Call: HORIZON-CL3-2022-FCT-01
Topic: HORIZON-CL3-2022-FCT-01-03
Funding Scheme: HORIZON Research and Innovation Actions

Countering Extremism in Digital Gaming Spaces

Content Moderation Issues and Measures

Grant Agreement no.: 101121345
Project Title: Gaming Ecosystem as a Multilayered Security Threat
Contractual Submission Date: 30.04.2025
Actual Submission Date: 30.04.2025
Responsible partner: P02: Agency for Security Research, Criminology and Criminal Policy

Funded by the European Union

GEMS Project has received funding from the European Union's Horizon Europe's research and innovation programme under grant agreement no. 101121345

Grant Agreement No.	101121345
Project Full Title	Gaming as a Multilayered Security Threat

Deliverable number	D4.3
Deliverable title	Countering Extremism in Digital Gaming Spaces: Content Moderation Issues and Measures
Type	R
Dissemination level	PU
Work package number	WP4
Work package leader	AgSKK - Prof. Dr. Dominic Kudlacek
Authors	Prof. Dr. Dominic Kudlacek, Dr. Jana Kudlacek, Lennard John, Tihomir Vrdoljak
Keywords	P/CVE; gaming; awareness

Contents

Abstract

This report explores the evolving challenges and approaches to content moderation within global gaming ecosystems, focusing on detecting and preventing radicalization. Gaming platforms have increasingly become complex social spaces, posing distinct governance challenges due to their immersive, real-time, and transnational nature. The study employs structured desk research, drawing from academic literature, policy documents, industry reports, and practical case studies to systematically map relevant stakeholders, assess regulatory and co-regulatory frameworks, and evaluate platform moderation strategies.

A comparative analysis highlights regional variations, particularly between the European Union, the United States, and East Asia, regarding regulatory responses, corporate policies, and enforcement practices. Special attention is given to the tension between safeguarding freedom of expression and ensuring user safety and the risks of content migration to less regulated platforms. Despite notable advancements, including the Digital Services Act in the EU and sector-specific initiatives such as trusted flagger programs, significant challenges remain in adapting moderation practices to new technologies, encrypted communication channels, and decentralized gaming environments.

The report concludes with policy recommendations aimed at enhancing the resilience and effectiveness of content moderation strategies. Emphasis is placed on the need for multi-stakeholder cooperation, context-sensitive regulatory approaches, technological innovation aligned with human rights standards, and the proactive empowerment of gaming communities to counter extremist narratives.

Funded by
the European Union
Grant agreement no. 101121345

1 Introduction

Digital gaming platforms have rapidly evolved from isolated entertainment media into highly networked social ecosystems. With millions of daily users engaging in multiplayer interactions, live-streaming, and user-generated content creation, gaming environments occupy a central role in contemporary digital culture. However, this expansion has also introduced significant governance challenges. Beyond traditional concerns such as toxicity and harassment, gaming platforms increasingly face issues related to the dissemination of hate speech, extremist narratives, and radicalization efforts. These dynamics are amplified by specific features of gaming spaces, including immersive participation, pseudonymity, cross-border accessibility, and the cultural normalization of transgressive behaviors (Wallner et al., 2025).

Against this backdrop, questions of content moderation, user protection, and safeguarding democratic values have become salient not only for platform operators but also for policymakers, researchers, and civil society actors. The task is further complicated by balancing effective intervention against harmful content with protecting fundamental rights, particularly freedom of expression, and preserving the participatory nature of gaming cultures.

This report systematically examines the issues and measures related to content moderation in gaming environments worldwide, focusing on countering radicalization risks. It aims to map relevant stakeholders, critically analyze regulatory and co-regulatory frameworks, evaluate platform policies and moderation technologies, and explore the effectiveness of practical interventions. The comparative dimension covers regions outside and within the European Union, enabling a differentiated understanding of varying legal traditions, institutional capacities, and societal expectations.

Methodologically, the study relies on structured desk research, synthesizing academic research, policy papers, industry transparency reports, and civil society initiatives. Emphasis is placed on identifying patterns of convergence and divergence across regional contexts and highlighting promising practices and persistent challenges.

The report provides evidence-based recommendations to strengthen content moderation strategies while preserving open, inclusive, and rights-respecting digital gaming spaces by integrating comparative legal analysis, policy evaluation, and a critical assessment of platform governance practices. In doing so, it seeks to contribute to the broader discourse on safeguarding democratic values in an increasingly interconnected digital society.

2 Methodology

The present study employs a structured desk research methodology to systematically gather, evaluate, and synthesize existing knowledge and practical measures related to content moderation and the prevention of radicalization within global gaming ecosystems. Emphasis is placed on practice-oriented reports, policy papers, and academic research. The geographical coverage encompasses both the European Union and regions outside the European Union to maintain a comparative perspective. The general approach was discussed with partners in the GEMS consortium to ensure uniformity with analyses of other relevant questions (Feta &

Armakolas, 2024; Halilovic-Pastuovic et al., 2024; Kudlacek et al., 2025a, 2025b; Moonshot, 2024).

The research process was initiated by establishing a comprehensive set of keywords and search combinations tailored to the project's objectives. These included terms such as "gaming radicalization“, "extremism prevention gaming“, "youth protection gaming platforms“, "content moderation gaming“, "platform policies extremist content“, "gaming trust and safety“, and related variations. The searches were conducted iteratively, a methodological approach that was instrumental in allowing for the refinement of the keyword set. This refinement was based on insights from the study, including recent developments and context-specific trends.

Information was systematically collected from multiple source categories. A comprehensive review of the extant literature reveals that academic research, including peer-reviewed journal articles, research papers, and scientific conference proceedings, has provided theoretical frameworks, empirical data, and analytical models. The understanding of regulatory and strategic approaches was informed by policy papers from governmental bodies, intergovernmental organizations and international agencies. Practice reports, including industry whitepapers, NGO publications, and reports from civil society initiatives, offered insights into the real-world implementation of moderation and prevention strategies.

The preliminary investigation was executed through an open web search and academic databases, encompassing EBSCOhost, Web of Science, and Google Scholar. In select instances, reputable news outlets with investigative coverage on gaming and radicalization topics were also consulted. The sources were classified according to their inherent nature, encompassing academic, policy, and practice categories. Additionally, the geographical scope was categorized as global, regional, or country-specific, and the relevance to core research themes was assessed, including moderation, prevention, stakeholder engagement, and regulatory measures. A source evaluation matrix was used to determine credibility, recency, and methodological soundness; however, the limited number of academically sound analyses led to a relatively broad inclusion of resources.

To ensure a robust and balanced analysis, the research approach emphasized comparative evaluation, identifying similarities and differences in strategies across countries and platforms. The study incorporated critical reflection by assessing tensions between security goals, freedom of expression, and user engagement. Moreover, the synthesis of best practices constituted a pivotal component, elucidating effective models of moderation and prevention.

Limitations inherent in desk research include potential language barriers, selective publication biases, and the dynamic nature of digital ecosystems. However, concerning the aim of this analysis, the combined examination of scientific papers and broader sources appears to offer the most sensible solution. The resulting recommendations aim to promote the development of more effective content moderation and prevention strategies in the global gaming sector.

3 Key Stakeholders in Combating Radicalization in Gaming

The effective moderation of extremist content in gaming environments relies on the coordinated actions of a wide range of stakeholders. Governments, platforms, developers, civil society actors, and gaming communities hold distinct but complementary responsibilities in shaping safer digital ecosystems. While governments possess formal regulatory authority, their actions alone are insufficient; a comprehensive response requires the active engagement of all actors across the gaming landscape (Wallner et al., 2025).

Government Regulators and Law Enforcement

Governments define the parameters within which platforms operate. These entities possess the authority to prohibit extremist content, allocate financial resources to prevention programs, and encourage companies to implement more stringent content moderation measures. Nevertheless, numerous governments have exhibited a slow response in acknowledging games as environments susceptible to radicalization. In the United States, for instance, the Senate Judiciary Committee only in 2023 called on gaming companies to explain how they address extremist recruitment (U.S. Senate Committee on the Judiciary, 2023). The Federal Bureau of Investigation (FBI) and the Department of Homeland Security (DHS) have initiated the dissemination of threat intelligence to digital platforms. However, a recent audit has highlighted the absence of strategic coordination in this process. Implementing legal authority alone is insufficient; governments must also proactively promote international cooperation and encourage gaming platforms to integrate proactive moderation standards into their operations (United States Government Accountability Office, 2024).

Gaming Platforms and Publishers

Predominant gaming platforms (e.g., Steam, Xbox Live, and PlayStation Network) control extensive virtual environments. However, these platforms have frequently exhibited a delay in addressing extremist risks, a tendency that contrasts with the promptness shown by social media companies in dealing with similar issues (U.S. Senate Committee on the Judiciary, 2023). For an extended period, digital distribution platforms such as Steam did not implement explicit bans on extremist groups, thereby enabling the proliferation of hate networks. A recent report by the Anti-Defamation League (ADL) has revealed a concerning trend regarding extremist and antisemitic activities on the gaming platform Steam (Anti-Defamation League, 2024). Specific platforms have initiated remedial measures in response to mounting public and political scrutiny. For instance, Roblox has recently implemented a policy prohibiting content that glorifies terrorism. However, the implementation of these practices exhibits significant variability. Most gaming services continue to show deficiencies in key areas, namely the absence of transparency reports and the dearth of dedicated trust and safety teams. These elements have become the norm in other sectors. Absent persistent investment in cultivating moderation staff, implementing artificial intelligence tools, and forging partnerships with external experts, even meticulously crafted policies are destined to fall short in safeguarding players. At the same time, platforms must balance moderation efforts with preserving open, vibrant player communities – a task that can trigger resistance if perceived as excessive control.

Game Developers and Designers

Developers are responsible for shaping the player experience from its inception. The decisions regarding chat systems, reporting tools, and community rules can influence whether games become safe environments or breeding grounds for hate. The concept of "safety by design“, which involves incorporating moderation features into the gameplay rather than adding them retroactively, is gaining traction in the field. Riot Games' "Summoner's Code" is an early example of using onboarding processes to foster positive behavior. Furthermore, developers play a pivotal role in curbing the misuse of user-generated content, such as displaying extremist symbols in custom avatars or maps, by implementing intelligent design decisions and integrating comprehensive review mechanisms. Developers should be encouraged to perceive moderation not as an afterthought, but as a design challenge with business benefits, such as player retention and brand protection (Olaizola Rosenblat & Barrett, 2023).

Civil Society Organizations (CSOs) and Researchers

Non-governmental organizations (NGOs), academic institutions, and independent researchers fulfill the role of critical watchdogs and advisors. Organizations such as the ADL and the Extremism and Gaming Research Network (EGRN) provide data on hate trends and offer recommendations to companies on how to respond (Extremism and Gaming Research Network, 2025). Researchers have identified a pattern of extremist groups migrating into gaming spaces after being banned in other online spaces. Civil society also assists in addressing deficiencies in moderator training by developing resources to aid gaming companies in recognizing emerging extremist codes and behaviors. Furthermore, CSOs advocate for transparency, compelling companies to disseminate public reports on moderation outcomes and to collaborate with independent audits. The risk to industry self-regulation from complacency is heightened without external oversight.

Gaming Communities and Influencers

Players themselves are at the forefront of content moderation. Gamers are the primary targets of extremist trolling and recruitment attempts. The reporting of these incidents and the level of trust in the moderation process can significantly impact the success of any anti-extremism effort. Empowering players to act as active bystanders has been demonstrated to be a potent instrument. Initiatives offering reporting incentives or accentuating positive community conduct have exhibited efficacy in other digital domains. Individuals such as influencers, streamers, and esports personalities are assuming an increasingly prominent role in this regard. By articulating their opposition to hate, they can contribute to the establishment of novel community norms. However, without substantial platform support, community-driven moderation alone cannot adequately address the challenges posed by these issues. Platforms must develop systems that facilitate reporting such content, ensure visibility, and explicitly communicate that extremist behavior is unacceptable. Many players may remain sceptical about reporting mechanisms, fearing unjust suspensions or the misuse of moderation tools for personal disputes (Wallner et al., 2025).

Funded by the European Union
Grant agreement no. 101121345

Conclusion

The content moderation process in the gaming context is not a solitary or monolithic endeavor. The unique and complementary roles of various entities in the gaming landscape should be acknowledged, including platforms, developers, governments, civil society, and players. Only through concerted efforts can extremist narratives be effectively challenged in gaming environments. Especially over the past ten years, several scientific research studies have been published (see for an overview: Thompson & Lamphere-Englund, 2024). The following chapter will therefore present salient issues regarding the role of the aforementioned actors in content moderation.

4 Governing Harmful Content in the Digital Age: Perspectives from Previous Research

The advent of digital platforms has precipitated a substantial democratization of expression, while concomitantly enabling novel modes of harm, including hate speech, extremist propaganda, and targeted harassment. As governments, corporations, and civil society actors contend with regulating online content, the scholarly debate has increasingly revealed the tensions between security imperatives, technological limitations, and the protection of fundamental rights.

From Self-Regulation to Legal Mandates: A Shifting Landscape

Initially, the governance of online platforms was predominantly characterized by self-regulation, with companies developing voluntary guidelines to address content deemed harmful while circumventing formal legal obligations. As Keller and Leerssen (2020) have observed, this hybrid model, which combines compliance-driven removals with broader community standards, produced significant inconsistencies across platforms and jurisdictions. The evident deficiencies in self-regulation, in conjunction with notable failures to effectively restrict terrorist content, have prompted a transition toward statutory regulation. This transition is exemplified by the European Union's Terrorist Content Online Regulation (TCO) and the Digital Services Act (DSA).

However, these legislative initiatives have even been criticized. Watkin (2024) contends that contemporary regulatory frameworks are disproportionately oriented towards content removal, thereby overlooking the psychosocial risks confronted by human moderators and the extensive societal ramifications engendered by regulatory overreach. Utilizing comparative insights from environmental and occupational safety regulation, she advocates for a social regulation model that prioritizes human rights, social inclusion, and preventative measures.

The issue of proportionality remains paramount. Bromell (2022) offers a cogent argument against the potential pitfalls of an overly zealous regulatory impulse, which, if unchecked, could potentially compromise the fundamental tenet of free speech in liberal democracies. Invoking the classical liberal principle that not all harm justifies legal restriction, he calls for a strong preference for counter-speech and societal resilience over censorship, except where incitement to violence can be established.

