

Prevent, Detect, and React: A Framework for Countering Violent Extremism on Gaming Surfaces

GIFCT Year 4 Working Group

February 2025



Table of Contents

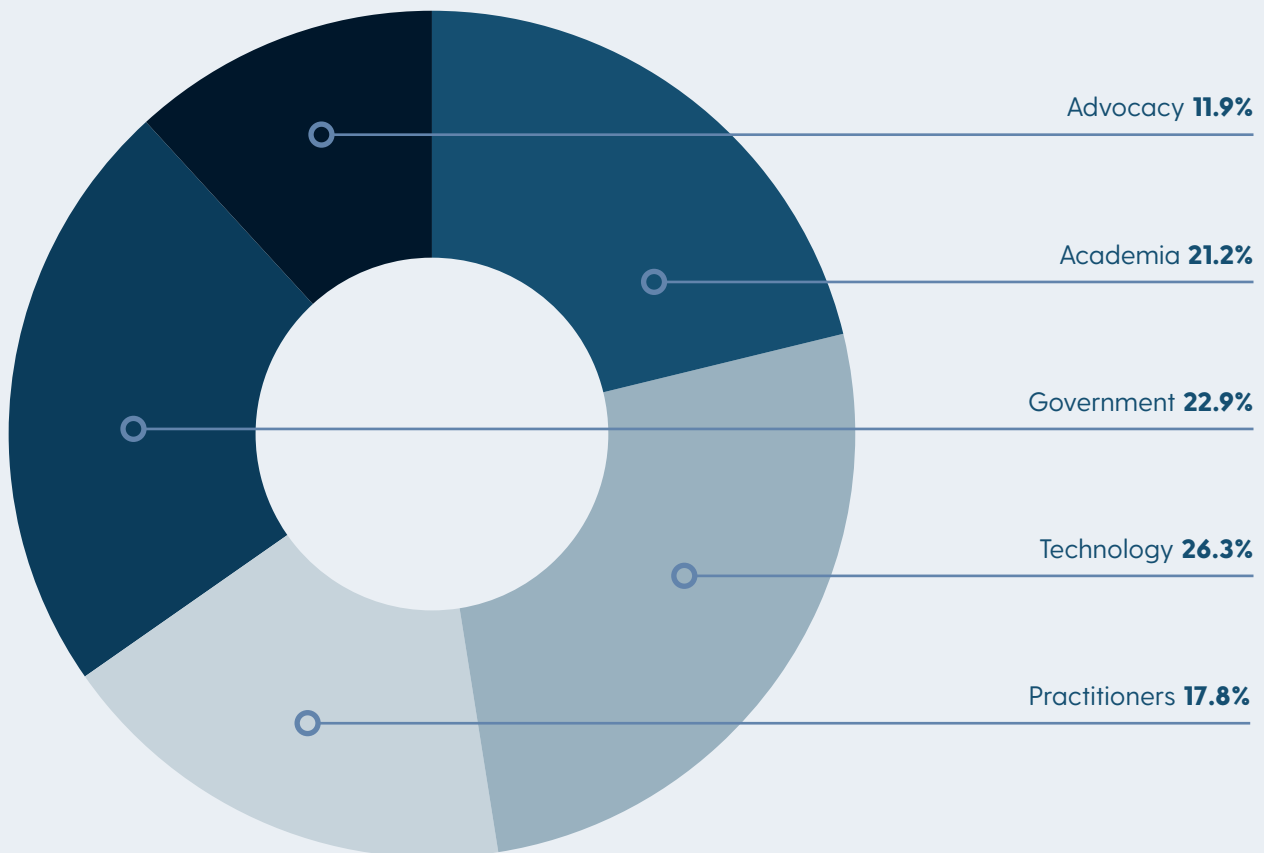
Introducing GIFCT Year 4 Working Groups	3
Overview	6
How to Use This Resource	6
Prevent	7
Denounce Terrorist and Violent Extremist Exploitation of Gaming Platforms	7
Design Games and Gaming Experiences With User Safety in Mind	8
Improve Knowledge and Awareness of Online Harms	9
Support Communities with Positive Interventions	12
Detect	15
Train Humans	15
Train Machines	15
Share Signals	17
React	19
Provide Easy User Reporting	19
Automate Reactive Player Mechanics	20
Pursue Post-Violation Behavioral Change	21
Make Mental Health, Psychosocial Support, and Law Enforcement Referrals	22
Conclusion	23
2024 GIFCT Working Group Participant Affiliations	26

Introducing GIFCT Year 4 Working Groups

In May 2024, GIFCT launched its Year 4 Working Groups to facilitate dialogue, foster understanding, and produce outputs to directly support our mission of preventing terrorists and violent extremists from exploiting digital platforms across a range of sectors, geographies, and disciplines. Started in 2020, GIFCT Working Groups contribute to growing our organizational capacity to deliver guidance and solutions to technology companies and practitioners working to counter terrorism and violent extremism, and offer multi-stakeholder perspectives on critical challenges and opportunities.

Overall, this year's three thematic Working Groups convened **145** participants from **32** countries across **6** continents with **51%** drawn from civil society (**12%** advocacy, **21%** academia, and **18%** practitioners), **23%** representing governments, and **26%** in tech.

Sectoral Breakdown of Working Group Participants



The 2024 GIFCT Working Groups focused on the following three topics:

Hash Sharing Working Group: Evolving Inclusion Parameters

GIFCT has managed and continually enhanced its [Hash-Sharing Database](#) (HSDB), which contains perceptual hashes of terrorist and violent extremist content, since 2017. The current inclusion parameters for the database have evolved through [close consultations with global experts](#). As technologies, content, and types of violent extremist and terrorist groups change, GIFCT aims to continuously review its definitions and parameters to evolve with the times.

