laws

MDPI

Article

AI Moderation and Legal Frameworks in Child-Centric Social Media: A Case Study of Roblox

Mohamed Chawki

Law Department, Naif Arab University for Security Sciences, Riyadh 11452, Saudi Arabia; mshaker@nauss.edu.sa

Abstract: This study focuses on Roblox as a case study to explore the legal and technical challenges of content moderation on child-focused social media platforms. As a leading Metaverse platform with millions of young users, Roblox provides immersive and interactive virtual experiences but also introduces significant risks, including exposure to inappropriate content, cyberbullying, and predatory behavior. The research examines the shortcomings of current automated and human moderation systems, highlighting the difficulties of managing real-time user interactions and the sheer volume of user-generated content. It investigates cases of moderation failures on Roblox, exposing gaps in existing safeguards and raising concerns about user safety. The study also explores the balance between leveraging artificial intelligence (AI) for efficient content moderation and incorporating human oversight to ensure nuanced decision-making. Comparative analysis of moderation practices on platforms like TikTok and YouTube provides additional insights to inform improvements in Roblox's approach. From a legal standpoint, the study critically assesses regulatory frameworks such as the *GDPR*, *the EU Digital Services Act*, *and the UK's Online Safety Act*, analyzing their relevance to virtual platforms like Roblox. It emphasizes the pressing need for comprehensive international cooperation to address jurisdictional challenges and establish robust legal standards for the Metaverse. The study concludes with recommendations for improved moderation strategies, including hybrid AI-human models, stricter content verification processes, and tools to empower users. It also calls for legal reforms to redefine virtual harm and enhance regulatory mechanisms. This research aims to advance safe and respectful interactions in digital environments, stressing the shared responsibility of platforms, policymakers, and users in tackling these emerging challenges.

Keywords: content moderation; Roblox; artificial intelligence; legal frameworks; online safety

check for updates

Academic Editor: Casey Watters

Received: 8 March 2025
Revised: 16 April 2025
Accepted: 22 April 2025
Published: 25 April 2025

Citation: Chawki, Mohamed. 2025. AI Moderation and Legal Frameworks in Child-Centric Social Media: A Case Study of Roblox. *Laws* 14: 29. https://doi.org/10.3390/laws14030029

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Context and Relevance of the Study

The rise of child-friendly digital platforms, particularly in the Metaverse, has raised significant concerns regarding content moderation, online safety, and legal accountability.

Roblox is one of the largest platforms in the growing Metaverse space. Created and operated by Roblox Corporation, an American publicly held company listed on the New York Stock Exchange (NYSE: RBLX), the platform is an in-world, user-created virtual world-based ecosystem in which players can create, publish, and socialize within 3D games and experiences. Originally founded in 2004 by David Baszucki and Erik Cassel, Roblox launched in 2006 and grew into an internet-based, social-gaming platform operating largely among kids and teenagers (Roblox Corporation 2024).

In contrast to other video games produced by studios, Roblox is both an in-game platform and game engine that enables its users—people who are largely not professional programmers—to build rich experiences within its Roblox Studio development software. Such created games are published in the platform and available for play to any user, gratis, typically funded by using in-game currency called Robux, which can be bought using real-life dollars (Baszucki 2021).

Roblox is among the largest virtual platforms, with over 380 million active users monthly (JetLearn 2024). It is primarily used by children and adolescents and is, therefore, a primary case study in examining AI-based content moderation, regulatory challenges, and emerging threats (Zhang et al. 2024).

Roblox primary legal jurisdiction is in the United States, and its operations—particularly content moderation and user accountability—are regulated by U.S. federal statutes, such as the Communications Decency Act (CDA §230) and Children's Online Privacy Protection Act (COPPA) (Roblox Corporation 2025). Yet because Roblox has millions of customers all over the world—large user communities in the European Union, United Kingdom, Brazil, and Southeast Asia—it also needs to adhere to foreign regulatory environments, such as the EU's General Data Protection Regulation (GDPR) and Digital Services Act (DSA). These overlapping jurisdictions cause regulatory fragmentation, particularly in managing concerns such as child data protection, moderation transparency, and virtual harm liability.

Roblox's moderation system is based on a hybrid of AI and human review. Its AI layer leverages natural language processing (NLP), image detection, and machine learning algorithms to scan for unwanted content in real time—such as bad chat messages, offending avatars, or policy-breaking user-created games. AI-based tools filter billions of interactions each day for specified risk indicators and bring those for human review (Roblox Corporation 2025). These automatic filters play an especially important role in chat moderation, in which real-time content scanning prevents hate speech, grooming, and adult content from spreading. But automatic systems by themselves are not enough. Roblox uses human moderators from all over the world to scrutinize flagged content, address user reporting, and analyze contextual offenses that AI can overlook—like sarcasm, coded language, and cultural context. A two-part system is necessary, considering the size of the platform and its demographic sensitivity (the average user is under 18). Even with such protections, Roblox's content moderation is still contentious because of slowness, AI error in categorization, and variability in its enforcement across regions (Roblox Corporation 2025).

Roblox's hybrid position, as part game creation platform and part social networking site, poses unique challenges to online safety, regulation of content, and legal responsibility. Children are participating in rich, avatar-mediated interactions at once entertaining, educative, and social. This intersection intensifies concerns about understanding moderation in this platform, especially in consideration of its susceptible user community and increasing impact in youth digital culture (King Law 2025).

Despite its popularity as a creative and interactive platform, Roblox has come under growing criticism for inadequate moderation systems that expose children to harmful material, cyberbullying, predation, and online exploitation (Mancuso et al. 2024). The sheer volume of user-generated material—over 60 billion messages a day—carries specific challenges regarding real-time moderation and regulatory compliance (Kumar 2024). This research examines both legal and technical dimensions of AI moderation and critically assesses existing safeguards and regulatory loopholes in virtual space.

With recent legislation such as the EU Digital Services Act (DSA 2022) and the UK Online Safety Act (2023), platforms such as Roblox are under mounting pressure

to live up to their responsibility to make their online community a safe space for children (Gray et al. 2024; Mujar et al. 2024). Current legal frameworks are not yet able to deal with the rapidly evolving Metaverse and pose fundamental legal and ethical questions on platform liability, algorithmic regulation, and virtual harm (Hinduja and Patchin 2024).

Existing legal frameworks—e.g., the EU's General Data Protection Regulation (GDPR), Digital Services Act (DSA), and U.S. Communications Decency Act (CDA §230)—are growing out of sync with the realities of new Metaverse environments such as Roblox. While these frameworks provide valuable principles surrounding privacy, platform responsibility, and content moderation, they were created for a Web 2.0 paradigm in which platforms host and retransmit static user-generated content. They are poorly suited to regulate the immersive, dynamic, co-creative, and highly social nature of virtual worlds such as Roblox, in which harm can be caused not just by speech, but also by behavioral and experiential design.

Section 230 of CDA grants U.S.-based platforms broad protection against liability for third-party speech. Under this approach, Roblox is largely insulated from legal liability for user-created harmful games or chats. Yet, contrary to platforms such as YouTube or Facebook, Roblox is not only a host for user content—including games, we should note—it offers the assets, game engines, and monetization platforms necessary for user-generated content. In this respect, Roblox performs less as an intermediary host and more akin to a co-developer or facilitator of virtual environments. As legal scholar Kyle Langvardt points out, "interactive virtual spaces challenge the safe harbor structure of Section 230 because platform design is not neutral when it influences conduct" (Langvardt 2020). Modern legal dogma does not draw a distinction among platforms hosting speech and platforms enabling social architectures in which users can create virtual nightclubs, casinos, or adult games accessible to minors (ibid.).

Similarly, data protection legislation such as the GDPR does not provide for the complexity of immersive digital identity. Under Article 8, the GDPR demands verifiable parental consent for data collection from individuals aged under 13–16, depending upon the member state. Roblox seems to satisfy this obligation, yet enforcement proves tricky in pseudonymous, avatar-based environments due to difficulties in age verification. Moreover, the GDPR stipulates that "data" is something passively gathered from user activity—names, emails, IP addresses. In the Metaverse, however, data are experiential in nature: choice of avatar, in-game conduct, in-game friendships, and in-game financial dealings constitute all different types of data used to determine who they become targeted toward, who they get ranked against, and who they receive special treatments from by platform algorithms. As Livingstone and Third (2017) have suggested, existing data protection frameworks can do little better than support "the lived data practices of children in play-oriented and hybrid environments". While Article 22 of the GDPR provides rights against automatic decision-making, few underage Roblox users—or their parents—are even aware AI moderation technology or recommendation platforms are influencing what they see online.

The Digital Services Act (DSA) takes us closer to grappling with these issues, particularly for platforms qualifying as Very Large Online Platforms (VLOPs). The DSA establishes systemic requirements such as transparency in algorithms (Article 42), independent examinations (Article 37), and child protection measures (Article 28). Yet, the DSA still defines harm primarily in terms of illegal content (e.g., hate speech, misinformation), rather than emergent harms specific to immersive platforms, such as in-game grooming, simulated sexual conduct, coercive peer play, or psychological manipulation through avatars. Roblox's blend of synchronous play, sandbox environments, and peer-to-peer play poses an enormously greater challenge. Legal scholar Mireille Hildebrandt has contended that regulation of algorithms needs to break free of checklists and confront "the performativity

of code"—i.e., how code in digital environments constructs user behavior and identity (Hildebrandt 2018). The DSA takes its first step toward grappling with this issue but does not include enforceable standards for experiential harm that resists categorization as speech or illegal content.

1.2. Research Questions and Objectives

The research seeks to address the following research questions:

- To what extent is artificial intelligence-based content moderation effective in protecting child users on platforms like Roblox?
- How do legal and ethical frameworks address the use of algorithmic systems to moderate harmful content and protect children in virtual environments such as Roblox?
- How does Roblox currently address virtual harm to children on its platform, and what regulations underpin this?
- How should virtual harm to children be conceptualized in law, and what regulatory mechanisms are needed to ensure that immersive platforms like Roblox are held accountable for failing to protect young users?
- Are avatars legal persons, and should liability be assigned in digital abuse cases in a specific way under UK law?

To address these questions, this research:

- Employs the theory of algorithmic governance developed by Kitchin (2017) and Hildebrandt (2018, 2020), emphasizing the consequences of decisional systems and automation for accountability, bias, and due process in regulating the digital world. When these theories consider AI systems to be more than just technical tools, they are instead regulatory agents that impact the formation of legal subjectivity, procedural justice, and ethical responsibilities within online platforms like Roblox.
- Reflects on legal frameworks for virtual worlds and considers whether existing regulations—such as the GDPR and DSA—offer adequate protection against digital harm. The evaluation is grounded in the context of child protection in the UK, drawing on national interpretations and enforcement practices to assess how effectively these frameworks address risks on platforms like Roblox.
- Reflects on avatars' legal personhood and whether they should have independent legal identities and liability arrangements in instances of digital wrongdoing.

1.3. Methodological Approach

This research employs a multi-disciplinary methodology that combines the following:

- **Legal Analysis**—The study conducts doctrinal legal analysis regarding major regulatory tools like the GDPR, the Digital Services Act (DSA), and the UK Online Safety Act. It makes systematic interpretations of provisions in statutes by applying traditional legal interpretational techniques like textual, purposive, and contextual interpretation to evaluate how current law governs online content moderation, platform responsibility, and user safeguarding in online spaces like Roblox.
- **Case Study Approach**—Analysis of actual cases on Roblox, including failures in moderation, legal controversies, and difficulties in content management.
- **Comparative Analysis**—Analysis of content moderation on Roblox, TikTok, YouTube, and other platforms to identify best practices and regulatory loopholes.
- **Engagement with Algorithmic Governance Theory**—Drawing on the works of Rob Kitchin and Mireille Hildebrandt, this research critiques the deployment of AI as a regulatory tool and considers ethical implications in automated decision-making.

This research integrates legal, technical, and ethical considerations and offers a holistic framework for optimizing AI content moderation, platform accountability, and regulatory control in child-friendly digital spaces.

2. Contributions of the Study

The research makes an important contribution to the topics of digital safety, regulation, and content moderation on child-focused and social platforms like Roblox by discovering significant gaps in regulation in current systems of moderation—most importantly related to algorithmic oversight, procedural justice, and protection and safeguarding of children. A comparative study of concurrent legal instruments (such as the *GDPR*, *the DSA*, *and the UK Online Safety Act*) leads to this paper making a unique contribution by suggesting specific reforms in the form of child-friendly appeal systems, context-sensitive AI moderation, and jurisdictional compliant data management practices. In this way, the research forwards academic and policy debates regarding how to apply rights-oriented digital regulation to virtual worlds and spaces.

The contribution of the study is as follows:

2.1. Comprehensive Analysis of Moderation Systems

The study has given a comprehensive review of Roblox's mechanism of content moderation with strong emphasis on how the site detects, assesses, and works to address harm targeting children. It entails an investigation into the technical architecture behind its human-AI blended system of moderating contents, grooming and cyberbullying detection in real-time multi-player spaces, and procedural transparency and efficacy of appeal procedures among youth users.

The study has also identified critical gaps, such as the AI system's inability to handle nuanced content and the challenges of scaling human moderation on a large user base.

The research provided some important findings regarding the limits of Roblox's existing content moderation system—especially in its inability to effectively shield minors from harmful material, cyberbullying, and online grooming. It pointed out inadequacies like algorithmic prejudice, lack of context sensitivity among AI systems, narrow appeal mechanisms available to minors, and latency in human moderation actions. These findings provide an impetus to enhance moderation practices on interactive platforms through bettering child-specific safeguards, improving real-time coordination between AI and humans, and alignment with upcoming regulatory frameworks like the DSA and the UK Online Safety Act.

2.2. Evaluation of Legal Frameworks

The research evaluates existing legal instruments, including the GDPR, EU Digital Service Act, and UK Online Safety Act, in the context of the Roblox platform.

Additionally, the study goes on to uncover how Roblox's regulatory and technical infrastructures respond to chief cornerstones of content moderation, user accountability, and platform protection. In doing so, it points to the unique difficulties presented by Metaverse platforms—namely those with persistent, interactive, and shared real-time virtual spaces inhabited primarily by minors. Among these are the challenge of moderating fluid user-created 3D spaces, the risk of grooming and predation through real-time multiplayer interactions, the limitations to existing algorithmic tools in reading sophisticated social signals, and the lack of standardized legal norms applicable to virtual harms. The study points to how these novel characteristics put pressure on traditional models of moderation and require a reconsideration of accountability and protection in virtual worlds catering to children like Roblox.

