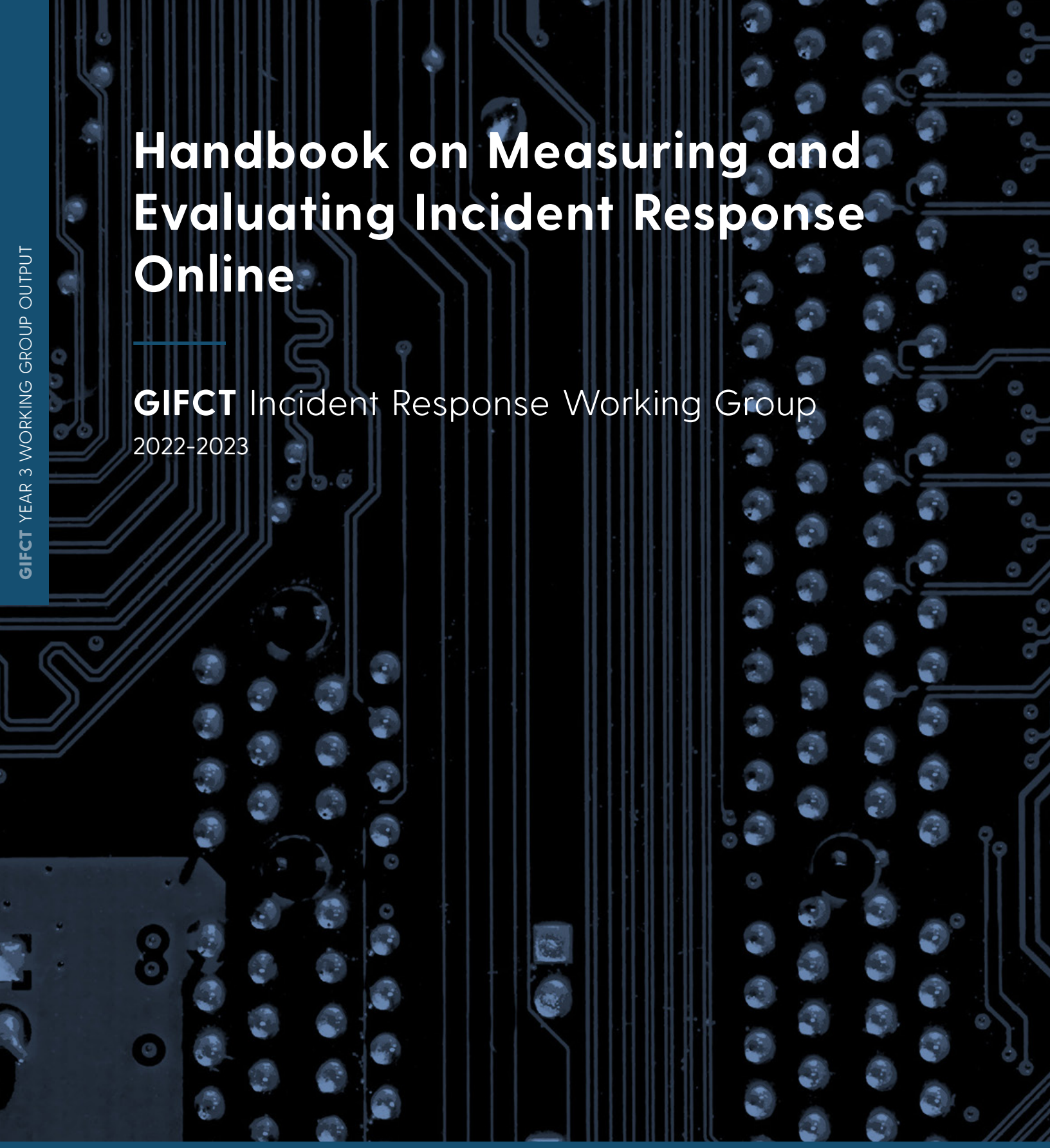
GIFCT YEAR 3 WORKING GROUP OUTPUT

Handbook on Measuring and Evaluating Incident Response Online

GIFCT Incident Response Working Group

2022-2023



GIFCT

Global Internet Forum to Counter Terrorism

About GIFCT Year 3 Working Group Outputs

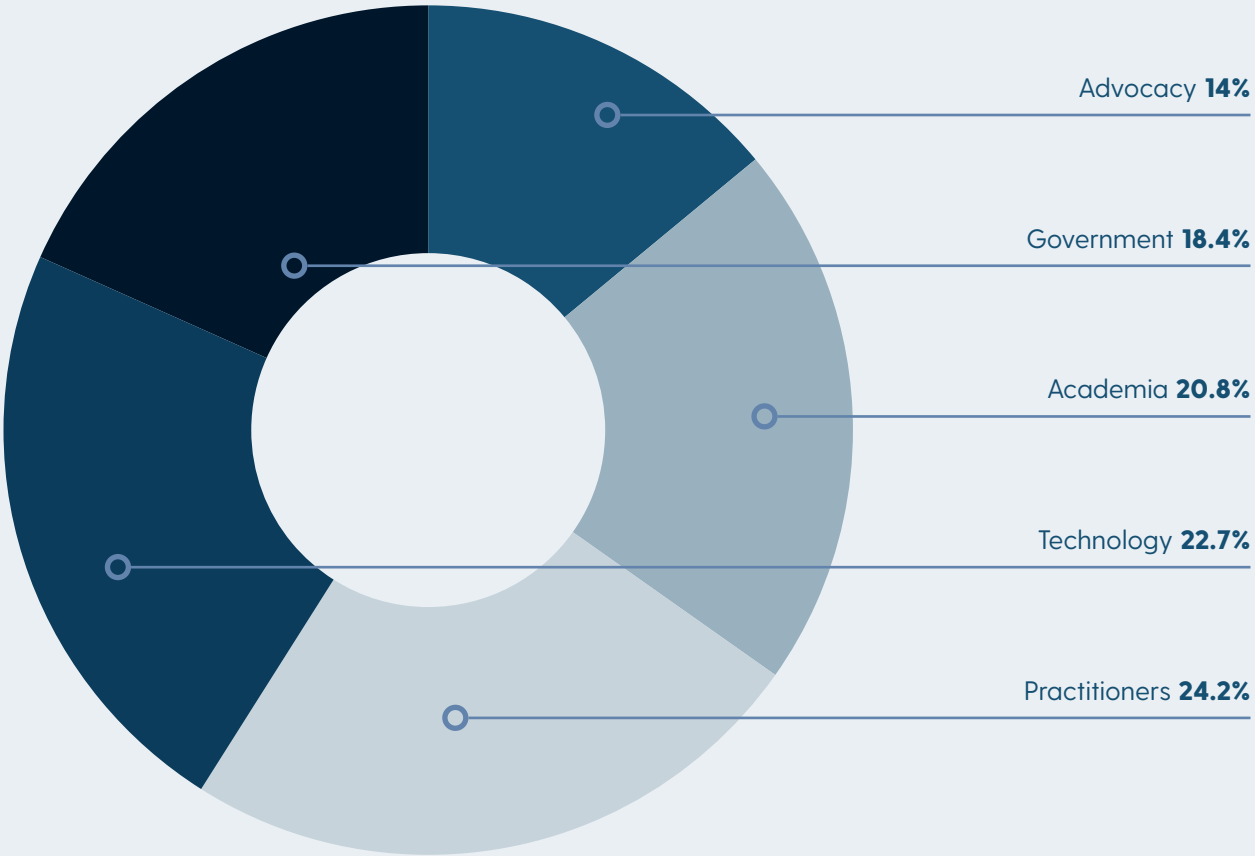
By Dr. Nagham El Karhili, Programming and Partnerships Lead, GIFCT

