

# Threat Surfaces in Games: Challenges and Best Practices

**GIFCT** Year 4 Working Group

January 2025

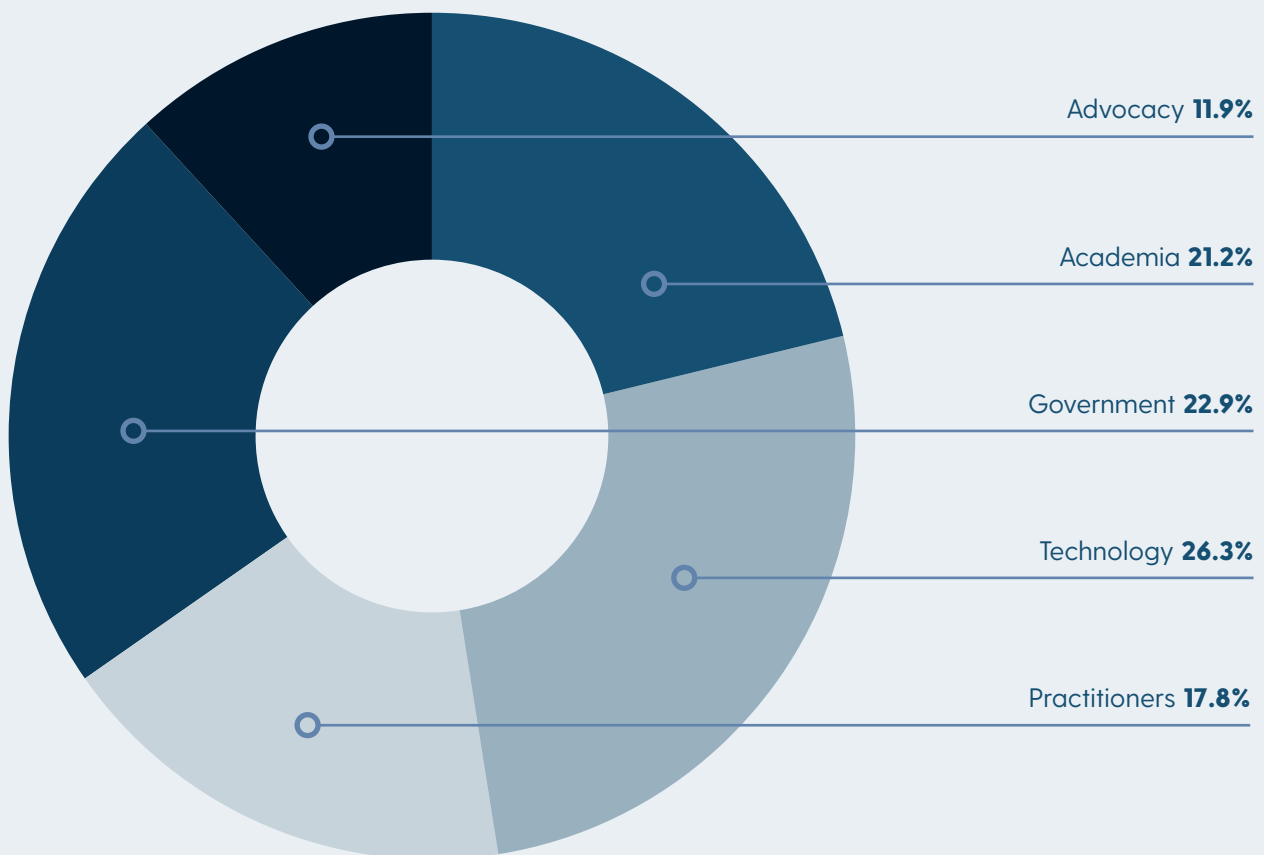


## Introducing GIFCT Year 4 Working Groups

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.

# Threat Surfaces in Games: Challenges and Best Practices

Rachel Kowert, Ph.D., and Sarah Chittick

## 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 Gaming Community of Practice](#) ran from May 2024 through January 2025, with the aim of providing a space to share insights and feedback on the ways in which game-play spaces online could evolve safety work, review safety policies, tools, and practices, and anticipate evolving safety risks.