Technological Approaches: Promise and Peril

Technological innovation, particularly the application of artificial intelligence (AI) and machine learning (ML), is often presented as a solution for scalable content moderation. Irfan et al. (2024) emphasize AI's transformative capacity in detecting radicalization pathways, optimizing counter-narratives, and enhancing platform reach. However, they also acknowledge the ethical dilemmas inherent in AI deployment, including the risks of algorithmic bias, opaque decision-making, and the reinforcement of existing power asymmetries.

In their 2023 study, Berjawi et al. thoroughly examine machine learning and graph-based methodologies for identifying online radicalization. Their findings indicate that, despite advancements in technological complexity, substantial shortcomings persist in accurately capturing culturally nuanced, context-dependent, and adversarial content. The extensive dynamics surrounding the development of new technological resources over the past months and years highlight the topic's current relevance. Despite the studies' recent nature, one should closely monitor the upcoming developments concerning technological approaches, especially with regard to AI and ML.

Furthermore, the efficacy of recommender systems in promoting extremist ideologies has been a subject of considerable scrutiny. Whittaker et al. (2021) demonstrate that engagement-optimized algorithms can amplify extremist material, creating echo chambers and accelerating radicalization dynamics. These findings provide empirical evidence supporting calls for transparency obligations and algorithmic accountability as integral components of content governance.

Gaming Platforms and the Emergence of New Radicalization Vectors

A particularly striking development is the emergence of online gaming ecosystems as sites of extremist mobilization, constituting the GEMS project's necessity (Feta & Armakolas, 2024; Halilovic-Pastuovic et al., 2024; Kudlacek et al., 2025a, 2025b; Moonshot, 2024). Drawing from the foundational studies of "Gamergate" and the culture wars surrounding gaming, recent research by Davey (2021) and Lamphere-Englund & White (2023) has documented the presence of far-right networks across prominent gaming platforms such as Steam, Discord, and Twitch.

As Bhatt and Mantua (2023) emphasize, these environments function less as deliberate recruitment hubs and more as spaces for bonding, cultural normalization, and the diffusion of extremist narratives. Extremist actors capitalize on gaming communities' affordances, anonymity, shared identity, and social cohesion to fortify ideological commitments.

Consequently, the evidence presented herein complicates simplistic narratives of direct grooming via gaming. However, as argued by Lamphere-Englund and White (2023), gaming spaces function as adjacent zones where extremist ideologies can flourish under the guise of ordinary leisure activities. This necessitates a more nuanced, community-centric approach to harm mitigation, transcending the mere removal of content and embracing proactive community management and resilience-building strategies.

Towards Situated and Reflexive Frameworks

A significant critique of contemporary moderation practices is their dearth of contextual sensitivity. As posited by Wilson and Land (2021), content moderation policies, particularly those on hate speech, frequently abstract messages from their socio-political contexts, resulting in overbroad and counterproductive regulation. Their work underscores the imperative of situating speech acts within their relational, historical, and power-laden contexts when assessing harm. Concurrently, Gruber and colleagues (2023) highlight the need for effective organizational change, including alterations in content governance regimes, to encompass cognitive burdens, individual differences, and systemic inertia considerations. A purely technocratic approach to moderation can potentially alienate users, foster distrust, and ultimately undermine the legitimacy of governance efforts.

Scholars such as Mattheis and Kingdon (2023) or Bovenzi (2024) have admitted that emerging decentralized platforms and metaverse environments will precipitate a fundamental challenge to prevailing regulatory methodologies. Consequently, implementing moderation strategies must be predicated on anticipating shifts in platform architectures, the strategy employed by actors, and the evolving communicative norms.

Conclusion: Towards Ethical, Adaptive, and Multi-Stakeholder Governance

The extant literature on the subject offers several critical insights. Firstly, one should acknowledge the limitations of purely technological solutions. Moderation must be deeply contextual, rights-respecting, and participatory. Secondly, regulatory interventions must judiciously balance security concerns with the preservation of democratic discourse and the mitigation of systemic bias. Thirdly, the emergence of digital subcultures, particularly within gaming and decentralized platforms, necessitates the development of customized, anticipatory strategies firmly rooted in community engagement.

Consequently, the development of future governance frameworks must prioritize the integration of technological innovation with ethical safeguards, foster collaborative efforts across diverse sectors, and nurture resilient digital communities that possess the capacity to resist the appeal of extremist narratives. As Bromell (2022) persuasively emphasizes, the challenge entails not only the regulation of speech but also the cultivation of the democratic conditions necessary for pluralistic societies to flourish.

5 Comparative Approaches Across Regions

Approaches to content moderation and anti-radicalization in gaming vary extensively around the world. The range of regulatory approaches is broad, encompassing a spectrum from pronounced centralization of government control in certain nations to instances of self-regulation within industry and civil-society partnerships in others. In the following examination, the several gaming regions are analyzed to describe their regulatory responses, platform policies, and moderation strategies for countering extremism in games.

5.1 European Union

The European Union has developed an integrated and multi-level approach to content moderation and the prevention of radicalization online, combining binding regulation, co-regulatory frameworks, and multi-stakeholder cooperation. The Digital Services Act is central to this effort, establishing extensive obligations for online platforms, including those relevant to gaming ecosystems. Under the DSA, huge platforms are required to assess and mitigate systemic risks such as disseminating illegal hate speech and extremist content, implementing effective notice-and-action mechanisms, and ensuring transparency in their moderation practices. These provisions are designed to improve content governance and safeguard fundamental rights, emphasizing freedom of expression and access to information (European Commission, 2022).

Complementing the DSA, the Code of Conduct on Countering Illegal Hate Speech Online, initiated in 2016, continues to play a pivotal role. Although initially voluntary, it has been progressively formalized within the EU's co-regulatory model, influencing platform policies towards more systematic and prompt responses to illegal hate speech (European Commission, 2025). Monitoring reports indicate that participating companies, including gaming-adjacent platforms, have substantially improved their removal rates. Nevertheless, concerns persist regarding the uniformity of implementation across different services and the potential chilling effects over-removal may have on legitimate discourse, particularly in complex environments such as multiplayer gaming.

The Regulation on addressing the dissemination of terrorist content online (Regulation 2021/784) mandates that platforms remove terrorist content within one hour of notification by competent authorities, under the threat of significant financial penalties. Europol's Internet Referral Unit (IRU) supports this process by identifying and flagging terrorist content, including extremist propaganda that may circulate within gaming-related forums and communication channels (Europol, 2022). This regulatory architecture aims to ensure a swift and coordinated European response to the risks associated with the viral spread of radicalizing materials, particularly in spaces that attract younger audiences.

Member states complement EU-wide regulations at the national level with domestic measures adapted to their specific legal traditions and policy priorities. Germany exemplifies a proactive regulatory stance: Its Network Enforcement Act (NetzDG), introduced in 2018, pioneered legally binding obligations for platforms to remove illegal content swiftly, effectively serving as a precursor to the DSA (Zipursky, 2019). Germany also enforces strict prohibitions on extremist symbols in digital media, including games, thereby shaping the content policies of developers and distributors operating within its jurisdiction (Echikson & Knodt, 2018). Beyond regulation, Germany supports preventive initiatives such as "Good Gaming – Well Played Democracy", which engages with gaming communities to counter extremist narratives through digital street work, online interventions, and educational workshops (Amadeu Antonion Stiftung, 2025).

France adopts a similarly assertive approach, emphasizing rapid response mechanisms through its centralized PHAROS platform, which enables citizens to report illegal online content directly to law enforcement (Council of Europe, n.d.). This facilitates swift interventions

and underscores a state-led model of content governance. Recent incidents, such as removing extremist-themed games from distribution platforms following public complaints and political pressure, illustrate the practical efficacy of this model. French authorities leverage existing criminal provisions against hate speech and glorification of terrorism to underpin their interventions, while actively participating in broader EU initiatives to harmonize standards and practices.

Despite these advances, several challenges remain within the European Union. The balance between effective content moderation and the protection of freedom of expression continues to demand careful calibration. Over-removal risks, where legitimate but controversial discourse may be inadvertently censored, pose a persistent concern, particularly given the EU's diverse linguistic and cultural contexts. The DSA's emphasis on due process and transparency, including requirements for platforms to provide users with avenues for appeal and justification of content decisions, seeks to mitigate this risk but requires vigilant enforcement and independent oversight mechanisms to be fully effective.

Moreover, the "platform migration" phenomenon, whereby extremist actors move from mainstream platforms to fringe services with less oversight, complicates moderation efforts. While the EU's strategy to extend obligations proportionally to smaller platforms and foster cross-platform collaboration addresses this challenge partially, significant enforcement and technical capacity gaps remain. Proactive intelligence-sharing, both among member states and with private sector actors, as well as continuous investment in technological tools for detection and analysis, are necessary to maintain the effectiveness of the EU's model.

In addition, the evolving technological landscape presents new challenges. The increasing popularity of decentralized platforms, encrypted communication channels within gaming environments, and emergent technologies such as virtual and augmented reality gaming experiences demand adaptive regulatory responses. Ensuring that existing frameworks remain fit for purpose in light of such developments will be critical to sustaining progress in the coming years.

The EU's approach represents a distinctive model characterized by legal enforceability, structured cooperation with platforms and civil society, and a commitment to upholding fundamental rights even under stringent security imperatives. While not without limitations, the European framework has contributed significantly to raising moderation standards and building systemic resilience against radicalization within online environments, including the increasingly influential sphere of gaming. Continued monitoring, adaptive policy-making, and inclusive stakeholder engagement will be essential to address emerging risks and uphold the delicate balance between security, innovation, and protecting open digital spaces.

5.2 United States

The United States is a prominent global gaming market and a significant hub for major gaming companies. The United States' strategy for regulating extremist content in video games is marked by a combination of industry self-regulation and government oversight, with the latter influenced by robust protections for free expression. There is no overarching legislation that directly regulates online game content (Kosseff, 2019). Instead, private platforms

exercise broad discretion over content moderation by the First Amendment and Section 230 of the Communications Decency Act. This suggests that the primary factors influencing moderation are the game companies' internal policies, frequently initiated in response to public pressure or the potential for reputational damage, rather than being compelled by legal mandates.

Until recently, many U.S.-based gaming platforms had only general rules against "harassment" or illegal activity, without explicitly mentioning extremist organizations or terrorist propaganda. This dynamic shifted as evidence of radicalization within the gaming community became increasingly evident. For instance, Roblox – a prominent platform in the United States – has revised its Community Standards to prohibit content that glorifies mass shootings, terrorist acts, or extremist groups. U.S. lawmakers commended the specific extremism policy of Roblox as a "welcome development“, observing that none of Roblox's major industry peers had yet implemented analogous terrorism-focused regulations (Roblox, n.d.). Concurrently, other companies have adopted more expansive anti-hate stances. Activision-Blizzard (2024) and Riot Games (n.d.) have established codes of conduct that prohibit hate speech. In 2022, Activision-Blizzard reported banning over 500,000 accounts and issuing millions of warnings to players for toxic chat, including racist or extremist language, in Call of Duty (Call of Duty Staff, 2024). Nevertheless, despite the efforts of companies to tighten their Terms of Service, enforcing these terms remains a persistent challenge. As indicated by the findings of a U.S. Senate report, implementing these policies does not guarantee the success of the platforms in question (U.S. Senate Committee on the Judiciary, 2023).

The United States gaming industry has embarked upon several interorganizational initiatives to align its endeavors (Thompson & Lamphere-Englund, 2024; Kowert & Kilmer, 2023a, 2023b). One such example is the Thriving in Games Group (previously known as Fair Play Alliance), a coalition of over 200 game companies formed to share best practices on reducing toxicity and promoting healthy communities. While the Fair Play Alliance has historically concentrated on issues such as harassment and cheating, experts have urged the organization to adopt a more pronounced stance on extremism (Thriving in Games Group, n.d.). Similarly, U.S.-based tech consortia such as the Global Internet Forum to Counter Terrorism (GIFCT), initially established by prominent social media companies, have started incorporating gaming-related entities. Notably, only a limited number of gaming companies have become members of GIFCT thus far, indicating the potential for increased engagement (Olaizola Rosenblat & Barrett, 2023). GIFCT facilitates the exchange of digital fingerprints (hashes) of terrorist content among firms, thereby assisting each other in identifying and removing known extremist material (Global Internet Forum to Counter Terrorism, 2024). Academic conferences at the Game Developers Conference (GDC) also highlight extremism in games (Game Developers Conference, n.d.). The industry is thus gradually undergoing a “belated reckoning with extremism“ (Olaizola Rosenblat & Barrett, 2023), transitioning from a state of denial to one of acknowledgement, recognizing the necessity for explicit standards and enhanced moderation instruments.

The federal government's regulatory oversight of online game content in the United States is restricted; however, authorities have become more vocal and involved by implementing soft pressure and support measures. The escalation in the prominence of violent incidents has been a contributing factor to this heightened level of scrutiny. In the aftermath of the 2022

Buffalo mass shooting, in which the perpetrator had been active on gaming platforms and had even cited games as an influence, U.S. Senate Majority Whip Dick Durbin sent letters to leading game companies – including Microsoft (Xbox), Activision-Blizzard, Riot Games, Epic Games (Fortnite), Take-Two (GTA), Valve (Steam), and Roblox – demanding information on their efforts to combat extremism. In these inquiries, Congress highlighted that the "online video game industry has been slow" to adopt measures that social media companies already use, such as dedicated trust and safety teams and transparency reporting on extremist content. The Senate has also identified specific instances of abuse, including the hosting of militant Nazi and hate content on platforms such as Steam, as well as the use of creative modes in Minecraft and Roblox by extremists to recreate concentration camps or Uyghur detention camps for propaganda purposes. While these letters lacked legal force, they exerted significant public pressure on companies to respond and improve (U.S. Senate Committee on the Judiciary, 2023).