In November 2022, GIFCT launched its Year 3 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. Initiated 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.

Overall, this year's five thematic Working Groups convened 207 participants from 43 countries across six continents, with 59% drawn from civil society (14% advocacy organizations, 20.8% academia, and 24.2% practitioners), 18.4% representing governments, and 22.7% in tech.

WG Participants

Sectoral Breakdown



Beginning in November 2022, GIFCT Year 3 Working Groups focused on the following themes and outputs:

1. **Refining Incident Response: Building Nuance and Evaluation Frameworks:** This Working Group explored incident response processes and protocols of tech companies and the GIFCT resulting in a handbook. The handbook provides guidance on how to better measure and evaluate incident response around questions of transparency, communication, evaluation metrics, and human rights considerations.
2. **Blue Teaming: Alternative Platforms for Positive Intervention:** After recognizing a gap in the online intervention space, this GIFCT Working Group focused on highlighting alternative platforms through a tailored playbook of approaches to further PVE/CVE efforts on a wider diversity of platforms. This included reviewing intervention tactics for approaching alternative social media platforms, gaming spaces, online marketplaces, and adversarial platforms.
3. **Red Teaming: Assessing Threat and Safety by Design:** Looking at how the tech landscape is evolving in the next two to five years, this GIFCT Working Group worked to identify, and scrutinizes risk mitigation aspects of newer parts of the tech stack through a number of short blog posts, highlighting where safety-by-design efforts should evolve.
4. **Legal Frameworks: Animated Explainers on Definitions of Terrorism and Violent Extremism:** This Working Group tackled questions around definitions of terrorism along with the impact that they have on minority communities through the production of two complementary animated videos. The videos are aimed to support the global counterterrorism and counter violent extremism community in understanding, developing, and considering how they may apply definitions of terrorism and violent extremism.
5. **Frameworks for Meaningful Transparency:** In an effort to further the tech industry's continued commitment to transparency, this Working Group composed a report outlining the current state of play, various perspectives on barriers and risks around transparency reporting. While acknowledging the challenges, the Working Group provided cross sectoral views on what an ideal end state of meaningful transparency would be, along with guidance on ways to reach it.

We at GIFCT are grateful for all of the participants' hard work, time, and energy given to this year's Working Groups and look forward to what our next iteration will bring.

To see how Working Groups have evolved you can access Year One themes and outputs **HERE** and Year Two **HERE**.

Table of Contents

Handbook on Measuring and Evaluating Incident Response Online

Introducing the Handbook

Erin Saltman and Nusrat Farooq, GIFCT Incident Response Working Group Leads

This year's GIFCT Incident Response Working Group (IRWG) explored incident response processes and protocols of tech companies and the GIFCT, resulting in this Handbook on Measuring and Evaluating Incident Response Online, referred to throughout as the "GIFCT Handbook on Incident Response" or "the Handbook." The Handbook aims to guide tech companies, GIFCT, and other entities running incident response protocols on how to better measure and evaluate incident response work and how those actions can be communicated to wider stakeholders. The IRWG included experts directly involved in various aspects of incident response work from international tech companies, governments, and civil society representatives. The Handbook is divided into sections reviewing considerations and best practices related to (1) communication during an incident response, (2) qualitative metrics, (3) quantitative metrics to track during an incident for transparency reporting, (4) human rights due diligence indicators, (5) potential implications, metrics, and considerations around bystander footage, and (6) virality of content in and around a violent extremist or terrorist incident. These six themes make up the six sections of the Handbook. All Working Group outputs are made available on the GIFCT Working Groups page.

The themes of this Handbook were chosen by group participants, made up of cross-sector experts, to explore measurement and evaluation approaches around incident response. Each Handbook section was led by a Working Group participant who synthesized the group's dialogues from the sessions alongside extra research, interviews, and surveys where appropriate. While the group attempted to find replicable and scalable best practices and advice, there was also the recognition that incident response protocols and the ability to track various metrics can be platform or organizationally specific, since different organizations and platforms have different functions and have different metrics or user data available to them.

There were, however, some common themes that ran across the different Handbook topics. The need to base protocols off of known incidents and review protocols before and after an incident occurs will strengthen the ability of protocols to evolve based on known threats. Roles and responsibilities should be clearly defined so that when an incident happens there are clear lines of command and teams can work together easily without duplicating efforts. Giving the conscious space to run through scenarios to play out a protocol, whether internally to a given company or within wider Table Top exercises with external stakeholders, will allow questions to emerge and strengthen protocols. Effective communication during incidents should stay concise and adapt to the audience and context. When there is an ongoing, real-world threat, there is always a trade off between getting out with messaging early, and ensuring facts are accurate. Accuracy is more important than speed, but there

might also be different levels of communication based on the audience. Lastly, there will always be the need to balance freedom of expression and other human rights when assessing actions taken in the midst of an incident, questioning whether those actions are principled, proportionate, and necessary.

This Handbook builds off of GIFCT's previous Working Groups. In 2021 the Crisis Response and Incident Protocols Working Group focused on reviewing the roles and expectations of different sectors, global protocols, and individual technology companies and governments when responding to an incident. In 2022, the group mapped roles and responsibilities across GIFCT, Christchurch Call, and other Incident Response Protocols, as well as holding a series of Table Top exercises, and producing a paper on human rights and the lifecycle of an incident.

