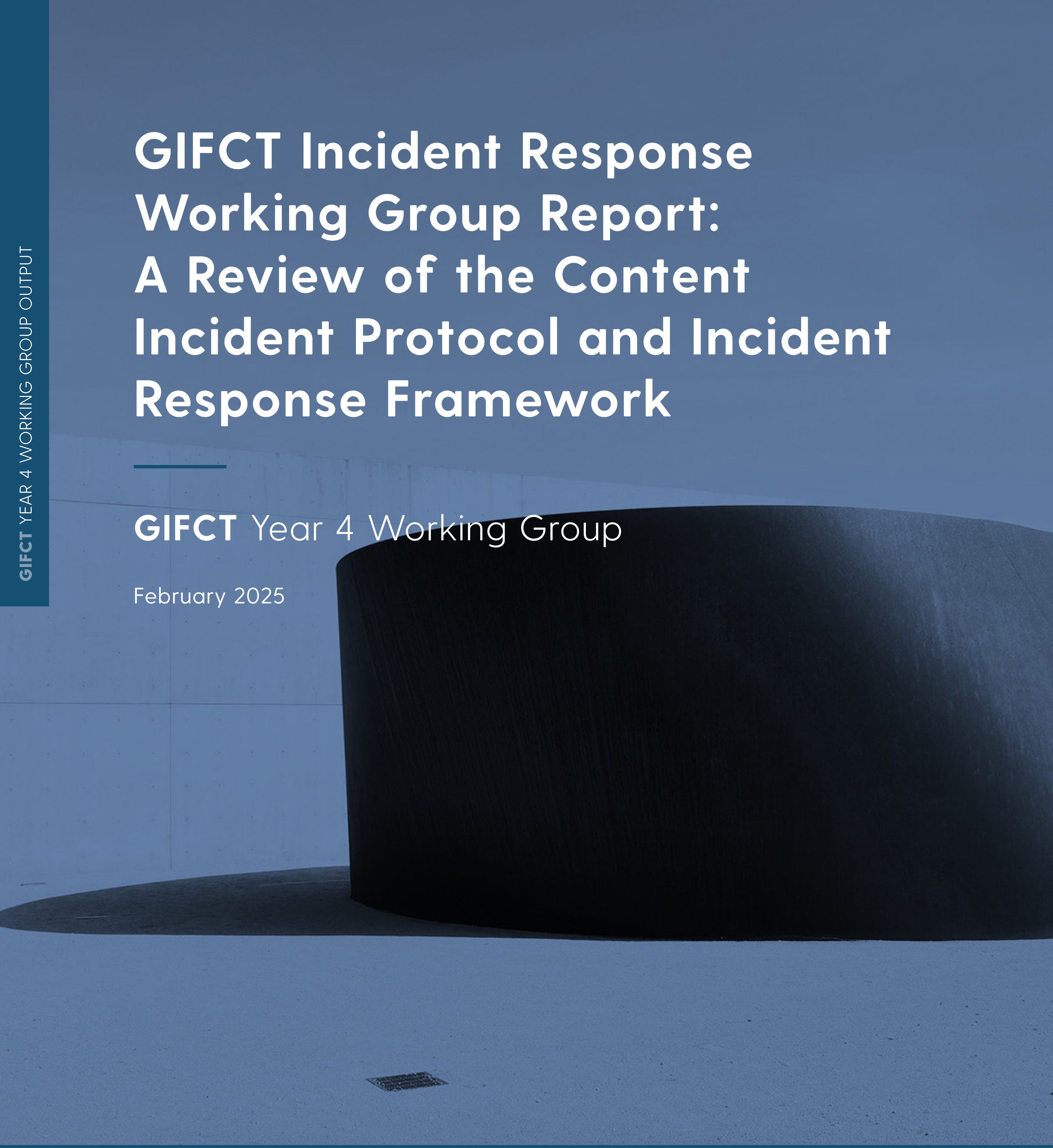
GIFCT YEAR 4 WORKING GROUP OUTPUT

GIFCT Incident Response Working Group Report: A Review of the Content Incident Protocol and Incident Response Framework

GIFCT Year 4 Working Group

February 2025



Middlebury Institute *of*
International Studies at Monterey

GIFCT
Global Internet Forum
to Counter Terrorism

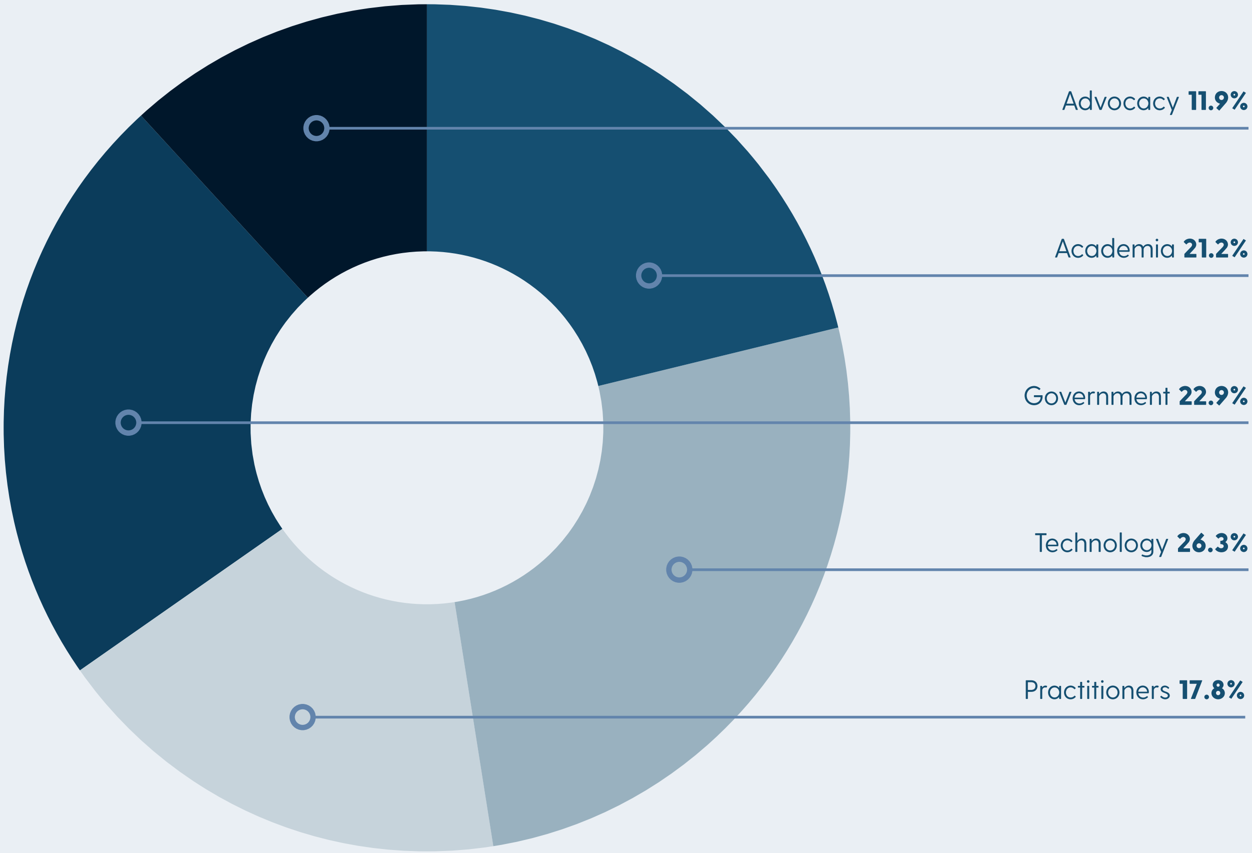
Table of Contents

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.

GIFCT's Foreword

The Global Internet Forum to Counter Terrorism's (GIFCT) Incident Response Framework (IRF) was created as an effort to mitigate harm during an offline terrorist or violent extremist (TVE) event with a significant online component. The IRF employs real-time coordination between GIFCT member companies to rapidly identify, respond, and hash terrorist content, helping prevent its viral spread across platforms.

As a part of GIFCT's 2024 Incident Response Working Group (IRWG), the Center on Terrorism, Extremism, and Counterterrorism (CTEC), serving in a consulting and advisory capacity, reviewed the efficacy of the current IRF and developed recommendations for improvement.[1] Specifically, CTEC evaluated societal harms around TVE attacks and mass violent events, assessed the IRF's effectiveness across regions, and examined emerging challenges like AI-generated and synthetic content.

Summary of Recommendations[2]

Revise IRF: The CTEC team found the nomenclature of the IRF unclear, particularly the names and focus of each level. To clarify the purpose and actions involved in each type of IRF activation, GIFCT should consider revising the IRF playbook, including altering the categories and renaming activation types. CTEC recommends a more intuitive naming system that better reflects the escalating nature of responses. CTEC also recommends GIFCT restructure the current IRF (and future iterations of the IRF) to include a greater focus for activation on the "incident" over the content.

Expand IRF Involvement: The CTEC team recommends that GIFCT engage with multiple stakeholders to review considerations such as its effectiveness and human rights impact. CTEC specifically recommends establishing an annual grant program or fellowships to increase academic partnerships, implementing independent reviews of each IRF case, and developing a robust after-action system to track specific metrics.

Improve IRF Communication: The CTEC team noted that communications around the IRF, particularly with external stakeholders, are in need of streamlining. CTEC recommends centralizing IRF information in a single location, implementing a dashboard functionality for data visualization, and establishing a baseline notification system for all stakeholders.

[1] Final editorial decisions rest solely with GIFCT and constituent elements may not necessarily reflect the views of CTEC nor the Middlebury Institute of International Studies (MIIS) and/or Middlebury College.