In addition, U.S. federal agencies have initiated measures to address this issue. In 2022, the Department of Homeland Security (DHS) allocated nearly $700,000 in grants to university researchers. These grants were intended to support the study of radicalization in gaming and the development of prevention strategies (Gault, 2022). The FBI and the DHS have been collaborating with gaming companies to exchange threat intelligence regarding domestic violent extremists. However, a 2024 review by the United States Government Accountability Office found that these efforts were still ad hoc and recommended developing a more structured strategy (United States Government Accountability Office, 2024). Notably, U.S. law enforcement has intervened in specific cases. In instances where credible threats or criminal plots are identified through gaming activities, law enforcement officials can take action, such as the arrest of individuals who are planning acts of violence via in-game chat functions. However, the proactive regulation of game spaces is primarily delegated to private entities due to resource limitations and concerns about civil liberties. American civil society also performs a regulatory-adjacent function. Organizations such as the ADL routinely provide policymakers with briefings and even testify in Congress (Anti-Defamation League, 2024).

The United States relies on voluntary platform moderation guided by government encouragement rather than direct censorship. The primary objective is encouraging the industry to self-regulate through Congressional oversight, inter-agency collaboration, and funded research. This approach endeavors to safeguard free expression. The government does not prohibit extremist speech in games, unless it constitutes criminal incitement or threat. Nevertheless, it acknowledges a state interest in curbing pathways to terrorism. The efficacy of this U.S. model and the trade-offs with free speech will be examined in a subsequent section of this report. At present, it is evident that U.S. companies have recognized the necessity to address the issue of "extremists and other malicious actors" leveraging gaming communities for radicalization and recruitment purposes. They recognize the imperative to take more proactive measures to prevent the exploitation of their platforms for such activities.

5.3 Canada

Canada and the United States share numerous similarities in their approaches to regulating game content. However, Canada has demonstrated a notable initiative in allocating funds to

preventive research and interventions. Public Safety Canada's Community Resilience Fund has provided financial support to research projects that examine extremism in gaming and develop counter-radicalization resources (Public Safety Canada, 2024). One such project, led by the Royal United Services Institute and the Extremism and Gaming Research Network, discussed the potential role of gaming communities in fostering radicalization and informing safety measures (Royal United Services Institute for Defence and Security Studies, 2025). This suggests that the Canadian government has acknowledged the necessity of addressing the issue of violent extremism in gaming spaces, irrespective of the absence of specific incidents within Canada. Law enforcement agencies within Canada, such as the Royal Canadian Mounted Police (RCMP), are also engaged in international initiatives to monitor extremist networks that traverse gaming and social media platforms. Many Canadian gamers utilize U.S. platforms, making them susceptible to adopting enhancements in U.S. industry policy. In Quebec, there has been discourse surrounding the potential for leveraging the province's robust community policing methodologies to engage youth at risk within gaming cafés. This initiative aims to transform these spaces into avenues for positive social interaction, thereby counteracting the potential for extremist recruitment. Canada's strategy is characterized by a focus on engagement and education, as opposed to the use of regulation, in its efforts to address the issue of radicalization among Canadian gamers. This approach involves collaboration with industry and community partners to identify and address potential radicalization before it becomes established within the gaming community (Public Safety Canada, 2025).

5.4 East Asia

China

In contrast, China employs a markedly divergent approach, characterized by stringent government control and censorship in the gaming sector to enforce content standards defined by the state. The Chinese government uses a rigorous regulatory framework for video games, encompassing pre-publication vetting and real-time content monitoring. All games should obtain licensing from the relevant regulatory authorities. These authorities possess the authority to impose bans on content deemed to be politically sensitive, terrorist, or extremist in nature, or to demand changes to such content. For instance, the depiction of "ethnic or religious hatred" or separatist ideologies is strictly forbidden. According to established government regulations, companies must implement measures that involve the censorship of chat and user-generated content. An illustration of this approach can be observed in the policies of Tencent, the largest gaming company in China: Tencent has established a comprehensive set of guidelines for streamers and players to promote "China's social values." A key component of these guidelines involves prohibiting "promoting or publishing content" related to terrorism, extremism, or politically sensitive subjects. The regulations also encompass any conduct that has the potential to "cause disputes or adverse social impact“, thereby affording censors a considerable degree of discretion (Chalk, 2019; Freedom House, 2022). In practice, Chinese gaming platforms collaborate with government censors to filter text chat for banned keywords (e.g., names of extremist groups or dissenting political slogans) and to surveil user communications. The enforcement of these regulations is rigorous; authorities have shut down live streams and even entire games that have contained disallowed content.

The Chinese government's primary objective is to prevent disseminating any content that could compromise social stability or undermine the authority of the Communist Party (Human Rights Watch, 2019). This encompasses the prohibition of jihadist propaganda, incitement to violence, and all forms of extremist ideology. A notable benefit of this model is that open extremist recruitment and hate speech are seldom observed in Chinese game servers. Such content is typically swiftly removed, and users who disseminate it may incur legal consequences. For instance, expressing support for a designated terrorist or separatist group in a Chinese game could result in the suspension of one's account and potentially lead to an interrogation by law enforcement. Nevertheless, the compromise entails a heavily censored environment. Legitimate political discourse and minority viewpoints have been known to be labeled "extremist" and subsequently removed. According to recent analyses, Chinese regulations have been found to conflate extremist content with broader violations of social values. This indicates that censorship extends beyond mere terrorism, encompassing other subjects such as discussions of democracy and human rights, which may also be subject to expunction. The absence of transparency and an appeal process for moderation decisions is compounded by intertwining these decisions with state directives. From the perspective of stakeholders, the government of China stands as the primary actor. Private companies and civil society possess minimal autonomy in their affairs. In terms of internet freedom, companies such as Tencent and NetEase have been observed to play a pivotal role in implementing government policies. Conversely, non-governmental organizations prioritizing internet freedom often operate in exile rather than collaborating with authorities.

China's strategy has proven effective in the purging of overt extremist content. This is mainly attributable to the country's pervasive surveillance and zero-tolerance enforcement policy. However, these measures are accompanied by a severe limitation of freedom of expression and user privacy. This model is distinctive and difficult to replicate in open societies; nevertheless, it exemplifies the most extreme reaches of the regulatory spectrum.

Japan

Japan has a robust and influential gaming culture, yet its regulatory approach to content moderation is relatively lax, particularly in contrast to China's stringent policies. No specific Japanese legislation targets extremist speech in online games. The onus of moderation primarily falls upon gaming companies and community managers. Japan's cultural emphasis on harmony and the avoidance of public conflicts may, indirectly, serve as a deterrent to overt extremist behavior in mainstream gaming spaces. On the other hand, gender-based violence is a documented and serious issue in Japanese digital media and games, particularly within the niche market of adult computer games. These often feature highly sexualized female characters and enable interactions that include simulated sexual violence, including rape scenarios, which critics argue contribute to the normalization of misogyny and sexual aggression (Galbraith, 2017; Díez Gutiérrez, 2014). Attempts to restrict such representations by law have had limited success. The government's involvement has predominantly transpired through the medium of rating boards (e.g., CERO) that possess the authority to impose restrictions or prohibit the dissemination of games characterized by extreme violence or politically sensitive content at the distribution level (Computer Entertainment Rating Organization, 2025). Moreover, the government has exercised its authority through general criminal laws about any individual inciting violence or propagating terrorist propaganda (irrespective of the medium

through which such content is disseminated). In practice, there have been few reports of incidents of organized extremist recruitment via Japanese games. This may be due to Japan's relatively low, although potentially increasing, prevalence of domestic extremist movements and the language barriers faced by international hate groups (Zeyu, 2019; Asano, 2022). Consequently, the content moderation strategies employed in Japan prioritize industry self-governance and community standards. It is evident that prominent Japanese developers, such as Nintendo, Sony Interactive Entertainment, and Square Enix, have instituted terms of service that proscribe harassment, the dissemination of obscene content, and any actions that contravene established laws. This includes, but is not limited to, terrorist activities. For instance, the user guidelines of the PlayStation Network explicitly prohibit hate speech and any activity that Sony determines, in its sole discretion, to be inappropriate (PlayStation, n.d.; Sony Interactive Entertainment, 2025). Although such clauses do not expressly mention extremism, they are broad enough to encompass a wide range of potentially problematic behaviors. In the Japanese gaming community, the prevalence of volunteer moderators in gaming forums and services has been observed to facilitate the swift enforcement of etiquette rules.

A notable feature in Japan is the use of automated filters to block specific terms in chat. For instance, many games in Japan filter slurs or politically offensive terms, and some Japanese online games do not offer open global chat at all, partly to preempt misconduct. A further consideration is the nation's recent legislative initiatives concerning general online hate speech. In 2016, Japan enacted the Hate Speech Act (non-binding), and in 2022, it augmented the penalties for online insults. Theoretically, these legal provisions could be invoked in cases involving the dissemination of extremist hate speech within a gaming context (Higuchi, 2020; Wakabayashi, 2020). However, enforcement of these laws is predominantly focused on cases involving social media platforms. The government collaborates with law enforcement to oversee far-right and terrorist activities on the internet. Japan's approach can be characterized as a moderation strategy facilitated by community culture and comprehensive platform policies, with the government assuming a relatively passive role. The prevailing inclination is toward safeguarding the freedom of user communication, with the caveat that such actions do not transgress legal or ethical boundaries. This stance is characterized by a reluctance to engage in proactive censorship. The fundamental limitation of this laissez-faire approach is that, if extremist content does appear, it relies on users or moderators to flag it. In an increasingly interconnected global landscape, Japanese platforms are not immune to the potential influence of extremist or ultranationalist groups. For instance, global services such as Steam and Discord, which are available in Japan, have been observed to host Japanese-language extremist or ultranationalist groups that may not be immediately apparent or easily detectable. Nevertheless, Japan has not exhibited the same degree of urgency regarding the radicalization of gaming, and as a result, its regulatory response remains limited and industry-driven.

Republic of Korea

The Republic of Korea (hereafter South Korea) adopts an intermediate stance between the approaches of China and Japan. The nation is a democratic society with vibrant gaming and esports communities. However, it also has a documented history of government intervention in media for security reasons. The Republic of Korea has enacted a series of stringent national security laws, most notably the National Security Act, which criminalizes the praise or propagation of the agendas of groups that are regarded as threats to national security, including

those from North Korea and terrorist organizations. Consequently, promoting overt extremist ideologies in any online forum, including gaming platforms, may result in legal repercussions (Global Online Safety Regulators Network, 2024). Concerning the regulation of gaming, South Korea has been recognized for its proactive measures aimed at regulating online gaming behavior. Illustrative of these measures are implementing a "real-name" verification system for specific online services and establishing curfews for underage gamers. While these measures were intended to address gaming addiction and bullying issues, the infrastructure for content oversight is already in place. Large Korean game companies, such as Nexon and NCSoft, have established teams responsible for content moderation (Baek, 2024). Popular games in Korea, including Lineage, MapleStory, and Overwatch, frequently incorporate advanced reporting systems and chat filters to maintain civil discourse among players. The dissemination of extremist content is typically addressed within the broader category of illicit or harmful content, which companies must remove upon receipt of a notice. The Korea Communications Standards Commission is authorized to issue takedown orders for online content infringing on applicable laws or disrupting public order. This authority may encompass game chat or user-generated content if it violates the standards above. In practice, South Korean authorities have prioritized the suppression of online sexual content and cybercrimes (Lee & Lee, 2025). However, they maintain a state of heightened vigilance regarding terrorism issues, a posture informed by historical threats. The following case is illustrative of the point being made: In 2021, media outlets reported that Korean police had investigated an online gaming clan suspected of spreading North Korean propaganda. This investigation demonstrates that law enforcement will act if gaming platforms are misused for extremist ends (though such cases seem to be rare).

South Korea has adopted a moderate strategy characterized by a collaborative regulatory approach. In this strategy, the government establishes specific expectations, requiring gaming providers to respond promptly to user reports of illegal content. Companies are then responsible for implementing these expectations within their operations while subjecting them to government oversight. A notable similarity between Korean and Japanese game communities is the tight management by publishers, particularly in esports titles where player behavior is subject to constant monitoring. A distinctive feature of the Korean cultural landscape is the prevalence of gaming cafés, known as PC bangs. These licensed and monitored physical venues serve as a crucial layer of moderation, ensuring minors are not exposed to potentially harmful content. This oversees public gaming spaces, contributing to a multifaceted approach to safeguarding individuals from undesirable content.

South Korea's strategy integrates industry-driven moderation with the government's willingness to intervene or initiate legal proceedings in instances of extremist activity. It endeavors to strike a balance between the principles of free expression and the imperative of security. To illustrate this point, consider the following scenario. While it is permissible to criticize the government within the context of a game, expressing praise for entities such as ISIS would undoubtedly result in a crackdown under the provisions of terrorism legislation. This approach has proven effective in maintaining the integrity of mainstream Korean gaming environments, as evidenced by the absence of significant incidents related to extremist recruitment. However, its efficacy hinges on the continuous implementation of vigilance measures, a challenge further compounded by online games' transnational nature.

5.5 Middle East and North Africa (MENA)

Several MENA countries exhibit high levels of gaming engagement, a notable example being the Gulf states. In this region, governments frequently implement a form of internet censorship that is also applied to games. For instance, countries such as Saudi Arabia and the United Arab Emirates (UAE), which enforce strict laws against terrorist propaganda and hate speech, do monitor gaming platforms, particularly chat rooms, for such violations. It is a recurring phenomenon that these states frequently impose restrictions, including outright bans, on games that contain content deemed politically sensitive or culturally "immoral" (Freedom House, 2022). Regarding extremist activities, their approach entails two primary concerns: the prevention of jihadist groups (such as ISIS or Al-Qaeda) from recruiting youth through online games, and the management of domestic dissent. During the period of peak ISIS activity in 2015, there were reports of militants utilizing PlayStation Network and other gaming network communication platforms for covert operations (Tassi, 2015). This prompted calls in some Middle Eastern countries for increased surveillance of gaming networks. In response to these developments, several governments have issued advisories urging parents to exercise caution and have collaborated with local telecommunications providers to monitor and analyze suspicious communications (Gordon, 2016). A hard-line approach generally characterizes the regulatory stance in MENA. Content deemed to support extremist activities is considered illicit and promptly removed. Users who engage with such content may face legal consequences, including arrest. For instance, in 2018, a Turkish man was apprehended on charges of utilizing an online shooter game's chat feature to disseminate terrorist propaganda in support of the PKK. Concurrently, regional gaming communities frequently depend on the moderation tools provided by gaming platforms. Popular games such as PUBG and Fortnite employ Arabic-language moderators and filters to identify extremist slurs or slogans (Anti-Defamation League, 2021). A distinctive feature of this phenomenon is the engagement of militant groups in conflict-affected regions in creating rudimentary video games for propaganda and training. However, these games are not hosted on mainstream platforms. A comprehensive analysis reveals that Middle Eastern and North African counter-terrorism strategies are congruent with broader counter-terrorism initiatives. These approaches encompass a high degree of state involvement, the censorship of extremist content, and an emphasis on religious and political conformity. However, these strategies prioritize security over free expression, particularly when such expression conflicts with security priorities.