In order to enhance the transparency and accuracy of GIFCT's HSDB, this Working Group reviewed the existing inclusion criteria, identified potential gaps, and put forward suggestions to enhance its use. Meetings included consultations with current GIFCT member companies and feedback sessions with global experts. The Working Group resulted in a final report mapping out recommendations and expectations on the future trajectory of GIFCT's HSDB taxonomy.

Incident Response Working Group: Future-proofing GIFCT's Incident Response Framework

GIFCT has continuously evolved its [Incident Response Framework](#) (IRF) since it launched in 2019 following the attacks in Christchurch, New Zealand. The IRF provides a centralized communications mechanism to share news of ongoing incidents that might result in the spread of violent content online, enabling widespread situational awareness and a more agile response among GIFCT member companies. Activations of the IRF allow GIFCT to heighten member awareness of ongoing incidents, circulate critical information regarding related online content, respond to member needs and requests regarding substantive or contextual information, and facilitate related uploads to the HSDB.

This Working Group reviewed and provided suggestions to future proof GIFCT's IRF. To do so, the Working Group evaluated the societal harms around terrorist and violent extremist attacks and mass violent events, examined case studies across different regions, and assessed different types of content, including AI-generated and synthetic materials, and their implications. Meetings included consultations with current GIFCT member companies and feedback sessions with global experts. The Working Group resulted in a set of recommendations regarding GIFCT's IRF. These inputs will inform GIFCT's ongoing efforts to assess lessons learned and good practices in strengthening the IRF and engagement with key stakeholders.

Gaming Community of Practice: Supporting Gaming Tech Safety

GIFCT established its Gaming Community of Practice (GCoP) to foster collaboration, knowledge sharing, and innovation among practitioners in the gaming industry and to enhance the development of best

practices to prevent terrorists and violent extremists (TVE) from exploiting games, gaming-adjacent services, and the gaming community.

This Working Group invited researchers, policy makers, and subject matter experts to support the GCoP by sharing their insights and feedback on the ways in which game-play spaces should evolve their safety work, review safety policies, tools, and practices, and anticipate evolving safety risks. Participants joined GCoP meetings in 2024 to contribute to specific themed discussions to help inform the Community of Practice's themes and goals such as positive intervention potentials across game-play services and sessions with international law enforcement bodies to understand threat signals. Outputs from this year's GCoP include Safety-By-Design one-pagers on best practices on specific gaming surfaces; a review of interventions approaches and research; and early concept work for expanding how terrorist and violent extremist signals can be shared across GIFCT platforms.

Overview

GIFCT hosts Working Groups annually to bring together international experts across sectors to work together in helping tech companies counter terrorist and violent extremist exploitation online. GIFCT's GCoP Working Group ran from May 2024 through January 2025, with the aim of providing a space to share insights and feedback on how gameplay spaces could evolve safety work, review safety policies, tools, and practices, and anticipate evolving safety risks.

The following document, *Interventions for Countering Violent Extremism on Gaming Surfaces*, is a series of explanations that detail various intervention strategies, structured across three stages of Prevent, Detect, and React.

Each stage outlines specific interventions that gaming platforms can implement, organized by clearly described objectives ("How") and supported by real-world examples with links to resources ("Case Studies").

Many thanks to the 2024 GIFCT GCoP members for sharing their expertise, particularly to Linda Schlegel (Peace Research Institute Frankfurt; PRIF) and Rachel Kowert, Ph.D. (Discord) for their helpful suggestions.

How to Use This Resource

Each section below presents a strategic goal (Prevent, Detect, or React), specific objectives that support that goal (such as "Design Games and Gaming Experiences With User Safety in Mind"), and practical case studies with linked resources. The examples given should be seen as illustrative, not exhaustive. GIFCT's academic research arm, the Global Network on Extremism and Technology (GNET), regularly shares the latest research related to gaming and gaming-adjacent platforms, which can be found here. Additionally, the Extremism and Gaming Research Network (EGRN), where GIFCT is a founding member, distributes resources here.

GIFCT offers tailored guidance and support to any online gameplay company seeking to prevent TVE from exploiting their platform.

To continue the conversation, GIFCT can be reached at outreach@gifct.org.

Prevent

Deterring terrorist and violent extremist exploitation of online gaming surfaces. Denounce Terrorist and Violent Extremist Exploitation of Gaming Platforms

How:

- 🌐 A fundamental first step should be explicitly stating that TVE activity is forbidden on gaming services and platforms.
- 🌐 This can be done by developing and implementing unequivocal policy guidance, community guidelines, and/or Terms Of Service language across gaming and gaming-adjacent platforms that prohibit their use for TVE purposes.
- 🌐 Additionally, policy language should be clear that it covers not only content-based violations but behaviors and gaming- and platform-specific functionalities.
- 🌐 While some gaming platforms have clearly taken this step, others still do not directly name violent extremism or terrorism in their policies (Lamphere-Englund & Hartgers, 2024). Notably, this is a requirement for all GIFCT members.

Case Studies and Resources:

- 🌐 Most major social media platforms and many gaming platforms have nuanced policies prohibiting violent extremist and terrorist use cases. For example:
 - 🌐 Meta: [Dangerous Organizations and Individuals | Transparency Center](#)
 - 🌐 YouTube: [Violent extremist or criminal organizations policy - YouTube Help](#)
 - 🌐 Microsoft: [Digital Safety | Policies](#)
 - 🌐 Call of Duty: [Call of Duty® Code of Conduct | FPS Game Terms](#)
 - 🌐 Roblox: [Roblox Community Standards – Roblox Support](#)
 - 🌐 Discord: [Violent Extremism Policy Explainer | Discord](#)
- 🌐 Policy resources include those from the ADL ([Addressing Extremism in Online Games Through Platform Policies](#)), GIFCT ([GIFCT and Member Resource Guide](#)), and the EU Internet Forum.

