forethought and resources. But established industry standards can help platforms translate existing best practices from two-dimensional media into three-dimensional experiences as well as develop new approaches that address the unique challenges their immersive services present. The Department of Commerce should work with platforms and other stakeholders to develop a set of voluntary guidelines for identifying, responding to, and reporting on harmful content in immersive experiences.

Key focus areas for such a framework should include:

- Definitions of how certain types of unwanted-but-not-illegal content, such as hate speech and harassment, present in immersive spaces
- Best practices for human moderation of immersive content, including conduct guidelines for in-world moderators and considerations for other forms of real-time monitoring
- Recommended user controls that can help platforms balance safety with privacy and free speech, such as blocking, muting, and filtering capabilities
- Guidelines on when and how to use different content and conduct moderation approaches, such as restricting content sharing or suspending users
- Guidelines for content moderation and monitoring in different use contexts, including how to adapt content policies to new and evolving uses for platforms

As more individuals use MUIEs in their social and professional lives, new challenges as well as best practices will likely emerge. Therefore, any voluntary framework should be iterative with clear processes to update it as new concerns from policymakers, platforms, users, and experts in key fields such as online safety and free speech emerge.

Create a Working Group to Develop Guidance on Intellectual Property and Copyright Protections

As discussed in this report, MUIEs present new considerations for intellectual property laws that dictate the parameters of content ownership and permissible use. In order to properly identify and address violating immersive content, MUIE providers will need a clear legal framework to work from. Congress should establish a working group in cooperation with the U.S. Copyright and Patent and Trademark Offices to develop guidance on key areas of intellectual property and copyright protections, such as:

- Ownership of digital versions of physical goods, both within and across different immersive platforms (e.g., what constitutes a "virtual knockoff" and what are the legal repercussions for these?)
- Ownership of virtual objects or environments created, shared, and sold by individual users or other third parties (e.g., who is allowed to replicate—and potentially profit from—digitally rendered materials?)
- Parameters of fair use of non-immersive trademarked items in immersive spaces (e.g., is reproducing a real-world object or figure in a cartoonish environment fair use?)
- Rights over digital layers on top of physical spaces in AR (e.g., who should be permitted to overlay virtual elements on privately owned physical spaces, images, or media displayed in the "real world"?)

In addition to the questions above, it is important to recognize that many MUIEs are driven by intra-platform economic exchange, such as individuals or companies selling virtual objects that users can interact with. And AR/VR enthusiasts hope that one day, users will be able to transfer virtual assets across platforms.[81] This raises important questions about cross-border exchange of digital goods. For example, when—if ever—should virtual exchanges be treated as imports or exports? Using the initial set of guidelines as a baseline, this working group should also consult with trade authorities including the USTR to address these questions related to international trade and offer recommendations as to whether and how immersive digital goods should be included in trade agreements.

This working group should prioritize building a framework that promotes innovation, fair compensation, and creative expression in this growing immersive digital economy. Members should solicit feedback from platforms, individuals and companies who are creating virtual items for immersive experiences, industry actors whose businesses depend on robust intellectual property protections, and civil society experts.

Ensure Intermediary Liability Protections Extend to Immersive Platforms

As online communications become more immersive and experiential, intermediary liability protections for service providers will be critical. MUIE providers should have the flexibility to try new approaches to content moderation in this new medium, and intermediary liability

protections—namely those provided under Section 230 of the CDA—provide the necessary legal shield to encourage this kind of experimentation. As policymakers consider proposals to amend Section 230 or otherwise alter the regulatory landscape for Internet intermediaries, they should include not only existing communications platforms, but also new mediums that are emerging such as MUIE, in their deliberations. For example, if new laws encourage over-moderation, this could have outsized impacts on users in immersive experiences where platforms can moderate not only the content they post, but also real-time gestures and interactions with other users or virtual objects. In order to encourage online speech in this new medium, Congress should not make significant changes to Section 230, and should instead focus on strengthening federal criminal law to address harms from malicious misuse of platforms operating in good faith—and punish immersive platforms that expressly engage in unlawful activity.[82]

CONCLUSION

As digital platforms have become increasingly valuable tools for interpersonal communication and free expression, they have also revealed just how challenging it can be to balance overarching goals of promoting online speech with protecting user safety and privacy. As MUIEs gain more widespread adoption, they will inherit many of these challenges—and present new considerations as users, developers, and policymakers explore what a more immersive future might look like. As these efforts continue, self-regulation and industry standards will be necessary to establish best practices for content policies in these immersive environments. But policymakers should also look to the lessons learned from existing platforms to establish necessary safeguards and ensure the companies developing these technologies face a regulatory environment that encourages—rather than dissuades—innovative approaches to content monitoring and moderation that balances user speech, safety, and privacy in immersive communications.