5.6 Middle and South America

The Latin American gaming industry is experiencing significant growth, accompanied by concerns regarding the potential infiltration of gaming spaces by gang-related and ideologically motivated extremism. A particularly noteworthy example is as follows: It has been noted that Mexican drug cartels have endeavored to recruit young people using online games. In 2021, a Mexican cartel approached a teenager via Grand Theft Auto V's online mode, employing a grooming strategy to prepare her to carry illicit drugs across the border eventually (Brewster, 2022). These incidents have prompted Latin American authorities to prioritize the issue. In response, the Mexican government and other authorities in Central America have initiated public awareness initiatives aimed at cautioning parents regarding the potential recruitment of their children by cartels through gaming. Regarding regulatory measures, Latin American

countries predominantly employ extant legislation to address terrorist or criminal organizations' activities within the digital domain. However, implementing specific policies tailored to the gaming sector remains limited. Brazil, for instance, has enacted comprehensive anti-terror legislation and has previously prohibited specific violent games. However, platform moderation or law enforcement intervention typically addresses extremist speech within the gaming environment when such speech constitutes a credible threat. A salient positive development is the involvement of (international) NGOs such as Tech Against Terrorism and local cybercrime units in training game moderators (for games popular in Latin America) to recognize recruitment tactics (Tech Against Terrorism, n.d.). Furthermore, many Latin American gamers utilize U.S.-based or international servers, such as those offered by Steam or Xbox Live. Consequently, these gamers are indirectly subject to the policies of these companies. A salient challenge in the region is the variability of law enforcement capacity across countries, compounded by the prevalence of multiple security crises in certain states. Consequently, while the government has historically prioritized social media, gaming emerges as a potential new domain of concern. The current response is predicated on a combination of platform policies, typically established by foreign companies, and reactive policing in the limited number of cases that come to light.

5.7 Australia and New Zealand

The Australian approach to content in games is chiefly through its Classification Board, which has the authority to refuse classification (effectively banning) games that promote or incite matters of terrorism or violence (Australian Classification Board, 2020). This has resulted in the prohibition of several games with content deemed unsuitable for all ages. In the context of online interactions, Australia has established an eSafety Commissioner who collaborates with various platforms to remove any content violating the law. Extremist material falls under this purview, and as such, if Australian authorities are alerted to terrorist propaganda circulating in a game's forum or chat, they can issue a takedown notice or facilitate a police investigation. Australia also engages in international cooperation, as evidenced by its participation in the Christchurch Call initiative, which New Zealand and France established in the aftermath of the 2019 Christchurch attack. This initiative aims to encourage technology companies, including gaming firms, to expedite the removal of violent extremist content. New Zealand, profoundly affected by the Christchurch terrorist's gamified livestream (in which he likened his attack to a first-person shooter), has emerged as a proponent of countering extremism in the online sphere (New Zealand Government, 2022). New Zealand's government, though diminutive in stature, has initiated initiatives on digital literacy that encompass the domain of gaming. Additionally, its security services have been known to meticulously monitor extremist forums that occasionally intersect with gaming subcultures. Australia and New Zealand tend to adopt a cooperative regulatory approach. They support global codes of conduct and anticipate that platforms will assume a primary responsibility for content regulation, intervening only when law enforcement is deemed necessary. A balance must be struck between this and their robust traditions of free speech. Any regulation is meticulously delineated to address the threat of terrorism, rather than being applied more broadly to content related to politics.

5.8 Comparison

A common theme is emerging across these regions: Governments outside Europe realize that action is required in the gaming sector. However, the methods employed vary based on the political context. In the context of video games, authoritarian regimes have historically demonstrated a preference for the implementation of direct censorship and surveillance measures. This approach has yielded notable successes in removing specific content, though it has come at the cost of compromising fundamental rights. In contrast, democratic governments have adopted a different approach, emphasizing industry self-regulation, complemented by targeted legislative measures. The effectiveness of these regulatory frameworks varies, reflecting the unique characteristics of each democratic system. In all cases, platform providers and developers remain the first line of defense. The ensuing discussion will address the efficacy of these measures in practice and how various actors balance the imperative to moderate harmful content with the values of freedom of expression and user engagement. Decisive governmental intervention can create safer digital spaces, but if unchecked or unbalanced, it may also curb legitimate dissent and creative expression.

6 Effectiveness of Measures and Balancing Challenges

6.1 Successes and Promising Practices

Despite the ongoing evolution of efforts to counter extremist content in gaming, notable successes have been achieved through interventions that have effectively enhanced the safety of gaming environments. One such case is the rapid response to live-streamed violence. In 2019, the Christchurch terrorist attack in New Zealand (which was live-streamed on Facebook) revealed the need for expedited content removal. Gaming-adjacent platforms have gained valuable insights from this phenomenon. In 2022, when a gunman in Buffalo, NY, attempted to live-stream his attack, Twitch swiftly removed the stream within two minutes of the initial shot (Stelter & Paget, 2022). This expeditious moderation effectively curtailed the video's dissemination, signifying a marked enhancement over previous shortcomings. The findings demonstrate that rigorous monitoring and explicit protocols, constituting a form of "crisis moderation“, can effectively curtail extremist propaganda in real time. Another success story originates from community moderation initiatives. Discord, a popular chat platform utilized by gaming communities, was found to host neo-Nazi groups in the mid-2010s. In the aftermath of the 2017 Charlottesville rally, which was, in part, organized on the platform Discord, the company took decisive action by purging alt-right and extremist servers *en masse*. This crackdown, which was executed in coordination with organizations such as the ADL, was lauded as effective, resulting in the dispersal of numerous extremist networks and their subsequent inability to regroup and operate on the same scale within the platform (Newton, 2017). This incident exemplifies the repercussions when an organization demonstrates a resolute commitment to its established policies.

From the game developers' perspective, design modifications have resulted in favorable outcomes, reducing toxic behavior (which frequently coincides with extremist harassment). Riot Games' initiative to incentivize positive player conduct (by allocating honor points and in-

game rewards for sportsmanship) has resulted in a notable decrease in abusive behavior within the context of League of Legends. By incentivizing pro-social behavior, they indirectly made it more difficult for hateful or extremist conduct to gain a foothold, since community norms shifted toward disapproval of toxicity. In addition, innovative user education has demonstrated potential to address this issue (Kou & Nardi, 2014; Lin, 2015). Some games now incorporate brief tutorials or messages to educate users about reporting bad behavior and respecting others. For instance, Activision's Call of Duty introduced a Code of Conduct pop-up that players must acknowledge, explicitly forbidding hate and extremism. Within a short time, the company reported millions of acknowledgments and a subsequent drop in in-game reports of hate speech, suggesting increased awareness (Activision Blizzard, 2022).

Cross-industry collaboration has also yielded notable achievements. The GIFCT hash-sharing database has facilitated identifying and removing many terrorist propaganda content, including ISIS videos, before their dissemination (GIFCT, 2021). Suppose a user attempts to upload a flag image or manifesto text associated with ISIS in a game forum that utilizes GIFCT data: In that case, the content can be automatically identified and removed. This proactive detection, which has gained widespread implementation on mainstream social media platforms, is now gaining traction within gaming platforms, particularly those owned by major technology companies.

Another area of progress is research-informed interventions. Due to collaborative efforts with academic institutions and non-governmental organizations, certain companies have initiated pilot programs that utilize "redirect" messages, which are defined as counter-narratives. For instance, an online war game may present a public service announcement promoting the rejection of real-world hate if a player searches within the game for specific extremist clan names. This strategy has been adopted from anti-extremism programs on YouTube. Although data concerning gaming-specific outcomes is still emerging, analogous programs have shown success in steering at-risk users towards help and away from extremist content on other platforms (Helmus & Klein, 2018). Consequently, the endeavor has culminated in the attainment of success in the realm of international information sharing. In response to legislators' urging, gaming companies might initiate a more proactive approach to their engagement with law enforcement authorities, particularly when credible threats have been identified.

The most effective measures to date have involved implementing early and active moderation, both human and AI-assisted, establishing explicit community standards with which players are made aware, promoting positive community engagement aimed at marginalizing extremists, and fostering robust cooperation between platforms and external experts or authorities. Although the battle has yet to conclude, these triumphs furnish a model for practical strategies, including prompt action, multifaceted approaches, and a commitment to prioritizing user safety.

6.2 Failures, Gaps, and Limitations

Notwithstanding a modicum of progress, substantial setbacks and persistent constraints have impeded the efficacy of gaming in the effort to counter radicalization. One of the most significant early failings was the industry's inactivity and refusal to acknowledge the issue. For an extended period, extremist factions within the gaming community proliferated with minimal

oversight. In recent years, digital platforms such as Steam have been observed to function as virtual spaces conducive to the proliferation of extremist and hateful content. A comprehensive investigation conducted in late 2024 revealed a substantial presence of such content on the platform (even if the prevalence mentioned in the report seems exaggerated based on methodological considerations). The identified content included various forms of extremist symbolism and rhetoric, such as Nazi swastikas displayed on user profiles, as well as groups that explicitly glorify acts of violence, particularly school shootings (Anti-Defamation League, 2024). The proliferation of extremist content on the Steam platform was partly attributable to the absence of an explicit policy prohibiting such content. Moderators lacked the mandate to remove material that promoted extremist ideology (D'Anastasio, 2024). It was not until mid-2022, when the public was exposed to the issue and a letter was sent by U.S. senators expressing disapproval, that Valve (the parent company of Steam) began to remove some of the most egregious hate groups. The delayed response enabled extremist subcommunities to establish a strong presence and even establish networks on the platform over the years. This case exemplifies a fundamental lesson in policy implementation, underscoring the critical role of regulatory gaps in fostering extremist exploitation. In this context, the absence of a governing rule translates into inaction, a factor that extremists quickly capitalize upon.

Another limitation has been the over-reliance on user reports for moderation. Many gaming companies initially adopted a reactive stance, opting to review content solely if a player submitted a report regarding it. This approach was ultimately deemed to be suboptimal. A paucity of studies has been conducted on this issue; however, preliminary findings indicate that only a small fraction of users report extremist or toxic incidents. There are several potential reasons for this reluctance: Players may fear retaliation, believe their complaints will not be addressed (particularly if they never receive feedback on their reports), or prefer to continue playing without interruption. Moreover, there is a gameplay-associated problem: Players could abuse the report buttons to remove players they do not like or who are better than them. Consequently, a significant proportion of hateful behavior went unaddressed. A 2023 analysis by Modulate itself revealed that conventional reporting mechanisms "addressed only a small fraction of violations", thereby leaving the most egregious instances of misconduct unidentified until the implementation of proactive artificial intelligence monitoring. The delayed adoption of proactive detection measures has enabled the proliferation of extremist rhetoric, which has remained largely unchecked (Takahashi, 2024).

It has been demonstrated that certain moderation attempts have been unsuccessful or have had unintended negative consequences. For instance, content filters can be circumvented. The use of coded language or innocuous terms by extremists to signal their views is a well-documented phenomenon, known as "dogwhistling." Early text filters, which were designed to block obvious slurs, were ineffective in identifying more subtle extremist memes or numerical codes such as "1488" or "14 words." It has been observed that individuals who engage in online trolling have devised innovative methods to circumvent filtering mechanisms by manipulating the spelling of prohibited terms (Mothershaw, 2020). This cat-and-mouse game revealed the limitations of basic technical remedies without continuous updates and context-aware filtering.

There have been several incidents where moderation efforts have been unsuccessful in preventing harm. The Buffalo shooter's transition to radical ideologies was characterized by

active engagement on the communication platform Discord. However, before the attack, his extremist postings on a private Discord server went unreported for months, indicating a failure of detection (though Discord did cooperate with investigators after the fact).

In another instance, members of a neo-Nazi group were known to convene in an online WWII shooter game's voice chat to socialize and recruit. The platform only learned of this when journalists brought it to light, highlighting the potential for covert extremist gatherings to persist undisturbed.

A substantial absence in this context is the paucity of transparency and data regarding moderation outcomes in gaming. In contrast to specific social media platforms, most game companies do not issue periodic transparency reports that specify the number of extremist accounts or chat messages that have been removed. This complicates the assessment of their effectiveness and the subsequent accountability of the companies in question. Furthermore, a salient concern is the potential for undisclosed failures to remain concealed until an external audit, such as an ADL survey or academic study. The dearth of autonomous audits signifies a critical vulnerability. Organizations frequently assert that incidents of extremism are "rare“, yet in the absence of empirical evidence, this assertion remains challenging to substantiate, potentially fostering a culture of complacency.

Moreover, even when companies enforce such rules, consistency remains an issue. It has been observed that specific digital platforms have resorted to prohibiting users for seemingly trivial infractions, while simultaneously overlooking those who disseminate overt expressions of extremism. This inconsistent enforcement can erode trust and deter users from reporting, should they perceive bias or randomness in the outcomes. Furthermore, this ambiguity enables extremists to operate with impunity within the gray areas of the law, leveraging the knowledge that enforcement is inconsistent and unreliable.

A more comprehensive evaluation reveals that government and societal responses are not without their deficiencies. In the United States, as previously mentioned, the Federal Bureau of Investigation and the Department of Homeland Security do not possess a unified strategy to disseminate threat information to gaming companies (United States Government Accountability Office, 2024). This discrepancy suggests that companies may lack awareness of specific extremist threats targeting their platforms or that law enforcement may overlook leads provided by companies. On a global scale, cross-border cooperation in the fight against extremist and terrorist activities is still in its nascent stages (United States Government Accountability Office, 2024). An individual banned from using a server in one country might migrate to another region's server due to a lack of global enforcement mechanisms (except in cases where companies choose to apply global bans).