GIFCT hopes that this Handbook and previous Working Group outputs can help inspire tech companies, governments, and other entities involved in incident response work to review and evolve their incident response protocols, including their ability to communicate and work with relevant cross-sector stakeholders.

Communications During an Incident Response

Nusrat Farooq, GIFCT

Executive Summary

This section of the Handbook offers lessons learned, metrics, and recommendations to improve the effectiveness of and measure the impact of communications during incident response. The output is based on discussions within the GIFCT Incident Response Working Group[1] knowledge about three existing international incident response protocols,[2] and learnings from GIFCT Incident Response (IR) tabletop communication exercises run with member companies.

To improve communication in the multi-stakeholder system, IR teams should provide communication and engagement training to their internal teams and those identified as relevant third-party stakeholders and partners. Where appropriate, conducting debriefs and establishing information-sharing agreements with close partners can be helpful if any sensitive information is shared. Metrics to measure communication during and after IR are qualitative and quantitative, which are discussed and elaborated on within the other contributions to this Handbook.

By following the recommendations provided in this report, IR teams can more effectively manage the complex communication needs of stakeholders during an incident, leading to better outcomes and improved collaboration.

Introduction

This section of the Handbook focuses on measuring the impact of communications during an IR, providing lessons learned, suggested metrics, and recommendations to improve the effectiveness of and measure the impact of communications.

The information is based on the following:

1. Discussions within the 2022-23 GIFCT Incident Response Working Group (IRWG), which included experts from tech companies, academia, government, international organizations, civil society practitioners, and advocacy groups.
2. Knowledge about three existing international incident response protocols: GIFCT IR,[3] European

1 GIFCT Incident Response Working Group Reports, GIFCT.com, 2022, https://gifct.org/year-two-working-groups/.

2 Crisis Response Protocols: Mapping & Gap Analysis, GIFCT Crisis Response Working Group, GIFCT.com, 2022, https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-CRP-MapGap-1.1.pdf.

3 GIFCT's Incident Response, number of activations, and number of events GIFCT and its members communicated about: https://gifct.org/incident-response/.

Union Crisis Protocol (EU-CP),[4] and Christchurch Call Crisis Response Protocol (CC-CRP).[5]
3. Learnings from GIFCT IR tabletop[6] communication exercises.

Existing Knowledge

Communication during IR is often complicated due to 1) the transnational and adversarial nature of terrorism and violent extremism and 2) the multi-stakeholder and cross-platform communication needs placed on tech companies. Therefore, robust communication within and across incident response teams in the IR ecosystem is essential to combat terrorism and violent extremism.

In terms of the multi-stakeholder nature of the IR ecosystem, there are three international incident response protocols that GIFCT has mapped: GIFCT Incident Response (GIFCT-IR), European Union Crisis Protocol (EU-CP), and Christchurch Call Crisis Response Protocol (CC-CRP). All three protocols operate in a multi-stakeholder setting with member tech companies, member states, and relevant CSO networks as part of a proactive communication process. National IR protocols also exist, such as the Australia Online Content Incident Arrangement (AU-OCIA), New Zealand Online Crisis Response Process (NZ-OCRP), and United Kingdom Crisis Response Protocol (UK-CRP). In addition to these international and national IR protocols, several tech platforms (based on the size of a company and human and tooling resources available) have established their own incident response protocols and processes.

In terms of the cross-platform nature of the IR ecosystem, tech platforms that are members of GIFCT work together to communicate and respond to terrorist and violent extremist (TVE) attacks that have an online component. They share open-source information that could help identify and respond to terrorist and violent extremist content (TVEC). For example, GIFCT-IR acts as a hub to communicate activations of incident response[7] levels with tech company members when an incident reaches incident activation[8] levels based on transparent and public criteria.

This output offers recommendations to develop effective and strong communication within and across IR teams with suggested metrics to measure the impact of such communication.

4 European Union Crisis Protocol (EU-CP): https://home-affairs.ec.europa.eu/document/8797f42c-e46f-4efd-8a3b-21e0d1dc242a_en.

5 Christchurch Call's Crisis and Incident Response: https://www.christchurchcall.com/our-work/crisis-response/.

6 GIFCT 2021 Tabletop Exercise: https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-CR-TableTop-1.1.pdf.

7 GIFCT's highest incident activation level in its Content Incident Protocol (CIP): see https://gifct.org/content-incident-protocol/.

8 Incident Activation means when an offline TVE event reaches certain thresholds for GIFCT to activate any of its Incident Response Framework levels. Refer to GIFCT Transparency Report-2022 for details on these levels (pg. 37): https://gifct.org/wp-content/uploads/2022/12/GIFCT-Transparency-Report-2022.pdf.

Lessons Learned

Who to Communicate with, What, When, Where and How

Given the complex and adversarial nature of TVE attacks, IR teams (whether from tech companies or governments) have to juggle communication expectations and the needs of their various stakeholders while responding to the threat. The question then arises: how can teams be efficient with communication needs, respond efficiently to the incident, and fulfill the expectations of the IR ecosystem in which they operate? The GIFCT IRWG discussed the 5W questions to help identify groups of stakeholders and communication needs:

1. **Who** should one communicate with during and after an incident?
2. **What** should be communicated?[9]
3. **When** is it appropriate to communicate?
4. **How** should it be communicated – the tone, sentiment, and messages of safety and precaution?
5. **Where** should the communications happen – social media, email, phone call, virtual call, public blog post, private chat app, mainstream media, etc.?

The communications during and after an incident have three layers: a) communications within the internal team responding to the incident, b) communications more widely within an organization, and c) communications and coordination with external stakeholders such as media, government, civil society, public, and researchers. At all three levels, each team involved in IR should keep in mind the above five questions.

Communication Priority Setting

It is difficult to satisfy the communication expectations and needs of all stakeholders in the IR ecosystem while also responding to the incident. Therefore, IR teams should establish communication priority levels: High, Medium, and Low.

High: In this level of communication priority, only communicate with those stakeholders who are needed to take action or share information and updates in responding to an ongoing incident. These teams will benefit greatly from information-sharing during incidents. This is usually done during an incident and in the 24 hours after.

Medium: In this level of communication, only communicate when the incident has concluded, or when crucial actions are taken that affect wider communities. This usually involves one or

9 Read more about what statistics and metrics to be transparent about, qualitatively and quantitatively, in other sections of this Handbook developed by Laura DeBenedetto and Emil Girdan respectively.

two touch points potentially during an incident with a select and a pre-identified group of stakeholders receiving more matured communication after 24-36 hours of the incident.