2.3. Insights into Emerging Risks

The Roblox study's initiative of documenting real-world instances of user-inappropriate content, predatory behavior, and cyberbullying sheds light on various risks that young users face in the digital space. The study further explores the different implications of this risk for children's mental health and privacy and also emphasizes the increased need for proactive measures to be met.

2.4. Comparative Study of Moderation Practices

The study compares content moderation practices on platforms like TikTok and YouTube not just on technical architecture and human-AI collaboration, but against the background of their obligations under legislation like the GDPR and DSA too. The analysis is mindful not to overplay the comparison, however: whereas TikTok and YouTube mostly deal with recorded and static content, Roblox poses unique issues through its real-time interactive and user-generated 3D worlds. The comparison is therefore used to underscore structure differences in demands on and exposure to regulation rather than to claim functional equivalence amongst inherently different platform designs.

2.5. Recommendations for Policy and Practice

The research presents actionable recommendations for strengthening child-centric platform safety protocols and legal standards. These include improving the combination of human supervision with automated moderation, changing the meaning of virtual harm under legal standards, and building cooperation worldwide to overcome issues related to legal authority.

This study identifies that the legal definition of 'virtual harm' differs greatly from one jurisdiction to another—for example, the EU DSA deals with prohibited content, whereas the UK Online Safety Act deals with psychological harm in specific contexts, and there is no defined category under U.S. law. Therefore, we advocate treaty cooperation—most probably under an international convention on platform liability—and harmonized industry self-regulation to deal with the fragmentation and to adopt uniform global protection in interactive virtual spaces.

2.6. Contribution to Academic and Practical Discourse

This study bridges the gap between academic research and practical implementation by combining theoretical analysis with real-world case studies. This fits within the emerging discussion on Metaverse safety and content moderation and will benefit future research as a single platform for studying these topics.

By addressing the intersection of technology, law, and user safety, this study offers a holistic perspective on the challenges and opportunities in moderating child-centric platforms. It paves the way for building safer and more open digital spaces that will improve vulnerable users' experiences while allowing free creativity development.

3. Brief Overview of Roblox

As the Metaverse continues to evolve and expand, Roblox has emerged as one of the most prominent platforms within this digital realm (Wang et al. 2022; Zhang et al. 2024). It is recognized as a sustainable and interconnected 3D virtual environment (Mancuso et al. 2024). Over the past few years, Roblox has witnessed extraordinary growth. The platform's user base grew from 12 million in 2018 to 42.1 million in 2021 and now exceeds 88.9 million daily active users as of 2024, with a global monthly active audience of 380 million (Singh 2025).

Data analysis conducted by Park and Kang introduced a ranking system based on metrics such as usage frequency and time spent on applications. In this ranking, Roblox

advanced from the 47th position in January 2020 to 29th by August 2020. Recent figures reveal that Roblox continues to dominate Metaverse traffic, with a significant portion of its audience being children and young users aged 5 to 16. Of its daily active users, approximately 32.4 million are under age 13, making up around 36% of the total user base (Park and Kang 2022).

Over 55% of Generation Z users in the United States actively engage with Roblox, positioning it as a vital communication platform within the Metaverse. It also strongly appeals to Generation Alpha (born after 2012), who spend more time on Roblox than any other platform. On average, children dedicate 2.6 h per day to Roblox, three times the time spent on YouTube, and seven times the time spent on Facebook (Figure 1). This underscores Roblox's significant role as a hub for immersive virtual interactions among younger generations in a sandbox environment game (Dwivedi et al. 2022; Mancuso et al. 2024).

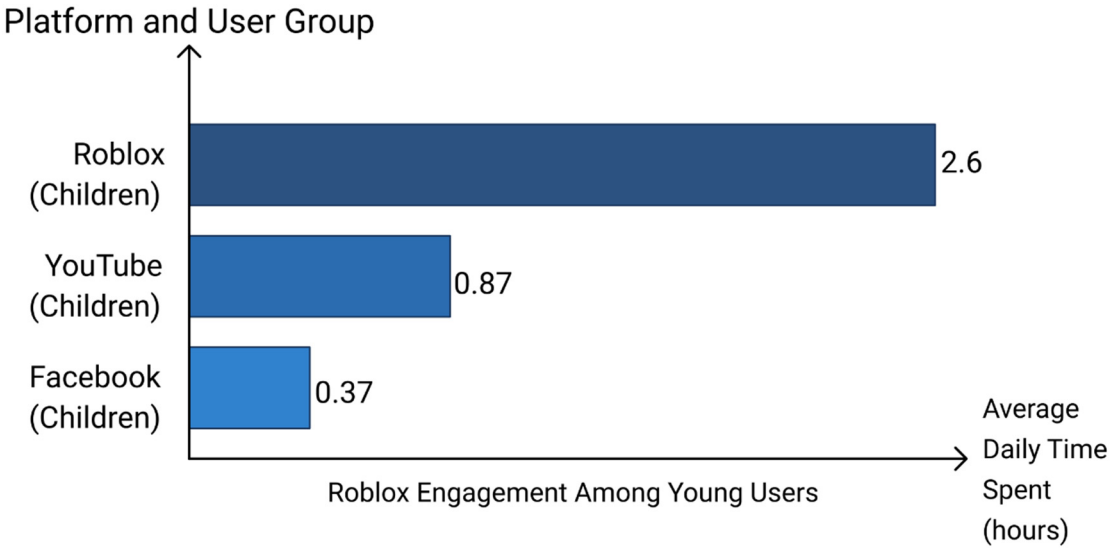


Figure 1. Comparative analysis of daily platform usage among children: average time spent on social media and gaming platforms [Credit: original figure created by the author].

Roblox is the biggest online game service with the most players who can create and play online sandbox games—open-ended, user-generated virtual worlds in which players can freely build, script, and interact in game worlds created by themselves or others. The company's service comprises virtual worlds, leisure communities, and self-built services. Roblox is a platform that enables users to build and explore their own virtual worlds; therefore, users can develop games and anything they desire in the world of Roblox. Roblox already has virtual worlds and games developed by hundreds of thousands of players. Users can go to Roblox to meet or game together, send nearly 60 billion messages daily in games created by other players, talk to people in the real world, buy, and create social networks in the 3D virtual environment. Roblox has nine key features of the Metaverse: an integrated network of 3D virtual worlds, identity, friends, immersion, accessibility, low friction, various content types, cost-effectiveness, and safety. These factors will help attract your audience and fuel contemporary creativity outputs (Dionisio et al. 2013; Lee et al. 2021).

4. Recent Incidents Highlighting the Risks in Roblox

The rapid growth and user engagement in Roblox have also brought attention to significant risks and challenges. Recent events in the Roblox sphere have raised several issues regarding safety, privacy, and moderation (Figure 2). Figure 2 outlines a layered strategy for digital platform safety, integrating technical innovations (algorithm upgrades, real-time monitoring), policy structures (authentication protocols, enforcement guidelines),

and content governance (automated filtering, user reporting systems). These components collectively aim to balance proactive risk mitigation with adaptive regulatory compliance, ensuring platform integrity and user protection. These concerns relate to trust, privacy, bias, disinformation, application of the law, and psychological aspects linked to addiction and its impact on vulnerable individuals (Dwivedi et al. 2022). This section presents an overview of such events to support the necessity of improving the level of security in the platform. Worth noting is that Roblox Corporation operates out of the United States, and thus, its operations—most importantly with respect to regulation and compliance—are mostly guided by U.S. federal law, encompassing the Children's Online Privacy Protection Act (COPPA) and Section 230 of the Communications Decency Act (CDA §230).

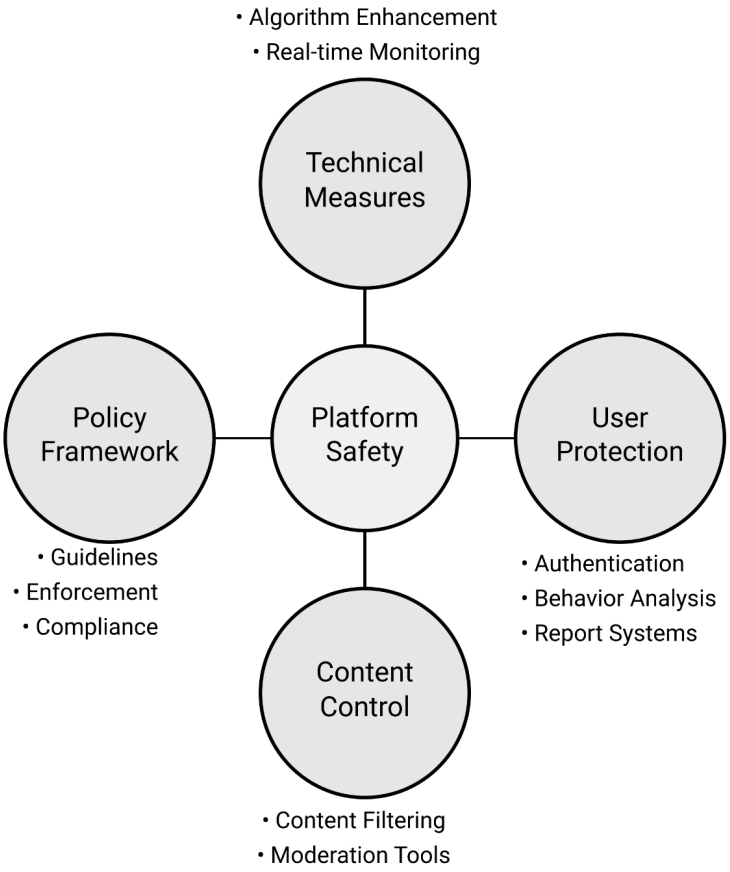


Figure 2. Platform safety framework: core components and interactions [Credit: Author, original figure created by the author].

4.1. Exposure to Inappropriate Content

Despite Roblox's moderation policies, there have been numerous reports of inappropriate content slipping through the cracks, revealing significant vulnerabilities in the platform's ability to protect its young users. The proposed usage of automatically generated filtering to include or exclude objectionable content and materials from Roblox has been found to pose certain limitations (BBC 2022). However, over the years, some of these shortcomings have come to the surface periodically, thus providing parents, guardians, and anybody concerned for children's welfare something to be worried about.

One particularly alarming incident occurred in 2021 when a game mimicking a sexually explicit experience was discovered on Roblox featuring a naked man wearing only a dog collar and a lead (Shen and Ma 2024). It caused much outrage and worry because children are subjected to material that is much beyond what is considered appropriate for their age. The game's existence also revealed a significant issue with how Roblox had organized its

auto-moderation mechanics: soon enough, the exact unpleasant content was uploaded into the game's channel and the visibility range of potential viewers. Such an event had quite a negative result: parents and guardians exerted more pressure on the platform, making it offer better protection measures and showed rather doubtful trust in the application protection (The Guardian 2024).

The 2021 incident was not an isolated case but part of a broader pattern of moderation failures (Wired 2021). Despite Roblox's ongoing efforts to improve its content filtering mechanisms, other instances of inappropriate content have continued to surface. These are games containing concepts, pictures, lewd content, and similar content unsuitable for minors. Every instance highlights the inherent challenges of using heuristically scoped algorithms to identify and manage billions of messages created on widely used social media interfaces.

Automated moderation systems, while essential for handling the massive content volume on a platform like Roblox, have inherent limitations stemming from user-generated virtual worlds (Kou and Gui 2023). Algorithms can be tricked or circumvented by savvy users who find ways to disguise inappropriate content, making it difficult for automated tools to catch every instance of violation. Further, most of the systems mentioned above do not possess the context awareness required to effectively evaluate the intent and significance behind particular content and thus are prone to generating false positives and failing to notice harm. For example, an algorithm may regard a practical joke as likely to damage society in some form. The algorithm will categorize the joke and other posts as 'offensive' and 'harmful' to society.

The need for more robust human oversight is evident. Human moderators can provide the contextual judgment that automated systems lack, making them essential for a comprehensive and effective moderation strategy. However, the scale of Roblox's audience and the tendency for new DAUs and content are challenges the workforce cannot control and address, mainly when creator information in personal profiles plays a significant role (Kang et al. 2024). This indicates a clear need to interface the candidate automated systems with specific predetermined options; however, the human element must also be considered, particularly about Metaverse technology or content creation for creators, which was insufficient (Kim and Rhee 2022).

Roblox has been developing its moderation processes to deal with such problems. This involves hiring additional human moderators and provisioning them with better technology such as machine learning-powered dashboards, natural language processing (NLP) filters, and real-time behavioral analytics to detect and filter out objectionable content more effectively. Roblox's Trust & Safety team reports that this technology can enable faster and better identification of grooming activity, objectionable language patterns, and contextual threats than in the previous system (Roblox Trust & Safety 2023 Transparency Report). All these measures aim to enhance safety for junior users by ensuring that most inappropriate content does not circulate through the network. Instead, middle-ranked users, rather than top-ranked ones, play a more critical role in fostering a safer environment of creativity (Shen and Ma 2024). Middle-ranked users are defined as active members with average visibility and engagement, and top-ranked users are high-visibility developers or influencers. All these efforts are intended to provide more protection to junior users, etc.

4.2. Cyberbullying and Harassment

Roblox has become a breeding ground for cyberbullying and harassment, significantly impacting the mental health and wellbeing of its young users; as constructive and integrating social and interactive models where players compose teams and cooperate within games, the platform, unfortunately, offers the chance to act negatively and inappropriately,

harming others (Patchin and Hindujaa 2020). Abuse on Roblox can be achieved verbally by denying others a chance to play in a group or with friends, dispersing rumors, and even threatening. Such actions can be calamitous to youthful users who have not fully developed their social and emotional regulations (Du et al. 2021).

Roblox's moderation system, while extensive, often struggles to keep up with the sheer volume of interactions co-occurring across its platform. For example, Roblox announced its plans to block users under age 13 from directly messaging other players (Guptaon 2024). The reliance on automated tools to flag inappropriate behavior can lead to situations where subtle or context-specific instances of bullying go unnoticed. Furthermore, the human moderation team, however hardworking and committed they are, will struggle to handle the many petitions they are presented within a day; hence, they will take time to address some critical issues (Han et al. 2023).

This gap in effectively managing cyberbullying is concerning, as timely intervention is crucial in preventing further harm and providing victims with the support they need. Stress arising from bullying results in anxiety, depression, and, in the extreme, suicidal thoughts (Arseneault et al. 2010). To young users, who are the most sensitive in this case, these experiences affect their health and further development (Kim and Kim 2023).