Threat Surfaces in Games: Challenges and Best Practices is a series of one-page documents that provide information and context around threat surfaces of in-game and game adjacent spaces online that have been and can be exploited by TVE. Each section focuses on a surface, defines the threat, shares actionable proactive and reactive strategies, identifies challenges moving forward, and provides illustrative case studies of solutions.










Thank you to the members of the 2024 GIFCT Gaming Community of Practice for sharing your expertise. Special thanks to Galen Lamphere-Englund (Extremism and Gaming Research Network), Suraj Lakhani (University of Sussex), Linda Schlegel (Peace Research Institute Frankfurt), and Michael Miller Yoder (Carnegie Mellon University) for their contributions. We hope that this project will help provide a foundation of knowledge across both big and small gaming studios and platforms and offer a significant step forward in standardizing some of the innovative and more effective best practices that have been developed.

## How to Use This Resource

These one-page sections identify commonly seen threats across four threat surfaces (voice communication, text-based communication, live streams, and user-generated content). These surfaces were chosen as they were identified by the working group as the most common and vulnerable threat surfaces related to game and game adjacent spaces. Each one-pager seeks to highlight key policy, operational, and design best practices that gameplay spaces might employ to mitigate these threats. 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](#).

GIFCT offers tailored guidance and support to any gameplay company seeking to prevent TVE from exploiting their platform. To continue the conversation, GIFCT can be reached at [outreach@gifct.org](mailto:outreach@gifct.org).

## Initial Questions for Gaming Companies to Consider

-  Do you have a policy prohibiting words or phrases linked to terrorist and/or extremist groups or individuals?
-  Are you proactively detecting words, phrases, or other content linked to terrorist and/or extremist groups or individuals? (e.g., utilizing the GIFCT Hash-Sharing Database or deploying keyword detection)
-  Is your system set up to capture the data points you need to flag terrorist and extremist content? To positively identify flagged terrorist and extremist content?
-  Are there automated processes that could be put in place to filter out the most egregious threats at scale? For text-based content? Voice content? User-generated content?
-  Do you currently employ network disruption strategies to undermine known terrorist and extremist actors?
-  What are your processes for addressing repeat offenders?
-  Do players know if they should and/or how they can flag terrorist and extremist content?
-  Do the reporting mechanics make it easy for a player to flag an immediate threat?
-  Is there a way to flag a more significant threat that requires immediate review?

## Threat Surface: Voice Communication

### Using voice communication to radicalize, organize, or mobilize

#### Identifying the Threat:

- 🎧 Using voice-to-transcription technology for in-game chat to identify keywords (e.g., use of violent language, terrorist and extremist terminology or narratives)
- 🎧 Extra-linguistic markers of conversations (e.g., silence after someone says something or marked increase in voice activity following a keyword flag)

#### Threat Examples:

- 🎧 [U.N. Study Examining the Intersection Between Gaming and Violent Extremism](#)
- 🎧 [The Online Gaming Ecosystem: Assessing Digital Socialisation, Extremism Risks and Harms Mitigation Efforts](#)

#### Proactive Best Practices:

- 🎧 Auto-mute for violative players
- 🎧 Clear guidelines in code of conduct about expectations and consequences
- 🎧 Education/awareness building about harms of extremist language

#### Reactive Best Practices:

- 🎧 Player reports
- 🎧 Automated word detection via keyword flagging with keyword searches and machine learning models (in-house or through third-party providers)
- 🎧 [Voice reporting](#) and [voice recording and evaluation](#)

#### Challenges:

- 🎧 Balancing freedom of speech and privacy regulatory and ethical challenges
- 🎧 Computational resources to process voice data in real time
- 🎧 Skepticism around voice moderation technology may negatively impact user trust
- 🎧 Determining “toxic” or violative words, across geographies and languages
- 🎧 Increased error rates with voice chat
- 🎧 Reporting barriers for players, such as ease of system use, lack of robust or systematic reporting categories

- 🎮 Language changes and unique style of language may vary across games (including the shifting landscape of subcultural language associated with extremist rhetoric)
- 🎮 Data collection restrictions (i.e., not collecting or retaining voice data)

**Solution Case Studies:**

- 🎮 [The Impact of AI Voice Moderation on the Call of Duty Player Experience](#)
- 🎮 [Anti-Toxicity/Disruptive Behavior Progress Report for Black Ops 6](#)
- 🎮 [Deploying ML for Voice Safety in Roblox](#)
- 🎮 [Riot Games Launches Voice Recording in Valorant](#)