[2] It should be noted that CTEC was not provided an opportunity to observe or participate in the implementation process of a Content Incident Protocol (CIP) activation despite multiple CIP activations occurring during the timeframe when CTEC conducted its review. While this did not prevent CTEC from making recommendations on improving the IRF, it likewise did not maximize the ability of CTEC to make the most informed recommendations possible. This limitation should be considered in the future with respect to external parties reviewing the IRF process.

Address Emerging Challenges: CTEC encouraged GIFCT to conduct a systematic review of global terrorist incidents to establish regional prioritization but postponing the inclusion of AI-generated content until more robust data is available. CTEC also recommends maintaining the exclusion of bystander footage due to practical implementation challenges and avoiding expansion into state-level incidents at this time.

CTEC Introduction

CTEC Framing

This report was prepared on behalf of GIFCT by CTEC serving in a consulting and advisory capacity. As such, final editorial decisions rest solely with GIFCT and constituent elements may not necessarily reflect the views of CTEC nor the Middlebury Institute of International Studies (MIIS) and/or Middlebury College.

As a preface to all conclusions and assessments within this report, CTEC was not provided an opportunity to directly observe or participate in the implementation process of a Content Incident Protocol (CIP) activation. While this did not prevent CTEC from making recommendations on improving the IRF, it likewise did not maximize the ability of CTEC to make the most informed recommendations possible. This lack of insight should be considered in future use of external parties to review the IRF process on behalf of GIFCT.

Executive Summary

The Global Internet Forum to Counter Terrorism's (GIFCT) Incident Response Framework (IRF) has enormous potential. Designed to provide robust and rapid impact on mitigating the initial and subsequent harms that occur online during a terrorist attack which utilizes digital mechanisms such as live streaming.

Branded as a real-time process by which GIFCT's member companies can exchange and hash terrorist content associated with a terrorist attack, the entities that the IRF is intended to service are well positioned to meaningfully mitigate harm as an attack unfolds and as its secondary impacts unfold online (e.g., viral sharing of footage of the incident).

Report Roadmap

This report is organized into four main sections:

- Background and Current Structure: Details the history, development, and current implementation of the IRF
- Efficacy of the IRF: Analyzes key challenges in authorization procedures and measurement
- Emerging Challenges: Examines geographic biases, AI-generated content, and bystander footage considerations
- Recommendations: Provides detailed suggestions for improving the IRF's effectiveness and scope

Background

History and Structure of GIFCT Incident Response Framework

Following the Christchurch attacks in March 2019, GIFCT established a formal Incident Response Framework (IRF) to facilitate communication among and coordinate responses by member companies during and in the immediate aftermath of TVE events. The attack was live-streamed on Facebook by its perpetrator, Brenton Tarrant, who took steps to ensure his video and 74-page manifesto would be widely circulated online.[3] In the wake of the attack, GIFCT member companies established methods for centralized communication during a violent event that included live or near-live posting of material by the perpetrator. Activations of the IRF allow GIFCT to raise member situational awareness, communicate timely information, enable faster decisions with respect to removing event-related material from online platforms, and facilitate incident-related uploads to the hash-sharing database (HSDB) to stop the spread of terrorist and violent extremist content (TVEC) online.

The Content Incident Protocol (CIP) was created as GIFCT's first IRF activation type in response to the Christchurch terrorist attack, establishing a framework to address live-streamed mass violence. Its origins significantly informed the type of TVE that IRF activation is inherently best positioned to address. Tarrant's meticulously planned attack created a latent cultural script for others to adopt and emulate. In subsequent years, the tactical and aesthetic decisions he made have been incorporated by over twenty terrorist attackers, spawning an entire terroristic "Saints" subculture that perpetuates this hyper-specific modality. Three of the four core criteria for a CIP reflect key elements of this attack pattern:

[3] Report: Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on March 15, 2019.

- A real-world terrorist, violent extremist, or mass violence event;
- Live-streamed or recorded video by perpetrator or accomplice;
- Depicting murder or attempted murder.

Three of the eight CIP activations to date have involved post-Tarrant militant accelerationist terrorism, demonstrating how this attack modality aligns with CIP thresholds (additional cases would undoubtedly fit the CIP criteria had they been successful in various aspects of their planned attack). The persistent efforts to exploit live-streaming capabilities continue as militant accelerationist networks actively promote Tarrant's blueprint in order to inspire further attacks.

The recent CIP activation in Eskişehir, Turkey, exemplifies this pattern. The attack—a live-streamed mass casualty event conducted by a lone perpetrator—aligned perfectly with the CIP's original design parameters. This stands in stark contrast to events like the October 7 Hamas-orchestrated attack, which involved multiple perpetrators, devices, and accounts in a large-scale coordinated act of violence. This comparison highlights how the CIP's design works well detecting and responding to individual, live-streamed attacks but complex, multi-actor operations add new layers of complexity and difficulty in verifying what is perpetrator content.

IRF Expansion

Since 2019, GIFCT has consulted several times with stakeholders from civil society, academia, governments, and tech companies to refine the IRF processes. This includes two multi-stakeholder tabletop exercises to test IRF efficacy and an annual Incident Response Working Group (IRWG) focused on GIFCT's incident and crisis response capabilities.

In 2021, GIFCT expanded the IRF to include two new activation types, the Content Incident (CI) and Incident (I). Each activation is associated with a different GIFCT and member response aimed at mitigating various risks, such as TVEC virality or identifying additional event-related content shared by a perpetrator. The (CI) focuses on responding to and hashing non-live-streamed perpetrator-produced media from an ongoing TVE event. An (I) activation entails responding to an ongoing TVE event that does not clearly involve perpetrator-produced content depicting mass violence but may require GIFCT to share information and resources with members to aid their operations.

How the IRF Works

CTEC noted that the current IRF structure has competing systems in explanations for how criteria and mechanisms are enacted. Publicly, the IRF is most routinely described as a three-tiered system of Incident, Content Incident, and Content Incident Protocol which buckets activity of concern into "escalating levels of online risk."[4] CTEC understands these three levels to be severity measures of a given incident, representing the significance of potential and real harms from violent extremist, terrorist, or mass violence activity that each category is meant to address. They are not progressive, but rather exist as a placement of the event to then guide how GIFCT and member companies triage a given incident. While they can evolve as events unfold, they are not intended to be a linear protocol.

Further complicating this three-tiered system, internal GIFCT documents describe the IRF through a different lens entirely: as a seven-stage approach designed to be fluid and cyclical. These stages are Identify, Validate, Assess, Activate, Notify, Conclude, and Debrief. No articulation is publicly available on how any of these stages are executed or evaluated. Adding to the framework's complexity, these seven stages were also identified as applying to all three incident levels of the current IRF structure despite each having very different action components.

What follows is CTEC's understanding of the three-stage IRF process from before an event triggers a response to its ultimate conclusion with a closing out (or deactivation) of the IRF due to the incident being triaged (having concluded).

In the first phase of activity, GIFCT is informed of (or detects) signal(s), which may indicate an incident associated with violent extremism, terrorism, or mass violence is imminent or ongoing. This signal can be solely digital, offline, or a mixture of the two mediums. Once that initial signal has been ingested and a preliminary investigation begins, GIFCT staff initiates a rolling process involving the critical assessment of the constituent elements of that signal against the criteria of the three levels of the IRF (Incident, Content Incident, and Content Incident Protocol) to determine which (if any) of those levels should be assigned to the nascent incident signal.

In the second phase of activity, GIFCT takes a leading role in shepherding the nascent signal and corresponding intelligence around a given incident into an actionable state of play that follows internal GIFCT playbooks.[5] It is also in this phase of activity that GIFCT staff make the determination as to whether or not the signal meets a definitional threshold for TVE activity and which incident level criteria of the IRF it best matches.

From a process standpoint, the activation of either the CI or CIP level of the IRF requires a vote from the Operating Board to formalize the designation as determined by GIFCT staff. Per GIFCT staff, the vote to formally invoke the CI or CIP response types within the IRF requires a majority of the Operating Board to vote in favor of an activation.

[4] Description of what constitutes the IRF is sparse on GIFCT's website, predominantly framing the IRF as the "collective capability to prevent terrorists and violent extremists from exploiting digital platforms." See https://gifct.org/incident-response/.

[5] These documents are not public.

In the third phase of activity, once the IRF has been activated at the CI or CIP level, digital material created by the perpetrator is added to the hash-sharing database, enabling companies to detect the material on their platforms.[6] Adding content to the database does not create an obligation for member companies to remove the material, although most do.

Upon activation of the CIP,[7] additional methods for handling violent events as they are happening are invoked, including sharing data points of signals associated with the incident, verifying perpetrator-produced content, and monitoring the evolution of and spread of live or near-live online content. The CIP also triggers enhanced communication about a given incident and its activation triggers alerts to GIFCT's broader stakeholder community beyond the Operating Board and member companies to include relevant civil society and government authorities that the activation has occurred.

Per GIFCT internal documents, the incident continues until the volume of posts related to the perpetrator's content subsides or the event is determined to be over. The internal nomenclature for the concluding of an IRF is a "deactivation," though the process may practically continue with the persistent application of hash sharing and adding to the hash-sharing database by GIFCT or member companies. After an incident concludes, "a formal debriefing process is launched to review GIFCT and member company responses and to identify any areas for improvement."[8] This typically involves stakeholders beyond the charter board members, such as the head of the Independent Advisory Committee (IAC) that guides GIFCT's Operating Board.[9]

Efficacy of the IRF: Process, Implementation, and Measurement Issues

To assess the efficacy of the IRF as it stands today, GIFCT provided CTEC with the following prompts:

- Does the current IRF accomplish the goal of providing member companies and GIFCT with an effective tool for activating incident response protocols and engaging in information sharing?
- How does GIFCT assess and evaluate on an ongoing basis whether the IRF and the workflows for evaluating incidents are applicable and appropriate to the evolving landscape of TVEC online and human rights concerns?