To address these challenges, Roblox must enhance its approach to handling reports of bullying and harassment. This could include hiring more human moderators for faster response time and more elaborate investigations (Gray et al. 2024). Moreover, more efficient training of the moderators, which will allow them to better comprehend the essence of bullying, is also helpful in achieving larger intercessions. Counselling or mental health services that are complementary or safer victim services have also reduced the impact of bullying as well.

4.3. Predatory Behavior

Perhaps the most alarming risk on Roblox is the presence of online predators, who exploit the platform's interactive features to target and groom young users (Mujar et al. 2024). Since most of its users are children, the platform offers a fitting ground for such disgusting individuals. They use the many profiles that the site has to interact with the minors, creating the impression of being fellow gamers of the children. This manipulation process will involve building a friendship with the child under consideration to use that child for evil ends, which is otherwise referred to as grooming (Whittle et al. 2013).

There have been multiple cases where predators have successfully used Roblox to groom and exploit children, exposing significant vulnerabilities in the platform's safety measures, especially when it is argued that its role as a learning environment can be maximized (Han et al. 2023). In such situations, these predators have, in one way or another, been able to outline the child, and they have used private messages, in-game communications, and innocuous items such as virtual gifts to gain the child's confidence. Once this trust is established, such an individual moves to groom the child into giving out more personalized information, sending out more improper pictures, or even creating a chance to meet this child physically, which will further harm the child.

In 2022, an alarming incident brought this issue to the forefront. A coordinated effort by law enforcement agencies led to the arrest of several individuals who were using Roblox to initiate contact with minors for illicit purposes (Schulten 2022). This specific operation showed that the predators were keen to avail themselves of all of the different facets of the site to get in touch with the troubled children. However, a probe into these people revealed that they had used all manner of subterfuge to infiltrate the platform, making individuals crave a better protection method that involves closer scrutiny of interactions (Schulten 2022).

This incident served as a wake-up call for Roblox and the broader online community, emphasizing the critical need for enhanced safety measures to protect young users. It elucidated what is wrong with automation that has already been implemented for moderation purposes; this model is inadequate in capturing these predators' exploiting behaviors.

The system typically depends on the detection of keywords and pattern-matching software, which are easily evaded by experienced cyber criminals—most notably, those with skills in exploiting neural networks and members of the technocratic class of online criminality (i.e., extremely skilled criminals with advanced knowledge of the infrastructure, manipulation of AI, and evasions that are used to circumvent moderation).

In response, Roblox has been working to implement more stringent safety protocols. This involves upgrading the current complex auto-moderation with advanced artificial intelligence and machine learning to identify suspect behavior patterns better. These technologies can show that grooming patterns have improved because interaction patterns, not just one-time situations, can be better ascertained (Roblox Corporation 2023).

Additionally, Roblox is increasing its human moderation team to provide more comprehensive oversight of user interactions. The human moderator is greatly needed to interpret the context of interlocutions and make appropriate judgments that an AS might fail to make. By growing this team, Roblox has addressed the goal of accepting and considering suspicious activity reports in a shorter time and using a more thorough approach (Roblox Blog 2021).

5. Technical Aspects of Moderation in Roblox

Moderation in Roblox uses automated tools and human review to keep the platform safe. It relies on innovative technology that scans user content for rule violations to balance safety and free expression (Hine 2023). These algorithms are trained to detect explicit language, inappropriate images, and harmful behavior patterns by analyzing text, images, and even in-game activities. Roblox also employs natural language processing because the application considers context and can identify potential issues. However, real-time filtering services work as a watchdog in that they ensure that students do not use abusive language and do not relay wrong information about each other, hence no more cyberbullying and other immoral behaviors. Still, due to the vast amount of content posted on Roblox, numerous human moderators watch or delete some material that did or did not violate the rules because machines can only do so much. AI also should be incorporated with human intelligence to achieve the best outcome, sufficiently expanding the reach without damaging understanding of the specificities needed to prevent specifically negative experiences for users on Roblox. Figure 3 presents a holistic approach to managing digital platforms with a multi-layered system primarily focusing on user-generated content. It combines a technology backbone (system layers), controls (risk management), content moderation (quality control), and regulatory compliance (legal adherence), demonstrating an integrated plan to address the dilemma between freedom of expression during creative peaks and systemic accountability. The mutually dependent layers are designed to generate safe and streamable showcases on this dynamic basis, anchored in evolving governance needs.

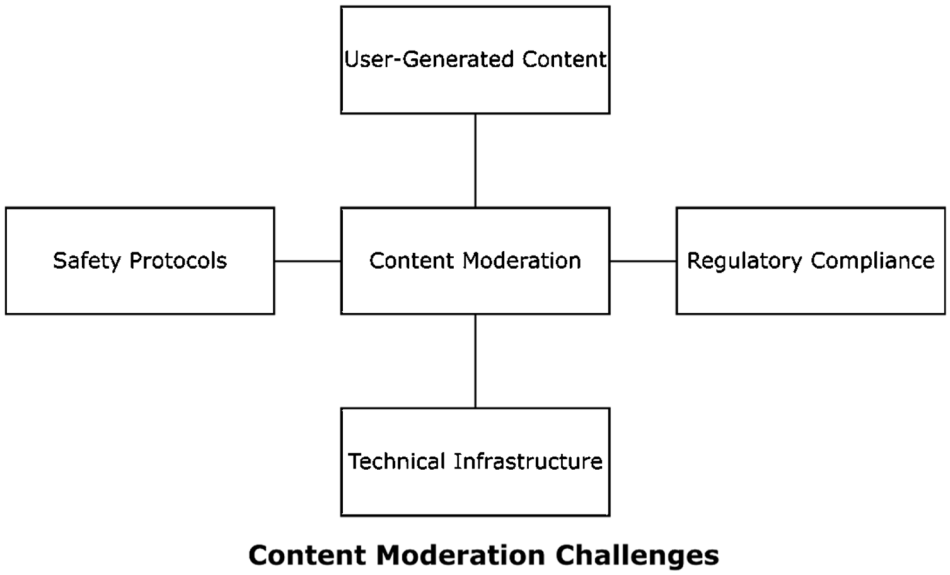


Figure 3. Systematic representation of interconnected components in content moderation framework [Credit: Author, original figure created by the author].

5.1. An Overview of Roblox's Moderation System

Roblox's moderation system is a complex and multifaceted approach designed to maintain a safe and enjoyable environment for its predominantly young user base. As a site containing millions of user-created games and being home to an active social community, the problem of eliminating obscene content and keeping communication appropriate and safe is overwhelming (INEQE 2025). Automated tools are primarily used in the system, and they also use moderators. Robot moderation is advanced and involves pattern recognition and artificial intelligence spreading over the given content, searching for inappropriate things such as naked images and violence. These algorithms are then adjusted for each changing content and the patterns of users' behavior. However, the threat cannot be wholly negated using only the system, which depends on the human moderator's staff. These moderators supervise the marked content, users' complaints, and issues that cannot be pre-coded due to complexity. However, several high-profile case scenarios reveal that the system lacks such opportunities even if it takes the above measures. For example, adverse reports of explicit games and content escape these filters once or twice. Some users have commented that it uses inefficient human moderators because it moderates over a billion average daily images. Moreover, the moderation system does not distinguish between contextual content; it becomes complicated for the program to understand which pictures are vulgar in one culture and acceptable in another (Roberts 2019).

Additionally, Roblox's moderation model has a report facility for users to report objectionable conduct or content. Although such a community approach is useful, its effectiveness fails when targeting children. It is possible that most young users do not explicitly comprehend what content should be reported on or that they will not report to classmates for fear of offending them or from confusion. Their poor understanding of subtle threats like grooming also undermines the protective function of the system. While there are chat filters installed, they tend not to have the contextual sensitivity necessary for the identification of manipulative tricks played on children.

There is adaptive chat filtration that hides obscene messages and the option to exclude phone numbers and addresses. This feature is logically inalienable for shielding the users from cyberbullying as well as all kinds of perversion. However, in this system, users find themselves entrapped with false positives and frustrating limitations to their communication. Roblox's economy is still facing moderation challenges. These use Robux,

which is non-NASDAQ game money, and are confronted by a cheat, a fraudster, and a fake player who realizes gameplay. Despite some possibilities to avoid and identify fraudulent transactions in the company, the users are sometimes criminals in such procedures; more actions should be taken. Hence, to respond to these continual issues, Roblox gives much money to equalize technology machinery and moderators' tools. It has also improved strict age checks and introduced other filtration layers to enable parents to regulate their kids' online activity. Education is used to help parents and children better understand the risks connected with internet usage, and the necessary informational materials are provided (Kou et al. 2024).

Despite these efforts, the rapidly evolving digital landscape means that Roblox's moderation system must constantly adapt to new threats and challenges, especially regarding harmful behavior design (Hine 2023; Kou et al. 2024). The issue of the moderation of content and the continuous fine-tuning of the balance between automated and manual moderation stays relevant so the site can offer proper safety, fun, and communication for young audiences and freedom of creation for the student artists. As much as the growth of Roblox is hence different and interesting, the firm will have to keep enhancing its moderation system to maintain the trust of its ever-expanding consumer base.

5.2. Comparative Analysis of Moderation Systems in Roblox and Other Platforms

The moderation systems of online platforms such as Roblox, TikTok, Facebook, and YouTube play a crucial role in maintaining user safety and content integrity. These platforms rely on technological help and employ moderators to monitor the massive flow of content the users produce (Table 1). Such differences are as helpful as the strengths and weaknesses of the modern contenders for avoiding online abusive content.

Table 1. Comparative analysis of content moderation systems across major digital platforms.

Feature	Roblox	TikTok	Facebook	YouTube
Core Approach	Automated + human oversight	Aggressive AI + human review	AI tools + human evaluation	ML algorithms + Content ID
Key Tech	• ML algorithms • Real-time filtering	• Computer vision • Audio analysis	• Pattern detection • Misinformation tools	• Content ID • Auto-scanning
Human Role	Complex case review	Contextual review	Flagged content review	Appeals handling
Main Challenges	Real-time monitoring of interactive content	Over-moderation, cultural bias	Global censorship concerns	Inconsistent enforcement
Strengths	Youth safety focus	Quick removal	Fact-checking	Copyright protection
Content Focus	Games + chat	Short videos	Mixed media	Long-form video
Timing	Real-time	Pre/post posting	Continuous	Pre/post upload

Note: ML = Machine Learning, AI = Artificial Intelligence. This comparison highlights key differences in moderation approaches across platforms, emphasizing their unique challenges and operational strengths in content management. **ML algorithms**: Adaptative systems that learn through data to automatically identify patterns of inappropriate or harmful content, and which continue to improve with exposure to flagged material. **Real-time filtering**: A moderation process that screens and assesses user-generated content (such as chat, uploads) as it is created or is being shared, blocking infractions prior to publishing. **Computer vision**: Artificial intelligence technology that allows platforms to examine and comprehend images or video frames to identify visual offenses, including nudity, violence, or copyright content. **Audio monitoring**: Automated monitoring of audio streams for explicit language, hate speech, copyrighted music, or other policies, commonly employed for video and livestream moderation. **Detection of patterns**: Applying statistical or machine learning models to detect patterns of abnormal or coordinated behaviors (i.e., grooming, spamming, coordinated manipulation) throughout the platform. **Misinformation platforms**: Mechanisms used to identify, mark, or suppress disseminations of disinformation or deceiving content based on cross-checking fact-checking databases or recognizing shared disinformation patterns. **Content ID**: A digital technology created by YouTube to screen videos for copyrighted material, automatically matching them with a database of rights-holding material and applying takedown or monetization policies. **Auto-scanning**: An umbrella term for automated systems which continually scan for content (images, text, video, etc.) containing words or characteristics matching an established list of rule infractions or targeted items.

Roblox's moderation system relies heavily on automated algorithms and human oversight. Its algorithms are more complex and, based on machine learning, can search for violations of the usage of obscene language and prohibited images and actions. These algorithms persistently update their performance in automated content identification depending on the flagged content. Furthermore, real-time filtering is in place to tackle cyberbullying and safeguard user privacy in chat interactions (Kou and Gui 2023). However, Roblox content usually floods such systems; such systems need a large team of moderators who analyze all patterns reported by users and the problematic cases that algorithms cannot consider. These two strategies are deemed synergistic to deliver a safe and enjoyable environment, especially to youthful customers. However, it has been seen that such attempts are not well suited to Roblox because the content is constantly changing, and the users put forward new games and their interactions, which are sometimes hard to frame or even monitor in advance (Kou and Gui 2023).

In contrast, TikTok employs a more aggressive approach to content moderation, leveraging AI and extensive human moderation. TikTok's algorithms are designed to identify and remove content that violates community guidelines, such as hate speech, nudity, and violent content. The platform's AI tools are particularly adept at analyzing video content, utilizing computer vision and audio analysis to detect inappropriate material. TikTok also has the services of many moderators who watch videos that users question, making better observations regarding the contextual content (Bonagiri et al. 2025). It is the case of moderation used by TikTok as a primordial, thus prophylactic course of action to delete prohibited content. This has assisted the platform in controlling the progression of the title population and the colossal quantity of content produced daily. However, it is subject to complaints of over-moderation, censorship, and moderation that is insensitive to social and cultural bias (Bonagiri et al. 2025).

Facebook's moderation system combines AI tools with human evaluation, focusing on identifying harmful content and misrepresentation. The platform's algorithms are trained to detect patterns in text, images, and videos that suggest policy violations (Gillespie 2020). Facebook has created more powerful dedicated instruments that can help control the distribution of fake news and toxic conspiracy theories, which have become widespread on its site. During mediation, humans, for all the content tagged as explicit by the people, evaluate specific content that may need more context. It also has fact-checking partners to demarcate the provenance of content and block the flow of misinformation. However, several times, the platform has been criticized for censorship of political speeches and user information, pointing out the realistic challenges of running an international social network with millions of registered users (Gillespie 2020).

YouTube employs machine learning algorithms to scan uploaded videos for inappropriate content, such as hate speech and violent extremism. The platform's Content ID system allows rights holders to manage their intellectual property by automatically identifying and acting on infringing content (Gorwa et al. 2020). YouTube also has moderators who analyze appeals and flag the videos to decide whether moderation was performed on them. Given the millions of video uploads per minute on YouTube, the platform can and does employ AI to respond to and stop the distribution of dangerous content even more efficiently. However, the platform has received backlash for its seemingly random policies on content removals and demonetization, which affect the earnings of many creators (Gorwa et al. 2020).

Comparatively, Roblox's moderation system is effective but faces challenges due to the platform's highly interactive and user-driven environment. Whereas TikTok, Facebook, and YouTube all have hybrid moderation solutions in place using AI and human review, those platforms' operating environments and challenges differ considerably from Roblox's.