Design Games and Gaming Experiences With User Safety in Mind

Provide Base-level Threat Awareness for Developers and Studios

How:

- 🌐 Training is especially needed for smaller studios that often lack robust knowledge of the types of risks present on gaming surfaces. These sessions can be delivered through formal university courses in game design and development, informal continuing education training such as the types that GIFCT provides, or via online classes provided for free through platforms like edX and Coursera. It is recommended that this training be framed inside the broader remit of “online harms” and “online safety,” thereby avoiding some of the immediate negative connotations and reluctance that framing around TVE can bring.

Case Studies and Resources:

- 🌐 A pilot masterclass for developers on "[Game Development – Recognizing Right-Wing Extremism, Countering Misanthropy](#)" by the German Federal Agency for Civic Education provides developers with tools for identifying and responding to far right extremism.
- 🌐 The [Swiss Safe Game Guide \(SSGG\)](#), supported by the Swiss Federal Police, educates game developers on identifying signs of extremism, understanding how it manifests, and recognizing which game design elements may be vulnerable to extremist exploitation.
- 🌐 [Digital Thriving Playbook](#) is a collaborative initiative between the Thriving in Games Group (TIGG) and the Joan Ganz Cooney Center at Sesame Workshop, with support from the Riot Games Social Impact Fund. It serves as a comprehensive resource for game developers, offering research-backed practical tools and methods to create online environments that promote positive user experiences and community well-being.

Ensure That Games and Platforms Go Through Red Teaming

How:

- 🌐 Just as major tech releases typically undergo adversarial red teaming to test for potential exploits by malicious actors, games can and should also be subject to red teaming that includes modeling TVE threats.
 - 🌐 For example, the exploitation of Far Cry V as [propaganda by far right actors](#) provides an informative example that could have been potentially mitigated by red-teaming the product prior to launch.
 - 🌐 Similarly, while moderators offer tremendous value for studios and enriching experiences for fans, they also pose substantial risks for manipulation by hostile actors.

- 🌐 Lastly, the rapid rise of *Among Us*, a breakaway hit from a small studio, drew congressional scrutiny. Ensuring a basic level of awareness and safety-by-design accommodations by the developers could have helped mitigate the unwanted attention the game received.

Case Studies and Resources:

- 🌐 [Year Three Working Groups | GIFCT](#), specifically outputs from the Red Teaming Working Group.
- 🌐 Examples of red teaming can be found in [Google's AI Red Team](#), OpenAI's [white papers](#), Apollo Research [outputs](#), and a general overview from HBR ([How to Red Team a Gen AI Model](#)).

Improve Knowledge and Awareness of Online Harms

There is a wide range of intervention modalities – platform-level campaigns, influencer-led initiatives, and game-based interventions – that aim to increase awareness of the harms present on gaming platforms and help change the attitudes and behaviors of users toward TVE.


Platform-Level Campaigns

How:

- 🌐 These might include running platform-wide campaigns to educate users about recognizing TVE content and online harms such as fundraising scams or adversarial activities. Alternatively, campaigns may provide resources and tools for users to enhance their digital literacy, critical thinking, and reporting skills. Other campaign strategies can include:
 - 🌐 Creating dedicated social spaces (e.g., groups, pages) for constructive dialogue and community engagement with specific audiences.
 - 🌐 Implementing features that encourage positive interactions and reward in-game behaviors (e.g., badges, recognition, skins), or using in-game advertising to nudge pro-social acts.
- 🌐 While highly flexible, the impact of these types of campaigns can be difficult to gauge.





Case Studies and Resources:

- 🌐 In 2021, Codemasters, the developer behind the car racing game *DiRT Rally 2.0*, updated the [in-game advertising banners](#) to showcase the UK National Health Service's (NHS) message, "Stay Home, Save Lives," as part of efforts to support COVID-19 prevention measures.



- 
 Tular Nalar is a Google.org-funded and Love Frankie Agency-developed program in Indonesia for digital literacy. It aims to enhance critical thinking across generations to combat misinformation and foster a healthy democratic climate by offering interactive online learning modules and comprehensive curricula tailored for various demographics (including first-time voters and the elderly). Since its inception, Tular Nalar has reached over 55,000 individuals across 38 provinces, effectively equipping participants with safe internet navigation skills.

Influencer-Centered Initiatives

How:


- 
 These typically entail collaborating with popular content creators/streamers to promote positive or alternative narratives.
- 
 The best-run initiatives often provide technical training and/or in-kind funding (such as ad credits for influencers) to create content that fosters empathy and counters extremism.
- 
 Content creation competitions are also sometimes used for micro-influencers to create content linked to P/CVE, while providing incentives to participate, such as training, recognition, and funding.
- 
 Influencer vetting is critical to ensure they feel organically connected to their story/brand. Given the reluctance to talk about TVE, many initiatives focus on adjacent themes that resonate, including memes, humor, and cringe/edgy approaches.


Case Studies and Resources:

- 
 The YouTube *Creators for Change* initiative developed influencer-driven initiatives aimed at preventing and countering violent extremism (P/CVE) on gaming platforms. Launched in 2016 in partnership with local CSOs and UN entities, this program leveraged the influence of popular YouTube creators to promote social good, counter hate speech, and address issues like extremism, discrimination, and misinformation.
- 
 The *Gaming Saves Democracy initiative* from the Friedrich Ebert Foundation partners with popular streamers who play with project staff and actively engage in discussions on topics like extremist activity on platforms.¹
- 
 Gaming-specific P/CVE efforts could partner with well-known gaming influencers and live-streamers who command large audiences on platforms like Twitch and YouTube Gaming. Influencers could embed P/CVE themes into gameplay discussions, commentary, or special event streams.

.....

¹ Linda Schlegel and Matthias Heider, *Extremism Prevention in the Gaming Sector: Approaches, Opportunities, and Development Possibilities*, RadiGame Project, Peace Research Institute Frankfurt (PRIF) (forthcoming).