## Threat Surface: Text-Based Communication

### Using text-based communication to radicalize, organize, or mobilize

#### Identifying the Threat:

- 🌐 Keywords (e.g., use of violent language, extremist terminology, or narratives)
- 🌐 Coded language (e.g., 88 being code numbers for Heil Hitler)
- 🌐 Use of [ASCII Characters](#) (i.e., character encoding standard from the American Standard Code for Information Exchange) to [depict](#) or [elicit](#) harmful content

#### Threat Examples:

- 🌐 [Extreme Right Radicalization of Children via Online Gaming Platforms](#)
- 🌐 [Far-Right Communities Rallied on Discord for the Unite the Right Rally](#)

#### Proactive Best Practices:

- 🌐 Remove chat functionality for violative players
- 🌐 Education/awareness and digital youth work (e.g., [Google's Interland](#), [YouTube's Hit Pause Campaign](#), [Roblox's partnership with the Simon Wiesenthal Center](#))
- 🌐 Clear guidelines in Code of Conduct about expectations and consequences for players
- 🌐 Raise awareness about Terms of Service, such as with player-facing messaging upon log-in
- 🌐 [Viewer controls](#) (e.g., blur or hide) for specific content types



#### Reactive Best Practices:

- 🌐 Player reports
- 🌐 Automated word detection via keyword flagging with keyword searches and machine learning models ([both commonly flagged words and custom words](#))
- 🌐 Linguistic processing (e.g., NLP algorithms)

#### Challenges:

- 🌐 Continued need to update keywords, memes, etc., to remain relevant
- 🌐 Language changes and/or unique style of language may vary across games
- 🌐 Ethical considerations around freedom of speech
- 🌐 Scaling interventions to address sheer number of text-based messages

**Solution Case Studies:**

-  [Modulate develops a violent radicalization category for Tox Mod](#)
-  [Riot Games' Valorant: Muted Word List](#) allows players to type in variations of words/phrases that they personally would not like to see appear in-game (additions are also used to improve automatic detection)

**Further Reading:**

-  [GIFCT Tech Trials: Combining Behavioral Signals to Surface Terrorist and Violent Extremist Content Online](#)
-  [Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience](#)
-  [Good Gaming: Text-Based Digital Streetwork](#)

## Threat Surface: Live-Streaming

Live-streaming TVE content; using live streams to share propaganda and radicalize; fundraising through live streams

### Identifying the Threat:

- 🌐 Key terms/extremist narratives and symbols in spoken word, on screen content, or in comments (e.g., use of violent language, extremist terminology, narratives)
- 🌐 Fundraising for terrorist and extremist groups through “tips” or streamer rewards
- 🌐 Known names or images of TVE actors or accounts appear in the stream

### Threat Examples:

- 🌐 [Extremist Activity on Video- and Live-streaming Platforms](#)
- 🌐 [Streaming Fraud Contributing to Money Laundering, Terror Financing](#)
- 🌐 [Investigative Report on the Role of Online Platforms in the Tragic Mass Shooting in Buffalo on May 14, 2022](#)

### Proactive Best Practices:

- 🌐 Automated takedowns using object-detection algorithms for weapon detection
- 🌐 Player reports
- 🌐 Allow for user moderation of specific live events (e.g., [Discord's Stages](#))
- 🌐 Remove live-streaming applications on platforms where it is not possible for the application or digital surface to reduce the threat surface
- 🌐 Utilize [GIFCT's Hash-Sharing Database](#) to identify and remove known TVE content
- 🌐 Strategic network disruption of groups known to share terrorist or extremist content

### Reactive Best Practices:

- 🌐 Takedowns of content while live and permanent removal post takedown
- 🌐 Banning/removal of streaming channel and/or streaming personalities

### Challenges:

- 🌐 Content must be monitored in real time
- 🌐 Potential virality makes it difficult to reduce the threat as content was likely to have been captured/shared even if taken down quickly

- 📹 Challenges in identifying violent behavior as there may be no obvious signal
- 📹 Relevant content may be hidden or muddled with other content
- 📹 Automated tools may not be tailored to specificity of gaming spaces (e.g., differentiating between sounds of gunshots in-game and IRL)
- 📹 Takedown of captured content across online ecosystems, such as other gaming-related spaces, archive sites, etc.