Perhaps most distinctively, Roblox is unique in requiring moderation of real-time activity in user-created 3D worlds, including live chat in games, avatar actions, and multiplayer game interaction (Roberts 2019). By contrast, TikTok and YouTube primarily moderate uploaded and recorded content. TikTok uses computer vision and audio-based analysis for short videos, and YouTube uses its Content ID system for copyright enforcement and retroactive video analysis. Facebook's Rights Manager manages copyright content in the same manner and enforces takedowns for asynchronous and static posts (Gillespie 2021).

In order to provide a better comparison, services such as Discord and Twitch—which also moderate live, real-time content—offer better analogies. Twitch uses AI moderation bots (such as AutoMod), per-channel rules, and community-based moderation for live streams. Discord uses automated keyword filtering and real-time flagging in combination with tiered human moderation to moderate live voice and chat in decentralized community servers (Gillespie 2021).

In contrast, Roblox has the special challenge of moderating live interactions among millions of players at one time participating in diverse, player-created virtual worlds. Such complexity necessitates a tiered system combining automated filtering using AI-based NLP and image identification, in-game behavior monitoring, and large-scale human moderation, applied to manage emergent, unpredictable sandbox environments.

Accordingly, Roblox can benefit from incorporating more proactive content removal strategies and enhancing its real-time moderation capabilities. Lessons from what TikTok, for example, or YouTube have done, such as building artificial intelligence applications and investing in rigorous manual moderation, can improve the user experience on Roblox (Firth et al. 2024). The frequency of the moderation systems is also helpful in attending to the new tendencies of the topic as well. Therefore, by applying new developments in AI and strengthening human control over processes, one can enhance the safety of platforms and safeguard the reputation of newly created territories, impacting psychological, cognitive, and social dimensions (Firth et al. 2024).

5.3. The Effectiveness and Challenges of AI in Roblox Content Moderation

The application of artificial intelligence (AI) in content moderation on platforms like Roblox has been central to dealing with large volumes of user-created content (Masi et al. 2024). AI-based moderation systems detect offensive language, objectionable photos, and toxic interactions in real-time and are, hence, vital for large-scale platforms (Masi et al. 2024). However, their efficacy is increasingly being questioned since their algorithms are not transparent, susceptible to bias, and not procedurally fair.

To better understand the implications of AI moderation for governance, one must place this debate in the context of algorithmic governance. Scholars such as Kitchin (2017) and Hildebrandt (2020) argue that AI systems used for governance purposes—such as automated moderation—are a form of "soft law" because decisions are not being made by policymakers but by algorithms. These delegations to AI systems have deep consequences for transparency, accountability, and procedures of due process in digital environments. For one, algorithmic opacity limits user comprehension of why content is being removed or demonetized, which erodes transparency. For another, accountability gets dispersed—when something goes awry, there is uncertainty as to whether blame rests with the platform, with developers, or with the algorithm itself (Hildebrandt 2020).Third, and often, automated moderation is not linked to accessible appeal processes, which raises concerns of due process, especially where decisions affect users' rights or ways of earning a living. These consequences extend beyond users and content creators to affect larger regulatory efforts, since algorithmic moderation is increasingly deciding what norms and boundaries govern digital environments (Hildebrandt 2020).

5.3.1. Algorithmic Bias and the Issue of Fairness

A significant problem with AI moderation is algorithmic bias. The historical data used to train AI systems usually bring their attendant biases and inequalities, resulting in disproportionate content takedowns on marginalized groups (Guo et al. 2024). For example, research on algorithmic censorship has found that AI moderation tools over-flag content from dialects, political positions, or communities and under-flag more subtle online abuse (Hildebrandt 2020). In Roblox terms, this means that certain objectionable content (i.e., predatory grooming or coded hate speech) is not caught while innocent speech is unfairly silenced.

In their work on algorithmic regulation, Kitchin (2017) and Hildebrandt (2020) identify the self-sustaining nature of algorithmic decision-making: once an AI system is trained to detect content as "harmful" or "safe", it has minimal feedback loops and limited capability for human intervention. This explainability—also termed "the black box problem"—undermines users' power to contest decisions made by AI. This especially concerns child-friendly online platforms like Roblox, where users may not be informed of reasons for having content removed or flagged.

5.3.2. Legal Accountability and Due Process in AI Moderation

Content moderation by AI is raising accountability and due process issues. Who is responsible when an AI program mistakenly removes acceptable content or misses objectionable content? While platform operators such as Roblox claim that AI moderation guarantees that community standards are maintained, legal experts argue that compliance through AI is without the process guarantees that are critical to due process (Siapera 2021).

For example, in tradition-based legal systems, individuals who are accused of rule-breaking have a right to appeal and redress. In AI-based moderation, on the other hand, users are subjected to untransparent decision-making with very limited opportunities for contestation (Hildebrandt 2020). Roblox's current process of appeal is untransparent and leaves users, especially children, without a meaningful channel through which to contest wrongful enforcement. This raises fundamental legal questions about whether AI-based moderation systems are held to the same due process standards as tradition-based legal systems.

Moreover, the EU Digital Services Act (DSA 2022) now mandates that huge online platforms (VLOPs) be more open about their content moderation policies, particularly on automated decision-making.

Under Article 14(1)(d), platforms must include in their terms and conditions "information on any restrictions imposed in relation to the use of the service in respect of content provided by the recipient of the service, including information on algorithmic decision-making and human review". Furthermore, Article 17(1) grants users the right to be informed of decisions to restrict content or suspend accounts, including the reasoning and whether the decision was made automatically. Critically, Article 17(3) ensures users have access to an internal complaint-handling system, allowing them to contest decisions, with Article 20 further requiring access to out-of-court dispute settlement mechanisms.

Obligations for VLOPs:

- **Risk Assessment:** VLOPs are required, according to Article 34(1) DSA, to conduct thorough assessments to identify and analyze systemic risks associated with their services, including the dissemination of illegal content, adverse effects on fundamental rights, and manipulation of services impacting public health or security.
- **Risk Mitigation Measures:** Based on risk assessments, VLOPs, per Article 35(1) DSA, must implement appropriate measures to mitigate identified risks. This includes

adapting content moderation processes, enhancing algorithmic accountability, and promoting user empowerment tools.

- **Independent Audits:** Under Article 37 DSA, VLOPs are mandated to undergo independent audits to evaluate compliance with DSA obligations. These audits ensure transparency and accountability in the platforms' operations.
- **Data Access for Researchers:** To facilitate public scrutiny and research, VLOPs, according to article 40 DSA, must provide data access to vetted researchers, enabling studies on systemic risks and the platforms' impact on society.

As of 23 April 2023, the European Commission had designated 19 platforms as VLOPs, including major entities like Amazon Store, Facebook, Instagram, TikTok, and YouTube. Platforms such as Zalando have contested their classification, arguing that their business models differ from those of typical VLOPs (European Commission 2023).

The designation of 19 platforms as VLOPs by the European Commission is an important regulatory milestone under the DSA. It brings these platforms—including Amazon Store, Facebook, Instagram, TikTok, and YouTube—under tighter obligations to address risks on a system-wide level, improve algorithmic transparency, and protect user rights, including access to appeal procedures and data for research purposes. It reflects a move toward increased accountability for leading digital platforms in the EU and affirms the EU's leadership in establishing international standards for platform governance and digital right protection.

- **Implications for Roblox:**

While Roblox has a substantial global user base, its classification as a VLOP under the DSA depends on its monthly active users within the EU. If Roblox meets the VLOP criteria, it would be obligated to

- Conduct comprehensive risk assessments related to content dissemination and user interactions (Article 34(1);
- Implement robust risk mitigation strategies, potentially overhauling existing content moderation systems (Article 35);
- Submit to independent audits, ensuring compliance with DSA mandates (Article 37);
- Provide data access to researchers, enhancing transparency and facilitating external evaluations (Article 40).

Failure to comply with these obligations could result in significant penalties, including fines of up to 6% of the platform's global annual turnover (Article 52(3).

5.3.3. A Technical Fix Is Not Enough: The Need for Ethical and Regulatory Reforms

Addressing these challenges requires more than a series of technical upgrades to AI moderation; legal and ethical reforms are required to regulate automated decision-making systems. Algorithmic systems cannot be seen as neutral tools but as political and legal actors within broader regulatory ecosystems, as argued by Kitchin (2017).

Instead of suggesting user-focused appeals mechanisms and hybrid human-AI oversight as novel solutions, one can better position them as necessary but insufficiently put into practice in Roblox. Such mechanisms have some sort of baseline existence in most large platforms, yet Roblox's specific setting of real-time, interactive user-created games requires specialized adjustments in order to function well. For example, Roblox's appeals process is highly opaque, especially for its largest demographic—adolescents and children—who do not have the digital literacy or self-confidence to use complicated feedback channels. Roblox should implement child-friendly, guided appeals interfaces relying upon visual indicators and streamlined workflows for its young users. Also, notifications for moderation actions must include transparent explanations written in plain terms, possibly

supplemented by AI-powered chatbots to respond to user queries in real-time. In the same manner, hybrid AI-human moderation would need to be context-based, not merely in surfacing surface-level breaks in moderation, but in understanding subtle actions unique to immersive multiplayer environments. For instance, instead of keyword flagging, machine learning algorithms should be trained to recognize sequences of actions signifying grooming, coercion, and bullying—specifically in team-based environments or chat-based scenarios. Here, human moderators can bring their attention to highlighted patterns as opposed to isolated content. In short, such changes are not new in theory but need specific refinement in line with Roblox's operational realities and demographic concerns. What is necessary is not merely applying generic moderation philosophy, but transferring its formulation into real-time responsive, game-specific, and age-sensitive frameworks.

5.4. Moderation Challenges

Roblox faces significant moderation challenges due to user-generated content's sheer scale and diversity. Ensuring a safe and appropriate environment is daunting, as millions of active users create and interact within the platform. One major issue is the challenge of monitoring real-time interactions and content across numerous games and social spaces while upholding individual privacy and dignity (Dolan 2001). Arguably, children who are the primary users of the services are in a precarious place that exposes them to a range of adverse outcomes—from the posting of obscene materials and bullying to assembling to be exploited. The automatic moderation programs and tools are extensive. However, if they remain too inefficient, they cannot adequately manage the surge of new content that arrives daily. The material that fuels hatred sometimes slips through the dynamic challenges of voice communication in multiplayer video games (Van Hoeyweghen 2024). Moreover, it implies that there exists little accountability for profiling by relying on user complaints; also, children know about the problem or cannot report it.

Though Roblox, YouTube, and TikTok are founded on content moderation based on AI, their approaches, effectiveness, and regulatory compliance are very diverse. The subsequent comparison in Table 2 evaluates significant moderation mechanisms like AI complexity, human oversight, regulatory compliance, and transparency measures.

Table 2. Comparison of content moderation strategies across Roblox, TikTok, and YouTube.

Moderation Variable	Roblox	TikTok	YouTube
AI Sophistication (Gillespie 2020)	Uses automated filters for chat, text, and images; limited ability to detect contextual harm	Highly advanced computer vision and NLP-based AI, trained for detecting hate speech, nudity, and misinformation	Uses Content ID and deep learning AI for identifying copyright violations and harmful content
Human Oversight (Roberts 2019; TikTok 2023)	Moderators review flagged content, but response times are slow due to scale	Large-scale human moderation team ensures AI decisions are reviewed quickly	Uses hybrid moderation, but reviewers mainly focus on appeals rather than proactive monitoring
Effectiveness of AI (Douek 2021)	Often fails to detect coded or evolving harmful content; users bypass filters with modified language	Proactive moderation efficiently removes harmful videos before mass exposure	Effective for copyrighted content, but struggles with misinformation and algorithmic bias
Real-time Moderation (Roblox Corporation 2023; YouTube Help 2023)	Real-time AI filtering for in-game chat and user interactions	Automated removals happen within minutes, limiting content spread	Delayed removals—AI flags content, but human moderation is often required for final takedown
Regulatory Compliance (European Commission 2023; UK Government 2023)	Partial compliance with GDPR, UK Online Safety Act, and DSA but lacks clear transparency reporting	Highly compliant with the DSA and GDPR; has been fined for violations but improved disclosure	Complies with GDPR and DSA, but faces criticism for poor transparency in algorithmic decision-making

Table 2. *Cont.*

Moderation Variable	Roblox	TikTok	YouTube
Transparency Mechanisms (TikTok 2023; Mozilla Foundation 2022)	Moderation decisions are opaque; lacks a clear appeals process for wrongful content removals	Provides detailed content removal reports, including policy rationale and country-specific enforcement	Transparency reports exist, but content removals are often inconsistent or politically contested
User Reporting and Appeals (European Commission 2022)	Users can report content, but appeal processes are slow and lack clarity	Users can appeal decisions, and TikTok has improved moderation response times	Users can dispute demonetization and content takedowns, but appeals take time
Handling of Virtual Harm (European Commission 2022)	Struggles to define 'virtual harm' legally; lacks mechanisms for addressing psychological distress and emotional harm	Has introduced psychological harm guidelines under EU DSA; enforces strict removal of harmful content	Limited focus on psychological harm, but misinformation regulation has improved
Platform Liability (European Commission 2022)	Claims Section 230 immunity in the U.S.; new regulations may force more accountability	Fined for AI failures in moderation but has proactive regulatory engagement	Criticized for evading liability under AI-driven content curation

Case studies show how broad these moderation failures are. In a game on Roblox, there was a Nazi concentration camp created from scratch that went unnoticed by the moderators until public outcry led to its removal. It forced children to watch videos with unsettling connotations, such as Nazi death camps and Holocaust imagery, leaving everyone questioning Roblox's moderation afterward (The Jewish Chronicle 2022). In a case of cyberbullying, a group of users harassed young players and inflicted severe emotional suffering. Such cases prove that, despite all their efforts to avoid such unauthorized contacts and efforts made by the Roblox company to apply keyword filters and introduce new, better reporting means, the problem of creating a safe environment for children remains (Kumar and Goldstein 2020).

The relative moderation tactics of Roblox, TikTok, and YouTube call for further context and justification. TikTok, for one, leverages sophisticated AI technologies in the form of computer vision, natural language understanding, and audio pattern identification to automatically detect and remove content harmful in nature—including hate speech, nudity, and disinformation—before mass dissemination. Based on TikTok's 2023 Transparency Report, in Q2 2023, 91 million videos were removed in total globally, and more than 95% were taken down ahead of any user reports being lodged, indicating a preemptive moderation mechanism based on real-time detection models and behavior indicators (TikTok 2023).