- 


[Games Done Quick](#) is a biannual event where gamers perform speedruns and live-stream to raise funds for organizations like Doctors Without Borders and the Prevent Cancer Foundation. In 2024, GDQ raised over \$2.5 million during a week-long marathon.
- 

Nearly 10 years ago, the [Abdullah-X campaign](#) featured a fictional animated character who engaged young UK Muslims through YouTube videos on topics like identity, faith, Islamophobia, and the conflicts in Syria and Iraq. Using raw animation, spoken-word performances, and references to pop culture and Islamic teachings, the campaign aimed to offer a relatable and credible voice for its target audience. At a time when ISIS's appeal among youth was at its peak, the initiative sought to reclaim the narrative by presenting counter- and alternative narratives that fostered critical thinking and emotional engagement. The campaign highlights the broader potential of influencer-based interventions, where both real and fictional figures can drive social and behavioral change.

Games-Based Interventions

How:

- 

Serious games, or games for good, can be highly effective educational tools when deployed in the correct context. These games can be built as standalone titles or as experiences inside games as platform services like Roblox or Fortnite, with the latter often proving more appealing and offering more viable paths to reach users. Game-based interventions can also be developed through game jams, where developers are encouraged to rapidly create prototype games around a specific topic. Linda Schlegel and Matthias Heider² define three main categories of these games.
- 

However, in a contested information environment where players are faced with fun, well-designed titles, simply releasing games as an intervention without a clearly defined target audience who have a reason to play them is rarely sufficient.

Case Studies and Resources (Games):³

- Games like [Through the Darkest of Times](#), [Tell me, Inge...](#), and [Voices of the Forgotten](#) create a “culture of remembrance” to deal with historical events (often the Holocaust).
- Games like [Cat Park](#), the [Bad News Game](#), [Gali Fakta](#), and [KonterBUNT](#) combat fake news, disinformation, and hate speech, typically through [inoculation theory](#).
- Games like [Hidden Codes](#), [Leon's Identity](#), and [DECOUNT](#) overtly talk about the potential risks involved in radicalization processes.

.....
 2 Extremism Prevention in the Gaming Sector: Approaches, Opportunities and Development Possibilities, RadiGame Project. Peace Research Institute Frankfurt (PRIF). Translated from German. Pending in January 2025.

3 Schlegel and Heider define three main categories of these games in Extremism Prevention in the Gaming Sector.

Case Studies and Resources (Game Jams):

- 🌐 Organized by Cinereach in collaboration with Anima Interactive, [Just Play: A Game Jam for Climate Futures](#) is a global competition that challenges developers to explore social justice themes through gameplay, aiming to inspire new perspectives on climate change.
- 🌐 Hosted by the University of Pittsburgh, Games for Social Impact is an annual event that encourages participants from diverse disciplines to design games addressing societal issues, fostering creativity and awareness among university students.
- 🌐 While not strictly a game jam, Games for Change Festival has brought together game creators and social innovators since 2004 to drive real-world impact through games and immersive media, addressing various social issues and promoting positive change.

Support Communities with Positive Interventions

Positive interventions can also be built for and among gaming communities, often in partnership with civil society, community groups, or mental health practitioners. These initiatives broadly fall into three categories: Digital Street Work, Direct Community Development and Mentorship, and Law Enforcement Outreach.

Digital Street Work

How:

- 🌐 In pilot initiatives, digital street work on gaming and gaming-adjacent platforms has seen social workers and P/CVE practitioners migrate traditional social work outreach practices to digital platforms, directly engaging with gamers in multiplayer servers, Discord servers, and elsewhere online.
- 🌐 These programs are typically run directly by psychosocial support or mental health organizations, but could also be run in-shop at platforms with the right (certified) contractors.

Case Studies and Resources:

- 🌐 [Good Gaming: Well Played Democracy](#) by the Amadeu Antonio Foundation in Germany employs trained social workers as part of a larger project to engage with youth directly on popular gaming platforms like Discord and Twitch. This digital streetwork allows social workers to meet young people in their everyday online spaces, fostering conversations on topics like belonging, conflict resolution, and diversity. By creating a low-barrier entry

point for dialogue, the project facilitates early prevention efforts that address both general socio-psychological issues and the social roots of radicalization.

- 🌐 [IPčko](#) is a Slovak non-governmental organization that provides comprehensive psychological support to children and young people in crisis situations. Operating a 24/7 anonymous helpline and offline physical spaces that include gaming corners, IPčko offers assistance through phone, online chat, video calls, and email, ensuring accessibility for those in need. The group also engages in digital street work, including through Twitch and Discord.

Direct Community Development and Mentorship

How:

- 🌐 In these interventions, organizations directly support existing gaming and e-sports communities to be more inclusive, less toxic, and more supportive spaces.
- 🌐 Other versions also create communities on gaming-adjacent platforms like Discord to build anti-extremist movements.

Case Studies and Resources:

- 🌐 [Raising Good Gamers](#) is an initiative launched by Games for Change and the Connected Learning Lab at the University of California, Irvine, focusing on promoting positive change among youth in gaming. Programs include partnerships with organizations like Cartoon Network to combat online bullying and collaborations with TED-Ed to empower young voices in creating pro-social gaming cultures.
- 🌐 The [Life After Hate Discord](#) server provides a secure, anonymous space where individuals seeking to leave right-wing extremist groups can connect with counselors, former extremists, and peer support. Launched in the US in 2024, this platform offers a space for those taking initial steps toward disengagement, allowing for discreet access to professional guidance and community support.

Law Enforcement Outreach

How:

- 🌐 In this layer of prevention work, community officers from local police directly engage with target audiences through gaming in informal, low-pressure interactions that promote dialogue and break down barriers.
- 🌐 This form of community engagement helps reduce distrust of law enforcement, build mutual understanding, and teach police officers about gaming ecosystems while kids learn about online harms.
- 🌐 These programs appear to work better in higher-social trust environments (e.g., the UK and

Netherlands) but seem to flounder in low-trust societies (e.g., the United States).