Low: In this level of communication, communicate with all relevant and wider stakeholders about response actions taken during and after the incident. Depending on the capacities of an organization (usually government or tech), you can choose to publish something more public such as a blog post, public thread, or media article to be transparent about IR actions and conclusions.

It is the responsibility of the IR teams to establish information-sharing agreements where necessary with partnered stakeholders. For different IR teams, the priority stakeholders may be different depending on the type of industry and response type (online, offline, or both), so their communication priority setting will be different for each stakeholder, be it internally or externally. More than one stakeholder may be in one or more of these communication levels.

Conduct Trainings

To improve communication in the multi-stakeholder system in which IR teams operate, it is useful to provide communication and engagement opportunities, and where possible training to internal teams and external partners with relevant stakeholders (especially teams in the High Priority Communication level discussed above). This will make communication within the IR teams more efficient, help identify gaps, and improve collaboration between various sectors during an IR. The communication training can focus on:

1. Having open-ended Q&A sessions to build relationships and develop trust in advance of an incident to understand the communication needs of different sectors;
2. Discussing the effects of online propaganda in and around incidents for awareness;
3. Discussing the relevance of social media, other platforms, and online news outlets in communication needs based on stakeholders;
4. Thinking about procedures, protocols, and training needs. These could include intersections between existing legislation and voluntary frameworks on IR;
5. Having clear points of contact for IR communication needs;
6. Training on the discipline and clarity of communication needs;
7. Developing short and medium-term objectives with communication; and
8. Developing strategic objectives and training on awareness of long-term effects.

As an example, based on recommendations from previous incident response-focused GIFCT Working Groups and a human rights impact assessment,[10] GIFCT conducts communication and engagement exercises and tabletop exercises each year with its multi-stakeholder community, catering to both the

10 GIFCT Human Rights Impact Assessment: https://gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf.

internal needs of member companies and the external needs of governments and civil society. Other examples include the UN Office of Counter-Terrorism offering a crisis simulation to train teams on crisis response.

Affected communities and journalists[11] might also be helpful to sensitively include in relevant training so that they can be part of the IR ecosystem where appropriate. Civil society, practitioners, and advocacy groups are vital in ensuring communications and follow-ups with affected communities.

Conduct Debriefs

Debriefs are feedback sessions with various stakeholders after the conclusion of an incident. They are helpful in bringing all stakeholders together to discuss what went well, what did not go well, and how an IR team can improve in the future. They also highlight adversarial shifts or new tactics by bad actors to be aware of for future incidents. For example, based on GIFCT's 2021 IRWG,[12] GIFCT began conducting debrief sessions with its multi-stakeholder community in the aftermath of an incident activation. To date, GIFCT has conducted three multi-stakeholder debriefs after activating its highest protocol, the Content Incident Protocol (CIP). These debriefs are immensely helpful in gathering nuanced feedback to constantly evolve the GIFCT IR Protocol.

Depending on the team and the sector, an IR team can divide a debrief process into multiple debriefs: one within their IR team, another with wider internal stakeholders from an organization or network, and a third with wider cross-sector stakeholders.

Once the debrief process is complete, the IR team should communicate potential next steps, relevant policy changes, or improvements being made to relevant stakeholders. What medium to communicate through is also a choice and dependent on the audience. This could include using email, social media, blog posts, or other combined methods.

Below are case studies of three debriefs GIFCT has conducted so far:

GIFCT Buffalo CIP Activation blog[13]
GIFCT Memphis CIP Activation blog[14]
GIFCT Louisville CIP Activation blog[15]

11 UNESCO Handbook on including Journalists in communications during Incident Response: https://unesdoc.unesco.org/ark:/48223/pf0000247074/PDF/247074eng.pdf.multi.

12 GIFCT's 2021 Incident Response Working Group: https://gifct.org/wp-content/uploads/2021/07/GIFCT-CrisisWorkingGroup21-AnnualOutput.pdf.

13 GIFCT Buffalo Content Incident Protocol Activation blog: https://gifct.org/2022/06/23/debrief-cip-activation-buffalo/.

14 GIFCT Memphis Content Incident Protocol Activation blog: https://gifct.org/2022/09/07/content-incident-protocol-activated-in-response-to-shooting-in-memphis-tennessee-united-states/.

15 GIFCT Louisville Content Incident Protocol Activation blog: https://gifct.org/2023/04/10/content-incident-protocol-activated-in-response-to-shooting-in-louisville-kentucky-united-states/.

Additional Lessons Learned from the IRWG Dialogues:

1. From a civilian perspective, there is a lot of misleading information or blatant misinformation about incidents that appear as a crisis is taking place and in the hours after an incident. IR teams need to be careful about what information they share with the general public, weighing timeliness and accuracy.
2. Once trust is established within a tiered communications system, more trusted sources and communities can be communicated with more quickly and effectively than the wider public.
3. Respond, do not react. Keep the response fact-based, strategic, and short-term. Keep in mind short-term and long-term impacts, such as reputation risk and relevance of symbolism.
4. Be transparent about what is communicated. (More information on transparency metrics and evaluation in IR can be found in other sections of the Handbook on Qualitative Indicators of Transparency and Quantitative Indicators of Transparency.)
5. Hold a strategic communications meeting after an incident where appropriate to convene a collective and participatory discussion on influencing narratives to fill the communication vacuum created after an attack. This is to avoid any false narratives going around if the government or media do not immediately say something appropriate in the aftermath of the attack.
6. Terrorist attacks are multi-dimensional and require preparation, partnership, response, agility, and foresight.
7. Know your vulnerabilities, which might help coordinate and anticipate challenges.

Metrics

Metrics to measure communication during and after IR are both qualitative and quantitative as well as internal and external in nature. Below are suggested metrics to track related to IR communications:

1. Track the amount of time it takes your team to respond to a TVE incident. From the time your team learned about the incident:
2. How long did it take to get the team(s) together to respond to the attack?
3. How long did it take for certain high-level decisions to be made so the IR team could move ahead with the response?
4. Track small, medium, and large communication blockers that your team faces during IR.
5. Track the types of communication platforms and methods your team uses.
6. Track how many different levels of communication your team has with different stakeholders.
7. Track the tone, outcomes of messaging, messages of safety, and considerations of human rights due diligence in communications with multiple stakeholders.

8. Track communication with the wider public. Include the extent of information from official sources (government/ tech companies/ international protocols) to the target audience.

Tracking these qualitative and quantitative indicators will give IR teams insight into the quality and effectiveness of communications during and after an IR.