[6] The hash-sharing database catalogs extremist content as a numerical code (hash) that has been labeled to allow companies to identify a piece of proscribed material without including the material itself in the database.

[7] In addition to the three criteria noted above, a fourth needs to be met to trigger a CIP: material is being distributed on GIFCT member platforms or so broadly online that such distribution appears inevitable. See https://gifct.org/content-incident-protocol/.

[8] https://gifct.org/content-incident-protocol/.

[9] https://gifct.org/governance/.

- How should GIFCT assess and evaluate on an ongoing basis whether the IRF and the workflows for evaluating incidents are applicable and appropriate to the evolving landscape of TVEC online and human rights concerns?

CTEC's review of internal GIFCT documents indicates that the IRF's CIP includes the most tools and mechanisms within GIFCT members' ecosystem to meaningfully counter or mitigate the effects of an ongoing terrorist, violent extremist, or mass violence event. Overall, CTEC assessed that the most articulated and observable advantage of the IRF is the robust sharing of information between GIFCT member companies that occurs when a CIP is declared. CTEC views the functionality of the hash-sharing database as a secondary benefit that does not present an obvious or immediate boost to the operational ability of GIFCT to identify, triage, and address the incident in question.

Stakeholder Assessment

To answer GIFCT's questions, CTEC held discussions with both GIFCT staff and representatives from multiple sectors, including technology and social media, government, and civil society. From these exploratory conversations, a general consensus emerged that GIFCT's IRF is a desirable tool that can be highly beneficial for information sharing around emerging and ongoing violent extremist or terrorist incidents.

Assessments of efficacy were met with varying levels of response, from satisfied to desiring significant changes to processes and scope. These responses generally congregated along two lines of preference, which aligned to the tech and government sectors respectively. The tech sector representatives by and large felt satisfied with the IRF process and only suggested or sought small-scale changes that could further provide their internal incident or crisis response processes more resources or utility. The government sector representatives consistently expressed the desire to more concertedly expand the scope and capabilities of the IRF, with firm suggestions to add bystander footage to be in scope. These representatives were also more resolute in their concerns around communication and notification of activations, seeking a more consistent approach moving forward.

A third sector in the form of academia and external stakeholders (such as non-governmental organizations focused on civil society engagement) was represented as well, which have a vested interest in improving counter terrorism responses while protecting human rights. This community expressed frustration over their comparative absence from the IRF and a desire to see even more transparency on the mechanisms of how the IRF is activated, which member companies are participating in IRFs, and other accountability measures.

Process and Implementation Issues

CTEC has identified several critical issues within the IRF authorization and implementation process.[10] First and foremost, within the IRF the second phase poses a crucial potential failure point or bottleneck. This phase takes the IRF process (which up to that point functions in an informal, free-flowing system) in the direction of a rigid and hierarchical authorization procedure. While there are numerous rational and practical reasons for this to be the case, CTEC has identified that this shift in the structural nature within the IRF as artificial and unnecessary in that it has no immediate or apparent benefit with respect to the purpose of the IRF. CTEC's concern with this shift lies in the unclear reason why a vote needs to be taken for either of these IRF response types if the criteria are met to the level that an incident is raised for a vote. There are two principal reasons why:

1. In the second phase of IRF activity, GIFCT has ostensibly verified that a given incident and the associated signal is, in fact, TVE in nature. Therefore, no additional verification needs to be sought to meet the criteria of the "content" aspect of the IRF's stated mission. In CTEC's estimation of the IRF purpose, GIFCT should be permitted to initiate any stage of the IRF without the additional hurdle of a procedural vote.
2. The IRF's content hashing is not mandatory once the IRF is activated for a given piece of content, incident, or other situation in which the hash-sharing database may be leveraged. Instead, it is a tool that GIFCT provides in conjunction with the member companies to more effectively address an emerging or ongoing incident that has the potential for spreading harm online. The authorization therefore does not lead to the necessary next step, making it contingent as well.

Overall, CTEC found that the language and details of this phase remain opaque, and the resulting actions taken appear to be dependent on the individual companies' internal processes, personnel, and capabilities. In part this is a logical outcome of an industry-designed mechanism that is sensitive to individual companies' business needs and internal processes, but it undercuts the importance attached to voting for particular IRF activation types. This largely derives from GIFCT's lack of authority to override or unify policy or enforcement of violative content for a given incident across member companies. There are numerous considerations as to why unified implementation would not occur, but principally because each member is its own autonomous entity with its own guiding policies. There is also the reality that member companies do not share a common infrastructural design (e.g., some members are social media companies while others are service-focused).

What stands out about the IRF's three levels is that the nature of the incident in question has already been determined to be terrorist, violent extremist, or a mass violence event. Each phase of the process

[10] Barriers such as privileged company information, user data privacy concerns, and other concerns limited CTEC's assessment of the IRF.

reviewed by CTEC demonstrates a baseline presumption that the identification, assessment, and validation of an incident has been completed. But there is no robust description of how those three steps are conducted, by whom, or on what timeline. This is a critical issue that has deep ramifications for potentially expanding the CIP to include items such as bystander footage or non-terroristic harms.

Measurement Issues

CTEC views the IRF's efficacy as tailored to the tech industry and its member companies. Due to this tailoring, the ability to measure its efficacy toward a larger mission is significantly impaired.

A key factor that weighed in CTEC's review of the efficacy of the IRF was the lack of documentation around the key performance indicators (KPIs) that it was intended to meet. Most benefits of the IRF (visible to CTEC's reviewers) were abstract goal statements rather than concrete, measurable metrics. Some exceptions were noted, such as the stated goal to complete the IRF activation process from detection to activation within one hour. However, there was no indication that a standardized process by which this goal is measured is currently being deployed or utilized.

KPI metrics provide unimpeachable accounts of efficacy and can inform longer-term change needs within an overall process, as these items provide a baseline or strong longitudinal data points over time. For example, in the case of the stated one-hour goal, a KPI assessing the conditions that may have allowed GIFCT to reach that goal sooner, on time, or beyond the desired time frame could inform where IRF efficacy is faltering and can be improved.

Notably, this type of measure does not need to be a point-forward metric. GIFCT can (and should) consider a retrospective evaluation of each IRF invocation, particularly given the relatively small number of activations throughout its history. For example, data could assist in understanding the efficacy of declaring a CIP by determinging what are the uptake levels of the hashed TVEC content that is identified and shared during a CIP.

In CTEC's assessment, deeper analytical evaluation of the data for incidents that fit the Incident criteria but do not rise to the levels of Content Incident or Content Incident Protocol could meaningfully inform how GIFCT could better build detection processes or tools for or with its member companies regarding the very early stages of an emerging incident. What remains unclear is the technical or tactical benefits of broadening this aperture, only to restrict it in the second level of "Content Incident." The second level introduces the perpetrator-shared content component of the IRF, indicating the escalating nature of the tactics employed in a given incident. However, this does not necessarily change the inherent harm that can emerge from either level, as the baseline presumption is that both are directly responding to an incident determined to be terrorist, violent extremist, or a mass violence event.

Emerging Challenges for the IRF: Current and Future Considerations

The Incident Response Framework faces further challenges in addition to those discussed in the preceding section. By its nature, the IRF is a reactive protocol that involves spur-of-the-moment decisions. Anticipating potentially complicated scenarios is therefore a critical part of the IRF process. While CTEC does not have robust data about the IRF consultation process, CTEC expects that it includes extensive discussion of borderline cases.[11] The information CTEC has about the CIP activations is more detailed (albeit with a small dataset). It depicts GIFCT taking a conservative posture toward incidents, preferring not to take action without significant reason. Some of the activations (e.g., Louisville and Perry) involved attacks that were neither identified as extremist nor especially viral, but the unavoidable real-time decision-making process involved in the IRF makes it inevitable that the protocol will sometimes be invoked for lower-profile incidents. These activation patterns suggest that GIFCT could be more proactive in certain areas, some of which are discussed below.

Incident Types and Classification

The criteria for any IRF level's activation state that the protocol should be invoked for a "real-world terrorist, violent extremist, or mass violence event" involving posted video and "murder or attempted murder." However, the nature of the IRF as a tool for reacting to developing situations necessarily means that decisions will be made before the nature of the incident is clear. This will inevitably result in action being taken on incidents that are not terrorism, such as the Memphis, Louisville, and Perry CIP activations. None of these incidents were found to be clearly motivated by terrorism or violent extremism. The seven activations over five years show no pattern of overreach (see Appendix).

Given this context and the low incidence rate so far, combined with the lack of any other obvious centralized coordination mechanisms, questions arise about GIFCT's scope. Should it be empowered to act outside of its ordinary remit of terrorism and violent extremism, or in cases that fall within that remit but do not involve imminent or ongoing violence? Additionally, the current CIP does not address violent content created by people who are not perpetrators (i.e., bystanders), content that might falsely depict violence that would otherwise qualify under the CIP guidelines, the archiving of certain content with evidentiary value, and misinformation more generally, whether specific to extremism or not.

While CTEC lacks access to data about other IRF consultations, high-profile events like the January 6 assault on the U.S. Capitol and the October 7 Hamas attack likely prompted significant discussion. Historical events that occurred before the establishment of GIFCT, like the 2015 Paris assault by ISIS or Anders Breivik's 2011 terrorist attack, would have undoubtedly fallen under consideration.

[11] See https://gifct.org/wp-content/uploads/2023/06/GIFCT-23WG-Borderline-1.1.pdf.

Bystander- or Victim-Produced Imagery

The four criteria for a CIP activation explicitly state that the content in question must be produced by a perpetrator or an accomplice, and all of the IRF activations have involved perpetrator-created content (typically video and text). Contextual information around the explanation of the IRF and GIFCT staff have communicated that activations are predicated on a perpetrator-produced basis. Thus the question of whether bystander- or victim-produced imagery should be included in the IRF is confounded by the mission focus of the IRF to curtail the harms of the perpetrator's actions in a crisis moment.