Case Studies and Resources:

- 🎮 The [Cops vs. Kids](#) initiative in North Yorkshire, UK, and the Dutch project [Gamen Met De Politie](#) both use video games as a tool to foster trust and communication between young people and local police.
- 🎮 The New York Police Department (NYPD) introduced a "[Game Truck](#)" in 2021. These mobile gaming units, equipped with PlayStation 5, Xbox Series S, and Nintendo Switch consoles, drew [intensive criticism](#) on launch.

Detect

Recognizing content-based and behavioral signals from terrorist and violent extremist actors.

Train Humans

How:

- 🌐 Detecting TVE behavior and content depends not only on well-crafted policies and Terms Of Service but on trust and safety teams, community moderators, and platform external personnel like law enforcement and social workers who know what to look for.
- 🌐 Research across platform staff, government officials, gamers, and law enforcement reveals considerable differences in knowledge of how TVE actors exploit gaming surfaces.
- 🌐 On-platform training should be targeted not only at trust and safety/policy teams but also at product teams and front-line staff like community moderators.
- 🌐 Off-platform training should also be available for law enforcement (especially at local levels), educators, parents/guardians, and social workers.

Case Studies and Resources:

- 🌐 The [EGRN](#) has developed and delivered a wide range of trainings on this issue set, ranging from 1-hour programs to multi-day in-depth workshops.
- 🌐 [GNET](#), a GIFCT partner group, has led workshops and short sessions on the subject as well.
- 🌐 The EU Radicalization Awareness Network (RAN), now the EU Knowledge Hub, also provided a range of training on the issue.
- 🌐 [Violence Prevention Network \(VPN\)](#) has an [online course](#) (in German) on right-wing extremism and gaming.






Train Machines

How:

- 🌐 Machine-learning-based systems for content flagging and moderation across gaming and gaming-adjacent platforms have advanced significantly in recent years. New systems offer automated tools to moderate voice chats, recognize user-generated content (UGC) with extremist or terrorist images, and more.
- 🌐 Both off-the-shelf private sector solutions, along with open-source moderation tools, are available.
- 🌐 Cultural, linguistic, and group-based identification capacities vary across these tools. While

most offer a substantial benefit and reduction in staff costs, the question of what to do with content once identified remains. Though it may be effective at [degrading VEO capacities over the](#) long term, removal and de-platforming can lead to [substantial backlash](#).

Case Studies and Resources:

-  [Modulate's ToxMod](#) is an AI-driven, proactive, voice-native moderation platform designed to enhance safety in online gaming and social environments by detecting and addressing toxic behavior in real time. A 2024 review of the impact of ToxMod on the [Call of Duty player experience](#) found a 25 percent reduction in toxicity exposure in Call of Duty: Modern Warfare II and Warzone, and a 50 percent reduction in toxicity exposure in Modern Warfare III. Additional moderation tools (text message filtering from Community Sift) in CoD are also discussed in the disruptive behavior report linked above.
-  Roblox's [implementation of machine learning](#) technologies to improve voice safety on its platform indicates significantly reduced harmful interactions in voice communication. The study underscores the role of ML in creating safer, more inclusive digital environments and details innovative solutions that empower moderation efforts to respond proactively to toxicity.
-  Similarly, [Xbox Voice Reporting](#) “allows players to capture and report inappropriate voice activity on any multiplayer game with in-game voice chat on Xbox Series X|S and Xbox One.”
-  [Altitude](#) is an open-source tool developed to assist online platforms in identifying and managing TVE content. It consolidates data from trusted sources, such as GIFCT and Tech Against Terrorism's [Terrorist Content Analytics Platform](#) (TCAP), providing moderators with a unified interface to review and act upon flagged content. This approach enables efficient content moderation, particularly for smaller platforms lacking extensive resources. In July 2024, Jigsaw open-sourced Altitude, transferring its stewardship to Tech Against Terrorism to ensure its continued development and accessibility. Code is [available on GitHub](#).
-  [Perspective API](#) is a free, open-source tool that utilizes machine learning to detect toxic language in online comments. Analyzing text for attributes like toxicity, it assists platforms in moderating content to foster healthier online conversations. Launched in 2017, Perspective has been adopted by various media companies to enhance their comment moderation processes.

Share Signals

How:

- 🌐 TVE actors, like most users, work across platforms online. Yet unlike most users, they will often directly obfuscate their efforts in more moderated settings and outlink or funnel other users from mainstream platforms (Facebook, Roblox, YouTube, etc.) to less moderated surfaces (DLive, Telegram, Signal, etc.).
- 🌐 Tracking TVE actors and behavioral signals across platforms is therefore critical.
- 🌐 However, existing solutions such as the GIFCT Hash Sharing Database and Tech Against Terrorism's TCAP are content-focused, and do not share user information or non-user identifiable behavioral signals.
- 🌐 Platform interviews regularly suggest a need to share such cross-platform behavioral threat patterns, whether tied to specific users and coordinated actions or simply based on generalized trends.
- 🌐 Organizations can share signals through formal data-sharing systems or through network-based associations dedicated to regular information exchange.

Case Studies and Resources:

- 🌐 GIFCT operates a [Hash Sharing Database \(HSDB\)](#) that enables tech companies to collaboratively detect and remove TVE content from their platforms. By creating unique digital "hashes" for harmful images, videos, and audio, the system allows participating platforms to identify and take down matching content quickly, supporting a collective approach to online safety and content moderation.
- 🌐 Similarly, Tech Against Terrorism's [TCAP](#) is a centralized system that identifies, verifies, and alerts tech companies about terrorist content found online. By providing real-time alerts with verified information on terrorist material, TCAP enables platforms in the system to remove harmful content while promoting transparency and accountability in counter terrorism efforts.
- 🌐 Going a step further, [Lantern](#), developed by the Tech Coalition, is a collaborative program that enables technology companies to share information about accounts and behaviors violating child safety policies, particularly concerning online child sexual exploitation and abuse (OCSEA). By facilitating the exchange of "signals"—such as email addresses, usernames, and URLs associated with harmful content—Lantern enhances the ability of participating companies to detect and address cross-platform abuse, thereby strengthening collective efforts to protect children online. Such a system does not exist for TVE content or actors.