Conclusion

This section of the Handbook provides lessons learned, metrics, and recommendations to improve the effectiveness of and measure the impact of communications during IR. Communication is crucial during IR due to its complexity. This complexity arises from the adversarial side of terrorism and violent extremism, which is transnational in both offline and online spaces, and the response side, where the IR ecosystem is multi-stakeholder, cross-platform, and multi-centric in nature. The report offers recommendations to develop effective and strong communication within and across IR teams and metrics to measure the impact of such communication.

Recommendations

1. **Assess your current communication strategy:** This helps to identify the areas that need improvement. This can be done by gathering feedback from stakeholders and analyzing the metrics of your previous IRs.
2. **Develop a crisis communication plan:** Plans should outline roles, responsibilities, and communication channels for all stakeholders involved in IR. The plan should be regularly updated and tested to ensure its effectiveness during a crisis.
3. **Establish communication priorities:** This can be done by using a communication priority matrix. This will help you to allocate resources efficiently and communicate effectively with stakeholders.
4. **Provide training:** Training is helpful for IR teams and stakeholders to improve communication and collaboration. This training should focus on the technical aspects of communication and building relationships and trust between stakeholders.
5. **Conduct tabletop exercises and crisis simulations:** These help train IR teams and stakeholders on effective communication and collaboration during an IR.
6. **Establish information-sharing agreements:** Where appropriate, establish clear information-sharing agreements with relevant stakeholders involved in IR. This will help to ensure that everyone is aware of their communication responsibilities and expectations in the event of an incident.
7. **Use a combination of communication channels:** This helps reach all relevant stakeholders during and after an IR. This can include social media, email, phone calls, Zoom calls, blog posts, chat apps, mainstream media, and other relevant channels.
8. **Establish a clear point of contact:** This will help to ensure that communication is streamlined and efficient during an IR. Points of contact should be updated regularly.

9. **Develop short-term and long-term objectives:** This will help ensure that communication efforts are aligned with IR goals and adapt to changing circumstances.
10. **Be transparent:** Be transparent about IR communication efforts and regularly communicate updates to stakeholders (including updates about Human Rights Due Diligence during IR). This can help to build trust and credibility with stakeholders during a crisis.

By following these recommendations, IR teams can effectively manage the complex communication needs of stakeholders during an incident response, leading to better outcomes and improved collaboration.

Qualitative Indicators of Transparency During an Incident Response

Laura DeBenedetto, Independent Researcher

Executive Summary

This year, the GIFCT Incident Response Working Group (IRWG) reviewed how government and technology organizations communicate their approaches to violent extremist incidents. This output aims to establish recommendations for qualitative transparency, which refers to the type and manner of information communicated externally regarding a violent extremist incident. It encompasses what information is communicated and how it is conveyed, with a focus on ensuring clarity, accuracy, and openness in order to foster public and platform user trust. This document will discuss six key qualitative indicators identified by the Working Group: Audience, Frequency, Outlets, Impacts, Feedback, and Iteration. This output complements the GIFCT members' resource guide[16] and builds on the work outlined in the GIFCT Transparency Report.[17] We limited the scope of this output to qualitative transparency indicators to external stakeholders during and after an incident; quantitative transparency indicators linked to key metrics will be addressed in a separate section in the Handbook.

In addition to the proposed indicators, the Working Groups have identified existing gaps in government and technology institutions' transparency, particularly regarding victims and their families, as well as the impact on individuals who share content related to an event in order to raise awareness. While the primary purpose of this output is to provide a set of qualitative indicators to assist both technology and government organizations in evaluating the effectiveness of their qualitative transparency, we also recognize broader opportunities for considering the human rights implications of content and account removal in such situations. This should come from a place of empathy to bridge the gap with more marginalized communities often left out of these conversations. It is important to note that fully addressing this issue goes beyond the scope of this output, but it serves as a crucial reminder for organizations as they assess their incident-related protocols.

Introduction

The IRWG came together to discuss the current state of qualitative transparency across the government and technology organizations. The forum took perspectives from those working in government, technology, academia, and civil society to address the question of how government and technology institutions should structure communication about violent extremist events to their audiences through the lens of qualitative transparency.

16 GIFCT Member Resource Guide.

17 GIFCT Transparency Reports can be found on GIFCT's Transparency page.

Technology and government organizations have worked to refine their processes over the years to ensure concise and accurate communication. Each institution has its own protocols based on the stakeholders they serve and are able to tailor information that is best suited to their audience. While government and technology companies will likely have different information to share, we have seen increased collaboration to ensure efficiency and consistency. The key methods of collaboration include a number of crisis protocols.[18] It is important to highlight that while there is close collaboration, governments and technology companies have different objectives during a violent extremism incident. Governments are responsible for public safety and security and should focus on official updates and sharing verified information. Technology platforms have a responsibility to address harmful content and maintain the safety and integrity of their platforms. While their objectives differ, collaboration across these two sectors is imperative to taking meaningful action on incident-related content during a violent extremist event and ensuring public safety and awareness.

The following themes emerged from the Working Group conversation in relation to the status quo:

- During an incident, effective communication should follow certain principles. It should be concise, iterative, and on a need-to-know basis. Organizations should clearly communicate their plans regarding the timing of statements, the frequency of updates, and the intended recipients of the information when possible.
- Organizations need to address the potential human rights implications associated with removing incident-related content that individuals share on technology platforms. While quantitative metrics can provide insights into the number of removed pieces of content or accounts, qualitative indicators are essential in assessing how these actions might impact public discourse.
- The level of detail and frequency of communication may vary based on the type of institution (government or technology company) and the intended audience, and modifications may be necessary to align with the specific context and audience requirements.

Existing Knowledge

GIFCT provides annual transparency reports[19] to its stakeholders to highlight what was done during a past incident and the lessons learned. The Working Group used this as a starting point to tackle some of the gaps in communication and considerations around human rights.[20] The conversation centered on reviewing existing best practices in this space and discussing how governments and tech companies can optimize qualitative communications for their audience by looking at specific indicators.

18 Examples include the EU Crisis Protocol, Christchurch Call Crisis Response Protocol (CRP), GIFCT Incident Response Framework (IRF), Australia Online Content Incident Arrangement, New Zealand Online Crisis Response Process, and United Kingdom Online Policy Unit Crisis Response Protocol.