Acknowledgements

The author would like to thank Ellysse Dick for her assistance with this report. Any errors or omissions are the author's responsibility alone.

About the Author

Daniel Castro (@castrotech) is vice president at ITIF and director of its Center for Data Innovation. He writes and speaks on a variety of issues related to information technology and Internet policy, including privacy, security, intellectual property, Internet governance, e-government, and accessibility for people with disabilities.

About ITIF

The Information Technology and Innovation Foundation (ITIF) is an independent, nonprofit, nonpartisan research and educational institute focusing on the intersection of technological innovation and public policy. Recognized by its peers in the think tank community as the global center of excellence for science and technology policy, ITIF's mission is to formulate and promote policy solutions that accelerate innovation and boost productivity to spur growth, opportunity, and progress.

For more information, visit itif.org.

ENDNOTES

1. Samuel Finley Breese Morse, October 1844, digital manuscript accessed from Library of Congress, http://hdl.loc.gov/loc.mss/mmorse.018001.
2. Jason Loviglio, *Radio's Intimate Public: Network Broadcasting and Mass-mediated Democracy* (University of Minnesota Press, 2005), p. xix.
3. Taylor Lorenz and Davey Alba, "'Zoombombing' Becomes a Dangerous Organized Effort," *The New York Times,* April 3, 2020, https://www.nytimes.com/2020/04/03/technology/zoom-harassment-abuse-racism-fbi-warning.html.
4. Lindsay Blackwell et. al., "Harassment in Social VR: Implications for Design," *2019 IEEE conference on Virtual Reality and 3D User Interfaces (VR)* (IEEE: 2019), https://doi.org/10.1109/VR.2019.8798165.
5. Jeremy Bailenson, *Experience on Demand: What Virtual Reality Is, How it Works, and What it Can Do* (New York: W.W. Norton and Company, 2018).
6. Ellysse Dick, "How to Address the Privacy Questions Raised by the Expansion of Augmented Reality in Public Spaces," Information Technology and Innovation Foundation, December 14, 2020, https://itif.org/publications/2020/12/14/how-address-privacy-questions-raised-expansion-augmented-reality-public.
7. Qiaoxi Liu and Anthony Steed, "Social Virtual Reality Platform Comparison and Evaluation Using a Guided Group Walking Method," *Frontiers in Virtual Reality* 2 (2021), https://doi.org/10.3389/frvir.2021.668181
8. Mark Roman Miller et. al., "Social Interaction in Augmented Reality," *PLoS ONE* 14, no. 5 (2019), https://doi.org/10.1371/journal.pone.0216290.
9. Jeremy Horwitz, "Enterprise AR Will Follow These 3 Paths in 2021," *VentureBeat,* December 9 2020, https://venturebeat.com/2020/12/09/enterprise-ar-will-follow-these-3-paths-in-2021.
10. Raph Koster, "Still Logged In: What AR and VR Can Learn from MMOS, video recording, accessed December 10, 2021, https://www.gdcvault.com/play/1024060/Still-Logged-In-What-AR.
11. Gergana Mileva, "How VR is Changing the Music Industry," *ARPost,* January 23, 2019, https://arpost.co/2019/01/23/vr-changing-music-industry.
12. Kate Klonick, "why the History of Content Moderation Matters," *TechDirt,* January 30, 2018, https://www.techdirt.com/articles/20180129/21074939116/why-history-content-moderation-matters.shtml.
13. Michael McLaughlin and Daniel Castro, "The Case for a Mostly Open Internet," Information Technology and Innovation Foundation, December 16, 2019, https://itif.org/publications/2019/12/16/case-mostly-open-internet.
14. Elisa D'Amico and Luke Steinberger, "Fighting for Online Privacy with Digital Weaponry: Combating Revenge Pornography," *NYSBA Entertainment, Arts and Sports Law Journal* 26, no. 2 (2015): 24-36.
15. Evelyn Douek, "Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation," Aegis: Security Policy In Depth, September 18, 2019, https://www.lawfareblog.com/verified-accountability-self-regulation-content-moderation-answer-special-problems-speech-0.
16. Ashley Johnson and Daniel Castro, "Overview of Section 230: What It Is, Why It Was Created, and What It Has Achieved," Information Technology and Innovation Foundation, February 22, 2021, https://itif.org/publications/2021/02/22/overview-section-230-what-it-why-it-was-created-and-what-it-has-achieved.
17. Ashley Johnson and Daniel Castro, "The Exceptions to Section 230: How Have the Courts Interpreted Section 230?" Information Technology and Innovation Foundation, February 22, 2021,

https://itif.org/publications/2021/02/22/exceptions-section-230-how-have-courts-interpreted-section-230.