A more complex scenario emerges when neutral third-party non-participants—true bystanders—record hate crimes or a terrorist attack that extremists subsequently weaponize. Such bystander video—even if deemed appropriate for a CIP—might be picked up by news media, creating still more complications. Additionally, some bystanders record and post video of a hate crime attack or an incident of non-extremist violence because they approve of the action, as demonstrated by many examples occurring in recent years.

Additional questions remain about whether bystander-created content (or even victim-created content) should be considered for addition to the hash-sharing database. Generally speaking, this would be a controversial step, although it's possible to imagine certain kinds of graphic content that would not present a problem.

Video of terrorist attacks and hate crimes are more clearly within GIFCT's remit. While perpetrator-created content meets the intended threshold for a CIP or other incident response, it's less clear whether bystander video of such an event should also be considered for interdiction. In part, this question may revolve around the nature of the bystander. For instance, if a perpetrator has enlisted a third party to document their crimes, and the person holding the camera can be heard clearly expressing their complicity, that would be a relatively simple determination. But if the third party does not explicitly endorse or participate in the attack, the correct course of action may be less clear. Additionally, it may not always be clear at the moment whether footage was created by a bystander or perpetrator, as happened during the October 7 Hamas attack.

Finally, video taken by non-participant bystanders and victims may provide much-needed evidence for public education. In light of this, the threshold for intervention with regard to non-perpetrator content should be high. Notably, there is no formal process to rectify mistakes (i.e., to reinstate content that was improperly hashed), although GIFCT informed CTEC that a hashing decision was reversed on at least one occasion. Given the fluid nature of the CIP and the concerns that accompany bystander footage, a formal process for appeal or reconsideration warrants consideration (see Recommendations below).

Misinformation and Generative AI Content

Given the preceding section, it is worth considering the misinformation landscape more broadly. Misinformation presents two challenges for GIFCT—one based on the use of misinformation by or on behalf of violent extremists, and one based on the use of misinformation by other actors.

Violent extremist misinformation comes in many varieties, most of which are not connected with a "real-world terrorist, violent extremist, or mass violence event" that includes "live-streamed or recorded video by [a] perpetrator or accomplice."[12] However, it is possible to imagine misinformation that would fit this category. For instance, extremists or other parties might stage a manipulated or manufactured video showing the purported execution of hostages. In 2004, three Americans created a hoax hostage beheading video that went viral and was picked up by global media.[13]

Though AI-generated content has not yet caused major disruptions, early warning signs have emerged through propaganda chatbots and misleading imagery. While current generative AI quality remains insufficient for viral content, GIFCT must prepare for inevitable improvements.[14] Current IRWG Participants noted that perpetrator-produced content likely will become harder to add and/or track as generative AI capabilities are leveraged adversarially. Perpetrators will gain the ability to more rapidly create variations of digital material that undermine the hash sharing technology and content moderation tools and systems.

Historical examples demonstrate that even low-quality hoaxes can achieve virality.[15] In 2005, a jihadist bulletin board posted an image purporting to be a U.S. soldier abducted in Iraq. The image was widely circulated online and picked up by global news media before being identified as a photo of a realistic-looking action figure.[16] While misinformation is much better understood today, an AI hoax need not be perfect (or even especially good) to go viral. Given the increasingly lax standards around misinformation on most major social media platforms and the growing level of political violence in the United States paired with a robust misinformation distribution infrastructure, a CIP-relevant AI event is likely to happen sooner than later.

The most dangerous scenario is the generation of AI content around a real-world event, such as a

[12] Participants in the current IRWG noted that this form of misinformation could cause significant psychological trauma to the audience (including anxiety, insomnia, ruminations, PTSD, etc.).

[13] Julian Guthrie, "Web Hoax Fools News Services / S.F. Man Fakes Own Beheading," SFGATE, August 8, 2004, http://www.sfgate.com/news/article/Web-hoax-fools-news-services-S-F-man-fakes-2702773.php.

[14] Participants in the current IRWG noted that generative AI will likely increase the difficulty of content moderation due to the ability for adversaries to produce rapid variations to real-world images.

[15] David Klepper and Ali Swenson, "Minutes after Trump shooting, misinformation started flying. Here are the facts," AP News, July 15, 2024, https://apnews.com/article/trump-assassination-biden-tiktok-misinformation-fact-check-4b7ab8e21c00aa6ef47f25ec76984fe6.

[16] Jim Krane, "Iraq rebel Web sites useful to U.S., as well," February 4, 2005, https://www.nbcnews.com/id/wbna6914713.

terrorist attack. So far, no truly notable efforts have emerged, although some have been made to promote Gaza war misinformation.[17] In the immediate aftermath of the July 13 assassination attempt on former U.S. President Trump, generative AI images of Trump smiling or his security detail smiling gained modest traction.[18] Both of the above scenarios reinforce the fact that, for now, low-tech and low-quality misinformation—so-called "cheapfakes"—continue to outperform AI-driven efforts.[19]

Regional Biases

The CIP activations indicate a geographic orientation toward the United States, where all of GIFCT's founding members and many of its constituent members are based, with only two of the eight CIP activations taking place outside the continental United States. This geographic bias exists despite numerous examples of live-streamed and near-live violence posted by perpetrators that have taken place around the world without IRF activation:

Mass Violence Attacks

- In February 2020, Jakrapanth Thomma killed 29 people and injured 58 in a multi-location mass shooting in and around Nakhon Ratchasima, Thailand, which included multiple posts online.[20]
- In December 2023, a local Ukrainian politician, Serhiy Batryn, threw a grenade into a local council meeting that was being live-streamed.[21]

Extremism-Linked Attacks

- In August 2021, a 15-year-old Swedish student with neo-Nazi views live-streamed a knife attack on a teacher. Three months later, a friend of his carried out another school attack without a video component.[22]

[17] David Klepper, "Fake babies, real horror: Deepfakes from the Gaza war increase fears about AI's power to mislead, AP News, November 28, 2023, https://apnews.com/article/artificial-intelligence-hamas-israel-misinformation-ai-gaza-a1bb303b637ffbbb9cbc3aa1e000db47.

[18] Klepper and Swenson, "Minutes after Trump shooting,"; Erin Flanagan and Anuj Chopra, "Misinformation Around Trump Shooting Lays Bare U.S. Political Fissures," Barron's, July 15, 2024, https://www.barrons.com/news/misinformation-around-trump-shooting-lays-bare-us-political-fissures-eb4ed5a0.

[19] Seana Davis, "We fact-checked some of the rumors spreading online about the Trump assassination attempt," Reuters, July 16, 2024, https://www.reuters.com/fact-check/misinformation-narratives-fact-checked-wake-trump-assassination-attempt-2024-07-15/.

[20] "Thailand shooting: Soldier who killed 26 in Korat shot dead," BBC News, February 9, 2020, https://www.bbc.com/news/world-asia-51431690; "One gunman, four locations, 29 dead: how the mass shooting in Thailand unfolded," Reuters, February 9, 2020, https://www.reuters.com/article/us-thailand-shooting-timeline-idUSKBN2030FQ/; "Mass shooter killed at Korat mall, 27 dead," Bangkok Post, February 9, 2020, https://www.bangkokpost.com/thailand/general/1853804/commandos-kill-soldier-after-korat-shooting-rampage-20-dead.

[21] "France attack: Three killed in 'Islamist terrorist' stabbings," BBC News, October 29, 2020, https://www.bbc.com/news/world-europe-54729957.

[22] Henrik Högström, "Skolattackerna i Skåne: Ville sprida sina åsikter-knivskar lärare," Allas, October 28, 2021, https://www.allas.se/samhalle/ville-sprida-sina-asikter-knivskar-larare/7980590; "Hur Många Gånger Måste Lärare Slå Larm Innan Något Görs?" Skolvärlden, October 29, 2020, https://skolvarlden.se/artiklar/hur-manga-ganger-maste-larare-sla-larm-innan-nagot-gors.

- In October 2023, Rajeh Husam Taha Abu Sneina, a Palestinian, live-streamed himself driving a bulldozer into a military roadblock.[23]

Other Live-Streamed Violence

- In February 2024, Indian politician Mauris Noronha murdered his chief rival during a live-streamed debate.[24]
- In China in May 2023, two children were attacked by a man with a knife during a live-stream. One of the children's hands was amputated.[25]
- In August 2023, a graphic video showed the torture, execution, and corpses of five men believed to have been killed by a Sinaloa drug cartel faction in Mexico.
- Also in August 2023, a Bosnian man murdered his wife during a live-stream.

These examples reflect wider patterns of live and near-live posting of violent content that suggests GIFCT may need a wider aperture for this problem.

Recommendations

Revise the IRF

CTEC recommends GIFCT clarify the IRF tier system to allow for more consistent application of IRF criteria. By better distinguishing the tiers, GIFCT may also address the requests from external stakeholders to broaden the inclusion criteria for incidents and content that may glorify violent extremist or terrorist content and activity, as well as those incidents that contribute to secondary and tertiary impacts on society (such as radicalization).

Expand and Rename the IRF Activations

The current naming convention for the IRF could benefit from additional explanatory language per tier. Little inherent distinction exists between the three levels. As such, CTEC recommends GIFCT rename and expand the three stages of IRF to better communicate their latent and inherent purposes.

[23] "Israeli Soldiers Kill a Palestinian in Hebron," IMEMC News, October 9, 2023, https://imemc.org/article/israeli-soldiers-kill-a-palestinian-in-hebron-6/.