It must be acknowledged that moderation is inherently challenging in scope and real-time, particularly in games (Takahashi, 2024). In contrast to text posts on Facebook, gaming interactions are frequently ephemeral voice chats or rapidly evolving text conversations, which pose significant challenges to automated scanning without raising privacy concerns. Extremists have the potential to exploit live voice communications, which have historically undergone minimal moderation. Until recently, the oversight of live voice in most games was essentially non-existent, constituting a significant oversight. Activision's 2023 initiative to

implement artificial intelligence voice moderation in Call of Duty marked a pioneering development for a major multiplayer title. However, this endeavor also exposed such practices' technological and ethical challenges. The discussion on balance issues will be addressed subsequently; however, one must acknowledge that technical failures (i.e., false negatives and false positives in AI detection) pose a persistent risk.

Even though extremist content in online gaming worlds is not widespread in terms of quantity (Moonshot, 2024), its relevance for radicalization processes and internal security is evident. The endeavor to combat extremism within the gaming milieu, however, has encountered many setbacks. These setbacks encompass policy lacunae that enable the proliferation of hate speech, under-resourced moderation teams grappling with overwhelming operational demands, and a delayed acknowledgment of the severity of the threat. These missteps have provided invaluable lessons, albeit lessons that have been painfully learned. To enhance the effectiveness of future initiatives, the following gaps must be addressed: the establishment of clearer policies, the implementation of proactive tools, the promotion of transparency, and the fostering of collaboration.

6.3 Balancing Moderation with Free Expression and User Engagement

A persistent challenge confronting all actors involved pertains to the delicate balance between the imperative for stringent moderation and the values of free expression, while concomitantly ensuring the maintenance of an engaging user experience. Games are fundamentally designed for recreation and fostering social connections. Understandably, players and creators are cautious about regulations that could impede creativity or authentic communication. Concurrently, unchecked extremism and harassment can destroy the open, enjoyable gaming spirit. Achieving an equilibrium is a continuous process of negotiation.

Freedom of Expression and Hate Speech

In democratic societies concerns have been raised that efforts to regulate extremist content may inadvertently result in the censorship of political speech or satire. In some cases, players employ games as a medium for addressing issues of a real-world nature. The distinction between a toxic extremist rant and a controversial yet legitimate opinion is not always clear-cut and often subject to national legislation. Moderators should exercise discretion and make judgment calls regarding the content of discussions. A pertinent question is whether a player's discourse on immigration constitutes the articulation of a policy perspective or if it is characterized by hateful extremist rhetoric. Errors have the potential to give rise to allegations of bias or censorship. Many companies have opted to emphasize behavior and harm instead of ideology in its fundamental sense to address this issue. For instance, the prohibition of hate speech, which encompasses slurs, direct harassment, and threats, is delineated in the established guidelines. However, banning an individual solely based on reciting a generic extremist slogan is not permitted unless it constitutes a component of harassing behavior – this approach endeavors to uphold a baseline of respectful discourse while eschewing the policing of mere opinions. Nonetheless, certain extremists endeavor to masquerade their propaganda as mere opinion, a stratagem engenders further complexity.

Players have expressed concerns regarding privacy and free speech, particularly in the context of the implementation of artificial intelligence monitoring tools. When Activision revealed its intention to utilize an AI system, dubbed "ToxMod“, to perpetually monitor voice chats for signs of hate or extremism in the Call of Duty gaming environment, a segment of the community voiced opposition, characterizing the measure as "invasive" and drawing parallels with eavesdropping (Takahashi, 2024). They contend that even private conversations between individuals would be scrutinized by algorithms. This prompts pertinent inquiries, such as ensuring moderation technology does not encroach upon surveillance practices. The developers of these tools emphasize that they prioritize the identification of potential violations and are configured to filter out content that is not problematic. Additionally, they underscore the imperative of safeguarding user safety, acknowledging that the period of unmonitored voice communications, which enabled widespread abuse, is now a thing of the past. The process of balancing entails two fundamental components: Transparency refers to informing users about the monitored aspects and the underlying rationale. Limits signify establishing boundaries, such as restricting scanning to public match chats while abstaining from private friend groups, to safeguard user privacy. Indeed, the ethical design of such tools is crucial. If players feel that their utterances are recorded and could be misinterpreted by an AI, they may disengage or move to external voice applications that are more difficult to moderate.

User Engagement and Community Trust

However, it is essential to note that overly aggressive moderation can lead to the alienation of the player base. Gamers have a rapid aptitude for discerning when the application of moderation appears disproportionate. Instances of overreach, such as the prohibition of users for seemingly innocuous comments or due to false positives (e.g., an artificial intelligence system misinterpreting a word as a slur), have the potential to incite community backlash on forums and social media platforms. This phenomenon has the potential to adversely impact a platform's reputation, thereby causing users to seek alternative platforms. Consequently, companies endeavor to achieve a balanced approach, characterized by a firm yet equitable stance, while avoiding unduly restrictive measures. To provide users with an opportunity for reform, many platforms implement an escalation system that utilizes a series of warnings, temporary suspensions, and bans. This approach allows users to rectify problematic behaviors or actions before imposing more severe consequences. This graduated approach acknowledges the potential for individuals to make errors or utter jokes that do not necessitate permanent exile. The objective is to balance enforcing rules and demonstrating leniency, to maintain positive relations.

Another aspect of engagement is ensuring that moderation itself does not disrupt gameplay. Reporting another player or undergoing a content check can, in certain circumstances, be excessively intrusive. This can have a deleterious effect on the gaming experience. Consequently, some developers seamlessly integrate reporting tools (a rapid method for flagging a comment) and postpone any action until after a match is complete, ensuring they do not interfere with the game in progress. Similarly, content filters can be customized in specific gaming contexts. Players can select a more stringent or less stringent profanity filter, thereby adjusting the level of censorship to their comfort level. These features indicate efforts to

empower users by giving them more control over their experience. This is achieved by finding the right balance between protection and autonomy.

Cultural Sensitivity and Inclusivity

Achieving global balance in moderation entails a concomitant respect for cultural differences in expression. The definition of "extremist" or "hateful" is not fixed and can vary according to the context in which it is used. For instance, using a username that pays homage to a historical figure may be interpreted as extremist glorification in one nation. Yet, it may not carry the same weight in another. Global platforms should navigate these differences, which sometimes necessitate maintaining region-specific moderation rules, as evidenced by the contrasting policies between Asia and the United States. In liberal societies, there is a risk of over-censorship if these standards are applied rigidly, as they are in settings characterized by heightened conflict. One proposed solution involves the establishment of a high-level, global standard (i.e., prohibitions against violence and defamation) and the subsequent implementation of local filters to address specific issues (e.g., the ban of Nazi references within Germany's servers by local legislation). This approach attempts to balance the principles of moderation and the prevailing local norms of free expression.

In the final analysis, efforts at moderation have sought to incorporate counter-speech and positive engagement to achieve equilibrium between suppression and free expression. In contrast to the mere prohibition of extremist rhetoric, specific communities have adopted a more nuanced approach. They encourage their members to engage with such rhetoric by refuting it with facts or humor, provided that the safety of the members is not compromised. This approach maintains the freedom of speech while steering the narrative away from extremist perspectives. To combat radicalism, companies have enlisted the services of prominent gamers to disseminate content that debunks extremist myths and fosters unity. This strategic initiative serves as a counterweight to radical voices, aiming to create a more balanced and inclusive environment within the gaming community. This approach is predicated on the notion that the most effective response to hate speech is not its deletion but promoting more positive speech.

The process of balancing moderation with freedom and engagement necessitates a constant calibration. This calibration involves establishing a sufficient level of moderation to address genuine threats and instances of harassment, while avoiding an excessive restriction that would impede the flow of ordinary conversation or stifle creativity. Engaging users as allies in the process is also imperative. To maintain their trust, transparency and fairness must be upheld. A variety of actors have delineated the parameters in diverse ways. For instance, China places significant emphasis on security, often at the expense of freedom of expression. In contrast, Japan and the United States demonstrate a greater inclination toward freedom of expression, albeit with the retention of certain forms of censorship. Neither of these approaches is optimal for achieving the objective of mitigating radicalization while preserving the element of enjoyment in gaming. The most effective equilibrium appears to be achieved through a moderate approach, characterized by establishing explicit guidelines that proscribe harm (extremist violence, targeted hate) and their consistent enforcement, while concurrently permitting a considerable degree of freedom for non-violent, non-hateful expression.

As the industry evolves, it should identify and preserve a state of equilibrium to ensure long-term success.

7 Strategies for Detecting and Countering Extremist Content in Gaming Environments

Many strategies and tools are currently used (or under proposal) to detect and counter extremist content in gaming spaces. These strategies encompass various approaches, including technical solutions and community-driven initiatives.

Automated Text Filtering

Most online games incorporating chat functionality utilize filters to identify and remove profane or obscene language. In real-time, these filters can identify and intercept common slurs, extremist slogans, and group names (e.g., "ISIS") and subsequently block or mask them. Contemporary filtering mechanisms leverage machine learning algorithms to identify leetspeak and deliberate misspellings employed by extremists to circumvent detection. These lexicons can be updated with new keywords as extremist slang evolves. Although not infallible, text filters are a primary defense mechanism, detecting hateful or extremist language in public chats and user-generated content titles.

AI-Powered Voice Moderation

A contemporary strategy that has emerged as a response to the challenges posed by extremist and abusive language in voice chats involves the deployment of artificial intelligence to monitor and analyze these interactions. Tools such as ToxMod employ machine learning models trained on extensive extremist vocabulary and hate speech datasets to identify and flag spoken content in real time. If a player's voice chat communication includes potential extremist slurs or threats, the system cannot notify human moderators or even implement automated actions, such as muting the player. This is a significant development given the popularity of voice communication in games. From a technical standpoint, this process is challenging due to the necessity of rapid processing and accuracy across languages. However, advancements in technology have rendered it increasingly feasible. This paradigm shift has been adopted by prominent industry leaders such as Activision, as evidenced by the integration advanced systems into their gaming titles. A notable example is the implementation of AI-driven voice moderation in public matches within the Call of Duty franchise. Implementing this system has enabled the identification of extremist harassment that, before its introduction, was not subject to any form of oversight. Privacy safeguards, such as the restriction of voice data retention to the minimum necessary and the limitation of monitoring to public channels, are commonly implemented in conjunction with these systems to address concerns within the gaming community. While AI moderation enhances detection capabilities, it risks infringing on privacy and misreading culturally specific or informal language patterns.

Image Recognition and Scanning

Many gaming platforms permit the integration of user-generated imagery or bespoke designs, encompassing avatars, clan insignia, skins, and cartography. Extremists have exploited this by uploading symbols such as swastikas, KKK hoods, and ISIS flags. Companies have adopted image recognition algorithms to scan user-generated images for extremist symbols in response to this challenge. The Global Internet Forum to Counter Terrorism furnishes a hash database of recognized terrorist images and videos. Gaming companies can employ this database to flag matches on their platforms automatically. Notwithstanding the absence of a hash match, computer vision systems are capable of detecting patterns (e.g., the shape of a Nazi iron cross) and transmitting an alert for manual review. It has been observed that certain video games, such as Call of Duty and Battlefield, have incorporated mechanisms designed to identify and eliminate imagery associated with National Socialism from custom player emblems. This strategy has been developed to address the visual propagation of extremist iconography in games.

User Reporting Systems

Notwithstanding their limitations, player reports remain a crucial component. Regardless of their particulars, games generally have in-game reporting tools that empower users to identify and report others for exhibiting toxic or extremist behavior. The strategy employed here is to facilitate and streamline the reporting process as much as possible. This is achieved by incorporating one-click report buttons and preset categories, such as "extremist content“, to facilitate the classification of reports. Subsequently, companies train moderation staff to prioritize and investigate these reports expeditiously. It has been demonstrated that specific platforms provide incentives for submitting reports. One such incentive is the notification of players when action has been taken based on their report, which has been shown to reinforce participation positively. Trusted flagger programs represent a further development in the realm of online content moderation. Several platforms, such as Roblox, partner with vetted community members or external experts. These individuals can directly report extremist content, expediting the review process. The community's collective visual and auditory faculties, through streamlined reporting, should be utilized as a crucial human element to complement automated detection.

Human Moderation Teams

It has become an imperative for all preeminent gaming companies to establish Trust & Safety or Moderation teams. These teams are comprised of personnel tasked with the responsibility of evaluating content that has been flagged, in addition to monitoring and assessing player behavior. These human moderators possess the capacity to consider context in a manner that algorithms cannot, thereby enabling them to exercise discretion in determining whether content constitutes extremist propaganda or an edgy joke. To address this need, some teams have adopted a 24/7 operational model across different regions, enabling the immediate management of incidents. In cases of a more intricate nature, the responsibility of the moderation of content may be transferred to analysts with specialized knowledge of the subject. Some companies have experts on extremism on call, or they may consult NGO databases of extremist trends. Human review is critical in voice chat incidents, where the AI has flagged

Funded by the European Union
Grant agreement no. 101121345

recordings for review to verify their authenticity. Additionally, human review is crucial in analyzing organized extremist group behavior. For instance, when a clan in a game is suspected of serving as a front for a hate group, human investigators will examine their chat logs, group descriptions, external forums, and other relevant data. The strategy encompasses implementing moderator training to equip staff members with contemporary knowledge concerning extremist symbols, codewords, and radicalization tactics. This training empowers staff members to identify subtler manifestations of extremist content effectively.

Tiered Moderation & Escalation

Several platforms employ a tiered system to manage content based on its severity. For instance, patently unlawful content, such as terrorist threats or manifestos, may result in an immediate prohibition and subsequent referral to law enforcement authorities. In less clear-cut cases, such as sharing extremist memes, a warning may be issued, and the content in question may be removed. Implementing a hierarchical response structure, namely warning, temporary suspension, and permanent ban, enables moderators to calibrate countermeasures according to the severity of extremist behavior. This strategy aims to correct and educate users early on if they unknowingly share content that could be considered borderline. However, users who exhibit a pattern of extremism will be decisively removed.