19 See GIFCT's Transparency page.

20 Human rights will be addressed more fully in a separate output.

During a violent extremist incident, both government and technology organizations strive to deliver accurate and timely information to their audiences. The timing of communication becomes crucial due to the immediate societal and media impacts of such incidents. However, a dilemma arises between the need to confirm information before widespread dissemination and the urgency associated with these events. It is worth noting that when organizations are forced to retract or correct information, it creates confusion among their audiences. Therefore, prioritizing accuracy over speed is essential in crisis communication to maintain clarity and avoid misinformation.

The GIFCT Incident Response Framework[21] was established to "address potential content circulating online resulting from an offline terrorist or violent extremist event." GIFCT reviews data from incident responses as noted in the Incident Response Framework (e.g., how quickly the protocol is activated and how much content is removed) to establish and fine-tune this framework after each activation. This Working Group sees this method of continuous improvement as critical to maintaining and improving the efficacy of the GIFCT Incident Response Framework and organizations' responses to violent extremist events.

Qualitative Transparency Indicators During Incident Response

The following list of qualitative indicators can help ensure that organizations address key qualitative concerns surfaced during the Working Group. The IRWG proposes that these be incorporated by organizations to gauge how well they are addressing key areas of concern:

1. **Audience**: Who the communication is for?
2. **Frequency**: When will there be updates on the incident?
3. **Outlets**: Where can individuals find updates?
4. **Impacts**: How the content is impacting society and what is happening to individuals who share incident-related content?
5. **Feedback**: How can individuals get in touch with relevant parties regarding new information about an incident?
6. **Iteration**: How can the process be improved going forward based on lessons learned?

When a violent extremist incident occurs, organizations need to execute their internal processes for managing and mitigating risk and determine their communication content and cadence to their external audiences as incidents evolve. Information shared across any channel needs to be verified for accuracy. This is the case for both government and technology companies. The key elements to address are the "who what when where why" - this can be summed up quickly and concisely for a broad audience with multiple stakeholders. Updates to communication are common and necessary as an incident evolves and should be done on a need-to-know basis. In some cases, organizations will also conduct a debrief on an incident to review what went well and what can be improved.

21 Found in the GIFCT Content Incident Protocol.

The IRWG looked at existing approaches and examined what is going well and where there are opportunities to improve communication. We looked at how organizations manage ongoing communication and incident impact and used that to highlight potential modifications that could help optimize how organizations structure communications.

Due to the varying nature of questions and time constraints, it is not always possible to address all inquiries simultaneously. Because of this, it is crucial to engage in iterative communication during and immediately after an incident. Once organizations have the opportunity to evaluate the outcomes of the incident, conducting debriefs becomes essential. These debrief sessions facilitate in-depth discussions about the event and identify areas that require improvement, thereby ensuring a continuous enhancement cycle. Some good examples of this include GIFCT's debriefs on recent incidents.[22]

One of the essential areas of managing an incident is controlling the spread of related content across tech platforms. Technology platforms are responsible for identifying incident-related content and mitigating its spread by employing an array of tools to detect and remove violating content. In the case of perpetrator-filmed content, tech platforms can use hash sharing and other forms of automation to mitigate the spread of imagery. While there might be several reasons that incident-related content is shared across platforms, it is important that platforms are transparent about their approach to removing the content. In the case of the Christchurch incident, we saw that some news outlets shared perpetrator content on their social media channels (specifically YouTube) as part of their coverage. Most tech platforms later removed it. There is an educational opportunity for individuals and media organizations who may share content to raise awareness versus those who share it to praise or support the act of violence. Governments are in a position to communicate that content should not be distributed, while technology companies should be clear that this content violates its policies and ensure that there are clear external community guidelines to ensure users understand why certain content is not permitted to be shared. As part of this, tech companies should have clear avenues to an appeals process in case something is incorrectly removed.

The Working Group also discussed the need to tailor information based on the audiences consuming incident information. This is especially relevant for governments. Depending on the organization, it makes sense to review what exactly a given audience needs to learn from a communication and determine the most useful information to convey. This scope and depth of information that organizations share will depend on the audience they aim to inform. While our Working Group can advise on some of the best practices, we also advise that each organization take stock of what has worked well for them and be open to feedback.

22 Examples of multi-stakeholder debriefs include GIFCT Memphis multistakeholder debrief and GIFCT Buffalo multistakeholder debrief.

Conclusion

Communication should continue to be brief and on a need-to-know basis. This is something that most organizations do well, but it is important to highlight the need to maintain this approach as part of the Working Group's recommendations.

Recommendations

We would like to highlight the following guiding communication themes for both governments and technology companies:

1. Based on the institution and the intended audience, there may need to be modifications in terms of the level of detail and frequency of communication. This can take the form of messaging that highlights the 5Ws and explicitly notes what information is relevant to different parties. Governments should be explicit about who the information pertains to when they release statements. When possible, we recommend parsing out communication for the following audiences:

    Impacted individuals - These are individuals directly affected by an incident, such as those present at the location or injured. Communication should include information about the incident's impact, affected areas, and available emergency services.

    People in the area of the incident - This refers to members of the public residing or present in the incident area. Communication should cover details about the affected area, safe spaces, available emergency services, and other pertinent information.

    Family members of impacted individuals - This communication aims to inform individuals who may not be in the incident area but have a connection to someone who is affected. It should provide contact methods, emergency services information, how to reach relevant authorities, expected update timings, and additional resources depending on the incident's nature.

    Media and general public - This communication should be the broadest and focus on answering the who, what, when, where, and why (if such information is available). Safety information should be prioritized (e.g., such as areas to avoid). General details about the incident, including perpetrators, response efforts, anticipated updates, and other relevant information, should also be included.

2. For governments, it is imperative that they understand the landscape of content being shared on platforms, so close collaboration with the technology sector is key. Governments should note how they view re-sharing content related to an incident in their communication (e.g., are there any instances where re-sharing is helpful or permitted?) and ensure alignment with platform policies.
3. Technology companies should be transparent about the implications of sharing content. Because their primary audience is platform users, tech companies should be clear about the implications

of incident-related content (e.g., there should be clear external guidelines around any context or parameters where technology companies allow sharing incident-related footage).

In addition to the themes above, it is recommended that both government and technology companies assess their communications using the six qualitative indicators derived from feedback received from the Working Group. The table below serves as a concise guide outlining these recommendations based on organization type. Addressing the questions under each qualitative indicator helps to gauge whether communication adequately addresses the primary areas of concern for different audiences during an incident.