[24] Manish K. Pathak and Megha Sood, "Mauris Noronha stole guard's gun to kill Abhishek Ghosalkar and self," Hindustan Times, February 10, 2024, https://www.hindustantimes.com/cities/mumbai-news/noronha-stole-guard-s-gun-to-kill-ghosalkar-and-slef-101707505322568.html.

[25] "Student's Hand Gets Hacked Off by Attacker in China During a Live Stream," The Straits Times, November 21, 2024, https://www.straitstimes.com/asia/east-asia/student-s-hand-gets-hacked-off-by-attacker-in-china-during-a-live-stream.

For example, one such revision might look like this:

Current Nomenclature	Recommended Nomenclature
Incident	Emerging Incident
Content Incident	TVE Incident
Content Incident	Crisis Response Protocol
Content Incident Protocol	Viral Content Protocol
Content Incident Protocol	Live-stream Protocol

The nomenclature provided is not a final recommendation but rather a starting point from which GIFCT can begin to move away from the current nomenclature that lacks intrinsically firm distinctions. Below CTEC has articulated the criteria for each of these hypothetical stages of a revised IRF:

- Emerging Incident: Declared by GIFCT when preliminary information about an event may fit the criteria for an escalated response (such as an ongoing mass casualty attack or attempted attack) is suspected to be associated with violent extremist or terrorist content. The primary feature of declaring an Emerging Incident is to initiate enhanced communication and situational awareness among relevant stakeholders.[26]
- TVE Incident: Declared when it is determined that an Emerging Incident is positively attributed to (or claimed by) a violent extremist or terrorist entity.[27]
- Crisis Response Protocol: Declared when a mass casualty or extremist campaign occurs. This category can include perpetrator-created content depicting violence from the event that has been posted to a member platform or is otherwise broadly available online. The goal would be to limit content that stems from a protracted civil disruption (e.g., 2024 Southport riots in the UK) and which intrinsically holds significant potential to produce or recycle content that has a radicalizing impact.
- Viral Content Protocol: This category would include perpetrator-created content potentially depicting violence or ideological components that promote a violent extremist or terroristic goal. The goal would be to limit content from spreading through an acute surge online

[26] Additionally, CTEC recommends that this enhanced footing initiated by GIFCT should be paired with a dedicated hashing effort that is held separate from the main hash-sharing database until the Emerging Incident is escalated to a positive attribution of violent extremist or terrorist activity, networks, or designated organizations. This approach will provide a layer of protection from errant inclusion of hashed content that is not proven or known to be violent extremist or terrorist content, such as activist or news media coverage, a main human rights concern.

[27] Hashing of content should move toward formal inclusion into the hash-sharing database in this phase as the content associated with the given incident has been positively identified as or attributed to violent extremist or terrorist activity, networks, or designated organizations.

within a rapid timeframe (for example, preventing a video's spread if a violent extremist or terrorist actor perpetrates and records an act of violence and then uploads it to the internet for dissemination). This could also apply to associated written content, audio, etc.

- Live-stream Protocol: Declared when the content includes a live-stream or other real-time posting by a perpetrator or an accomplice that is on a member platform or otherwise broadly available online. The goal would be to shut down the stream and prevent the spread of its viewership.

All three protocols activate aggressive measures by GIFCT and member companies to monitor, hash, and otherwise interdict the sharing of relevant content.

Incident (I):

An **emerging** terrorist, violent extremist or mass ~~violence~~ **casualty** event, threat, or attempt; AND

Content related to the terrorist event but unclear whether depicting **killing**, attempted **killing**, violence, or bystander footage from a **suspected terrorist, violent extremist or mass casualty event** OR

Gaining ~~international~~ media attention AND appearing to have a significant online element *on any technology or social media platorm.*

Revise the Current IRF Tiers

CTEC recommends GIFCT restructure the current IRF (and future IRF iterations) to include a greater focus for activation on the "incident" over the content. These suggestions are provided on the assumption that the current tier labels do not change; the language below should be adapted if tier labels are changed. Recommended language has been included below in bold text, with new text included in italics.

CTEC suggests framing the Incident level of the IRF as "emerging" to better position GIFCT's response capability in an ambiguous and nascent stage of triaging and assessing on-the-ground and real-time activity. Categorizing this stage as emerging also allows for low-level information sharing to begin with the implicitly communicated goal of determining either escalating to a more concrete footing (e.g., CIP) or excluding the event from the IRF's more stringent responses.

CTEC recommends that the term "mass violence" be renamed "mass casualty" across all three tiers. In the current approach, the inherent meaning of this term lacks explanatory power as to the purpose of the category. The current mass violence definition per GIFCT is "An event in which four or more victims are murdered, or had serious bodily harm inflicted in attempted murders, in one or more locations in close geographical and temporal proximity." Most definitions of mass casualty track along the above definition, and by converting from an abstract label of 'violence' to casualty, the data point gains broader appeal and potential for uptake as it aligns closer with other fields of academic research and civil society reporting on offline harms.

CTEC recommends that the term "killing" be utilized to further lean into the behavioral element of the IRF. Murder is an inherently legalistic notion, while killing is an inherently behavioral notion. If the purpose of the IRF is to capture behavior, then prioritizing universal concepts over legal systems concepts would be the logical posture to adopt within the criteria. The use of killing over murder would also place that specific component of the criteria in line with the next component, as violence is also a universalized behavior and not a legalistic notion.

CTEC recommends removing "international" from the discussion of media as it is a higher bar that likely will exist at this level of detection. Most emerging incidents will upcycle from local and regional media reporting to international reporting and attention over time. Notably, this is a threshold of attention that may delay the stated one-hour activation of the IRF by GIFCT.

CTEC recommends that at the Incident level of the IRF, monitoring should extend beyond GIFCT member companies' platforms to track potentially dangerous content across the broader internet. This wider scope would help identify emerging threats before they become internet-wide phenomena. While this could present scale issues, there are numerous ways in which GIFCT could leverage off-the-shelf tooling, stakeholders, and member companies' capabilities to augment that scaled aperture.

CTEC recommends that hashing be conducted at each level. At the Incident level, CTEC recommends that GIFCT begin hash collecting and storing them in a sequestered temporary access space that can be available to member companies. Once the emergent nature of the Incident has been formally assessed as a TVE event, GIFCT should migrate the temporarily stored hashed content into the full hashed content database. This process will allow GIFCT to prevent inclusion into the permanent database of emergent or unconfirmed content and keep GIFCT's processes in line with respecting human rights.

Content Incident (CI):

A **confirmed** terrorist, violent extremist or mass ~~violence~~ **casualty** event; AND
Other content (ex. photo, audio, or text) **produced** by perpetrator or accomplice; AND
Depicting **killing**, attempted **killing**, or violence from the **perpetrator**; AND
On a member platfom (or so broadly available online that it will inevitably be shared on member platforms).

CTEC recommends using "confirmed" over "ongoing" as it clearly demarcates the distinction between an Incident and higher levels of the IRF.

CTEC recommends using "produced by perpetrator or accomplice" to further and explicitly state the need for content to be perpetrator-produced in the current IRF framework.

With specific respect to activation criteria, CTEC recommends that GIFCT no longer votes to activate elements of the IRF. There is very little benefit in a crisis response protocol to rapidly shift organizational dynamics into a hierarchical or parliamentary process and then back to a dispersed, fluid dynamic.

Content Incident Protocol (CIP):

An ongoing terrorist, violent extremist or mass ~~violence~~ **casualty** event; AND
Live-streamed or recorded video by perpetrator or accomplice; AND
Depicting **killing** or attempted **killing**; AND
On a member platform; AND
So broadly available online that it will inevitably be shared on member platforms.

This wording aligns the elements of the CIP with the other two elements of the IRF.

Expand IRF Involvement

In line with previous recommendation reports, CTEC recommends that GIFCT expand the parties involved in IRF activations as well as IRF reviews. The current IRF structure is highly insular and heavily favors tech and social media companies' involvement while limiting the involvement of non-tech or social media stakeholders to a second-class or after-the-fact awareness. Tech and social media companies are uniquely situated to conduct information and data sharing toward the hashed content, which undergirds much of the IRF's obvious benefits. But as GIFCT staff have extolled throughout the review process conducted by CTEC, information sharing is a crucial component and benefit of the IRF—information that necessarily extends beyond hashed content.[28]

Specifically, CTEC recommends that GIFCT deepen its established partnerships within academia and think tanks to ensure it is capturing the full range of detection, collection, and attribution for a given incident being evaluated. Based on current IRWG participants' input, CTEC recommends that GIFCT consider an annual grant program or grant fellowships which could increase its direct partnerships with academic entities and networks to encourage and facilitate academic evaluation of the IRF. Topics could include how the tech and social media sectors engage with or utilize this mechanism in their pursuit of their counterterrorism goals.

CTEC also recommends that GIFCT engage external parties to conduct independent reviews of each IRF case, as well as an annual review of all cases for that year. CTEC recommends a cross-section of IAC members, researchers, civil society, and government representatives to conduct an independent review of the IRF. GIFCT should also work with the IAC to craft a balanced checklist of factors to have the independent review consider, ranging from human rights considerations and individual company roles and contributions to effectiveness metrics. To support these reviews, GIFCT should enhance its current IRF debriefs into a robust after-action system that tracks specific metrics, including content hashing uptake rates, information sharing patterns between parties, and data on incidents that meet Emerging Incident criteria but do not rise to full IRF levels. These metrics would provide crucial insights into the IRF process's effectiveness and its broader societal impact.

Improve IRF Communications

CTEC recommends that GIFCT adopt a more expansive explanatory posture on the specifics related to the IRF's activation criteria, thresholds, and response actions. While GIFCT has provided extensive justifications across its annual reports, blog posts, and other documentation, this information remains scattered rather than consolidated in one accessible location. Additionally, much of this content merely recycles established talking points and framework criteria without providing deeper insights into how GIFCT and its member companies actually implement the IRF during specific incidents. This lack of additional information creates accountability and transparency gaps for the IAC, external stakeholders, and civil society.