18. Not all exceptions to intermediary liability protections are created equal. For example, advocates have raised concerns that SESTA/FOSTA unnecessarily restricts online speech and endangers the very individuals it was meant to protect by encouraging over-moderation of content that could potentially put platforms at risk of liability and forcing sex workers off of platforms that provided them with valuable information as well as a level of personal safety and security that would not otherwise be available.
19. 17 U.S.C. §â€¯512(c)(1)(A) (1998).
20. Cyber Civil Rights Initiative, "Related Laws," accessed December 10, 2021, https://www.cybercivilrights.org/related-laws.
21. Cyber Civil Rights Initiative, "48 States + DC + One Territory Now Have Revenge Porn Laws," accessed December 10, 2021, https://www.cybercivilrights.org/revenge-porn-laws.
22. Shannon Vavra, "Deepfake Laws Emerge as Harassment, Security Threats Come Into Focus," *CyberScoop,* January 11, 2021, https://www.cyberscoop.com/deepfake-porn-laws-election-disinformation.
23. Ashley Johnson and Daniel Castro, "How Other Countries Have Dealt with Intermediary Liability," Information Technology and Innovation Foundation, February 22, 2021, https://itif.org/publications/2021/02/22/how-other-countries-have-dealt-intermediary-liability.
24. United Nations Human Rights Council, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, A/HRC/38/35 (April 6, 2018), https://www.undocs.org/A/HRC/38/35.
25. Evelyn Douek, "More Content Moderation is Not Always Better," *WIRED,* June 2, 2021, https://www.wired.com/story/more-content-moderation-not-always-better.
26. Network Enforcement Act (Netzwerkdurchsetzungsgesetz, NetzDG) (Federal Republic of Germany, 2017).
27. UN A/HRC/38/35
28. Spectrum Labs, "What is Content Moderation?" Accessed December 10, 2021, https://www.spectrumlabsai.com/content-moderation.
29. Evelyn Douek, "Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability," *Columbia Law Review* 121, no. 3 (2021), https://columbialawreview.org/content/governing-online-speech-from-posts-as-trumps-to-proportionality-and-probability.
30. Spandana Singh, "The Limitations of Automated Tools in Content Moderation," in *Everything in Moderation: An Analysis of How Internet Platforms are Using Artificial Intelligence to Moderate User-Generated Content,* New America, July 22, 2019, https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-moderation.
31. Renee DiResta and Tobias Rose-Stockwell, "How to Stop Misinformation Before It Gets Shared," *WIRED,* March 26, 2021, https://www.wired.com/story/how-to-stop-misinformation-before-it-gets-shared.
32. Duncan Steward et. al., "From Virtual to Reality: Digital Reality Headsets in Enterprise and Education," Deloitte, December 7, 2020, https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2021/vr-immersive-technologies.html.
33. Divine Maloney et. al., "Social Virtual Reality: Ethical Considerations and Future Directions for an Emerging Research Space," *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW),* 2021, https://doi.org/10.1109/VRW52623.2021.00056.