Organization Type	Governments	Tech Companies
Audience(s)	· Impacted individuals · People in the area of an incident · Family of impacted individuals · General public/media outlets	· Platform users
Frequency	· How often will updates be available? · Why that cadence? · How has the incident impacted how governments are communicating?	· Highlight what is being done and when updates will be available
Outlets	· Government sites · Traditional media outlets	· Company blogs · Newsroom posts, media
Impacts	Discuss how tech companies are addressing incident-related content	Highlight potential harms to users and the type of content being shared
Feedback	Review communication cadence and efficacy	Review accuracy in moderations and appeals of content
Iteration	Discuss how to improve processes	Discuss how to improve processes

Although there may be variations in qualitative indicators between government and technology companies, it is crucial to prioritize coordination in order to achieve effective and inclusive communication with the public. By combining qualitative indicators with quantitative measures and taking into account human rights considerations, sustained collaboration can ensure meaningful transparency for all those affected by violent extremist events. Therefore, it is essential for government and technology institutions to maintain their cooperation, strengthen their collaboration, and leverage their respective expertise in order to continuously develop comprehensive strategies that prioritize meaningful transparency.

Quantitative Indicators of Transparency During an Incident Response

Emil Gîrdan, Romanian Ministry of Internal Affairs

Introduction

This section of the Handbook aims to provide guidance on quantitative metrics and measurements for incident response teams to consider, which (alongside other measures) will enable greater transparency around the impact and success of incident response efforts, including facilitating cooperation between actors involved in incident management. Measuring the impact of incident response is critical to improving its effectiveness, and transparency is a key component of that.

With reference to online assets, Incident Response (IR) work should focus on the three different phases of an incident – before, during, and after. During each phase, an IR team should perform specific activities that require different approaches and involve different ways of measuring performance. An IR team might be internal to a tech company, and could include law enforcement agencies, wider government entities, or international government bodies or networks. The guidance below consists of a list of quantitative indicators to measure the activities carried out among relevant stakeholders involved in IR based on the top challenges expressed by technology companies' representatives during the GIFCT Incident Response Working Group (IRWG). For GIFCT member companies, this includes the time taken to activate the GIFCT Incident Response Framework, the activation level, and the number and type of hashes added to the GIFCT database. Then there is data from member or other platforms that could be tracked on the number of pieces of content removed or accounts suspended (as a result of hash sharing versus internal tools and systems such as machine learning, trusted flaggers, referrals, or user reports). Other metrics for consideration include time taken to action content or accounts, the number of content flagged to human moderators, the type of users sharing terrorist and violent extremism (TVE) content, the number of public messages about the incident, and the number of instances of false positives.

This section of the Handbook ends with conclusions and recommendations for how the impact of IR teams during an incident can be assessed in terms of communication, management of violent extremist terrorist content, type of content that is managed, response time, accuracy of artificial intelligence tools, best practices, and lessons learned.

Methodology

The research strategy consisted of a coherent and efficient mix of methods, techniques, and tools used to achieve the assumed objectives. Descriptive research was used to establish the conceptual framework in the context of TVE events. Interview-based research and discussions within the IRWG assisted in determining how social media data is used by major international law enforcement agencies and national institutions with crime-fighting responsibilities and how it is exploited by terrorists and

violent extremists (for example, by using online content around an incident to attract new followers or spreading fear to the wider public).[23]

The main approach used in the research was comparative legal analysis of the national and international crisis response protocols.[24] Interviews complemented the information available from open sources to identify and collect new information on how technology companies can work in conjunction with national and international IR protocols, governments, law enforcement agencies, and civil society organizations to provide potential quantitative transparency metrics. Also investigated were the challenges that technology companies, governments, and national and international protocols face during the GIFCT Content Incident Protocol activations. At the end of the chapter, a list of transparency metrics is presented based on the lessons learned from previous GIFCT activation protocols and Working Group discussions.[25]

Quantitative Transparency Metrics in Incident Response

GIFCT recognizes the importance of transparency in its mission to prevent terrorists and violent extremists from exploiting digital platforms.[26] This chapter aims to provide guidance on quantitative measures for IR teams to consider during different phases of a TVE incident, which alongside other measures will enable greater transparency around the impact and success of the entire system. Cross-platform IR mechanisms are a relatively new initiative that is constantly evolving and improving. Still, due to the efforts of professionals, it has proven its usefulness on many occasions.[27] However, to further enhance reporting, there is a need for agreed-upon tools or frameworks to measure the impact of the IR work when a streaming TVE incident occurs. This Handbook attempts to provide guidance on the various ways an incident can be measured and evaluated, with strong safeguards to protect fundamental rights in line with UN Guiding Principles.

Although the GIFCT's Incident Response Protocol[28] is activated after an incident has occurred with specific online criteria, incident management begins even before the live-streaming of the TVE event,

23 Emil Gîrdan, "The role of SocMInt for national and international law enforcement agencies," in Intelligence Analysis in Social Media (2nd ed.), ed. Emil Gîrdan (Sitech Publishing House, 2023), 93-119.

24 These included the EU Crisis Protocol, Christchurch Call Crisis Response Protocol (CRP), GIFCT Incident Response Framework (IRF), Australia Online Content Incident Arrangement, New Zealand Online Crisis Response Process, and the United Kingdom Online Policy Unit Crisis Response Protocol. See the GIFCT Working Group 2021 output Crisis Response Protocols: Mapping & Gap Analysis, https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-CRP-MapGap-1.1.pdf.

25 See Ben Brody, "Nobody is coming to help Big Tech prevent online radicalization but itself," Protocol.com, May 16, 2022, https://www.protocol.com/policy/buffalo-shooting-social-media-policy and "Update: Content Incident Protocol Activated in Response to Shooting in Louisville, Kentucky, United States," GIFCT.com, April 10, 2023, https://gifct.org/2023/04/10/content-incident-protocol-activated-in-response-to-shooting-in-louisville-kentucky-united-states/.

26 See GIFCT.com, "Transparency," https://gifct.org/transparency/.

27 See for example GIFCT's CIP activations to date reporting on incidents in Buffalo (https://gifct.org/2022/06/23/debrief-cip-activation-buffalo/) and Louisville (https://gifct.org/2023/04/10/content-incident-protocol-activated-in-response-to-shooting-in-louisville-kentucky-united-states/).

28 The GIFCT's Incident Response Protocol has seven stages – Identify and validate, Incident detection, Information gathering and validation, Assess, Activate and notify, Prepare and act, and finally Conclude.

so the impact of the IR work should focus on three phases: before, during, and after incidents. IR teams therefore should perform specific activities in each phase, which require different approaches and involve different performance measurement methods (see Figure 1).