The information on how the IRF works should be more readily and easily available. CTEC recommends that GIFCT centralize and streamline the communications around the IRF to better facilitate a common understanding of its functionality, scope, and utility to the broader public. CTEC suggests that GIFCT seek

[28] For example, hashed content is largely consigned to single snapshots of TVE content and documents (e.g., manifestos) and can be severely limited in its utility (as each hash is generated for a single iteration of the TVE content). These types of content items are ephemeral and can be easily manipulated to circumvent moderation techniques.

to consolidate the location of information from the annual reports, blog posts, and other sources into a single location on the GIFCT website. Additionally, the creation and use of a dashboard functionality could aid in the visualization and export of the data from the activations, ancillary data associated with the incident that prompted the IRF activation, and after-action reviews. In this respect, GIFCT would be serving as a resource and issuing consistent data points that third parties could leverage for academic or media purposes.[29]

CTEC recommends that GIFCT continue its movement away from designation list-based approaches to the IRF activation criteria. In line with previous reports provided to GIFCT, CTEC recommends that GIFCT adopt a formal and publicly detailed articulation of what constitutes violent extremist and terrorist content. As the BSR report stated, GIFCT should endeavor to "develop a common understanding of terrorist and violent extremist content." Feedback from GIFCT has asserted that this would complicate their ability to meaningfully integrate tech companies, social media companies, and GIFCT membership broadly into GIFCT processes. However, both within individual nations and in international bodies, contrasting definitions of terrorism have not prevented meaningful cross-agency and international collaboration toward combating the threat. With this in mind, CTEC recommends that GIFCT continue to embrace the IRF and ascribe clear, unambiguous definitional conditions for the content it seeks to address through its crisis response framework.

CTEC recommends that GIFCT work to increase clarity around how and when the parties alerted to the initiation of an IRF are notified. Multiple stakeholders expressed confusion and inconsistency with when or whether they were notified of an IRF initiation. Regardless of which level of the IRF has been declared, CTEC recommends a baseline level of notice should go out alerting stakeholders.

Specifically, CTEC recommends a baseline alerting for the activation of the IRF from a set list-serve notification system such as MyEmma, Mailchimp, Constant Contact, or an equivalent service. Services that provide email-to-phone SMS alerting should also be considered for GIFCT IRF activation. While these may seem intrusive, they ensure that the appropriate parties are getting near instantaneous alerting for any actions taken by GIFCT in a window that is incredibly short for crisis response protocols. CTEC also notes that the Slack channel, while useful for tech, is an isolated space and could be tiered to provide access to other stakeholders within GIFCT's partnership network. This approach would allow GIFCT to centralize communications across multiple sectors while also respecting the privileged and sensitive nature of the various sectors engaged in the IRF process with GIFCT.

[29] Any public dissemination of these data points should carefully consider individual member company sensitivities but should pursue, at minimum, a neutral aggregate to keep public and external stakeholders apprised of the IRF's effectiveness.

Address Emerging Challenges

CTEC recommends that GIFCT commission a report to evaluate the empirical evidence of social harms caused by digitized terrorist content. While GIFCT stakeholders and partners—including the Accelerationism Research Consortium (ARC) and various governmental bodies—have identified the weaponization of live-streaming features by terrorist groups like the Terrorgram Collective as a grave concern, there is currently limited empirical research examining these evolving threats. Such a report would help bridge this knowledge gap and provide evidence-based guidance for addressing these harms.

CTEC recognizes that correcting a geographic bias in IRF activations is not a simple task and is severely impaired by the limited resources currently available. GIFCT is not currently structured to allow for a comprehensive global focus for the IRF, nor well positioned to incorporate a global footing in a feasible and scalable manner without sacrificing effectiveness and resources in areas that could be more meaningfully improved in the short term.

CTEC recommends that GIFCT conduct a robust assessment of what steps it would need to take to reach a global focus within a 1–3-year timeline. Specifically, GIFCT should:

1. Conduct a systematic review of all terrorist incidents that have occurred in the past 24 months and categorize them by regional occurrence. With this baseline data, establish a regional prioritization of incident triage capability toward the IRF, prioritizing the highest frequency region and adding the next highest in 6-month increments. This approach would steadily and intentionally build GIFCT's abilities to meaningfully and consistently maintain a global approach to the IRF.
2. Pair this strategic roll-out with expanded membership to include companies from prioritized regions, along with financial backing to staff a greater capacity for global awareness. While member companies may be better situated to alert GIFCT staff to emerging incidents, this capacity is not currently in place.
3. Secure additional fiscal investment into staffing or tooling, along with fiscal buy-in from member companies, tied to a corresponding push to include new technology companies in prioritized regions.

CTEC does not recommend GIFCT expand the IRF to intentionally address AI-generated content at this time. Lacking more robust data sets and extant literature reviews of the emergent threat from AI-generated violent extremist and terrorist content, it remains difficult to prescribe any specific actions as guidance for GIFCT. CTEC has observed a consensus among external stakeholders and member companies that the emergent AI-generated violent extremist and terrorist content phenomenon has yet to yield a sophistication found in other threat actors' arenas. Simultaneously, it is undoubtedly the case that a CIP-relevant AI event is likely to happen sooner rather than later. As such, CTEC recommends that GIFCT and its member companies leverage GIFCT's academic partnerships to engage in robust study and evaluation of the threat to help establish a unique protocol tailored to the threat.

CTEC does not recommend adding bystander footage as a component of the IRF. CTEC does not believe that GIFCT is currently positioned to incorporate bystander footage in a feasible and scalable manner without sacrificing effectiveness and resources in areas that are more meaningfully able to be improved in the short run.

While the majority of stakeholders supported including bystander footage, practical concerns exist around determining if content is perpetrator/accomplice-produced within GIFCT's short response window. Individual member companies' capabilities vary, making systematic enforcement difficult. This also creates significant challenges around error tolerance, as incorrectly hashing bystander content could restrict legitimate human rights and activist-based recording. While CTEC acknowledges concerns about the weaponization of bystander footage, the IRF's role as a de facto content moderation mechanism means errors could have outsized impacts. After consulting with stakeholders, CTEC assesses that these risks currently outweigh potential benefits.

Moving forward, CTEC recommends that GIFCT conduct a targeted, robust study of the impact of bystander footage, specifically seeking clarification as to what extent bystander footage comprises the principal vector of spread for violent extremist and terrorist content online.

Regarding state-level incidents, CTEC does not recommend that GIFCT pursue the inclusion of state-level actors/actions in the current IRF, as GIFCT's infrastructure is not situated to accommodate such incidents. Instead, CTEC recommends that GIFCT conduct a thorough assessment of the feasibility for including state-level actors/actions, potentially creating a separate playbook and IRF dedicated to these incidents, as the varying conditions would likely not align well with non-state actor violent extremism and terrorism.

Bibliography

Activist Post. “Biden Gives ‘Five Eyes’ What It Always Wanted: Access to Everyone’s Social Media.” *Activist Post*, July 27, 2021.

Amarasingam, Amarnath, Marc-André Argentino, and Graham Macklin. “The Buffalo Attack: The Cumulative Momentum of Far-Right Terror.” *CTC Sentinel 15*, no. 7 (2022): 1–10.

Anti-Defamation League. “Footage of Buffalo Attack Spread Quickly Across Platforms, Has Been Online for Days.” *ADL Blog*, May 20, 2022.

Armstrong-Scott, Georgia, and James Waldo. “Guns, Incels, and Algorithms: Where We Are on Managing Terrorist and Violent Extremist Content Online.” *Science, Technology, and Public Policy Program Papers*, 2023.

Badii, Fanny. “Human Rights Lifecycle of a Terrorist Incident Online.” *GIFCT Crisis Response Working Group*, 2022.

Bell, James G., and Barbara Perry. “Outside Looking In: The Community Impacts of Anti-Lesbian, Gay, and Bisexual Hate Crime.” *Journal of Homosexuality* 62 (2015): 98–120.

Bilic, Vesna. “Violence among Peers in the Real and Virtual World.” *Pedijatrija Danas: Pediatrics Today* 9, no. 1 (2013): 78–90.

Boyd, Robert W., and W. S. Swanson. “The Evolution of Virtual Violence: How Mobile Screens Provide Windows to Real Violence.” *Pediatrics* 138, no. 2 (2016). doi:10.1542/peds.2016-1358.

Campbell-Verduyn, Malcolm, and Moritz Hütten. “Locating Infrastructural Agency: Computer Protocols at the Finance/Security Nexus.” *Security Dialogue* 54, no. 5 (2023): 455–474.

Chang, Wendy. “Buffalo Shooting Demonstrates the Limitations of Existing Content Moderation Protocols.” *Tech Policy*, May 17, 2022.

Chernik, Itamar. “Social Media Expert: It’s a Mistake to Blame Technology for Our Problems.” *The Jerusalem Post*, October 10, 2019.

Christensen, Larissa S., and Jodie Woods. “‘It’s Like POOF and It’s Gone’: The Live-Streaming of Child Sexual Abuse.” *Sexuality & Culture* 28, no. 4 (2024): 1–15.

Clifford, Bennett. “Moderating Extremism: The State of Online Terrorist Content Removal Policy in the United States.” Program on Extremism. Washington, DC: George Washington University, December 2021.

Cohen, Julie E. “Infrastructuring the Digital Public Sphere.” *Yale Information Society Project and Yale Journal of Law and Technology* 25 (2023).

Conway, Maura. “Online Extremism and Terrorism Research Ethics: Researcher Safety, Informed Consent, and the Need for Tailored Guidelines.” *Terrorism and Political Violence* 33, no. 2 (2021): 367–380.

“Data on Online Spread of Livestreamed Germany Shooting Kept Secret.” *Iran Daily*, October 19, 2019.