In contrast, YouTube's Content ID system, although technologically sophisticated, exists mainly for copyright enforcement purposes and not for other purposes of content moderation. It enables rights holders to submit reference files, which are used to automatically match against newly uploaded videos. Although this system has been found to be potent in intellectual property protection—automating blocks, monetization, or tracking content matches—it does not directly confront problems such as hate speech and cyberbullying (Google 2023).

Furthermore, other such mechanisms for platforms include Facebook's Rights Manager and TikTok's Audio Fingerprinting System, indicating that YouTube's system is not unique, but part of an overall trend in automatically enforcing copyright tools across platforms. A better contrast would involve an examination of how such moderation platforms operate differently in practice, in terms of speed of reaction, scale of enforcement, and contextual understanding. For instance, although TikTok is good at removing content in advance, we have criticized its over-moderation and cultural bias in removing political or minority-themed content. YouTube's moderation approach is reactive and community-flag driven, and commonly tied to appeals, resulting in slow enforcement, notably in regards to novel harms such as misinformation or psychological distress. Roblox's challenge is not in detecting static content, but in moderating user interaction within dynamic, user-

created 3D worlds in real time, an environment in which in-advance review is impossible (Gillespie 2021).

6. Existing Legal Framework

This section aims to critically scrutinize the capability of current legal frameworks in addressing sexual violence within Roblox. The findings of this paper are based on comprehensive legal research of the legal systems in the United States, the European Union, and the United Kingdom (Table 3). The law in these areas has evolved to address sexual violence, preferably in the natural sense. However, the shift to virtual realms, especially Roblox, poses several legal issues for discussion.

Table 3. Comparative analysis of legal frameworks.

Jurisdictional Framework	Legislative Instruments	Regulatory Mechanisms
United States	*Child Pornography Prevention Act (CPPA), 1996* • *Prosecutorial Remedies and Other Tools to End the Exploitation of Children Today (PROTECT) Act, 2003* • *Communications Decency Act §230* • *State-specific statutory provisions*	Prohibition of virtual child exploitation content • Regulation of digitally manipulated minor imagery • Platform immunity doctrine • State-level virtual conduct governance
European Union	*Council of Europe Convention on Cybercrime* • *EU Directive on Sexual Abuse and Exploitation* • *Digital Services Act (DSA), 2022*	• Expansive child exploitation content definition • Very Large Online Platform (VLOP) obligations • Systematic risk assessment protocols • Monetary penalties framework ($\leq$6% global revenue)
United Kingdom	*Online Safety Act 2023*	• OFCOM regulatory oversight • Platform safety compliance requirements • Mandatory transparency protocols • User protection framework

Note: This table synthesizes the primary legal mechanisms and their operational constraints across three major jurisdictions addressing virtual sexual violence in platform environments.

The key problem is the application of laws which have been designed for scenarios of tangible contact being applied in the context of sexual violence in a virtual space, where physical and psychological trauma may look quite different.

6.1. United States

Studying the current legislation of the United States, we are concerned with how current legal provisions relating to sexual violence might apply to the landscape of Roblox. We focus on exploring how the US legal system deals with sexual offenses within the course of physical contact and, in the digital realm, Roblox. Concerning virtual child pornography, the United States' first legally expanded foray came in the form of the *Child Pornography Prevention Act of 1996* (CPPA).

This legislation included sections pertaining to displaying minors engaged in sexually explicit conduct or displaying computer-generated images that looked like they contained minors.

Thus given the CPPA's broad definition—including virtual depictions of minors—its provisions could potentially apply to user-generated content on Roblox, especially in cases such as sexual portrayal of children using fake or morphed images. However, in *Ashcroft v Free Speech Coalition*, the Supreme Court (Legal Information Institute 2002) was able to strike out those as unconstitutional under an overbroad provision of the free speech clause that prohibited depictions that did not involve actual kids. The Court also disagreed

with the proposition that they should be classified as obscene because examples may pose a potential danger to child abuse in real social life (ibid.). It claimed that provoking danger alone is insufficient; speech must meet the threshold of imminent harm to be legally prohibited. That is all. It is not sufficient to bring it under the prohibited section. Tied to the provision was a violation of free speech, constitutional freedom as provided for by the First Amendment. To this, the US responded by passing the *Prosecutorial Remedies and Other Tools to end the Exploitation of Children Today* (PROTECT) Act in 2003. While not as broad as CPPA, this law involved 'digitized' images of a minor or those that had been further modified to imitate confident kids. Several factors serve to undermine the success of the Act, such as in capturing and addressing the beliefs that underpin sexual violence in Roblox; this and this paper are yet to be answered; thus, the call for legal infrastructure models that can address what is likely to happen in the social-interactional web of virtual systems. There are also state laws concerning virtual sexual conduct in the United States, but these also encounter conflicts between control and free speech, a cardinal principle in American law. Leveling some criminality charges like California Penal Code Section 653.2 shows how hard it is to acquire enough proof of intent during virtual encounters in Roblox. However, to prove the same, it is necessary to satisfy the requirement that the messages were sent with the intention of imminently causing at least that other person to receive an unpleasant message or violence through touching, harm, or sexual misconduct—touching the client in some manner, harming the client—or it would be challenging to quantify harassment in Roblox interactions (Hildebrandt 2020).

Section 230 of the Communications Decency Act (CDA) further complicates the issue of accountability in Roblox. This clause clears the Roblox platform from user-generated content, releasing service providers from legal liability for third-party content. This lack of liability and accountability raises considerable questions regarding the interfaces' interventions and prevention of sexual assault within the spaces they host. Clause (c) (Firth et al. 2024) of the Act prohibits any provider or user of an interactive computer service to be regarded as the publisher or speaker of any information of another content provider. Not being taken as publishers protects service providers from outcomes against users' uploaded content. Although the authors have not posted any defamatory statement on the Roblox platform that appears to have been presented before the courts, the author can conclude with relative certainty that the Roblox platform, being interactive computer services, qualifies as part of Section 230 immunity any other service provider.

The juxtaposition of these legal frameworks with the new confrontation, which Roblox opened, identifies a gap. The US legal system, while offering a framework for addressing sexual violence, is inadequate in fashioning appropriate solutions for the challenges experienced in virtual environments. There is much need for legal change, considering that Roblox creates a relatively active and engaging experience for the user to save and apprehend offenders in these cyberspaces. This situation asks for reconsidering the existing legal regulations and, in some cases, creating new ones: legal standards that more effectively can explain the nature of sexual violence in Roblox.

6.2. European Union

The European Union (EU) legislation analysis shows that, like all other jurisdictions, no law is designed entirely for Roblox. This absence implies the colossal role of monitoring such large and complex digital spaces within which such legislation functions. Ruling out the possibility of a single flagship Metaverse is somewhat uncommon, as the Metaverse is, in fact, complex, and its architecture evolves. The same applies to the internet: there is no Metaverse law, as mentioned above.

However, considering Roblox's specificity, the EU approach addresses issues of concern more in formulating an all-encompassing 'Metaverse law'. Key focus areas include competition, data protection, and intellectual property. All of these domains are rather complex when it comes to Roblox. For example, concerns about competition have aspirations regarding how virtual platforms engage, the monopolization of virtual arenas, and the reasonable usage of virtual markets. Privacy especially is becoming more challenging to protect whenever personal and behavioral information can be gathered in more detail than on typical web-based applications. Situations of legal rights in Roblox encompass ownership and infringement of virtual commodities, products, or unique exclusively generated content.

Certain aspects of the current framework might assist in dealing with sexual harassment cases in Roblox. "Child pornography", as per the Council of Europe Convention on Cybercrime, means any representation, for actual or simulated sexual conduct, of a minor. There is another and even broader definition of "child pornography" under the EU Directive on Combating Sexual Abuse: the protection of children and child pornography where the term means "realistic images of a child engaged in sexually explicit conduct or realistic images of the sexual organs of a child for primarily sexual purposes". These are rather loose definitions, mainly because of the term "realistic images". This permits the potential of including virtual avatars and pornography in Roblox into their domain. In addition to this, the EU has also addressed intermediaries' legal obligations in their role through the Digital Services Act (DSA). Through the DSA, intermediaries have to act upon illegal content upon being notified and have to put in place clear content moderation processes, thus limiting their liability only when they are non-compliant with such obligations.

However, as of now, any rigid and comprehensive CO that comprehensively addresses all the issues connected with sexual crimes taking place in Roblox with no exceptions is currently absent in the legal environment of the European Union. However, attempts have been made to provide for this gap in the past. The European Union adopted the Digital Services Act ("DSA"), legislation regarding illegal and dangerous content on the internet, in April 2022. Applying to intermediaries, the law enlists tech companies to prevent hosted content or face fines of up to 6 percent of global revenue. The obligations have been proportional to the size of the tech company depending on the type of service offered and the number of users (European Commission 2022). Higher standards are set for large online platforms or VLOPs, as they are known. These platforms will be expected to assess systemic risks that they generate and perform risk mitigation assessments. Of the dangers that can be highlighted, we have risks to and on fundamental rights, gender, and sexual violence, the impact of the product on the health and wellbeing of the users, and the impact on minors, among others. Although the specifics of how the DSA will work cannot be seen until it is implemented, this solution can be applied to Metaverse platforms. In the following years, such platforms will become VLOPs due to increased users.

Another reason for them to become VLOPs could be the upcoming trend in many essential services offered in the Metaverse. A large number of great companies started their stores virtually as well as offices on the Metaverse. As this wave continues, the services these platforms offer will expand systematically and increase in popularity. The more contact in the world that is posed with enacted pretend reality would contain a bad side as to sexual harassment, verbal abuse, and even fraud. Instead, legislation like the DSA would inform the platforms that they have duties to regularly process the transmitted content and establish ways to mitigate the associated risks.

6.3. United Kingdom

The United Kingdom's introduction of the Online Safety Act 2023 represents a significant stride in digital regulation, particularly relevant to emerging virtual environments like Roblox. Existing legislative instruments, in particular this one, encouraging the protection of users from various threats when using the internet and targeting child sexual abuse, require reflecting on its consequences for establishing the Metaverse. The legislation provides new tasks for OFCOM, whose positioning has embraced a formal attitude towards child sexual abuse on the internet. This concerns Roblox, which offers an open chance, unlike the previously mentioned domain. However, the nature of user interactions in such immersive environments—characterized by anonymity, lack of physical boundaries, and real-time communication—can increase the likelihood of harassment, sexual abuse, and bullying.

When the idea is taken into the Metaverse, a more significant and complicated virtual space challenges addressing virtual crimes by contrasting fraud; this law can become ambiguous, and the information presented to consumers in these environments often suffers from issues such as lack of transparency, difficulty verifying sources, and limited regulatory oversight, which are not as prevalent in conventional internet media (Karapatakis 2025).

One source of dispute in the law is its handling of misinformation that has drawn much debate. This is mostly based on how it tries to balance regulation of dangerous or inaccurate content with the promotion of the constitutional right to free expression.

The Metaverse generates an extra conflict as a temporality introduced through rich self-generated outcomes to various material and stylistic websites. The distinction between the negative form of fake news and permissible freedom of speech in such a complicated cyberspace is difficult for the law to attempt.

The implementation strategy of the legislation puts much work into the hands of the tech firms, which is both essential and challenging. These firms are forecast to transform safety management significantly, the service required to reflect the safety need, and enhance the users' control over experience. However, as these expectations are noble, their implementation is somewhat doubtful in the context associated with the Metaverse; the change in technologies and user interactions is not simple. The efficiency of these measures in a virtual context is relatively high nowadays. Further, the law-phased approach, including regulating the censorship of illegal content, child ends of safety, and user empowerment, presupposes a rigid, systematic approach and heavy regulation. However, what is its applicability within the Metaverse? Since the type of user engagement is subtle and the kind of content being produced constantly changes, it needs much more consideration. That is why publishing transparency reports is a step forward, although applying the same in the dynamically developing area of the Metaverse poses challenges.

The Online Safety Act provides substantial fines and adds to OFCOM's concrete enforcement attitude (OFCOM 2023). However, guaranteeing compliance across the board, especially for small businesses with constrained funds, and keeping up with the ever-quickening Metaverse development present several challenges (Reuters 2024). The practicality and flexibility of the varied difficulties faced by the Metaverse will play an essential role in it. People want to balance protecting users, especially the young, and respecting everybody's rights and freedoms. The successful regulation of this emerging form of virtual social interaction that involves participants in the Metaverse is crucial to the digital frontier.

7. Proposals for a New Legal Framework Applied in the Metaverse Platforms

A clear regulatory gap has been identified based on analyses of current legal frameworks and their applicability to the Metaverse. This gap demands a specialized legal approach tailored to address the distinct challenges posed by virtual environments. Based

on the discussions presented in this paper, it is possible to indicate several legislative proposals and recommendations that can contribute to creating proper Metaverse legislation. One of the components of this framework relies on the deconstruction of the topic of consent concerning the Metaverse because the legal definition of the concepts does not allow describing the phenomena taking place in virtual space. Consequently, new rules of noncontact communication are needed to ensure that individuals offer explicit and sufficient consent.

While current definitions in law might not always encapsulate Metaverse phenomena, this is a principled stance of technology neutrality in legislating. This helps facilitate flexibility but might not adequately address the specific virtual environment dynamics, especially concerning consent and presence.

The global and borderless nature of the Metaverse presents significant jurisdictional challenges. Solving these necessitates creativity, inter alia, in the form of an international legal convention for providing structures for cooperation in the detection and prosecution of offenses in the Metaverse. This also means that the offenders cannot escape justice because they are in a different area of the country. Furthermore, users and communication partners again remain anonymous, and interactions in the Metaverse are comparably temporary regarding enforcement. Such measures should also represent a fair approach toward user privacy and the need for identification. Furthermore, as with financial transaction records, there appears to be a legal concern that requires storing the content of digital interactions within the Metaverse for investigation and legal procedures.

Without clear jurisdictional boundaries, it becomes difficult to determine which country's laws apply, particularly in cross-border interactions involving multiple users and platforms. This legal ambiguity can delay investigations, complicate evidence collection, and ultimately hinder justice.

Metaverse platforms, as service providers, are expected under emerging legal frameworks (the *European Union's Digital Services Act (DSA, Regulation (EU) 2022/2065)*, UK *Online Safety Act (2023)*) to implement robust moderation procedures, verify user identities, and ensure effective oversight of digital content to prevent the spread of illegal or harmful material. This corresponds to legal obligations, including the UK's Online Safety Act; therefore, the platforms should indicate how and by whom abuses that occur within these domains will be addressed and penalized. Finally, the user power in the Metaverse has to be enabled. Of course, there are legal avenues for reporting abuse, seeking assistance and protection, and some measures to manage social media. Simplified mechanisms for templates with accounts of violation and procedures for obtaining remedy should be inherent in the regulation framework.