- 🌐 Other networks⁴ that facilitate information sharing include:
 - 🌐 [EGRN](#), an international association of over 150 members from the fields of research, civil society and prevention work, international organizations, national and European policy makers, security and police authorities, technology, and gaming.
 - 🌐 The [Christchurch Call Foundation](#), launched in 2019, has over 50 countries, along with many CSOs and platforms, as part of its Call Community.
 - 🌐 [Games for Change](#), a non-profit organization that aims to support positive social change through video games and video game communities.





.....
⁴ Additional networks, especially in Germany, are outlined by Schlegel and Heider in Extremism Prevention in the Gaming Sector.

React




Respond in a timely, proportional, and human rights-compliant manner to TVE content and exploits once detected.

Provide Easy User Reporting

How:

-  While user reporting of violative content has improved across gaming platforms and services, easily reporting TVE along with experiences of hate and harassment remains inconsistent in the sector.
-  Multiple reports have found that reporting harmful content remains challenging. A recent 2024 EGRN consortium study led by RUSI of over 2,000 gamers in seven countries found that 43 percent of gamers polled did not report any harmful content in the last year, while one in ten respondents did not know how to report at all.
-  Follow-up after reporting is also perceived to be inadequate: only 26 percent of gamers felt their voice was fully heard when reporting incidents, while 40 percent reported feeling only somewhat heard, and 34 percent not heard at all.⁵ This aligns with 2021 polling by the [United Nations Office of Counter Terrorism \(UNCOT\)](#), which found that players felt most reports resulted in no action against the offender, leading to the most common reaction to these behaviors of just ignoring them.
-  Improving easy text and voice reporting systems (especially for consoles) should be a **priority**, building on some of the ML tools noted above in the Detect section.

Case Studies and Resources:

-  The [Anti-Defamation League](#) emphasizes that accessible and efficient reporting mechanisms should enable users to easily flag abusive behavior and receive prompt assistance, include comprehensive reporting flags across categories including terrorism, violent extremism, hate, and harassment, and offer protection against retaliation.
-  The Fair Play Alliance (FPA) emphasizes the importance of well-designed user reporting systems in gaming to foster healthy communities and address disruptive behaviors effectively. Key best practices recommended by the FPA in its [Disruption and Harms in Online Gaming Framework](#) include:
 -  Accessible Reporting Mechanisms: Ensure that reporting tools are easy to locate and

.....

⁵ Jessica White, Claudia Wallner, Galen Lamphere-Englund, Love Frankie, Rachel Kowert, Linda Schlegel, Alexandra Phalen, Alex Newhouse, Ashton Kingdon, Gonzalo Saiz Erasquin, and Petra Regeni, Exploring Gendered Socialisation and Radicalisation in Gaming Spaces, RUSI, 2024, <https://rusi.org/explore-our-research/projects/examining-radicalisation-gaming-spaces-through-gender-lens>; Jessica White, Claudia Wallner, Galen Lamphere-Englund, Love Frankie, Rachel Kowert, Linda Schlegel, Ashton Kingdon, Alexandra Phalen, Alex Newhouse, Gonzalo Saiz Erasquin, and Petra Regeni, "Radicalisation through Gaming: The Role of Gendered Social Identity," RUSI, December 17, 2024, <https://rusi.org/explore-our-research/publications/whitehall-reports/radicalisation-through-gaming-role-gendered-social-identity>.

use within the game interface, allowing players to report issues without interrupting their gameplay experience.

- 🌐 Detailed Reporting Categories: Provide specific categories for different types of misconduct, such as harassment, cheating, or hate speech, to facilitate precise reporting and appropriate responses.
 - 🌐 Timely and Transparent Feedback: Communicate with players about the status and outcomes of their reports to build trust in the moderation process and demonstrate accountability.
 - 🌐 Educational Components: Incorporate educational elements that inform players about community guidelines and the impact of their behavior, promoting a culture of respect and understanding.
- 🌐 **Overwatch 2:** Through the Defense Matrix initiative, Blizzard has implemented measures to prevent and discourage disruptive behaviors. This includes text chat changes, voice chat detection and actioning, and improved reporting systems to maintain a positive player experience.

Automate Reactive Player Mechanics

How:

- 🌐 Beyond user reporting, platforms can also take proactive steps against violative and disruptive in-game actions. These can include automated penalties or systems that issue warnings or penalties to players exhibiting disruptive behavior, deterring future incidents.

Case Studies and Resources:

- [Riot Games' Vanguard Anti-Cheat System](#): Vanguard, used for Valorant, features a kernel-level driver that detects and prevents cheating behaviors such as aimbots and wallhacks. Since its implementation, Vanguard has significantly reduced cheating incidents, enhancing fair play and maintaining game integrity. The system's success is evident in how most remaining cheats are reportedly now limited to less impactful methods like triggerbots.
- Blizzard enhanced the [World of Warcraft \(WoW\) reporting system](#) to streamline the process of reporting disruptive behavior. Players can right-click on a player's name in chat, select 'Report Player,' choose the appropriate reason, and add additional details. This modernization aims to facilitate quicker and more accurate responses to violations. Additionally, players who are close to being actioned receive warning messages, providing them an opportunity to correct their behavior before penalties are applied.

Pursue Post-Violation Behavioral Change

How:

- 🌐 Seeking to change player behavior is critical, as bans are often circumvented with new accounts, while companies are also loathe to lose paying users. As such, actually improving toxic or extremist-supporting behavior that falls short of outright illegality should be a priority.
- 🌐 Take This has [highlighted the financial and reputational costs](#) of not addressing toxicity in online gaming. While player safety is a critical concern for game developers, revenue-driven arguments often carry greater weight in driving policy change. To support their argument, a survey of 2,408 players examined how experiences with hate and harassment affect player trust, satisfaction, and spending behavior. The findings underscore that toxic in-game environments not only harm player well-being but also negatively impact player retention and in-game purchases, presenting a compelling business incentive for companies to prioritize trust and safety measures.