Davis, Aaron L. “Artificial Intelligence and the Fight Against International Terrorism.” *American Intelligence Journal* 38, no. 2 (2021): 63–73.

DeBenedetto, L. “Qualitative Indicators of Transparency During an Incident Response.” *GIFCT Incident Response Working Group*, September 20, 2023.

Donato, S., Eslen-Ziya, H., and Mangone, E. “From Offline to Online Violence: New Challenges for the Contemporary Society.” *International Review of Sociology* 32, no. 3 (2022): 400–412.

Drejer, C., Riegler, M. A., Halvorsen, P., Johnson, M. S., and Baugerud, G. A. “Livestreaming Technology and Online Child Sexual Exploitation and Abuse: A Scoping Review.” *Trauma*, *Violence*, *& Abuse* 25, no. 1 (2024): 260–274.

Fishman, B. “Online Platforms’ Responses to Terrorism.” *Lawfare*, November 14, 2023.

Ganesh, B. “How to Counter White Supremacist Extremists Online.” *Foreign Policy*, January 28, 2021.

Graña, J. L., Cruzado, J. A., Andreu, J. M., Muñoz-Rivas, M. J., Peña, M. E., and Brain, P. F. “Effects of Viewing Videos of Bullfights on Spanish Children.” *Aggressive Behavior* 30, no. 1 (2004): 16–28.

Hadero, H., and Swenson, A. “A Beheading Video Was on YouTube for Hours, Raising Questions About Why It Wasn’t Taken Down Sooner.” *Associated Press*, February 2, 2024.

Hinduja, Sameer, and Justin W. Patchin. “Offline Consequences of Online Victimization: School Violence and Delinquency.” *Journal of School Violence* 6, no. 3 (2007): 89–112.

Hinduja, Sameer, and Justin W. Patchin. “Connecting Adolescent Suicide to the Severity of Bullying and Cyberbullying.” *Journal of School Violence* 18, no. 3 (2019): 333–346.

Honzel, Hannah T. "Confronting White Supremacist Activity Online: Examining Internet Speech in the Post-Christchurch Era." 2020.

Iganski, Paul, and Spiridoula Lagou. "Hate Crimes Hurt Some More Than Others: Implications for the Just Sentencing of Offenders." *Journal of Interpersonal Violence* 30, no. 10 (2015): 1696–1718.

Ingelevic-Citak, Maja, and Zuzanna Przyszlak. "Jihadist, Far-Right and Far-Left Terrorism in Cyberspace—Same Threat and Same Countermeasures?" *International Comparative Jurisprudence* 6, no. 2 (2020): 154–177.

Irwin-Rogers, Keir, James Densley, and Christopher Pinkney. "Gang Violence and Social Media." In *The Routledge International Handbook of Human Aggression*, edited by Jane L. Ireland, Philip Birch, and Carol A. Ireland. Routledge, 2018.

James, Emily. "Teenage Girl, 17, and Medical Assistant Are Two of Four Victims Killed in Memphis by Ex-Convict, 19, in 22-Hour Shooting He Livestreamed on Facebook." *The Daily Mail Online*, September 8, 2022.

Llanos, Jorge. "Transparency Reporting on Terrorist and Violent Extremist Content Online: An Update on the Global Top 50 Content Sharing Services." *OECD Digital Economy Papers*, no. 313, 2021.

Marsden, Chris, and Trisha Meyer. "A Coregulation Model to Advance the Standards: Introduction." In *Red Lines and Baselines: Towards a European Multistakeholder Approach to Counter Disinformation*, edited by Louk Faesen, Alexander Klimburg, Simon van Hoeve, and Tim Sweijs. The Hague Centre for Strategic Studies, 2021.

Metwally, Ahmed. "The Governors' Advisors: Experts and Expertise as Platform Governance." *Yale Journal of Law and Technology* 24 (2022): 510.

"Microsoft, Other Tech Industry Leaders Team Up with an International Coalition of Governments for a Multi-Stakeholder Solution." *Microsoft on the Issues*, September 23, 2019.

"Microsoft Submission to the Parliamentary Joint Committee on Intelligence and Security Inquiry on the Telecommunications Legislation Amendment (International Production Orders) Bill 2020." Submission 29 to Parliamentary Joint Committee on Intelligence and Security, 2021.

Naher, Jannatul, and Md. Rafiqul Minar. "Impact of Social Media Posts in Real Life Violence: A Case Study in Bangladesh." *arXiv preprintl*, arXiv:1812.08660 (2018).

Noelle, Meaghan. "The Ripple Effect of the Matthew Shepard Murder: Impact on the Assumptive Worlds of Members of the Targeted Group." *American Behavioral Scientist* 46, no. 1 (2002): 27–50.

O'Callaghan, Jody. "Videos Retraumatise March 15 Survivors." *The Press* (Christchurch, New Zealand), May 17, 2022.

Pandey, Pratnashree. "One Year Since the Christchurch Call to Action: A Review." *ORF Issue Brief* no. 389 (2020).

Paterson, Jennifer L., Rupert Brown, and Mark A. Walters. "The Short and Longer Term Impacts of Hate Crimes Experienced Directly, Indirectly, and Through the Media." *Personality and Social Psychology Bulletin* 45, no. 7 (2019): 994–1010.

Patton, Desmond U., Jun Sung Hong, Melissa Ranney, Sheehan Patel, Caitlin Kelley, Robin Eschmann, and Tanjala Washington. "Social Media as a Vector for Youth Violence: A Review of the Literature." *Computers in Human Behavior* 35 (2014): 548–553.

Pearson, Elizabeth, Joe Whittaker, Till Baaken, Sara Zeiger, Farangiz Atamuradova, and Maura Conway. "Online Extremism and Terrorism Researchers' Security, Safety, and Resilience: Findings from the Field." *Vox Pol*. 2023.

Perry, Barbara, and Shahid Alvi. "'We Are All Vulnerable': The 'In Terrorem' Effects of Hate Crimes." *International Review of Victimology* 18, no. 1 (2012): 57–71.

Rasmussen, Nicholas J. "The Dynamic Terrorism Landscape and What It Means for America." Testimony before the House Committee on Homeland Security, 117th Congress, February 2, 2022. U.S. Government Publishing Office.

Saltman, Erin. "Countering Terrorism and Violent Extremism at Facebook: Technology, Expertise and Partnerships." In *Tackling Insurgent Ideologies in a Pandemic World*, edited by Maya Mirchandani. Observer Research Foundation, 2020.

Saltman, Erin, and Noor El Karhili. "LEVEL UP: Policies, Practices, and Positive Interventions to Counter Terrorism and Violent Extremism in Gaming Spaces." In *Gaming and Extremism: The Radicalization of Digital Playgrounds*, edited by Laura Schlegel and Rachel Kowert. Taylor & Francis, 2024.

Sandhu, Amardeep, and Daniel Trottier. "The Criminal Selfie: Conveying Grievance While Recording and Live Streaming Antisocial Behavior." *European Journal on Criminal Policy and Research* 29, no. 3 (2023): 423–436.

Sasso, Rebecca. "Tech Giants' Role in Countering the 'Media Jihad.'" International Team for the Study of

Security. Verona, 2022.

Scheuerman, Morgan Kay, Jieyu (Jesse) A. Jiang, Casey Fiesler, and Jed R. Brubaker. "A Framework of Severity for Harmful Content Online." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (2021): 1–33.

Schlag, Gabi. "Bilder, die verletzen—Die Regulierung von Gewalt in sozialen Medien zwischen Politisierung, Responsibilisierung und Verrechtlichung." In Sprache und Politik, edited by Manuel Fröhlich. Nomos Verlagsgesellschaft mbH & Co. KG, 2023.

Schwieter, Christian. "Online Crisis Protocols – Expanding the Regulatory Toolbox to Safeguard Democracy During Crises." Institute for Strategic Dialogue, 2022.

Speech in the Social Media: Summary of Scientific Research. Centre for Internet and Human Rights. Frankfurt, Germany, 2018.

Sticca, Fabio. "Bullying Goes Online: Definition, Risk Factors, Consequences, and Prevention of (Cyber) Bullying." (PhD diss., University of Zurich, 2013).

Stubbs, Jessica E., Laura L. Nicklin, Lucy Wilsdon, and Jennifer Lloyd. "Investigating the Experience of Viewing Extreme Real-World Violence Online: Naturalistic Evidence from an Online Discussion Forum." *New Media & Society* 26, no. 7 (2024): 3876–3894.

Taylor, Josh. "Australia Quietly Shuts Down Anti-Terror Taskforce Set Up After Christchurch Attack." *The Guardian*, May 16, 2022.

Thorley, Thomas G., and Erin Saltman. "GIFCT Tech Trials: Combining Behavioural Signals to Surface Terrorist and Violent Extremist Content Online." *Studies in Conflict & Terrorism* (2023): 1–26.

UK Home Office. "Annex 1: Statement on Preventing and Countering Violent Extremism and Terrorism Online." Policy paper, 2021.

Walters, Mark A., Jennifer L. Paterson, Rupert Brown, and Laura McDonnell. "Hate Crimes Against Trans People: Assessing Emotions, Behaviors, and Attitudes Towards Criminal Justice Agencies." *Journal of Interpersonal Violence* 35, no. 21-22 (2020): 4583–4613.

Watkin, Amy Louise, and Maura Conway. "Building Social Capital to Counter Polarization and Extremism? A Comparative Analysis of Tech Platforms' Official Blog Posts." First Monday 27, no. 10 (2022). https://doi.org/10.5210/fm.v27i10.12345.

Ybarra, Michele L., Diener-West, Markow, Donna, Leaf, Philip J., Hamburger, Merle, and Paul Boxer. "Linkages Between Internet and Other Media Violence with Seriously Violent Behavior by Youth." *Pediatrics* 122, no. 5 (2008): 929–937. https://doi.org/10.1542/peds.2007-3377.