34. Nick Yee and Jeremy Bailenson, "The Proteus Effect: The Effect of Transformed Self-Representation on Behavior," *Human Communications Research* 33, no. 3 (2007), https://doi.org/10.1111/j.1468-2958.2007.00299.x.
35. Devon Adams et. al., "Ethics Emerging: The Story of Privacy and Security Perceptions in Virtual Reality," Proceedings of the Fourteenth Symposium on Usable Privacy and Security, August 2018, https://www.usenix.org/system/files/conference/soups2018/soups2018-adams.pdf.
36. Guo Freeman and Divine Maloney, "Body, Avatar, and Me: The Presentation and Perception of Self in Virtual Reality," *Proceedings of the ACM on Human-Computer Interaction* 4, no. 3 (2020), https://doi.org/10.1145/3432938.
37. Renee Gittins, "Social Virtual Reality Best Practices," International Game Developers Association (IGDA), July 30, 2018, https://igda.org/resources-archive/social-virtual-reality-best-practices-2018.
38. Ellysse and Ashley Break The Internet, "How Section 230 Promotes Competition, with Jessica Ashooh," podcast recording, Information Technology and Innovation Foundation, accessed December 10, 2021, https://itif.org/podcast-how-section-230-promotes-competition-jessica-ashooh.
39. "AltspaceVR," accessed December 10, 2021, https://altvr.com.
40. At the time of this writing, the Mac OS version of AltSpace VR was in beta. See: AltspaceVR, "No Headset? No Problem," accessed December 10, 2021, https://altvr.com/get-altspacevr/#desktop-mode.
41. GitHub user vryunji and Harrison Ferrone, "AltspaceVR Community Standards," Microsoft Technical Documentation, April 5, 2021, https://docs.microsoft.com/en-us/windows/mixed-reality/altspace-vr/community/community-standards.
42. Qian Wen, "Frequently Asked Questions about Accounts and Avatars," Microsoft Technical Documentation, September 30, 2021, https://docs.microsoft.com/en-us/windows/mixed-reality/altspace-vr/faqs/account-avatar-faq; AltspaceVR, "Introducing Safe Bubble," July 13, 2016, https://altvr.com/introducing-space-bubble.
43. "Rec Room," accessed December 10, 2021, https://recroom.com.
44. Rec Room, "Comfort and Safety," accessed December 10, 2021, https://recroom.com/comfortandsafety.
45. Rec Room, "Junior Accounts," Rec Room Support Center, June 23, 2021, https://recroom.happyfox.com/kb/article/19-junior-accounts.
46. "VRChat," accessed December 10, 2021, https://hello.vrchat.com.
47. VRChat, "Community Guidelines," accessed December 10, 2021, https://hello.vrchat.com/community-guidelines; Anne Hobson, "Phantoms, Crashers, and Harassers: Emergent Governance of Social Spaces in Virtual Reality," Center for Growth and Opportunity, July 2020, https://www.thecgo.org/wp-content/uploads/2020/09/Phantoms-Crashers-and-Harassers-Emergent-Governance-of-Social-Spaces-in-Virtual-Reality.pdf.
48. Meta, "Introducing Horizon Workrooms: Remote Collaboration Reimagined," Meta Newsroom, August 19, 2021, https://about.fb.com/news/2021/08/introducing-horizon-workrooms-remote-collaboration-reimagined.
49. Meta Quest, "Conduct in VR Policy," accessed December 10, 2021, https://support.oculus.com/articles/accounts/privacy-information-and-settings/conduct-in-vr-policy.
50. "Spatial," accessed December 10, 2021, https://spatial.io.
51. Spatial, "Create Your Avatar: Build Your Metaverse Presence," accessed December 10, 2021, https://spatial.io/create-an-avatar.
52. Spatial, "Download Spatial," accessed December 10, 2021, https://spatial.io/download.
53. Brianna Scully, "Spatial Community Guidelines," Spatial Support, August 11, 2021, https://support.spatial.io/hc/en-us/community/posts/4405055325076-Spatial-Community-Guidelines

54. Spatial, "Host Tools in Spatial," Spatial Support, last updated December 4, 2021, https://support.spatial.io/hc/en-us/articles/360057390011-Host-Tools-in-Spatial-.

55. Ellysse Dick, "Current and Potential Uses of AR/VR for Equity and Inclusion," Information Technology and Innovation Foundation, June 1, 2021, https://itif.org/publications/2021/06/01/current-and-potential-uses-arvr-equity-and-inclusion.

56. Anti-Defamation League (ADL), Hate in Social VR: New Challenges Ahead for the Next Generation of Social Media," accessed December 10, 2021, https://www.adl.org/resources/reports/hate-in-social-virtual-reality.

57. Bailenson, *Experience on Demand*

58. Julia Alexander, "'Ugandan Knuckles' is Overtaking VRChat," *Polygon,* January 8, 2018, https://www.polygon.com/2018/1/8/16863932/ugandan-knuckles-meme-vrchat.

59. Jessica Outlaw, "Social VR Bystander Intervention—and an Invitation to a Free Training," Medium post, June 28, 2018, https://jessica-outlaw.medium.com/social-vr-bystander-intervention-and-an-invitation-a-free-training-5bf38e5616c0.