Pre-Moderation of User-Generated Content

Certain companies implement a pre-moderation process in games that permit users to generate content, such as maps, scenarios, and modifications. This process involves the review of content by moderators or artificial intelligence filters before its publication or availability for search purposes. For instance, a user-created level in a game uploaded to a public repository might be scanned for titles or imagery violating the established guidelines; If an individual attempts to upload a map bearing the name "Jihad Training Camp" and extremist imagery, the moderators are equipped with the capability to prevent its display. Pre-moderation is a process that is implemented to ensure that content that is deemed extremist in nature never reaches an audience. However, it should be noted that this process can be resource-intensive, particularly in cases where a significant volume of user creations must be addressed.

Behavioral Monitoring and Analytics

In addition to keyword scanning, sophisticated techniques are employed to examine user behavior patterns to identify extremist activity. It has been observed that extremist recruiters often exhibit behaviors that diverge from those of ordinary players. These behaviors may include excessive time spent in chat rather than playing, sending unsolicited friend requests that contain propaganda, or the congregation of extremists in specific private lobbies. In this context, companies are developing analytical tools capable of identifying anomalous behaviors. For instance, if a newly created account rapidly communicates with numerous recipients via a standardized message (potentially a recruitment script or an extremist link), it can be identified as a likely "spammer" and subjected to further scrutiny. Social network analysis tools can map relationships within the game. If many known extremist accounts join the same clan or frequent the same server, new cluster members can be monitored more closely. Such

pattern analysis can identify coordinated efforts to radicalize or establish extremist communities within games.

Counter-Extremism Chatbots and Counter-Speech

An unconventional yet intriguing strategy involves the utilization of bots or AI personas to counter extremist narratives in real-time. In principle, if an extremist were to disseminate propaganda via a game chat interface, an AI chatbot could intervene by providing factual corrections or posing derailing questions. Researchers have tested this hypothesis in some online forums. In practice, counter-speech by members of the human community is encouraged. In the context of online communities, specific moderation teams have adopted a nuanced approach by selectively empowering positive community leaders. These leaders are then tasked with engaging with and challenging extremist viewpoints that emerge, as opposed to the more immediate censorship of such content. This assertion is supported by counter-radicalization research, which demonstrates that peer disapproval can deter individuals from further extremist expression. A related measure that has been implemented is the highlighting of community values. Numerous games currently display messages such as "Be respectful. The sentiment that "hate has no place here" functions to reinforce a culture that counterbalances extremist ideology implicitly.

Integrating Educational Content

Developers have begun incorporating educational and resilience-building content into video games to address these issues proactively. This phenomenon can be subtle, as illustrated by a popular military shooter that includes a narrative of cultural cooperation into its storyline, thereby implicitly countering extremist "us versus them" messaging. Alternatively, the message could be conveyed directly. Some games incorporate loading screen tips that instruct players on reporting hate speech or caution them against disclosing personal information to strangers (to prevent grooming). There are documented instances of non-governmental organizations creating games or game modifications to instruct players on the methods employed by extremist recruiters. This pedagogical approach, called "inoculating" players, aims to equip individuals with the knowledge to identify and avoid potential recruitment tactics. These in-game education strategies aim to fortify players against the manipulation of extremists by enhancing their awareness and critical thinking skills.

Collaboration with Extremism Experts

In light of this, numerous companies have initiated consultations with external counterterrorism and hate group experts to enhance their moderation strategies. For instance, they may collaborate with organizations such as the ADL's Center on Extremism to revise their lists of extremist symbols, or with Moonshot CVE, a counter-extremism firm, to develop intervention strategies. To enhance their internal capabilities, some organizations have engaged the services of advisors with backgrounds in intelligence. The collaboration strategy is predicated on the principle of informed moderation, whereby the latest understanding of extremist tactics informs approaches. To illustrate, the resurgence of a particular hate symbol on social media can indicate an imminent attack. A game platform may implement a preemptive filter for that

symbol in such cases. The dissemination of information across various digital platforms is a concern that has garnered significant attention in recent times. Suppose an extremist group is prohibited from utilizing a prominent social networking platform and is known to resort to gaming platforms as a migration path. In that case, pertinent intelligence can be conveyed to gaming companies. Gaming companies can then utilize this intelligence to monitor specific usernames or group names within the gaming environment.

Industry Coalitions and Standards

As mentioned, establishing industry-wide standards and coalitions can be considered a strategy. The unification of companies has the potential to facilitate a collective agreement on baseline moderation practices for extremism, such as the prohibition of all terrorist organization symbols. Additionally, this unification could enable the sharing of technological resources among the participating companies. The Fair Play Alliance and GIFCT's gaming-focused working groups serve as forums for discussing and disseminating these strategies. The strategy is that by speaking with a unified voice, the industry can make it more difficult for extremists to gain a foothold in the mainstream. If all major platforms prohibit a particular hate slogan, extremists will face more difficulty moving from one platform to another. This initiative also assists smaller companies that lack the necessary resources in gaining access to tools, such as hash databases or trained classifiers, through coalition efforts within the gaming community.

Legal Enforcement and Reporting Pipelines

While not a platform strategy in the strict sense of the term, it is an integral component of the ecosystem, facilitating the apparent escalation of serious extremist threats to law enforcement. Gaming companies have established protocols that mandate reporting certain behaviors to law enforcement authorities. This includes situations where a user makes a direct terrorist threat or there is evidence indicative of imminent violence. The strategy entails the incorporation of this functionality into the moderation workflows. To illustrate, a flag for "possible imminent threat" is to be implemented, which, upon confirmation by a moderator, will initiate a notification process to the company's legal team and the relevant authorities. Pertinent data, such as chat logs and user IP addresses, will accompany this notification. Formal partnerships are sometimes established among nations to facilitate this process. For instance, the United Kingdom and France have established police units that liaise with technology platforms in the event of urgent incidents. The efficacy of this strategy is predicated on the understanding that moderation must extend beyond the mere removal of content. It is essential to proactively forestall potential attacks by enlisting law enforcement to intervene when a gamer appears to be on a trajectory toward real-world violence.

Community Moderation and Peer Networks

Decentralized moderation through the community itself has the potential to be a formidable tool. In many cases, the administration of games is entrusted to volunteer moderators or clan leaders, who oversee their respective communities. To illustrate, within an MMO (massively multiplayer online game), the presence of "player guides" or moderators on each server is notable. These individuals possess the capability to mute or report players engaging in

extremist bullying behaviors. Peer enforcement through clan rules represents a distinct approach. Certain gaming clans or guilds implement stringent no-hate policies, prompting the expulsion of members who advocate extremist views. This self-regulating mechanism within the community purges individuals who do not align with the prevailing norms and values. Game companies frequently provide support for these volunteer efforts by offering moderation tools, such as the capacity to initiate a vote-kick against a disruptive player. This strategy leverages the fact that players often understand their respective community norms and can act more expeditiously than official moderators in informal contexts.

Shadow Banning and De-amplification

In the interest of nuance, some platforms employ shadow banning (rendering a user's posts invisible to the general public) or de-ranking content to mitigate extremist influence without confrontation. When a suspected extremist has not yet crossed a clear policy line, a platform may limit their reach. For example, the content shared via chat messages might be restricted to their acquaintances, reducing recruitment potential. Similarly, if an in-game forum thread exhibits extremist tendencies, moderators may elect to discontinue its promotion or discreetly remove it, thereby allowing the thread to dissipate without undue attention. This strategy is contentious, yet it is employed in certain instances to circumvent the martyrdom of extremists (who might utilize a ban as "evidence" of repression). Instead, it merely seeks to diminish the extremists' audience.

Regular Audits and Penetration Testing

A frequently disregarded yet significant strategy entails proactively auditing the platform for extremist content. Companies periodically have their staff or hire external "red teams" to act as malicious users and ascertain what extremist content they can inject or find. For instance, an audit might generate test accounts to attempt to join extremist groups with known names or search for key terms, thereby mapping the available content. The results of this study provide a foundation for identifying areas where the moderation process may be enhanced. This is analogous to penetration testing in cybersecurity, wherein the objective is to assess the efficacy of content controls. Some organizations (e.g., the ADL) effectively fulfill this role as third-party auditors, disseminating the results publicly to pressure companies (as evidenced by the incident with Steam). To maintain a competitive advantage, platforms should prioritize this initiative internally.

User Empowerment Tools (Ignore/Block/Mute)

Equipping players with practical tools for filtering and managing their experience constitutes a defensive strategy. The vast majority of online games currently available offer users the capability to mute communication from specific players, decline friend requests from unknown individuals, and implement additional customizable settings. While this does not prevent the dissemination of extremist rhetoric, it safeguards potential targets from exposure. In essence, it serves as a counterbalance to extremist content by reducing its reach to a niche audience of those who are tolerant of such content. In specific gaming environments, users can disengage from global chats or utilize "whitelist" modes, which restrict a user's visibility

to messages from a predefined list of approved friends. Educating users about these features constitutes a component of the strategy. To illustrate, when a player voices concerns about harassment, the game's interface prompts the player to utilize the chat settings' mute function. The objective is to mitigate the probability that a single toxic radical can adversely impact the experience of numerous others.

Conclusion

Collectively, these strategies constitute a comprehensive set of tools for addressing extremist content in gaming environments. One must note that these methods are most efficacious when employed in conjunction. For instance, implementing artificial intelligence detection facilitates identifying content that requires human moderators' review. After this, user reports assist in identifying content that AI has overlooked, and community engagement reinforces established norms. The multi-layered defense system has been developed to prevent extremist actors from identifying and exploiting vulnerabilities. Nevertheless, as extremist groups evolve tactics, these strategies must be continuously refined. The subsequent section will synthesize the findings outlined above into concrete policy recommendations for governments and industry stakeholders to strengthen these approaches in the future.

8 Policy Recommendations

To effectively address the issue of radicalization in the context of online gaming, regulators, industry leaders, and communities worldwide should demonstrate a sustained commitment and collaborate closely. The preceding analysis indicates that significant deficiencies persist while certain advancements have been observed, including implementing more stringent policies and introducing novel instruments. To enhance prevention strategies and improve moderation while upholding user rights, this report offers the following policy recommendations, which are aimed at both government authorities and gaming platform providers:

8.1 For Government Regulators and Policymakers

Facilitate a Global Multi-Stakeholder Initiative for Safer Gaming

It is incumbent upon governments to assume a leadership role in convening international dialogues centered on the issue of extremism in gaming. GEMS's Active Exchange Events can deal as a blueprint for these endeavors. Policymakers around the globe should sponsor forums that facilitate the convergence of governments, technology companies, civil society, and researchers. The purpose of these forums is to encourage the sharing of data and the development of cohesive cross-border strategies. A global approach is imperative in light of the transnational nature of online games. Extremists can operate on servers across various jurisdictions. The formation of a multi-stakeholder task force, potentially under the auspices of the United Nations or another international body, is imperative to address the nexus between gaming platforms and violent extremism. This initiative can augment existing networks, such as the Extremism and Gaming Research Network, extending their reach. Implementing such an initiative would contribute to the standardization of expectations within the industry,

thereby mitigating the discrepancy between, for instance, U.S. and Asian platforms in their approach to extremism.

Develop Clear National Guidelines or Codes of Conduct for Platforms

Absent substantial regulatory oversight, governments retain the capacity to promulgate guidelines or voluntary codes that delineate optimal practices for content moderation in games. For instance, the U.S. Department of Homeland Security or the Department of Justice could publish guidelines for gaming companies on countering extremism. These guidelines could cover recommended policy elements (e.g., "prohibit terrorist propaganda"), reporting mechanisms, and cooperation with law enforcement. Similarly, countries such as Japan or Australia could incorporate gaming platforms into their existing online safety codes. While these guidelines are not mandatory, they establish a benchmark and implicitly pressure companies to comply. If self-regulation proves inadequate over time, regulators may consider establishing baseline requirements. For instance, large gaming platforms could be obligated to publish annual transparency reports on extremist content removal, similar to the requirements stipulated in the EU's Digital Services Act. However, it should be noted that the primary focus of the EU legislation pertains to activities occurring outside the territorial boundaries of the European Union. They must be meticulously crafted to ensure that such measures do not infringe upon free expression. Hence, it is incumbent upon governments to emphasize process and transparency rather than dictating the removal of specific viewpoints. The objective is to guarantee that companies possess procedures to address extremist content in a manner that safeguards users.

Support Research and Information-Sharing (Public-Private Partnership)

It is incumbent upon governments to allocate financial resources and expand existing programs to foster a comprehensive understanding of radicalization in gaming. Moreover, governments should prioritize the enhancement of information-sharing with relevant industry sectors. This encompasses grants for research, such as the funding of studies on gaming and extremism by Canada or the grants provided by the U.S. Department of Homeland Security. These investments yield insights that can inform effective interventions. Furthermore, federal agencies like the FBI, DHS, and Interpol must establish formal information-sharing frameworks with gaming companies. It is recommended that a dedicated unit or liaison be established for the gaming sector within counter-terrorism task forces. This unit would routinely disseminate declassified threat intelligence (e.g., recent extremist narratives or individuals of concern) to platform trust and safety teams. In exchange, it would procure relevant anonymized data from companies. As recommended by the U.S. Government Accountability Office, establishing goals and protocols for this exchange will ensure a systematic approach to cooperation, thereby replacing the ad hoc nature of current practices. Governments can incentivize corporations to disseminate data to independent academics, contingent upon implementing privacy safeguards, to assess the efficacy of various methodologies. Enhancing the extant evidence base enables public policy to be more precisely calibrated and responsive to emergent threats. For instance, if research indicates a surge in extremist recruitment within specific game genres, resources can be apportioned accordingly.

Strengthen Legal Frameworks for the Worst Offenders (Last Resort Enforcement)

Although the implementation of extensive censorship is not recommended, it is incumbent upon governments to establish robust legal frameworks that address particularly egregious actions. Such actions include the utilization of gaming platforms to orchestrate terrorist attacks or disseminate extremist content, such as instructional manuals on bomb fabrication, or the exploitation of children in conjunction with extremist networks. Numerous countries have already instituted legal frameworks to address these issues, and it is essential that these frameworks explicitly encompass virtual spaces. Law enforcement agencies are strongly encouraged to pursue legal action against cases of extremist recruitment or threats made via gaming platforms, as this will serve as a deterrent. Furthermore, governments may wish to consider implementing targeted legislation to close loopholes. For instance, the requirement for companies to retain chat logs for a limited time (under strict privacy rules) could ensure the preservation of critical evidence during ongoing investigations. A balance must be struck between legal obligations and civil liberties in the context of surveillance and data collection from gaming communities. Oversight and due process become paramount in such scenarios. Governments should also clearly communicate these laws to the public, ensuring gamers understand that "virtual" extremist actions can carry real penalties.