Ybarra, Michele L., Kimberly J. Mitchell, and J. D. Korchmaros. "National Trends in Exposure to and Experiences of Violence on the Internet Among Children." *Pediatrics* 128, no. 6 (2011): e1376–e1386. https://doi.org/10.1542/peds.2011-0118.

Zeiger, Sara, Paola Regeni, Fatima Atamuradova, and Damir Suljic. "Defining and Classifying Terrorist Content Online: Leveraging National Countering Violent Extremism Strategies and Action Plans." *Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps* (July 2021): 12.

Appendix: IRF Activations Overview

This appendix enumerates each time the IRF has been activated by type.

CI Activations

Table 1: Overview of CI activations

Date	Location	Elapsed time from event start to CI activation (approx.)	CI duration	Type of media	Items added to hash-sharing database	Extremist ideology
6/28/2022	Udaipur, Rajasthan, India	N/A	23'10"	N/A	54	N/A
8/23/2024	Volgograd, Russia	N/A	24'7"	Image and recorded video	252	Islamic State

CIP Activations

Table 2: Overview of CIP activations

Date	Location	Elapsed time from event start to CIP activation (approx.)	CIP duration	Type of media	Items added to hash-sharing database	Extremist ideology
10/9/2019	Halle, Germany	7'15"	22'8"	Live video and manifesto	36	Antisemitic
5/20/2020	Glendale, AZ, USA	2'35"	24'34"	Recorded videos	198	Incel
5/14/2022	Buffalo, NY, USA	2'22"	25'39"	Live video, manifesto, attack plan diary	870	Great replacement
9/7/2022	Memphis, TN, USA	3'16" *	17'58"	Live video	50	None

4/10/2023	Louisville, KY, USA	3'32"	15'58"	Live video	3	None (anti-gun)
1/4/2024	Perry, Iowa, USA	7'15"	23'53"	Image with music	2	None (possibly Columbine-inspired)
3/16/2024	Levittown, Pennsylvania, USA	7'41" *	42'26"	Recorded video	23	Far right
8/12/2024	Eskişehir, Türkiye	N/A	19'21"	Live video and manifesto	910	Far right

* From start of live-stream or post.

Seven documented CIP activations have taken place between the initiation of the protocol and the date this report was written. These provide enough data to draw a rough narrative, but readers should keep the small sample size in mind and avoid drawing any sweeping conclusions.

- A typical CIP incident lasted around 24 hours, with the shortest coming in around 16 hours and the longest running 42 hours and 26 minutes.
- The shortest elapsed time from the start of an incident (i.e., the beginning of streaming/posting) to CIP activation was 2 hours and 22 minutes, and the longest was 7 hours and 41 minutes. There were no clear trends or data that might suggest why one incident took longer to prompt action than another.
- The number of items added to the hash-sharing database during CIP activations was widely variable, ranging from 2 to 870.
- The fastest response and the highest number of items both applied to the 2022 Buffalo shooting. This may reflect the fact that the shooter was more methodical in planning the promotion of his attack and manifesto, the contents of which were designed to activate online communities oriented to accelerationism, Saints Culture, and the Great Replacement conspiracy theory.
- Of the seven incidents, four were aligned with far right extremist ideologies, while three had no definitive and unambiguous ideological elements.
- Six of the seven incidents took place in the United States. While this is likely driven in part by the American epidemic of gun violence, it may also suggest that GIFCT needs to build capacity in other parts of the world.

- Four of the seven CIP activations involved live-streamed video. Three involved the posting of recorded material during or close to the time of the attack.

A brief look at each CIP activation follows.

Halle, Germany, 2019

On October 9, 2019, Stephan Balliet, 27, attempted to carry out a live-streamed mass shooting at a Jewish community center and synagogue in Halle, Germany, emulating previous mass shootings (including specifically the Christchurch attack). The attacker posted a manifesto and live-stream via Twitch, per a public statement from the company.[30] The live-stream lasted 35 minutes and was viewed live by five people. A recorded version was available for about 30 minutes on Twitch and was viewed by 2,200 people before the company actioned it. The company shared a hash of the video with GIFCT. Twitch's statement credited GIFCT with assisting in the incident, but did not cite the CIP.[31] The CIP was activated seven hours and 15 minutes after the event and continued into the next day, during which time according to GIFCT 36 visually distinct versions of the video were circulated.

Glendale, Arizona, USA, 2020

On May 20, 2020, 20-year-old Armando Hernandez, a self-professed incel, shot three people in a mall in Glendale, Arizona. He recorded the shooting on his phone and subsequently posted multiple videos to Snapchat. The event received relatively little media coverage. The incident took place around 6:25 p.m. PT.[32] The CIP was activated at 9 p.m., roughly two and a half hours after the incident began, and concluded at 9:34 p.m. May 21, during which time according to GIFCT 198 visually distinct versions of the video were posted.

Buffalo, New York, USA, 2022

On May 14, 2022, 18-year-old Payton S. Gendron shot and killed 10 people and injured three more, targeting Black people based on his understanding of the Great Replacement conspiracy theory.[33] He

[30] Daniel Koehler, "The Halle, Germany, Synagogue Attack and the Evolution of the Far-Right Terror Threat," CTC Sentinel 12, no. 11 (December 2019).

[31] Thread on Twitch's official Twitter account. October 9, 2019. https://twitter.com/Twitch/status/1182036266344271873, retrieved July 12, 2024.

[32] Matt Rodewald, Anita Roman, Brian Webb, Nicole Garcia, Justin Lum, and Matt Galka, "Glendale police: Suspect in Westgate shooting said he was bullied, wanted to gain respect," Fox 10 Phoenix, May 21, 2020, https://www.fox10phoenix.com/news/glendale-police-suspect-in-westgate-shooting-said-he-was-bullied-wanted-to-gain-respect; "Westgate shooting suspect accepts plea deal, faces 19 to 44 years in prison," 12 News Phoenix, April 1, 2022, https://www.12news.com/article/news/local/arizona/westgate-shooting-suspect-accepts-plea-deal-faces-19-to-44-years-in-prison/75-94659330-e39e-4b9c-8071-e7732ba5af18.

[33] Amarnath Amarasingam, Marc-André Argentino, and Graham Macklin, "The Buffalo Attack: The Cumulative Momentum of Far-Right Terror," CTC Sentinel 15, no. 7 (July 2022): 1–10.

documented his attack plans, posted a manifesto, and live-streamed the attack for 30 minutes using GIFCT various member platforms. The live-stream was seen by 22 people, one of whom recorded and reposted it, resulting in it being seen by more than 3 million people as of May 16.[34] About 870 visually distinct videos and images were identified. The shooting began around 2:30 p.m. ET, and the CIP was activated at 4:52 p.m., around 2 hours and 20 minutes later. The CIP concluded May 15, 2022, at 6:31 p.m.

Memphis, Tennessee, USA, 2022

On September 7, 2022, Ezekiel Kelly, 19, went on a shooting spree in Memphis. The shootings started at 12:33 a.m. ET, paused, then resumed at 4:30 p.m. Around 5:50 p.m., Kelly began streaming his violence on a GIFCT member platform. A total of six people were shot, three of whom were killed. About three hours and 16 minutes later, a CIP was activated, with 49 distinct videos and one image added to the hash-sharing database. No motive was conclusively identified for the violence. The CIP concluded at 3:04 p.m. on September 8.

Louisville, Kentucky, USA, 2023

On April 10, 2023, 25-year-old Connor Sturgeon shot and killed five people and injured eight more while live-streaming the attack on a GIFCT member platform. Evidence collected by police pointed to mental illness as a cause of the attack, along with a suggestion that he wanted to prove a point about gun violence. The attack began around 8:30 a.m. ET and was reported to police at 8:38 a.m. The CIP was activated at 12:02 p.m., three and a half hours after the event began. Three visually distinct versions of the video were identified. The CIP concluded April 11, 2023, at 4 a.m.

Perry, Iowa, USA, 2024

On January 4, 2024, Perry High School student Dylan Butler, 17, shot five students and three school employees, causing injuries that killed two of them. He posted an image related to the attack on TikTok. The attack began around 8:30 a.m. ET, and the CIP was activated at 3:30 p.m., about seven hours later. Two visually distinct items related to the TikTok post were added to the hash-sharing database. The CIP concluded at 3:23 p.m. on January 5, 2024.

Levittown, Pennsylvania, USA, 2024

On January 30, 2024, Justin Mohn, 32, killed his father in Levittown, Pennsylvania, sometime before 5 p.m. ET. Around 5 p.m., Mohn posted a video to a GIFCT member platform showing his father's decapitated head and calling for the killing of news media and government employees. A CIP was

[34] Drew Harwell and Will Oremus, "Only 22 Saw the Buffalo Shooting Live. Millions Have Seen It Since," The Washington Post, May 16, 2022.

activated at 12:41 a.m. ET on January 31, 2024, about seven hours and 41 minutes later, and 23 distinct items were added to the hash-sharing database. The CIP concluded at 7:07 p.m. on February 1, 2024.

Eskişehir, Türkiye, 2024

On August 12, 2024, Arda Küçükyetim, 18, committed a stabbing attack in Eskişehir, Türkiye. Küçükyetim live-streamed his attack and published a manifesto that expressed adherence to militant accelerationist ideology and advocated violence against Syrian immigrants. A CIP was activated at 6:22 p.m. ET on August 12, 2024, and was deactivated at 1:43 p.m. on August 13, 2024. While this CIP was active, GIFCT members shared 910 hashes of perpetrator-produced content through the hash-sharing database.

2024 GIFCT Working Group Participant Affiliations[35]

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

[35] 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				

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 ✉ outreach@gifct.org