7.1. Redefining Virtual Harm: Legal Protections in the Metaverse

The overlap between virtual and physical realities in the Metaverse is a significant challenge in conceptualizing harm in both spheres. With virtual spaces so integral to daily life, digital harm is no longer limited to loss of reputation and money; it now includes psychological distress and emotional suffering as well as social harm. Legal systems have traditionally been developed to address physical and psychological harm in offline life and are hence ambivalent in defining and offering legal remedies for virtual harm (Jang and Suh 2024).

A fundamental question is posed: How can virtual harm be defined and quantified in law? Is it only confined to material loss, reputation harm, and identity theft? Or is it to encompass psychological distress, emotional harm, and impact on digital dignity and agency? Legal scholars and ethicists like Luciano Floridi (Information Ethics) and Danielle Keats Citron (Cyber Civil Rights) argue that digital harm is a fundamental category of

harm in and of itself. Floridi's scholarship highlights that one's digital self is an extension of one's identity and that harm in virtual space—such as online harassment, cyberbullying, or deepfake exploitation—is as significant as harm in physical space. Citron's scholarship highlights discriminatory and gendered forms of online abuse and advocates for a civil rights model of digital harm regulation (Dolan 2001).

The existing legal regimes such as *GDPR*, *the EU Digital Services Act*, and *the UK Online Safety Act* acknowledge that digital platforms need to be regulated, and certain types of online harm should be addressed. They are insufficient in dealing with the legal definition and enforcement of responsibility for virtual harm.

Although the *DSA*, *the GDPR*, *and the Online Safety Act* all embrace regulation of online harm, they are deficient in specifying 'virtual harm' and determining legal responsibility. Based in traditional forms of online architecture, they are poorly equipped to respond to immersive environment-specific dynamics. Because of this, there is a lack of clear enforcement, particularly where nonphysical but injurious interactions in virtual worlds are involved.

Inconsistent enforcement and varying interpretations of what constitutes actual harm in a virtual environment stem from a lack of a particular legal definition. Cyberbullying and online harassment are criminalized as offenses in certain jurisdictions, and no liability is placed on platforms that host such harmful activities in others, leaving victims with minimal avenues for redress (Jang and Suh 2024).

7.1.1. GDPR and Paradox of Protecting Children in Platform Economics

The General Data Protection Regulation (GDPR) is the EU's rights-based data protection approach, especially about sensitive information related to minors. Article 8 of the GDPR in combination with Recital 38 provides additional protection for children, demanding parental consent and age-appropriate information for disclosures. In theory, this is commendable. In practice—particularly for Roblox—it raises regulatory tensions, which are technical and structural.

Verification Deadlock: Parental Consent vs. Privacy

Roblox needs to obtain verifiable parental consent for kids under age 13 (or age 16 in certain countries). Yet, stringent age verification mechanisms—like biometric tests, ID scans, or facial identification—create new privacy hazards and can themselves contravene data minimization rules (Art. 5) and purpose limitation rules (Art. 6) of GDPR. So, platforms such as Roblox become stuck in an enforcement dilemma: rigorously verify and risk over-collecting; verify lightly and risk non-compliance against consent requirements.

This is of special concern in an onboarding-based platform using gamified sign-up—ideally, the easier it is for players to sign in, the quicker Roblox can bring in and retain players. Mandating such high-friction consent regimes risks causing user attrition, which affects engagement stats and virtual goods-driven revenue (Robux purchase), ad inventory, and behavioral data.

Children's Datafication and Algorithmic Exploitation

Roblox is not just a game, it is a datafied one. Every aspect, from avatar customization to game design to chat interactions, creates traceable behavior data. In terms of GDPR logic, they are not just user interactions—their combinations are personally identifiable data when connected to specific accounts. What this implies is Roblox is legally compelled to justify

- Behavioral profile-based content recommendations (usually automatic and non-transparent)
- Application of in-game marketing based on engagement metrics

- Targeted content filtering (e.g., for moderation of chat), potentially analyzing sensitive phrases

GDPR restricts such profiling (Article 22) in cases in which automatic decisions have "legal or similarly significant effects". These can be exclusionary bans or moderation blocks affecting whether a child can join in content or communicate with other users—a fundamental aspect of Roblox's business. A lack of explainability in moderation can violate the GDPR's transparency principle (Art. 12–14) in cases in which automatic moderation blocks or shadow-bans content without human oversight.

Compliance Design Dilemma

The GDPR demands platforms to exercise "data protection by design and by default" (Art. 25), in other words, mechanisms for protecting privacy should be incorporated into the platform's design. In Roblox's case, this can mean

- Limiting behavioral tracking by default
- Providing granular opt-ins (instead of opt-outs) for data gathering
- Turning off targeted content recommendations for minors in the absence of parental approval.

In reality, this might necessitate rebuilding the user interface and retraining the personalization engine—problems not just technical, but financial. It attacks at the root of Roblox's platform logic, which rests upon behavioral loops and network effects.

7.1.2. Digital Services Act (DSA): From Content Moderation to Systemic Risk Governance

The Digital Services Act (DSA) constitutes a regulatory turning point: it reconfigures moderation from an ad hoc enforcement practice to an integral part of risk prevention, transparency, and governance. If Roblox were to have status as a Very Large Online Platform (VLOP) under Article 33, it would be subject to the toughest set of requirements under any platform regulation system in the world.

Systemic Risk Identification: From Damaging Games to Social Manipulation

According to Article 34 of the DSA, VLOPs have to carry out extensive risk assessments not only for illicit content (e.g., hate speech and CSAM), but also

- Breach of fundamental rights (such as freedom of expression, protection of minors)
- Recommender systems for manipulative purposes (e.g., sensationalist or harmful game amplification)
- Implications for social cohesion and mental health.

This requires Roblox to move beyond policing discrete chunks of content and begin to audit its platform's architecture: how do its game design affordances, content recommendation algorithms, and social features structure environments to facilitate grooming, propagate toxicity, or subject kids to in-game manipulation?

It challenges Roblox to address the structural causes of online harm: Roblox abuse is not just an accident—it can be encouraged or endorsed by the design elements themselves that optimize for stickiness and virality.

Algorithmic Transparency and Research Access (Articles 42 and 40)

Roblox's moderation and recommendation AI would have to be auditable and transparent. This involves

- Revealing whether harmful content is prioritized or deprioritized
- Error reporting rates for automatic flagging
- Providing authorized scientists with access to platform information.

This constitutes an explicit challenge to claims of platform secrecy and proprietary algorithms. It would expose rules for making decisions at Roblox, demonstrate vulnerabilities in NLP-based moderation (particularly in non-English speaking environments or coded speech), and provoke ethics related to false positives, appeals, and shadow banning.

Human Empowerment, Rather than Just Protection

The DSA brings new rights for users in the form of transparent appeals, opt-out from personalization, and control of their algorithmic experiences. For Roblox, this can translate to

- Providing non-recommendation-based discovery choices (e.g., content organized by ratings or recency and not by likelihood of engagement)
- Providing mechanisms for kids and parents to appeal moderation choices through transparent, understandable channels
- Incorporating age-relevant explanation of algorithmic choices.

For a platform-based business with such a youth constituency, user empowerment is not a UX feature—it is a requirement under law. Non-compliance is not merely reputational risk—under Article 52, penalties can be up to 6% of worldwide annual turnover.

7.1.3. Section 230 CDA: Shield or Sword for Platform Inertia?

The Communications Decency Act (CDA) 230 offers arguably the most contentious protection for platforms in American law. Roblox is not liable for user-created content, even in cases where said content is abusive, harmful, or traumatizing, according to §230(c)(1).

Scope of Immunity: It is Strong but Not Absolute

This wide berth of immunity allows Roblox to host billions of user-generated messages, games, and social interactions without incurring publisher liability. In practical terms, this means

- Vulnerable children who have been groomed, bullied, or exposed to adult content in Roblox have minimal legal remedies available against the site
- Roblox can moderate at will without suffering punishment for missing content or incorrect identifications.

But as in *Fair Housing Council v. Roommates.com*, when a site actively contributes in a meaningful way to creating criminal content (e.g., by creating tools to aid in its creation), it can lose its immunity. This raises legal questions around

- Whether Roblox's game-creation platform and monetization features facilitate the creation of exploitative content
- Whether inaction in moderation can turn into enabling behavior, particularly whenever reported content is disregarded or recurrence is systematic.
- **Good Samaritan Clause: Voluntary Protection, Rather Than an Obligation**

Section 230(c)(2) urges—but does not compel—platforms to police "objectionable" content. That leaves Roblox free to decide to what degree they would like to police. But note this paradox: the more they do, the greater legal and ethical obligations run in its direction. If, for instance, they create AI filters, employ censors, and construct reporting mechanisms, courts can ask: *Why didn't you still prevent harm?*

This can be seen in cases such as *Doe v. Roblox Corporation* (2021), in which Roblox was accused of negligence in not stopping child sexual exploitation.

Platform Ethics vs. Legal Minimalism

Section 230 allows Roblox to do the bare minimum, legally speaking. But reputational, commercial, and ethical imperatives—particularly in a kids' context—require much

more. Parents, teachers, and regulators do not just demand legal safety; they demand moral responsibility.

This establishes a twofold structure of accountability: legal immunity on one side, and growing moral responsibility on the other. Roblox's ability to navigate this—especially as it expands in Europe and faces scrutiny under the DSA—will ultimately shape its global credibility.

Furthermore, to address the challenges of Roblox, legal scholars propose that virtual harm is not only defined in terms of reputation or economic harm but is conceived as a violation of one's digital rights, agency, and dignity (Citron 2014; Floridi 2013; Peloso 2024). This would be a step towards recognizing that virtual harm consists of

- **Emotional and Psychological Distress**—Persistent cyber harassment, online grooming, and digital manipulation have a very detrimental psychological impact on victims, particularly children. These must be criminalized under cyber protection law (Citron 2014).
- **Violations of Digital Identity**—Exploitations of deepfakes, abuse of avatars, and online impersonations directly undermine a person's digital life and autonomy and require greater legal recognition and enforcement capabilities (Floridi 2013).
- **Reputational and Economic Harm**—Current defamation law already safeguards against harm to reputation, and new frameworks need to be created to tackle virtual slander, doxxing, and financial exploitation in the Metaverse (Peloso 2024).

A redefinition of virtual harm needs to be complemented by a clearly defined legal regime for redress with compensatory remedies for victims, platform liability, and regulatory enforcement. The UK's Online Safety Act and the European Union's Digital Services Act (DSA) are starting points in that very large online platforms (VLOPs) are made liable for systemic harm. These regulations need to be further developed to encompass more specific digital harm that considers the interactive and immersive nature of Metaverse interactions.

With more interactive and realistic online interactions, legal notions must extend beyond material harm and pecuniary loss to accommodate the psychosocial effects of online experience. Not redefining digital harm would deprive victims of protection and enable perpetrators to exploit jurisdictional loopholes (Dolan 2001).

7.2. Ensuring Safe and Respectful Interactions in the Metaverse: The Need for Effective Consent Mechanisms

The Metaverse requires consent mechanisms to ensure that its users engage in consent-based interactions with other people's domains. Such propositions need to have roots in fundamental legal precepts, including the principle of legality, which mandates no punishment of a behavior that could not have previously had a law prohibiting it; the materiality of fact, which regulates the admissibility and materiality of evidence, especially in online environments; the principle of proportionality, which guarantees actions taken are reasonable and not disproportionate in the context of the offense; and the principle of presumption of innocence, which ensures no user is held guilty until due process is observed (Dolan 2001; Peloso 2024). These should start with raising user awareness about the fundamentally different dynamics of this new technological environment. When users are generally in the Metaverse, the meta key concepts regarding the behaviors people should follow in a civilized society must be present. To this end, platforms could require all new users and those users complained about taking some tutorial on the acceptable behavior on the platform. One of the modes that should inform these tutorials was the provision of 'straightforward and in-word' ways of seeking and receiving affirmative responses from avatars. This could mean straightforward verbal queues or contracts before any contact might be considered violating boundaries or personal. It eliminates cases of one

party feeling harassed or coerced into a particular interaction or engaging in an interaction they did not want because all parties agree to specific engagement protocols. Moreover, users should be able to select and set the degree of activity in different environments presented within the Metaverse. This means letting users define their personal space and regulate an avatar's interaction.

An example is in meta introducing the 'Personal Boundary' feature. This feature puts a four-foot wall around an avatar through which no one can approach, thus minimizing the experience of harassment. Therefore, elements like these allow users to control their environment proactively, preserving the quality of the virtual experience. These consent features are a way to contribute to a safer Metaverse and are one step from the Metaverse teaching consent to its citizens.

As the Metaverse continues evolving, these mechanisms must adapt and respond to new challenges and user needs. Consent discussions must be constantly monitored and updated so that all online interactions remain mutually appreciated and safe for everyone involved. There could be other forms of interaction that would be useful, or the ability to make one's avatar uncomfortable. Users may also be allowed to change the symbolic portrayal of the their avatar (cartoon-like, silhouetted form, realistic form) based on what the user would prefer in a particular context. It would enable the individuals to participate in a targeted way and simultaneously ensure they lack recognizable data.

Metaverse users should have utilitarian approaches to these relationships—for example, regulating the visibility of other avatars and 'panic buttons' to instantly eject users from one's House/Office. These features must be implemented to minimize unnecessary contact and stop harassment. Metaverse platforms must provide comprehensive and easily scalable safety features, such as the 'Safe Zone' mode, which automatically restricts the distance between users, comprehensive gestures for signaling alarms, and rigid tutorials introducing community rules backed by moderators. As has already been mentioned, it is preferable to apply all of the strategies above simultaneously to achieve the highest level of safety. However, addressing harassment in the internet space goes beyond the adoption of platform-specific policies to regulate content, where state legislation is still not very sophisticated, and issues of sexual harassment in the context of virtual interactions are still not clearly defined by law. There has also been an evolution observed in legal regulations relating to digital avatars. Like many developing frontiers of social communications, the Metaverse, in particular, poses some emerging legal issues largely unaddressed by current statutory laws. Improving the safety of online experiences requires a more complex approach, implying the cooperation of corporations, non-governmental organizations, and governmental structures. These stakeholders can then collaborate to create organizational policies and standards for better technology use. Such an endeavor is essential for safeguarding the user, and it changes with time, as the Metaverse exemplifies. The general safety characteristics of a platform, general concepts, and, at best, industry initiatives in the ICTS sphere may contribute to creating safer and more dignified conditions for participants in cyberspace.