Promote Digital Literacy and Resilience Programs

A collaborative effort between regulators and education ministries is imperative to develop digital literacy campaigns incorporating gaming. As schools educate youth about social media safety, they should also address the issue of safe gaming (see also D4.2 on citizen awareness campaigns). Young gamers must develop the ability to discern grooming or extremist manipulation in games. Furthermore, it is crucial to equip them with the capacity to report concerns and to cultivate the ability to critically evaluate information, even when it emanates from a fellow player. Integrating this approach into counter-radicalization initiatives in certain countries is a potential avenue for further research. For instance, a government could collaborate with an esports organization to disseminate brief talks or materials at gaming events concerning rejecting hate and extremist propaganda. The inoculation of players with knowledge and skepticism is a strategy employed by governments to mitigate the vulnerability of individuals to the recruitment efforts of extremists. Civil society can provide financial support to facilitate this process. Non-governmental organizations could facilitate workshops or generate content (e.g., videos, comic strips) meticulously designed to appeal to gamers on these subjects. In essence, players with heightened awareness complement moderation by engaging in self-policing and resisting extremist narratives within the gaming domain.

8.2 For Gaming Platforms, Publishers, and Developers

Embed Safety by Design and Positive Community Features

It is incumbent upon game developers to incorporate safety and counter-extremism considerations during the design phase of games, a process referred to as "safety by design." This objective necessitates the development of games that inherently mitigate opportunities for exploitation. For instance, when designing a new multiplayer game, developers can include

robust player reporting tools and moderation interfaces from the outset, rather than as an afterthought. Design choices that discourage toxic behavior may include implementing matchmaking algorithms that systematically separate consistently disruptive players from the general population or incorporating cooperative game modes that promote teamwork and empathy. Research has demonstrated that developers can incentivize prosocial behavior by strategically integrating game mechanics. Riot's honor point system in League of Legends is a prime illustration of a system that reinforces positive play with tangible rewards. In a similar vein, games can penalize or deemphasize antisocial behavior subtly. For instance, a game could lower the voice chat volume of a player yelling slurs, making it less likely that others will hear them. This is an example of soft moderation that is built into the game. Another design aspect that merits consideration is content curation. If a game incorporates user-generated content, developers may elect to implement an approval queue or a community voting system that functions as a natural filter for extremist submissions. In narrative-driven games, developers are encouraged to include storylines or characters that model diversity and conflict resolution. Such inclusion can counter extremist narratives that thrive on "us versus them" mindsets. In conclusion, the recommendation is for developers to adopt a proactive stance: to anticipate potential misuse of features, such as those related to hate or recruitment, and to implement modifications to mitigate associated risks without compromising the enjoyment of the game. Governments can promote this phenomenon through the implementation of incentives. One possible incentive would be to provide recognition or tax benefits to games that incorporate robust safety-by-design features. These features are analogous to certifications for privacy or security. It is posited that, over time, a culture of safety-conscious game design will engender a more resilient ecosystem by counteracting the exploitation of extremists.

Empower and Educate the Gaming Community

It is incumbent upon platform providers to collaborate with community stakeholders to cultivate a culture of active bystander intervention and inclusivity. Providing tools and encouragement for regular players to assist in moderating and establishing a conducive atmosphere is critical to this process. Specifically, companies can augment their recruitment and training programs to include volunteer moderators, or "game guardians“, from within the player community. These individuals are then empowered to flag or moderate content in real time, with the necessary oversight to ensure the program's integrity. Furthermore, they can organize workshops or create content, such as brief video recordings, featuring prominent gaming influencers discussing the significance of rejecting hate and standing up to extremist rhetoric. By leveraging respected voices within the community, platforms can shape norms. For instance, if a prominent streamer conveys that "racism and extremist propaganda are not acceptable here“, it has the potential to resonate with players. This phenomenon has been observed in various gaming contexts, such as integrating community guidelines into the framework of e-sports tournaments and broadcasts. Furthermore, companies should implement feedback loops to players, whereby users are informed of the subsequent actions taken in response to their reports of extremist incidents, reinforcing the efficacy of reporting mechanisms. One should acknowledge the contributions of players who exhibit exemplary conduct. One potential strategy to further incentivize such positive behavior is recognizing and rewarding players who contribute positively. This could be achieved by showcasing top community

helpers or by providing in-game bonuses for those with unblemished records. The overarching objective is transforming the community into an ally by implementing "crowdsourcing" to moderate and counter-speech content. When players collectively assume responsibility for maintaining a welcoming environment, extremists encounter significant challenges in gaining traction. This recommendation is consistent with the principle that the most effective safeguard for free expression is not the allowance of abuse, but rather the establishment of mutual standards of respect that enable all individuals to express themselves without fear of retaliation. It is incumbent upon gaming platforms to allocate resources toward hiring community management personnel who will engage directly with players, execute campaigns (e.g., diversity celebrations within the game), and showcase positive community creations. This will engender an environment in which extremist views are rendered utterly invalid.

Adopt Clear and Comprehensive Anti-Extremism Policies

All prominent gaming platforms and online games should implement an explicit policy prohibiting extremist content and behavior. This policy should encompass the promotion of violent extremist ideologies, recruitment activity, glorification of terrorist organizations or acts, and the utilization of extremist symbols in usernames or user content. These regulations must be articulated in a manner accessible to the general user and integrated into the player experience. They should not be obscured within the labyrinth of legalese. In the corporate realm, businesses can emulate the success of those who have effectively navigated this terrain. Roblox, for instance, has instituted explicit community standards that proactively preclude the presence of any content associated with terrorism or extremism. Similarly, Riot Games has adopted a distinctive approach by presenting its rules as an engaging "Summoner's Code“, a strategy that effectively fosters a positive relationship with its user base. The recommendation is to make these policies public, prominent, and easily comprehensible, ensuring all users know the boundaries. Implementing regular reminders, such as pop-ups, loading screens, or email notifications, has effectively enhanced awareness. A critical component of this process is the establishment of precise definitions for the terms "extremist" and the subsequent articulation of the actions that will be taken in response. This clarity serves a dual purpose: it deters would-be offenders by establishing clear boundaries and empowers moderators to act decisively. It is incumbent upon companies to update these policies with regularity, in consultation with extremism experts, to address the emergence of novel trends (e.g., the proliferation of hate symbols or codes).

Implement Robust Moderation Mechanisms (Blend of AI and Human)

Platforms should allocate resources toward implementing a multi-layered moderation system that integrates automated detection mechanisms with the active involvement of skilled human oversight personnel. From a technical perspective, the implementation of advanced artificial intelligence filters for text, voice, and image analysis, as previously delineated, is imperative. These models should be capable of operating in multiple languages and should be continuously trained with real gaming context data to improve accuracy. The utilization of tools such as the GIFCT hash database is strongly advised for identifying known terrorist content. From a human resources perspective, it is incumbent upon companies to ensure that they have the necessary number of adequately trained moderation teams. It is incumbent

Funded by the European Union
Grant agreement no. 101121345

upon the moderators to review content flagged promptly and consistently, ensuring the established rules are enforced. In cases of particularly egregious content, it is essential to implement clear escalation pathways to address the situation effectively. Platforms should allocate resources towards the provision of continuous training for moderators. This training should encompass the identification of contemporary extremist symbols and memes, including numerical codes and dog whistles. Speed and consistency in addressing extremist incidents cannot be overstated. Policies should be applied uniformly to all users, without exception, to ensure the equitable mitigation of extremist activities. Platforms such as Twitch and Discord have demonstrated the efficacy of swift and decisive bans of extremist groups. Gaming companies would be well-advised to exercise stringent enforcement measures when rules are transgressed. Concurrently, companies should examine the features of their platforms that might be exploited by malicious actors and adjust them accordingly. For instance, they should limit the mass-messaging capabilities that recruiters abuse or review game modes that could be repurposed for extremist role-play.

Increase Transparency and Collaboration with Experts

To establish trust and enhance outcomes, gaming platforms must demonstrate a commitment to transparency regarding their efforts to moderate extremist content. This includes the regular publication of reports containing data such as the number of reports of extremist content received, the number of accounts or uploads that were actioned for extremist violations, and examples of content that have been removed (in anonymized form). The advent of social media has led to a paradigm shift in the realm of transparency, rendering it a universal standard. This shift is poised to facilitate enhanced accountability and identify industry-wide trends. Additionally, it is incumbent upon companies to furnish independent researchers with access to data in a manner that safeguards privacy. The dissemination of anonymized chat logs or incident datasets to academic institutions and non-governmental organizations can facilitate external evaluation of the prevalence of extremist content and the efficacy of moderation measures. This external scrutiny can shed light on potential blind spots and validate the veracity of corporate claims that extremism is rare on their platform. It is noteworthy that several governments have contemplated the implementation of mandates for data access in the realm of social media. In anticipation of potential regulatory measures, the gaming industry has the opportunity to collaborate with researchers through voluntary partnerships proactively. Furthermore, companies are advised to regularly consult with extremism experts and civil rights groups when revising policies. This multifaceted input will facilitate a balanced approach that ensures individuals' safety while safeguarding free expression rights and protecting marginalized communities from abuse.

Participate in and Strengthen Industry Coalitions

Gaming companies, particularly those providing major platforms and publishing games, should engage in industry coalitions and standards initiatives prioritizing online safety and counter-extremism. Organizations such as the Global Internet Forum to Counter Terrorism and the Fair Play Alliance provide frameworks for collaborative efforts. However, historically, gaming firms have been underrepresented in these bodies. Companies such as Microsoft (Xbox) and Roblox are already in GIFCT, and it is recommended that other companies (Valve,

Sony, Nintendo, etc.) follow suit and contribute to collective solutions. Establishing these coalitions enables the industry to develop shared definitions and best practices, such as establishing consensus on the meaning of extremist content in games. Furthermore, these coalitions can facilitate the creation of shared resources, including inter-company hotlines for imminent threat situations. The Fair Play Alliance, which convenes gaming companies to address player behavior issues, should explicitly expand its scope to encompass extremism and toxicity. It is recommended that members establish a common baseline standard. For example, all FPA companies could pledge to prohibit promoting extremist groups. This would eliminate any loopholes that offenders might exploit by hopping between games. Additionally, industry coalitions can establish independent audits or peer reviews. Companies may consent to periodic assessments by a neutral panel (including civil society) of their moderation systems, with the findings and recommendations disseminated among them. Implementing collective accountability mechanisms can potentially enhance the overall standards within the sector. The gaming industry's ability to engage with governments is contingent upon its ability to present a unified front. Demonstrating a commitment to self-regulation can engender trust in the gaming industry's ability to manage content without external laws.

Maintain a Balanced Approach Respecting User Rights

In pursuing the objectives mentioned above, platform providers must continuously calibrate their methods to ensure respect for user privacy and freedom of expression, thereby facilitating legitimate discourse. This necessitates the incorporation of privacy safeguards. For instance, when employing artificial intelligence to monitor voice chats, their use must be transparent to limit the scope of monitoring to what is essential for safety. Established protocols must manage any data collected for moderation purposes to ensure its security and appropriate retention periods. Furthermore, companies should develop formal appeal mechanisms for users who feel they have been unfairly moderated. A separate team should review these appeals to ensure procedural integrity and prevent potential biases from influencing the review process. By upholding these rights, platforms demonstrate their commitment to fostering a secure environment without the imposition of unwarranted censorship. They should publicly articulate that their moderation targets behavior, not ideology. Individuals are free to engage in discourse concerning various views; however, they are not permitted to engage in harassment or threats against others. Achieving this equilibrium is paramount and using precision instruments over broad sweeps is strongly advocated. To illustrate, rather than imposing a universal prohibition on all political discourse, which might be considered excessively broad, it would be more effective to identify and proactively eliminate specific harmful behaviors, such as calls for violence or hate speech, while ensuring the continued viability of civil discourse. In the corporate realm, entities such as Blizzard have articulated their codes with the phrases "Gameplay First" and "Play Nice; Play Fair“, thereby emphasizing conduct. Adopting a similar ethos, with explicit inclusion of anti-extremism, maintains a focus on actions and impacts rather than suppressing ideas wholesale. Ultimately, protecting user rights is instrumental in safeguarding the gaming experience. Gamers place a high value on creativity and social connection. Therefore, effective moderation should be implemented to enhance these aspects by filtering out merely disruptive content.

In conclusion, the challenge of radicalization in gaming spaces can be met with determined action and smart policy. As this report has detailed, various countries and companies are experimenting with solutions. The most effective approach is emerging as a collaborative effort; it is becoming evident that success is unattainable for governments and industry alone. Governments outside the European Union should acknowledge the significance of online games as social platforms and incorporate them into national strategies to counter extremism. This approach should be accompanied by a meticulous preservation of the freedoms that contribute to the enjoyment of these games. Gaming companies must conceptualize safety not as an onerous obligation but as an integral component of their responsibility to players and the sustainability of their communities. The implementation of the recommendations above, ranging from the establishment of clearer policies and the provision of enhanced tools to the empowerment of communities and the promotion of transparency, can substantially mitigate the presence of extremist content on these platforms.

One must accentuate a balanced perspective, recognizing that gaming engenders numerous positive benefits, including fostering friendship, facilitating learning, and providing entertainment. These benefits should be safeguarded from the corrosive influence of extremist abuse. The objective of moderation is not to sanitize or politicize gaming; instead, it aims to prevent a select group from exploiting games to cause harm in the real world. When moderation is executed effectively, it has the potential to augment freedom of expression, enabling users to engage without the concern of harassment or manipulation. Successful cases, including a prompt response to live-streamed violence and the purging of overt hate groups, evidence the efficacy of such measures. These examples demonstrate that progress can be achieved without compromising the open nature of games.