Case Studies and Resources:

- 🌐 [Blizzard Entertainment's Endorsement](#), built for Overwatch 2, encourages positive player interactions as part of the studio's broader [Defense Matrix Initiative](#). Players can endorse teammates or opponents for exhibiting good sportsmanship, teamwork, or leadership. Accumulating endorsements grants players rewards, incentivizing constructive behavior and fostering a more positive community experience.
- 🌐 Riot Games has implemented several initiatives to address and reform toxic player behavior based on internal behavioral science studies:
 - 🌐 [The Tribunal System](#): Introduced in 2011, the Tribunal allowed players to review reports of negative behavior and vote on appropriate penalties. This community-driven approach aimed to promote accountability and collective enforcement of community standards. It was disabled in 2014 in favor of other behavioral controls.
 - 🌐 New [automated behavioral systems](#) were created in 2023 to detect and penalize toxic behavior in real time. By utilizing machine learning models, these systems can swiftly identify disruptive actions and issue appropriate penalties, reducing the prevalence of negative behavior.

Make Mental Health, Psychosocial Support, and Law Enforcement Referrals

How:

- 🌐 For offenders: Beyond in-game behavioral change, off-platform referrals for repeat and particularly aggravated TVE offenders can be made to mental health and psychosocial support systems. Those in direct violation of law or who demonstrate an imminent threat to their life or the lives of others should be referred to law enforcement (such mechanisms are already in place at most large gaming platforms and services).
- 🌐 For victims/survivors: Links to support services can also be provided as a follow-up to those who have been repeatedly abused or harassed on the platform. Consolidated support services are not easily accessible at present.

Case Studies and Resources:

- 🌐 [The Redirect Method](#), of which one version was developed by Moonshot in collaboration with Jigsaw, utilizes targeted advertising to direct individuals searching for harmful content toward constructive alternatives. This approach has been applied to counter violent extremism, violent misogyny, disinformation, and other online harms. Gaming services have yet to be publicly used for Redirect work, but they could easily be.
 - 🌐 Recent adaptations of the Redirect Method have integrated mental health support into their interventions. In the Redirect Method USA initiative, Moonshot [conducted experiments](#) to assess whether individuals engaging with violent far right or jihadist content were more inclined to interact with mental health resources. The findings indicated that users seeking to join or engage with violent far right organizations were 48 percent more likely to click on mental health advertisements compared to a control group.
- 🌐 The [VPN](#) in Germany specializes in countering violent extremism and promoting deradicalization through tailored interventions. Their approach includes providing mental health referrals for individuals at risk of radicalization or those seeking to disengage from extremist ideologies. By collaborating with mental health professionals, VPN ensures that participants receive comprehensive support addressing psychological factors contributing to radicalization. This approach could also be tailored to gaming-linked referrals.
- 🌐 The [Games and Online Harassment Hotline](#) was a free, confidential, text-based emotional support hotline that was established to assist individuals in the gaming community facing harassment or emotional distress. Users could text "SUPPORT" to 23368 to connect with trained agents for guidance and resources. The hotline ceased operations in October 2023.
- 🌐 [Cyber Civil Rights Initiative \(CCRI\)](#) offers support and resources for victims of online harassment, including non-consensual image sharing and cyberstalking. Their services encompass a crisis helpline, information on legal rights, and connections to pro bono legal






assistance. While not specific to gaming, it provides a template that could be expanded on.

Conclusion

Gaming ecosystem(s) represent an evolving frontier for interventions aimed at preventing and countering violent extremism (P/CVE). This paper highlights a broad spectrum of strategies, tools, and partnerships that can equip stakeholders to meet the challenges posed by TVE exploitation of gaming surfaces. From prevention-focused initiatives that prioritize user safety and awareness to detection mechanisms leveraging advanced human and AI moderation and reactive measures emphasizing proportional and restorative responses, the interventions presented above are designed to address the multifaceted nature of online extremism and terrorist exploitation of gaming systems.

We have provided a structured approach to addressing these challenges through the three interconnected stages of **Prevent**, **Detect**, and **React**. Together, these stages form a cohesive framework for safeguarding gaming spaces while fostering resilience and inclusivity against violent extremism.

At the **Prevent** stage, the emphasis is on proactive measures that prioritize user safety and awareness. Strategies such as red teaming, influencer campaigns, and including safety-by-design principles in game development aim to mitigate risks before they manifest. By equipping developers, studios, and gamer communities with the tools to identify and counteract extremist exploitation, this stage underscores the importance of foresight and education. This includes:

-  **Denouncing Terrorist and Violent Extremist Exploitation:** Clear and unequivocal policy language that explicitly prohibits such behaviors is fundamental. Community guidelines, Terms of Service, and platform policies must reflect this commitment.
-  **Designing Games and Gaming Experiences With User Safety in Mind:** Developers must integrate safety-by-design principles, supported by red-teaming methodologies to preempt potential exploitation by extremists.
-  **Providing Threat Awareness for Developers and Studios:** Especially for smaller studios, robust training programs—whether formal or informal—can bridge knowledge gaps regarding risks on gaming surfaces.
-  **Improving Knowledge and Awareness of Online Harms:** Platform-level campaigns, influencer-centered initiatives, and games-based interventions offer avenues to educate users, nurture positive interactions, and combat extremist narratives.
-  **Supporting Communities with Positive Interventions:** Digital street work, direct community development, mentorship programs, and law enforcement outreach are critical for engaging vulnerable or targeted populations, fostering trust, and promoting a sense of belonging.