7.3. Navigating Jurisdictional Challenges in the Metaverse: The Need for International Cooperation and Legal Frameworks

Due to the inherent global location of the Metaverse, the attempts to settle disputes present commendable legal issues of jurisdiction. The overlap of national laws can lead to competing jurisdictional claims, complicating the legal landscape. For instance, the General Data Protection Regulation (GDPR) "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016" is an example. Based on GDPR, any company, irrespective of its location, which has developed websites collecting and analyzing data originating from users is subject to those rules if this user is an EU citizen. Concerning

the Metaverse, which connects the US to EU citizens, this imaginary case shows how Metaverse platforms cause jurisdiction controversies under US and EU laws regarding selecting appropriate legal forums and laws governing Metaverse cases.

Since no identifiable sovereign state is recognized universally as having jurisdiction over virtual arenas like the Metaverse, cooperation is needed across states. They could agree on some model legal approach to regulating such critical issues as the right to consent, and the anonymity of users in the Metaverse, The United Kingdom devotes a separate section associated with internet safety in the framework of the Online Safety Act that can act as an example for other states. However, since the internet is complex and global and does not respect territorial borders, it is possible to define the specific authority within the applicability of AI and virtual worlds.

This complexity makes the efforts for online safety not simply a national responsibility but an international obligation. Although having a virtual jurisdiction system regulated mainly by an existing legal authority that deals primarily with Metaverse offenders is unique, its serving reality is questionable.

While the notion of states surrendering aspects of jurisdiction might appear to be a romantic notion, it indicates the increasing awareness of the reality that in a highly connected digital environment, pragmatic and strategic types of collaborative governing are needed to deal with transboundary problems—not born of blind trust, but of mutual interest.

A more articulate approach would be to create an international organization responsible for supervising and dispensing standardized rules and regulations for state implementation. This body could aid in standardizing regulation across jurisdictions, reducing the distribution of overlapping jurisdictions. It would rationalize legal undertakings in the Metaverse, making society more coordinated and contributive in overseeing its future by synthesizing and harmonizing essential activities mandated by various legal systems of different sovereign states. Legal norms must protect individuals in response to the growing virtual component, focusing on contracts between Metaverse enterprises and their customers, antitrust and competition restraints, copyright protections, biometric data protections, rights of publicity, and safeguarding customer expression in Metaverse environments (Garon 2022; Filipova 2023).

7.4. Legal Subjectivity in Virtual Worlds: Avatars' Status and Liability in Cyber Misbehavior

As increasingly sophisticated virtual worlds like Roblox and the Metaverse emerge, legal subjectivity in virtual space is gaining significance. Traditionally, legal systems allocate rights and obligations based on recognized categories of natural persons (human beings) and legal persons (corporations, organizations, and, in some cases, AI systems) (Solum 1992). However, the advent of avatars and digital identities introduces novel legal questions: Do avatars possess legal standing? Who is liable for digital misconduct: the user, the platform provider, or the avatar itself? These inquiries necessitate a reassessment of liability frameworks, digital personhood, and legal accountability in virtual spaces.

7.4.1. Should Avatars Have Legal Status?

On most virtual platforms, avatars function as extensions of users, facilitating interactions, transactions, and even contractual agreements within the Metaverse. Some legal scholars advocate for recognizing avatars as a form of digital legal personhood, akin to how corporations are granted legal personality (Garon 2022). This recognition would allow avatars to

- **Enter contracts**—A user's avatar could purchase virtual goods, lease virtual property, or engage in smart contracts with legally recognized rights and obligations.

- **Hold digital assets**—As digital economies expand, avatars could be granted ownership rights over virtual property and NFTs, much like corporations possess assets.
- **Be legally accountable**—If avatars were granted legal personhood, they could be held liable for online harassment, digital fraud, or virtual trespassing, subjecting them to enforceable penalties.

However, granting avatars legal personhood presents significant ethical and legal challenges. Unlike corporations with boards and shareholders who can be held accountable, avatars are directly controlled by users. If an avatar engages in harmful actions (e.g., cyber harassment, financial fraud, or identity theft), should responsibility fall on the avatar itself or the user behind it? (Garon 2022).

A compromise approach could involve granting avatars a quasi-legal status, enabling them to function as proxies for users while maintaining a distinct legal identity. This would ensure that contracts executed through avatars are legally binding and that users remain liable for misconduct committed under their avatar's identity.

7.4.2. Liability in Cases of Cyber Misbehavior: Users, Platforms, or Avatars?

A significant challenge in virtual governance is determining liability in cases of digital wrongdoing. Virtual environments present unique difficulties in attributing responsibility:

- **User Liability**—In most legal systems, users are directly accountable for their online behavior. If a person engages in online harassment, fraud, or illicit transactions through their avatar, they can be prosecuted under existing cybercrime laws (Peloso 2024).
- **Platform Liability**—Platforms like Roblox, Meta, and Decentraland have partial legal immunity under laws such as Section 230 of the U.S. Communications Decency Act (CDA), which protects them from liability for user-generated content. However, emerging regulations like the EU Digital Services Act (DSA 2022) impose stricter responsibility standards. Platforms that fail to moderate harmful behavior may be held liable for facilitating digital misconduct.
- **Avatar Liability?**—Some scholars propose treating avatars as "digital agents". If an avatar is involved in fraudulent contracts, virtual property theft, or harassment, legal frameworks could attribute liability to the avatar as a separate entity—like how corporations are held accountable independently of their owners (Floridi 2013).

The challenge with this model lies in enforcement. If an avatar is found guilty of misconduct, what penalties would apply? Would virtual imprisonment be introduced, like platform bans? Would fines be imposed or digital assets be seized? Without precise enforcement mechanisms, this model remains largely theoretical.

7.4.3. The Role of Smart Contracts and Blockchain in Virtual Liability

A potential mechanism for avatar accountability is the integration of blockchain-based smart contracts into digital governance. Platforms could require avatars to be registered on blockchain networks, incorporating identity verification, behavioral records, and compliance mechanisms (Dolan 2001). If an avatar engages in misconduct, penalties could be automatically enforced through innovative contract protocols, such as (Dolan 2001):

- **Reputation scoring systems**—Avatars violating platform policies could have their digital reputation downgraded, limiting their access to specific virtual spaces.
- **Token-based penalties**—Virtual fines could be deducted from an avatar's digital assets as an economic deterrent against misconduct.
- **Contract enforcement mechanisms**—If an avatar engages in fraudulent agreements, smart contracts could automatically execute restitution payments to affected parties.

While these technological solutions could enhance legal accountability, they also raise privacy concerns. Should all digital actions be permanently recorded on the blockchain? Who controls the governance of avatar identities? These questions require further legal and ethical evaluation.

7.4.4. Towards a Legal Framework for Virtual Personhood

As digital identities become increasingly integrated into online platforms, legal frameworks must develop structured approaches to avatar liability and virtual legal personhood. Potential solutions include

- **A tiered recognition system**—Avatars used for personal interactions remain legally tied to users, while those engaging in commercial transactions or digital contracts have distinct legal identities.
- **Mandatory avatar registration**—Platforms could require avatars to be linked to verified user accounts to reduce the risk of anonymous misconduct.
- **Hybrid liability frameworks**—A balanced approach that combines user accountability, platform responsibility, and limited avatar legal personhood to ensure fair enforcement of virtual laws.

As virtual platforms evolve, legal scholars and policymakers must engage in ongoing discussions to ensure that digital personhood, liability, and enforcement mechanisms align with technological advancements and fundamental legal principles (Hildebrandt 2020).

8. Discussion

The findings of this study highlight the significant challenges and opportunities associated with moderating content on child-centric platforms like Roblox. Given the sheer scale of user-generated content and real-time interactions, the platform's Reliance on AI-driven moderation systems is indispensable. These systems are highly efficient at filtering explicit content, identifying harmful behaviors, and enforcing community guidelines at scale. However, their limitations are equally pronounced. AI often lacks the contextual understanding to differentiate between harmful and benign content. However, it is argued that integrating AI and blockchain may significantly enhance the security and inclusiveness of the system Metaverse (Floridi 2013). For example, they may regard friendly play between children and other kids in their group as bullying incidents, or they may not perceive grooming conduct by online offenders. This gap suggests there is no such alignment of this study between mechanical algorithms and their moderators formalized. However, there is still a need for human interaction, and human curation faces the problem of scale because there are billions of communications, for instance on the Roblox platform. Due to such issues, it is necessary to enhance the training paradigms and the evolution of technology to enable human moderators to address those issues while using AI tools.

The legal frameworks examined, such as the *GDPR*, the *Digital Services Act*, and the UK's *Online Safety Act*, provide essential safeguards for users by establishing platform accountability and regulatory oversight. However, these frameworks often fall short of addressing the unique challenges posed by Metaverse platforms. For instance, a weak distinction between virtual harm and geographic location constrains the authorities. That

these legal loopholes are possible with Roblox's user base worldwide is particularly concerning because there seems to be no way to deal with it when many places are involved. Thus, there remains a need to act internationally to set standard legal norms and respond to changes in such virtual reality specialty niches. However, another structure would be required for these ideas or to express them more precisely: virtual cruelty, information protection, and network responsibility.

Comparative insights from platforms like TikTok and YouTube reveal valuable lessons Roblox can adopt to strengthen its moderation practices. TikTok, proactive content moderation, and YouTube and machine learning algorithms present the best examples of enhancements of real-time moderation. However, these platforms have drawbacks: they want censorship; they make biased decisions. These examples highlight the importance of social moderation with AI-based approaches, emphasizing propriety and equity on their platform. Roblox must take lessons from these examples to draw insights, as a collaboration between public health experts, policymakers, and behavioral scientists will be required to develop evidence strategies for the challenges that an ever-changing digital space brings (Kang et al. 2024; Nagyova 2024).

User empowerment emerges as a critical aspect of creating safer digital environments. Providing tools such as personal boundaries, customizable privacy settings, and accessible reporting mechanisms can significantly enhance user safety and foster a sense of control over one's digital experience. There is also the need to publicize the activities that would lift user competencies and make them fully responsible for risk probabilities in the computing environment, particularly on new media generation and all their trappings. Hence, for measures of platforms, educators and policymakers should focus on creating information security awareness, gaining a safety culture, and reducing risks.

To complement the analysis of moderation challenges and regulatory shortcomings, Figure 4 introduces a cyclical model of digital user safety and empowerment. This framework conceptualizes the dynamic and interdependent processes necessary to promote child protection and agency in immersive environments such as Roblox. The model comprises six interconnected components: tool provision, awareness creation, stakeholder education, activity dissemination, control fostering, and safety enhancement. These stages reflect a holistic approach to user empowerment that goes beyond reactive moderation, aiming instead to build proactive, user-centric safeguards. By emphasizing the continuous interplay between education, technological tools, and participatory control mechanisms, the figure underscores the need for systemic reforms that embed user empowerment as a foundational principle in platform governance. It also aligns with broader calls for regulatory models that prioritize safety-by-design and involve users, particularly minors and their guardians, as active participants in digital safety ecosystems.

Balancing innovation with safety remains one of the most complex challenges for Roblox. The platform thrives on user creativity and interactivity, but these elements make moderation particularly difficult. Similarly, over-regulation risks diminishing the company's competitiveness, making Roblox unique or insufficiently protected, and harming users (Zhang et al. 2024). This study notes moderation, flexibility, and adaptiveness as the opposite. For that reason, Roblox's moderation practices, users, and technology can support the company's adherence to continuous innovation for free and safe creation.



Figure 4. Digital user safety and empowerment cycle. Note: A cyclical model illustrating the six interconnected components of user empowerment in digital environments: tool provision, awareness creation, stakeholder education, activity dissemination, control fostering, and safety enhancement. [Credit: Author, original figure created by the author].

9. Conclusions

The rapid growth of child-centric platforms like Roblox highlights the dual-edged nature of technological innovation, offering unparalleled opportunities for creativity and interaction while presenting unique challenges for content moderation and user safety. However, as this research has demonstrated, it is inevitable, and even a disadvantage, that these largely automatic processes preclude a great deal of current, unmodified user input. Although moderation is achieved at the required scale and speed for this platform size, errors will be made because an AI system does not understand the context. The problems of the deficiency of human supervision resources also show that highly complicated moderation models involving young users' moderators are needed beyond AI capabilities.

Legal frameworks play a critical role in holding platforms accountable, but current guidelines do not address the complications of virtual spaces like Roblox. Frameworks such as the GDPR, the Digital Services Act, and the UK's Online Safety Act offer valuable starting points for user protection, yet their applicability to the Metaverse remains limited. Virtual harm is not defined, and the legal issues that international platforms raise for countries present a significant problem for similar enforcement. This study, therefore, provides the need to compile legal rules to set standard laws across the globe and make sure they sync appropriately in their reporting systems in Metaverse platforms—such an approach would fit when accessing existing laws and brand new threats associated with activity within the internet space.

Roblox's challenges are not unique but emblematic of broader issues that other platforms like TikTok and YouTube face. Studying moderation practices, this work reveals the best practices, including regular content deletion and using the most efficient AI algorithms, that may help Roblox improve its safety measures. However, increasing user power, for example, in reporting systems, options for privacy, and materials, is also essential. Accepted facts are checked and verified by the user, such as the kids and their caregivers, and some of the safety aspects can be enhanced to allow better and safer use of spaces by the members.

In conclusion, the findings of this study highlight the shared responsibility of platforms, regulators, and users in addressing the challenges of content moderation and user fortification in child-centric digital spaces. Roblox and other gaming platforms will have to develop advanced technologies for moderation, enhanced human supervision, and adaptation to existing and novel legal trends to mitigate new risks successfully. In endorsing the safety of the users, the study understands that users should act on level ground so that they do not risk undue harm as they develop creativity. Thus, platforms can maintain their constant use and many children's attention irrespective of the frequency of dangerous facets.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflict of interest.

References

Arseneault, Louise, Lucy Bowes, and Sania Shakoor. 2010. Bullying victimization in youths and mental health problems: 'Much ado about nothing'? *Psychological Medicine* 40: 717–29. [CrossRef] [PubMed]

Baszucki, D. 2021. The Metaverse Is the Future of Human Experience. *Roblox Blog*.

BBC. 2022. Roblox: The children's Game with a Sex Problem. Available online: https://www.bbc.com/news/technology-60314572 (accessed on 15 April 2025).

Bonagiri, Akash, Lucen Li, Rajvardhan Oak, Zeerak Babar, Magdalena Wojcieszak, and Anshuman Chhabra. 2025. Towards Safer Social Media Platforms: Scalable and Performant Few-Shot Harmful Content Moderation Using Large Language Models. *arXiv* arXiv:2501.13976.