**Solution Case Studies:**

- 📹 [How Twitch took down Buffalo shooter's stream in under two minutes](#)

**Further Reading**

- 📹 [Deploying a Community-Driven Moderation Intervention on Twitch](#)
- 📹 [Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting](#)
- 📹 [Build a Weapon Detection Algorithm using YOLOv7](#)

## Threat Surface: User-Generated Content

Using custom objects, third-party modifications, or bespoke games to radicalize or flag/congregate like-minded users.

### Identifying the Threat:

- 🌐 User-Generated Content (UGC) can include the creation of usernames, third-party modifications, custom objects (e.g., skins), comments, and bespoke games, but can also refer to the removal of content, such as removing all women, people of color, etc.
- 🌐 UGC can also include violent words and phrases, extremist symbols and art, relevant references to historical events and prominent figures (e.g., Hitler), and lesser-known ideological references appropriated by TVE groups
- 🌐 Individual lists of UGC hosted by civil society, academia, and/or government (these should be considered within the context of individual company policy)
- 🌐 Sharing of UGC in channels known to spread TVE content by cross-platform efforts like GIFCT (e.g., terrorist or extremist groups on Discord or Telegram)
- 🌐 Developers of bespoke games with ties to terrorist or violent extremist organizations, “branding” by extremist organizations and advertisement for the game, title, content, and comment threads that point to extremist gameplay
- 🌐 Computer vision models can be trained to identify markers of UGC extremism
- 🌐 UGC detection can be used to proactively surface accounts for review and feed classifiers

### Threat Examples:

- 🌐 [Use of Mod Platforms by Extremist Actors](#)
- 🌐 [30 Years of Trends in Extremist Games](#)
- 🌐 [Violent Political and Antisemitic Content on Fortnite](#)
- 🌐 [Steam-Powered Hate: Top Gaming Site Rife with Extremism & Antisemitism](#)

### Proactive Best Practices:

- 🌐 Inability to build certain assets (e.g., forbidding the creation of a swastika shape)
- 🌐 Create P/CVE content in UGC spaces to start conversations with players who are attempting to create terrorist or extremist assets
- 🌐 Add this threat as a violation of Terms and Conditions
- 🌐 Ban certain words in usernames and create symbol-based hashing/name-based wordlists

- 🌐 Check (automatically or otherwise) for markers of terrorism or violent extremism
- 🌐 Collaborate with third-party sites to ensure their services are not used for TVE UGC
- 🌐 Strategic network disruption of groups known to share terrorist or extremist content

#### **Reactive Best Practices:**

- 🌐 Player reporting tools
- 🌐 [Logo detection](#)
- 🌐 Word list detection (for avatar, game, and modification names)
- 🌐 Deplatform/block accounts creating prohibited UGC

#### **Challenges:**

- 🌐 Self-made visual content or third-party modifications may be too subtle, ambiguous, and context-specific for automated detection or even manual flagging
- 🌐 Large amount of content that relies on manual human review (putting onus on players to report content if not reviewed prior to publishing)
- 🌐 Restricting UGC may create backlash from players (freedom versus security)
- 🌐 Third-party sites may not be eager to collaborate
- 🌐 New usernames will constantly surface and terminology is constantly changing
- 🌐 Repurposing of UGC in “Let’s play” videos
- 🌐 Difficult to identify UGC in bespoke games considered prestige products for terrorist or extremist organizations if/when shared within closed communities
- 🌐 Little research on UGC to draw from

#### **Solution Case Studies:**

- 🌐 [Roblox De-platforms Patriot Front Content](#)
- 🌐 [More than Sports: Building Resilience against Extremism in Esports](#)

#### **Further Reading**

- 🌐 [Hateful Usernames in Online Multiplayer Games](#)
- 🌐 [Computer Scientists Build, Test, and Present Model to Curb Online Ban Evasion](#)

## 2024 GIFCT Working Group Participant Affiliations<sup>1</sup>


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

<sup>1</sup> 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)		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: Dr. Rachel Kowert and Sarah Chittick, Threat Surfaces in Games: Challenges and Best Practices (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](http://www.gifct.org)



[outreach@gifct.org](mailto:outreach@gifct.org)