The **Detect** stage focuses on leveraging both human expertise and machine-driven solutions to identify signals of violent extremism across gaming platforms. We categorized enhanced training programs for moderators, law enforcement, and social workers, alongside the deployment of advanced AI moderation tools through three categories:

- 🌐 **Training Humans:** On-platform training for trust and safety teams, product developers, and community moderators, as well as off-platform training for law enforcement, educators, and social workers, enhances the ability to identify risks and respond effectively.
- 🌐 **Training Machines:** Leveraging advanced AI and machine learning systems enables proactive content moderation, voice safety measures, and recognition of extremist behaviors, significantly reducing reliance on manual oversight.
- 🌐 **Sharing Signals:** Platforms must collaborate across the digital ecosystem to track and share behavioral patterns and content trends. Mechanisms like the GIFCT HSDB and Tech Against Terrorism's TCAP demonstrate the value of cooperative frameworks in addressing cross-platform exploitation.

Finally, the **React** stage emphasizes timely, proportional, and rights-respecting responses. Mechanisms such as streamlined user reporting, automated reactive player mechanics, and off-platform referrals for offenders and victims are designed to address incidents effectively while prioritizing accountability and rehabilitation. Reactive responses should focus on:

- 🌐 **Providing Easy User Reporting:** Simplified reporting mechanisms for gamers, coupled with transparent feedback loops, build user trust and enhance the efficacy of moderation systems.
- 🌐 **Automatic Reactive Player Mechanics:** Platforms can deploy automated systems to flag and penalize disruptive behaviors in real time, deterring future violations.
- 🌐 **Post-Violation Behavioral Change:** Rehabilitation-focused approaches—such as engagement with offenders to promote constructive behaviors—help reduce recidivism while maintaining platform inclusivity.
- 🌐 **Mental Health, Psychosocial Support, and Law Enforcement Referrals:** Both victims and offenders benefit from referrals to appropriate support services. Mental health initiatives like the Redirect Method can address underlying issues, while involving law enforcement ensures a measured response to imminent threats.

The 2024 GIFCT GCoP Working Group, along with EGRN, have laid a strong foundation for advancing P/CVE within the gaming industry, showcasing the power of cross-sector collaboration. The Prevent, Detect, and React framework not only offers a roadmap for addressing current challenges but also

sets the stage for ongoing innovation and adaptability in response to emerging threats.

By prioritizing inclusivity, transparency, and respect for human rights, stakeholders can foster gaming environments that are not only resistant to exploitation but actively promote resilience, empathy, and positive engagement. As gaming platforms evolve, this structured approach ensures that they have a pathway to becoming inclusive and safe spaces of community and play.


2024 GIFCT Working Group Participant Affiliations⁶

Academia	Advocacy	Practitioner & Researcher	Government & Intergovernmental	Tech
ACUNS, ISACA	ADL	Brookings Institution	Aqaba Process, Jordan Government	Amazon Web Services (AWS)
American University	All Tech is Human (ATIH)	Centinel	Australia, Department of Home Affairs	Discord
Center for Cyber Strategy & Policy, School of Public and International Affairs, The University of Cincinnati	ARTICLE 19	Digital Security Group	Christchurch Call	Dropbox
Central University of Gujarat	ASEAN Coalition to Stop Digital Dictatorship	Fem AI	Department of Internal Affairs NZ (Digital safety and illegal harms)	ExTrac AI
Collaboratory Against Hate, Carnegie Mellon University and University of Pittsburgh	Association of British Muslims	Global Disinformation Index	eSafety Commissioner Australia	Giphy
Columbia University School of International and Public Affairs (SIPA)	Internet Society	Hedayah	European Commission	GoDaddy.com
Extremism and Gaming Research Network (EGRN)	KizBasina (Just-a-Girl) NGO	Jihadoscope	Federal Bureau of Investigation (FBI)	Insikt AI and Dataietica.org
Georgetown University	Koan Advisory	Moonshot	Federal Ministry of the Interior and Community, Germany	ISACA Kenya
Hesse State University of Public Management and Security	Moroccan Observatory on Extremism and Violence	Online Safety Exchange	Netherlands Ministry of Justice and Security	Meta
Macquarie University	Policy Center for the New South	Peace Research Institute Frankfurt (PRIF)	New Zealand Classification Office	Microsoft

⁶ This table highlights participants across all Year 4 Working Groups.



Royal Holloway, University of London	Search for Common Ground	Swansea University	Ofcom	Mozilla Corporation
RUSI	Southern Poverty Law Center	Tech Against Terrorism	OSCE Secretariat, Action against Terrorism Unit	Nexi Group
Sapienza University of Rome (Italy)	Take This	The Millennium Project (South Asia Foresight Network) [SAFN]	Public Safety Canada	Niantic Labs
Swansea University		Tremau	U.S. Department of Homeland Security	Resolver, a Kroll business
Trinity College Dublin			U.S. Department of State	SoundCloud
University of Cambridge			UK Home Office	SpaceYaTech and Africa ICT Alliance
University of Essex, Department of Government			UNICRI - United Nations Interregional Crime and Justice Research Institute	Twitch
University of Paris Cité (France)			United Nations	X
University of South Wales			United Nations Office of Counter Terrorism (UNOCT) / United Nations Counter- Terrorism Centre (UNCCT)	Xbox
University of Sussex			Virginia State Police, USA	YouTube
University of Waterloo				Yubo
Vox-Pol Institute				



Copyright © Global Internet Forum to Counter Terrorism 2025

Recommended citation: Galen Lamphere-Englund, Interventions for Preventing and Countering Violent Extremism on Gaming Surfaces (Washington, D.C.: Global Internet Forum to Counter Terrorism, 2025), Year 4 Working Groups.

GIFCT is a 501(c)(3) non-profit organization and tech-led initiative with over 30 member tech companies offering unique settings for diverse stakeholders to identify and solve the most complex global challenges at the intersection of terrorism and technology. GIFCT's mission is to prevent TVE from exploiting digital platforms through our vision of a world in which the technology sector marshals its collective creativity and capacity to render TVE ineffective online. In every aspect of our work, we aim to be transparent, inclusive, and respectful of the fundamental and universal human rights that TVE seek to undermine.



www.gifct.org



outreach@gifct.org