Citron, Danielle Keats. 2014. *Hate Crimes in Cyberspace. Vol. 52, Osgoode Hall Law Journal*. Cambridge: Harvard University Press.

Dionisio, John David N., William G. Burns Iii, and Richard Gilbert. 2013. 3D Virtual worlds and the metaverse. *ACM Computing Surveys* 45: 1–38. [CrossRef]

Dolan, Lisa. 2001. The Legal Ramifications of Virtual Harms. Master's thesis, Vilnius University European Master's Programme in Human Rights and Democratisation, Vilnius, Lithuania.

Douek, Evelyn. 2021. Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability. *Columbia Law Review* 121: 759–833. [CrossRef]

Du, Yao, Thomas D. Grace, Krithika Jagannath, and Katie Salen-Tekinbas. 2021. Connected Play in Virtual Worlds: Communication and Control Mechanisms in Virtual Worlds for Children and Adolescents. *Multimodal Technologies and Interaction* 5: 27. [CrossRef]

Dwivedi, Yogesh K., Laurie Hughes, Abdullah M. Baabdullah, Samuel Ribeiro-Navarrete, Mihalis Giannakis, Mutaz M. Al-Debei, Denis Dennehy, Bhimaraya Metri, Dimitrios Buhalis, Christy M. K. Cheung, and et al. 2022. Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management* 66: 102542. [CrossRef]

European Commission. 2022. DSA Articles 17–20 on User Redress Mechanisms. Available online: https://eur-lex.europa.eu/ (accessed on 15 March 2025).

European Commission. 2023. More Responsibility, Less Opacity: What It Means to Be a "Very Large Online Platform"—Statement by Commissioner Breton. Available online: https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_23_2452 (accessed on 20 February 2025).

Filipova, Irina A. 2023. Creating the metaverse: Consequences for economy, society, and law. *Journal of Digital Technologies and Law* 1: 7–32. [CrossRef]

Firth, Joseph, John Torous, José Francisco López-Gil, Jake Linardon, Alyssa Milton, Jeffrey Lambert, Lee Smith, Ivan Jarić, Hannah Fabian, Davy Vancampfort, and et al. 2024. From "online brains" to "online lives": Understanding the individualized impacts of Internet use across psychological, cognitive and social dimensions. *World Psychiatry* 23: 176–90. [CrossRef]

Floridi, Luciano. 2013. *The Ethics of Information*. The Information Society. Oxford: Oxford University Press, vol. 32.

Garon, Jon. 2022. Legal implications of a ubiquitous metaverse and a web3 future. *SSRN Electronic Journal* 106: 163. [CrossRef]

Gillespie, Tarleton. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7: 205395172094323. [CrossRef]

Gillespie, Tarleton. 2021. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Google. 2023. How Content ID Works. YouTube Help Center. Available online: https://support.google.com/youtube/answer/2797370 (accessed on 17 April 2025).

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7. [CrossRef]

Gray, Joanne E., Marcus Carter, and Ben Egliston. 2024. Content harms in social VR: Abuse, misinformation, platform cultures and moderation. In *Governing Social Virtual Reality: Preparing for the Content, Conduct and Design Challenges of Immersive Social Media*. Cham: Springer Nature, pp. 11–22.

Guo, Keyan, Freeman Guo, and Hongxin Hu. 2024. Moderating embodied cyber threats using generative AI. *arXiv* arXiv:240505928.

Guptaon, Ishika. 2024. Roblox Limits Meesaging for Under-13 Users Amid Safety Concerns. Available online: https://www.medianama.com/2024/11/223-roblox-limits-messaging-for-under-13-users-amid-safety-concerns/ (accessed on 7 February 2025).

Han, Jining, Geping Liu, and Yuxin Gao. 2023. Learners in the metaverse: A systematic review on the use of roblox in learning. *Education Sciences* 13: 296. [CrossRef]

Hildebrandt, Mireille. 2018. Algorithmic Regulation and the Rule of Law. *Philosophical Transactions of the Royal Society A* 376: 20170355. [CrossRef]

Hildebrandt, Mireille. 2020. *Law for Computer Scientists and Other Folk*. Oxford: Oxford University Press.

Hinduja, Sameer, and Justin W. Patchin. 2024. Metaverse risks and harms among US youth: Experiences, gender differences, and prevention and response measures. *New Media & Society*, 1–22. [CrossRef]

Hine, Emmie. 2023. Content moderation in the metaverse could be a new frontier to attack freedom of expression. *Philosophy & Technology* 36: 43. [CrossRef]

INEQE. 2025. Roblox: A Parents Guide to Protecting Children from Harmful Content. Available online: https://ineqe.com/2022/01/19/roblox-parents-guide-and-age-restrictions/ (accessed on 12 February 2025).

Jang, Yujin, and Youngmeen Suh. 2024. Cyber sex crimes targeting children and adolescents in South Korea: Incidents and legal challenges. *Social Sciences* 13: 596. [CrossRef]

JetLearn. 2024. Roblox Statstics: Users, Growth and Revenue. Available online: https://www.jetlearn.com/blog/roblox-statistics#:~:text=Roblox%20has%20an%20estimated%20380,the%20United%20States%20and%20Canada (accessed on 15 January 2025).

Kang, Young-Joo, Ui-Jun Lee, and Saerom Lee. 2024. Who makes popular content? Information cues from content creators for users' game choice: Focusing on user-created content platform "Roblox". *Entertainment Computing* 50: 100697. [CrossRef]

Karapatakis, Andreas. 2025. Metaverse crimes in virtual (Un)reality: Fraud and sexual offences under English law. *Journal of Economic Criminology* 7: 100118. [CrossRef]

Kim, Kang-Ho, and Dae-Woong Rhee. 2022. The necessity of content development research for metaverse creators—Based on the analysis of roblox and domestic academic research. *Journal of Korea Game Society* 22: 81–88. [CrossRef]

Kim, Soyeon, and Eunjoo Kim. 2023. Emergence of the Metaverse and Psychiatric Concerns. *Journal of the Korean Academy of Child and Adolescent Psychiatry* 34: 215–25. [CrossRef] [PubMed] [PubMed Central]

King Law. 2025. Is Roblox Safe for Kinds? Available online: https://www.robertkinglawfirm.com (accessed on 12 December 2024).

Kitchin, Rob. 2017. Thinking critically about and researching algorithms. *Information, Communication & Society* 20: 14–29. [CrossRef]

Kou, Yubo, and Xinning Gui. 2023. Harmful design in the metaverse and how to mitigate it: A case study of user-generated virtual worlds on roblox. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. Edited by Daragh Byrne, Nikolas Martelaro, Andy Boucher, David Chatting, Sarah Fdili Alaoui, Sarah Fox, Iohanna Nicenboim and Cayley MacArthur. New York: ACM, pp. 175–88.

Kou, Yubo, Yingfan Zhou, Zinan Zhang, and Xinning Gui. 2024. The ecology of harmful design: Risk and safety of game making on a metaverse platform. In *Designing Interactive Systems Conference*. Edited by Anna Vallgårda, Li Jönsson, Jonas Fritsch, Sarah Fdili Alaoui and Christopher A. Le Dantec. New York: ACM, pp. 1842–56.

Kumar, Harish. 2024. Virtual worlds, real opportunities: A review of marketing in the metaverse. *Acta Psychologica* 250: 104517. [CrossRef] [PubMed]

Kumar, Vidhya Lakshmi, and Mark A. Goldstein. 2020. Cyberbullying and Adolescents. *Current Pediatrics Reports* 8: 86–92. [CrossRef]

Langvardt, Kyle. 2020. Regulating Platform Architecture. *Georgetown Law Journal* 109: 1353–88.

Lee, Lik-Hang, Zijun Lin, Rui Hu, Zhengya Gong, Abhishek Kumar, Tangyao Li, Sijia Li, and Pan Hui. 2021. When creators meet the metaverse: A survey on computational arts. *arXiv* arXiv:2111.13486.

Legal Information Institute. 2002. Ashcroft V. Free Speech Coalition (00-795) 535 U.S. 2002, vol. 234. Available online: https://www.law.cornell.edu/supct/html/00-795.ZO.html (accessed on 2 February 2025).

Livingstone, Sonia, and Amanda Third. 2017. Children and young people's rights in the digital age: An emerging agenda. *New Media & Society* 19: 657–70.

Mancuso, Ilaria, Antonio Messeni Petruzzelli, Umberto Panniello, and Chiara Nespoli. 2024. A microfoundation perspective on business model innovation: The cases of roblox and meta in metaverse. *IEEE Transactions on Engineering Management* 71: 12750–63. [CrossRef]

Masi, Vincenzo De, Qinke Di, Siyi Li, and Yuhan Song. 2024. The metaverse: Challenges and opportunities for AI to shape the virtual future. Paper presented at 2024 IEEE/ACIS 27th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Beijing, China, July 5–7; pp. 31–38.

Mozilla Foundation. 2022. YouTube Regrets Report. Available online: https://foundation.mozilla.org/ (accessed on 2 February 2025).

Mujar, Jose Miguel A., Denise Rhian R. Partosa, Leeann Kyle J. Porto, Derik Connery F. Guinto, John Roche Regero, and Ardrian D. Malangen. 2024. Perspective of senior high school students on the benefits and risk of playing roblox. *American Journal of Open University Education* 1: 26–35.

Nagyova, Iveta. 2024. Leveraging behavioural insights to create healthier online environment for children. *European Journal of Public Health* 34: ckae144.49. [CrossRef]

OFCOM. 2023. Online Safety Act 2023: OFCOM's Powers and Enforcement Framework. UK Government. Available online: https://www.legislation.gov.uk/ukpga/2023/50 (accessed on 12 April 2025).

Park, Daehee, and Jeannie Kang. 2022. Constructing data-driven personas through an analysis of mobile application store data. *Applied Sciences* 12: 2869. [CrossRef]

Patchin, Justin, and Sameer Hindujaa. 2020. Tween Cyberbulling Report. Cyberbulling Research Center. Available online: https://www.developmentaid.org/api/frontend/cms/file/2022/03/CN_Stop_Bullying_Cyber_Bullying_Report_9.30.20.pdf (accessed on 1 April 2025).

Peloso, Caroline. 2024. The metaverse and criminal law. In *Research Handbook on the Metaverse and Law*. Edited by Larry A. DiMatteo and Michel Cannarsa. Cheltenham: Edward Elgar Publishing, pp. 350–60.

Reuters. 2024. UK Won't Change Online Safety Law as Part of US Trade Negotiations. Reuters. Available online: https://www.reuters.com/world/uk/uk-wont-change-online-safety-law-part-us-trade-negotiations-2025-04-09/ (accessed on 15 March 2025).

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.

Roblox Blog. 2021. Keeping Roblox Safe and Civil Through AI and Human Review. Available online: https://en.help.roblox.com/hc/en-us/articles/360029134331-Roblox-Blog (accessed on 12 December 2024).

Roblox Corporation. 2023. Roblox Transparency Report—Trust & Safety. Available online: https://corp.roblox.com/trust-safety/transparency-report-2023/ (accessed on 26 March 2025).

Roblox Corporation. 2024. Q4 Shareholder Letter. Available online: https://ir.roblox.com (accessed on 2 March 2025).

Roblox Corporation. 2025. About Us. Available online: https://corp.roblox.com/ (accessed on 1 April 2025).

Schulten, K. 2022. Roblox and the Risks of Online Child Exploitation. *The New York Times*, August 19. Available online: www.nytimes.com (accessed on 27 December 2024).

Shen, Haiyang, and Yun Ma. 2024. Characterizing the developer groups for metaverse services in Roblox. Paper presented at 2024 IEEE International Conference on Software Services Engineering (SSE), Shenzhen, China, July 7–13; pp. 214–20.

Siapera, Eugenia. 2021. AI content moderation, racism and (de)coloniality. *International Journal of Bullying Prevention* 4: 55–65. [CrossRef]

Singh, Shubham. 2025. How Many People Play Roblox in 2025. Available online: https://www.demandsage.com/how-many-people-play-roblox/?utm_source=chatgpt.com (accessed on 25 March 2025).

Solum, Lawrence B. 1992. Legal Personhood for Artificial Intelligences. *North Carolina Law Review* 70: 1231. Available online: https://ssrn.com/abstract=1108671 (accessed on 12 February 2025).

The Guardian. 2024. Pushing Buttons: With the Safety of Roblox Under Scrutiny, How Worried Should Parents Be? Available online: https://www.theguardian.com/games/2024/oct/16/pushing-buttons-roblox-games-for-children (accessed on 2 February 2025).

The Jewish Chronicle. 2022. Children's Game Roblox Features Nazi Death Camps and Holocaust Imagery. Available online: https://www.thejc.com/news/childrens-game-roblox-features-nazi-death-camps-and-holocaust-imagery-ddzzz1lg (accessed on 7 February 2025).

TikTok. 2023. Transparency Report. Available online: https://rmultimediafileshare.blob.core.windows.net/rmultimedia/TikTok%20-%20DSA%20Transparency%20report%20-%20October%20to%20December%202023.pdf (accessed on 11 November 2024).

UK Government. 2023. Online Safety Bill Factsheet. Available online: https://www.gov.uk/ (accessed on 12 April 2025).

Van Hoeyweghen, Sarah. 2024. Speaking of Games: AI-Based Content Moderation of Real-Time Voice Interactions in Video Games Under the DSA. *Interactive Entertainment Law Review* 7: 30–46. [CrossRef]

Wang, Yuntao, Zhou Su, Ning Zhang, Dongxiao Liu, Rui Xing, Tom H. Luan, and Xuemin Shen. 2022. A Survey on Metaverse: Fundamentals, Security, and Privacy. *IEEE Communications Surveys & Tutorials* 25: 319–52. [CrossRef]

Whittle, Helen, Hamilton-Giachritsis Catherine, and Collings Beech. 2013. A review of online grooming: Characteristics and concerns. *Aggression and Violent Behavior* 18: 62–70. [CrossRef]

Wired. 2021. On Roblox, Kids Learn It's Hard to Earn Money Making Games. Available online: https://www.wired.com/story/on-roblox-kids-learn-its-hard-to-earn-money-making-games/ (accessed on 10 March 2025).

YouTube Help. 2023. Overview of Content ID. Available online: https://support.google.com/youtube/ (accessed on 1 April 2025).
Zhang, Zinan, Sam Moradzadeh, Xinning Gui, and Yubo Kou. 2024. Harmful design in user-generated games and its ethical and governance challenges: An investigation of design co-ideation of game creators on roblox. *Proceedings of the ACM on Human-Computer Interaction* 8: 1–31. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.