60. Queenie Wong, "Facebook's and Social Media's fight Against Fake News May Get Tougher," *Cnet,* December 27, 2018, https://www.cnet.com/tech/services-and-software/facebooks-and-social-media-content-may-become-harder-to-police-in-the-future.

61. Joseph Pisani, "Nike Files to Sell Digital Sneakers, As It Seeks Downloadable Kicks," *The Wall Street Journal,* November 2, 2021, https://www.wsj.com/articles/nike-files-to-sell-digital-sneakers-as-it-seeks-downloadable-kicks-11635873070.

62. Ryan Calo, Statement before the U.S. Senate Committee on Commerce, Science, and Transportation hearing on "Exploring Augmented Reality," November 16, 2016, https://www.commerce.senate.gov/services/files/D8EFA7CA-4FCC-4196-BB5D-E64FE937D01F.

63. Crystal Nwaneri, "Ready Lawyer One: Legal Issues in the Innovation of Virtual Reality," *Harvard Journal of Law and Technology* 30, no. 2 (2017), https://jolt.law.harvard.edu/assets/articlePDFs/v30/30HarvJLTech601.pdf

64. Kimberly Ruth, "Opportunities and Pitfalls for Multi-User AR Experiences," *ARPost,* July 15, 2020, https://arpost.co/2020/07/15/opportunities-pitfalls-multi-user-ar.

65. Hobson, "Phantoms, Crashers, and Harassers."

66. The Fortnite Team, "Your First Drop Into Party Royale: Getting to the Main Stage," Epic Games, September 21, 2021, https://www.epicgames.com/fortnite/en-US/news/your-first-drop-into-party-royale-getting-to-the-main-stage.

67. Patrick Klepek, "'Fortnight' Streams Panel on Race, While Some Players Throw Tomatoes," *VICE,* July 8, 2020, https://www.vice.com/en/article/dyzdkq/fortnite-streams-panel-on-race-while-some-players-throw-tomatoes.

68. Jessica Outlaw and Beth Duckles, "Why Women Don't Like Social Virtual Reality: A Study of Safety, Useability, and Self-Expression in Social VR," The Extended Mind, 2017, https://www.extendedmind.io/why-women-dont-like-social-virtual-reality.

69. Ibid.

70. Franziska Roesner et. al., "Security and Privacy for Augmented Reality Systems," *Communications of the ACM* 57, no. 4 (2014), https://doi.org/10.1145/2580723.2580730.

71. Similar forms of misuse have been reported with existing multi-player, location-based entertainment. For example, when the location-based mobile game Pokémon Go first launched in 2016, there were reports of criminals targeting "Pokéstops," or geographic locations where players would gather. See for example: Alan Yuhas, "Pokemon Go: Armed Robbers Use Mobile Game to Lure Players Into Trap," *The Guardian,* July 11, 2016, https://www.theguardian.com/technology/2016/jul/10/pokemon-go-armed-robbers-dead-body.

72. Ruth, "Opportunities and Pitfalls for Multi-User AR Experiences."

73. Alexander Lee, "Why Metaverse Builders Want to Create Safe, Consensual Worlds," *Digiday,* September 2, 2021, https://digiday.com/marketing/metaverse-builders-want-to-create-safe-consensual-virtual-worlds.
74. Hobson, "Phantoms, Crashers, and Harassers.
75. Lee, "Why Metaverse Builders Want to Create Safe, Consensual Worlds."
76. Ben Lang, "In 'Horizon' Facebook Can Invisibly Observe Users in Real-Time to Spot Rule Violations,"
77. Meta Quest, "Conduct in VR Policy."
78. Gittins, "Social Virtual Reality Best Practices."
79. Matt Wood, "Virtual Reality Could Transform Pornography—But There Are Dangers," *The Conversation,* May 22, 2017, https://theconversation.com/virtual-reality-could-transform-pornography-but-there-are-dangers-78061.
80. Mike Isaac and Kate Conger, "Google, Facebook and Others Broaden Group to Secure U.S. Election," *The New York Times,* August 12, 2020, https://www.nytimes.com/2020/08/12/technology/google-facebook-coalition-us-election.html.
81. Theo Priestly, "Metanomics: Building the Economy of the Metaverse," *AR Insider*, August 25, 2021, https://arinsider.co/2021/08/25/how-will-the-metaverse-economy-materialize.
82. Ashley Johnson and Daniel Castro, "Proposals to Reform Section 230," Information Technology and Innovation Foundation, February 22, 2021, https://itif.org/publications/2021/02/22/proposals-reform-section-230.