As the global gaming community continues to expand, reaching billions of individuals, implementing these comprehensive, rights-respecting prevention strategies is a moral imperative and in the self-interest of all stakeholders. A proactive approach to radicalization in the gaming sector is imperative to ensure the safety of individuals and maintain the integrity of virtual worlds as platforms for healthy competition, creativity, and camaraderie, rather than as breeding grounds for extremism.

9 References

Activision Blizzard. (2022). *Call of Duty: Modern Warfare II and Warzone 2.0 – Code of Conduct*. https://www.callofduty.com/news/2022/09/modern-warfare-II-code-of-conduct

Activision Blizzard. (2024). *Activision Blizzard code of conduct: the right way2play*. https://www.activisionblizzard.com/code-of-conduct/launch

Amadeu Antonion Stiftung. (2025). *Good Gaming – Well Played Democracy*. https://www.amadeu-antonio-stiftung.de/projekte/good-gaming-well-played-democracy/

Anti-Defamation League. (2021). *Hate is No Game: Harassment and Positive Social Experiences in Online Games 2021*. https://www.adl.org/resources/report/hate-no-game-2021

Anti-Defamation League. (2024). *Hate is No Game: Hate and Harassment in Online Games 2023.* https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2023

Asano, T. (2022). Ideological Extremism and Political Participation in Japan. *Social Science Japan Journal, 25*(1), 125-140.

Australian Classification Board. (n.d.). *Australian Classification.* https://www.classification.gov.au

Baek, Byung-yeul (2024). *NCSOFT, Nexon, Krafton, Netmarble adopt AI to boost efficiency in game development*. https://www.koreatimes.co.kr/business/tech-science/20240310/ncsoft-nexon-krafton-netmarble-adopt-ai-to-boost-efficiency-in-game-development

Berjawi, O., Fenza, G., & Loia, V. (2023). A comprehensive survey of detection and prevention approaches for online radicalization: Identifying gaps and future directions. *IEEE Access*, *11*, 120463-120491.

Bhatt, S., & Mantua, J. (2023). The transnational threat of radicalization through the use of online gaming platforms. In C. W. Gruber & B. Trachnik (eds.), *Fostering innovation in the intelligence community: Scientifically-informed solutions to combat a dynamic threat environment* (pp. 113-131). Springer.

Bovenzi, G. M. (2024). *Content moderation in (decentralized) metaverses*. https://catedrametaverso.ua.es/wp-content/uploads/2024/07/Content-moderation-in-decentralised-metaverses-BOVENZI.pdf

Brewster, T. (2022). *How cartels recruit teenagers through online games*. https://www.forbes.com/sites/thomasbrewster/2022/02/17/how-cartels-use-online-games-to-recruit-kids/

Bromell, D. (2022). *Regulating free speech in a digital age: Hate, harm and the limits of censorship*. Springer.

Call of Duty Staff. (2024*). Anti-Toxicity / Disruptive Behavior Progress Report – Modern Warfare III Season 2.* https://www.callofduty.com/blog/2024/01/call-of-duty-ricochet-modern-warfare-iii-warzone-anti-cheat-progress-report

Chalk, A. (2019). Tencent imposes new regulations on streamers in China. https://www.pcgamer.com/tencent-imposes-new-regulations-on-streamers-in-china

Computer Entertainment Rating Organization. (2025). About CERO. https://www.cero.gr.jp/en/publics/index/3/

Council of Europe. (n.d.). *Octopus Cybercrime Community: France*. https://www.coe.int/en/web/octopus/-/france

D'Anastasio, C. (2024). *White Supremacist, Nazi Content Spread on Steam Game Service*. https://www.bloomberg.com/news/articles/2024-11-14/white-supremacist-nazi-content-spread-on-steam-gaming-platform

Davey, J. (2021). *Gamers who hate: An introduction to ISD's gaming and extremism series*. https://www.isdglobal.org/isd-publications/gamers-who-hate-an-introduction-to-isds-gaming-and-extremism-series/

Díez Gutiérrez, E. J. (2014). Video games and gender-based violence. *Procedia - Social and Behavioral Sciences, 132*, 58-64.

Echikson, W. & Knodt, O. (2018). *Germany's NetzDG: A key test for combatting online hate*. https://cdn.ceps.eu/wp-content/uploads/2018/11/RR%20No2018-09_Germany's%20NetzDG.pdf

European Commission (2025). *The EU Code of conduct on countering illegal hate speech online*. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

Europol. (2022).*EU Internet Referral Unit Transparency Report 2021*. https://www.europol.europa.eu/cms/sites/default/files/documents/EU_IRU_Transparency_Report_2021.pdf

Extremism and Gaming Research Network. (2025). *Extremism and Gaming Research Network*. https://extremismandgaming.org/

Feta, B., & Armakolas, I. (2024). *Gaming Stakeholders Mapping Report*. https://www.projectgems.eu/post/the-initial-two-gems-reports-have-been-released-to-the-public

Freedom House. (2022). *Freedom on the Net 2022*. https://freedomhouse.org/sites/default/files/2022-10/FOTN2022Digital.pdf

Galbraith, P. W. (2017). Adult computer games and the ethics of imaginary violence: Responding to Gamergate from Japan. *U.S.-Japan Women's Journal, 52*, 67-88.

Game Developers Conference. (n.d.). *Game Developers Conference*. https://gdconf.com/

Gault, M. (2022). *DHS to Spend Almost $700,000 Investigating 'Radicalization in Gaming'*. https://www.vice.com/en/article/dhs-to-spend-almost-dollar700000-investigating-radicalization-in-gaming

Global Internet Forum to Counter Terrorism. (2021). *Transparency Report*. https://gifct.org/wp-content/uploads/2021/07/GIFCT-TransparencyReport2021.pdf

Global Internet Forum to Counter Terrorism. (2024). *Our Impact in 2023*. https://gifct.org/2024/04/04/our-impact-in-2023/

Global Online Safety Regulators Network. (2024). *Regulatory Index. Comparing international approaches and perspectives to online safety regulation*. https://www.ofcom.org.uk/siteassets/resources/documents/about-ofcom/international/other/global-online-safety-regulators-network-regulatory-index.pdf?v=383839

Gruber, C.W., Kirby, C., Dalpe, S., Silverstein, L., & Frey, S. (2023). Ubiquitous Technical Surveillance: A Ubiquitous Intelligence Community Issue. In C. W. Gruber & B. Trachik (eds), *Fostering Innovation in the Intelligence Community: Scientifically-Informed Solutions to Combat a Dynamic Threat Environment* (pp. 1-17). Springer.

Halilovic-Pastuovic, M., Wylie, G., & Vukic, N. (2024). *Towards a Sociology of Gaming and Radicalisation: A Report on the State of the Art.* GEMS Project Publication.

Helmus, T. & Klein, K. (2020). *Assessing Outcomes of Online Campaigns Countering Violent Extremism: A Case Study of the Redirect Method*. https://www.rand.org/content/dam/rand/pubs/research_reports/RR2800/RR2813/RAND_RR2813.pdf

Higuchi, N. (2020). When hate becomes illegal: Legislation processes of the anti-hate speech law in Japan. In M. Kang, M.-O. Rivé-Lasan, W. Kim & P. Hall (eds.), *Hate Speech in Asia and Europe* (pp. 112-126). Routledge.

Human Rights Watch. (2019). *China's Algorithms of Repression*. https://www.hrw.org/report/2019/05/01/chinas-algorithms-repression/reverse-engineering-xinjiang-police-mass

Irfan, M., Giannotta, F., Gagliardone, I., & Prucha, N. (2024). *AI and online radicalisation: Between ethics and efficacy*. United Nations Interregional Crime and Justice Research Institute.

Keller, D., & Leerssen, P. (2020). Facts and where to find them: Empirical research on internet platforms and content moderation. In N. Persily & J. A. Tucker (eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 220–252). Cambridge University Press.

Kosseff, J. (2019). *The Twenty-Six Words That Created the Internet*. Cornell University Press.

Kou, Y. & Nardi, B. (2014). *Governance in League of Legends: A Hybrid System*. https://www.researchgate.net/publication/309738489

Kowert, R. & Kilmer, E. (2023a). *Extremism in Games: A Primer*. https://www.takethis.org/wp-content/uploads/2023/10/ExtremismInGames_APrimer_FINAL-1.pdf

Kowert, R. & Kilmer, E. (2023b). *Toxic Gamers Are Alienating Your Core Demographic – The Business Case for Community Management*. https://www.takethis.org/wp-content/uploads/2023/08/ToxicGamersBottomLineReport_TakeThis.pdf

Kudlacek, D., Kudlacek, J., John, L. & Vrdoljak, T. (2025a). *(Non-EU) Citizen Awareness Campaigns, Youth Protection, and Gaming*. GEMS Project Publication.

Kudlacek, D., Kudlacek, J., John, L. & Vrdoljak, T. (2025b). *Preventing/countering violent extremism and youth protection in gaming outside the European Union*. GEMS Project Publication.

Lamphere-Englund, G. & White, J. (2023). *The Online Gaming Ecosystem: Assessing Socialisation, Digital Harms, and Extremism Mitigation Efforts*. https://gnet-research.org/2023/05/26/the-online-gaming-ecosystem/

Lee, S. & Lee, M. (2025). *South Korea's Approach to Age Assurance*. https://www.techpolicy.press/south-koreas-approach-to-age-assurance

Mattheis, A. A., & Kingdon, A. (2023). Moderating manipulation: Demystifying extremist tactics for gaming the (regulatory) system. *Policy & Internet*, *15*(4), 478-497.

Moonshot. (2024). *Extremism across the Online Gaming Ecosystem*. GEMS Project Publication.

New Zealand Government. (2022). *New Zealand's Countering Terrorism and Violent Extremism Strategy*. https://dpmc.govt.nz/sites/default/files/2021-10/New%20Zealands%20Countering%20Terrorism%20and%20Violent%20Extremism%20Strategy.pdf

Newton, C. (2017). *Discord bans servers that promote Nazi ideology*. https://www.theverge.com/2017/8/14/16145432/discord-nazi-ban-white-supremacist-altright

Olaizola Rosenblat, M. & Barrett (2023). *Gaming The System: How Extremists Exploit Gaming Sites And What Can Be Done To Counter Them*. https://bhr.stern.nyu.edu/wp-content/uploads/2024/01/NYUCBHRGaming_ONLINEUPDATEDMay16.pdf

PlayStation. (n.d.). *Community Code of Conduct*. https://www.playstation.com/en-us/support/account/community-code-of-conduct/

Public Safety Canada. (2024). *Government of Canada announces funding to study potential for radicalization to violence across gaming platforms*. https://www.canada.ca/en/public-safety-canada/news/2024/03/government-of-canada-announces-funding-to-study-potential-for-radicalization-to-violence-across-gaming-platforms.html

Public Safety Canada. (2025). *Community Resilience Fund: Funding Project Descriptions*. https://www.publicsafety.gc.ca/cnt/bt/cc/fpd-en.aspx

Riot Games. (n.d.). *League of Legends Code of Conduct*. https://www.leagueoflegends.com/en-us/event/league-of-legends-code-of-conduct/

Roblox. (n.d.). Content Moderation on Roblox. https://en.help.roblox.com/hc/en-us/articles/21416271342868-Content-Moderation-on-Roblox

Royal United Services Institute for Defence and Security Studies. (2025). *Extremism and Gaming*. https://www.rusi.org/explore-our-research/projects/extremism-and-gaming

Sony Interactive Entertainment. (2025). *Online Safety*. https://sonyinteractive.com/en/impact/online-safety/

Stelter, B. & Paget, B. (2022). *Twitch says livestream of Buffalo mass shooting was removed in less than 2 minutes*. https://edition.cnn.com/2022/05/15/business/twitch-livestream-buffalo-massacre/index.html

Takahashi, D. (2024). *How ToxMod's AI impacted toxicity in Call of Duty voice* chat. https://venturebeat.com/games/how-toxmods-ai-impacted-toxicity-in-call-of-duty-voice-chat-case-study

Tassi, P. (2015). *On Game Consoles, Terrorism And Missing The Point*. https://www.forbes.com/sites/insertcoin/2015/11/16/on-game-consoles-terrorism-and-missing-the-point/

Tech Against Terrorism. (n.d.). *How we work*. https://techagainstterrorism.org/how-we-work

Thompson, E. & Lamphere-Englund, G. (2024). *30 Years of Trends in Terrorist and Extremist Games*. https://gnet-research.org/wp-content/uploads/2024/10/GNET-47-Extremist-Games_web.pdf

Thriving in Games Group. (n.d.). Who is Thriving in Games Group? https://thrivingingames.org/about/

U.S. Senate Committee on the Judiciary. (2023). *Durbin Calls on Online Video Game Industry to Do More to Identify & Remove Extremist Content from their Platforms*. https://www.judiciary.senate.gov/press/dem/releases/durbin-calls-on-online-video-game-industry-to-do-more-to-identify-and-remove-extremist-content-from-their-platforms

United States Government Accountability Office. (2024). *Countering Violent Extremism (GAO-24-106262)*. https://www.gao.gov/assets/gao-24-106262.pdf

Wakabayashi, T. (2020). Hate Speech and Legal Restrictions in Japan. *Zeitschrift für Japanisches Recht, 19*(38),249-263.

Wallner, C., White, J., & Regeni, P. (2025). *Extremism in Gaming Spaces: Policy for Prevention and Moderation*. https://www.rusi.org/explore-our-research/publications/policy-briefs/extremism-gaming-spaces-policy-prevention-and-moderation

Watkin, A. L. (2024). Moderating the moderators: Rethinking regulatory approaches to online harm through a public health lens. In J. Schlegel & R. Kowert (eds.), *Theories of digital games and radicalization* (pp. 37–49). Routledge.

Whittaker, J., Looney, C., Reed, A., & Votta, F. (2021). Recommender systems and the amplification of extremist content. *Internet Policy Review, 10*(2).

Wilson, R. A., & Land, M. K. (2021). Hate speech on social media: Content moderation in context. *Connecticut Law Review, 52*(3), 1029-1076.

Zeyu, L. (2019). Towards an Understanding of Online Extremism in Japan. In *WI '19: IEEE/WIC/ACM International Conference on Web Intelligence*,7-13.

Zipursky, R. (2019). Nuts About NETZ: The Network Enforcement Act and Freedom of Expression. *Fordham International Law Journal, 42*(4), 1325-1374.