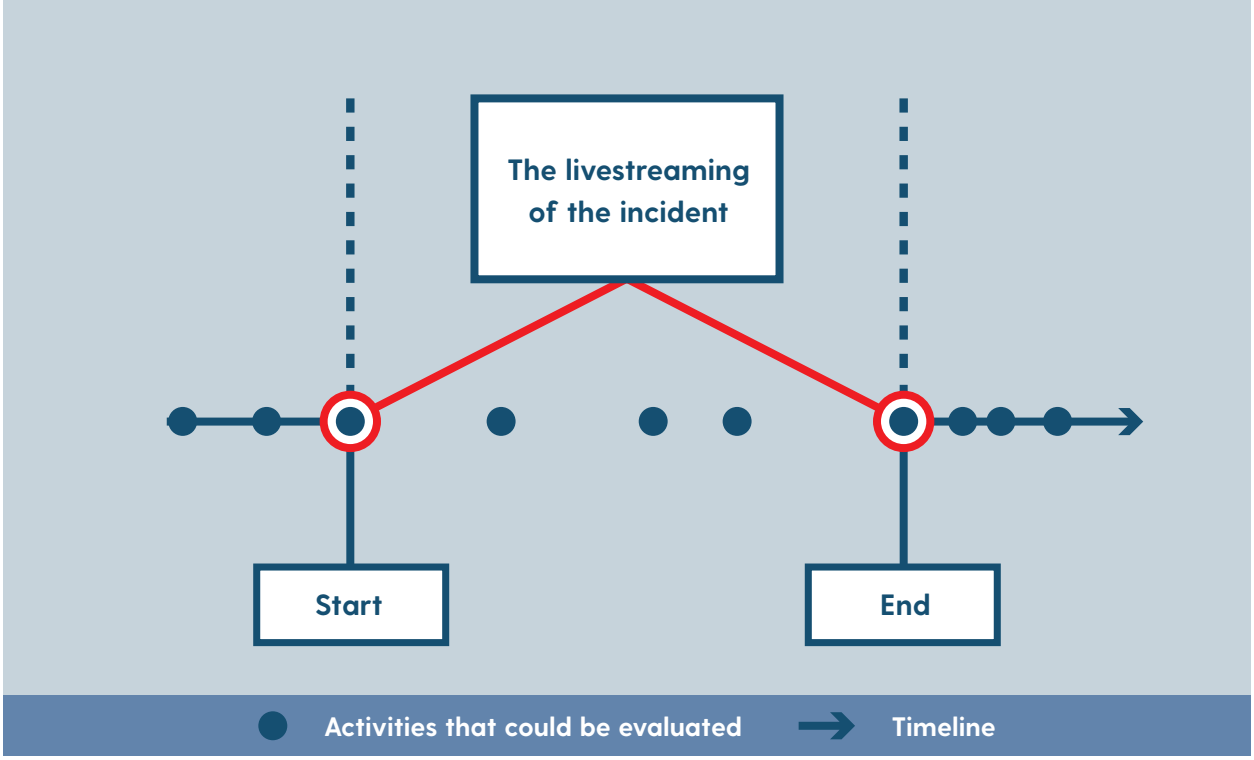


Fig. 1. Phases and activities related to live-streaming of an incident

During the 2022-2023 meetings, the GIFCT Incident Response Working Group agreed that an effective response depends primarily on the relationships among technology company representatives and among tech companies, government representatives, and relevant stakeholders. How these groups communicate with each other before, during, and after an incident should be based on a respect for human rights, accuracy, transparency, and the prioritized needs of various stakeholders. Similar conclusions were drawn from reviewing times that GIFCT CIPs were activated.[29]

In addition, the IRWG considered the results of the tabletop exercises (TTX)[30] on GIFCT's centralized communication mechanism and their impact on the human rights of online users, victims, and others. The IRWG determined that during an incident, IR teams within tech companies (which might include moderation teams, law enforcement outreach teams, and public policy leads) play a key role with

29 GIFCT publishes blog posts in the aftermath of activating a CIP; see GIFCT's CIP activations reporting on incidents in Buffalo (https://gifct.org/2022/06/23/debrief-cip-activation-buffalo/) and Louisville (https://gifct.org/2023/04/10/content-incident-protocol-activated-in-response-to-shooting-in-louisville-kentucky-united-states/).

30 See "Introducing 2022 GIFCT Working Group Outputs," GIFCT.com, 2022, https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-CR-Table-Top-1.1.pdf.

respect to the following:

- The level of communication among GIFCT and its member companies during an incident;
- Removals of TVE content, including measures taken by technology companies to ensure the transparency and accountability of their takedown processes;
- Identifying forms of perpetrator content (video, image, PDF, URL, etc.);
- Potential harm to users (as indicated by the prevalence and distribution of frequency of user exposures to the content, reach, and underlying causes for exposures);[31]
- GIFCT's response time to requests from government / civil society / technology companies;
- Measuring the accuracy of the AI tools and the humans involved in online content moderation tracking false/positives or false/negatives; and
- Best practices and lessons learned as a result of the activation of the incident protocols.

During IRWG meetings, technology company representatives also listed challenges that can often be difficult to measure:

- The spread velocity of the original TVE online content;
- The communication between technology companies and society during the incident (especially because for technology companies and national government bodies, accuracy is always valued before providing a quick answer); and
- Society's need to be informed versus the reluctance of technology companies and national government bodies to provide information.

Considering the above-mentioned aspects, the IRWG developed a list of potential metrics to measure the impact of incident response teams within tech companies (though some are also applicable to governments):

List of quantitative indicators for tech companies to consider tracking to measure transparency and effectiveness in relation to the phases of a TVE incident:

a. Transparency indicators that measure the impact of the IR team's *pre-incident* activities:
    - the number of suspended accounts for TVE activity;
    - the average time taken to suspend social media accounts sharing TVE online content;

31 "Metrics & Transparency: Data and Datasets to Track Harms, Design, and Process on Social Media Platforms," Integrity Institute, August 22, 2021, https://static1.squarespace.com/static/614cbb3258c5c87026497577/t/617834d31bcf2c5ac4c07494/1635267795944/Metrics+and+Transparency+-+Summary+%28EXTERNAL%29.pdf.

- the type of content that was removed (manifestos, manuals, in-action videos, propaganda images, etc.) in percentages; and
- the number of instances of false positives in takedown processes (legitimate content misidentified as TVE content).

b. Transparency quantitative indicators that measure the impact of the IR team's activities *during the incident*:
   - time to activation and activation level;
   - the amount of TVE content removed;
   - the time needed by social media administrators to remove the flagged content;
   - the number of users that saw or interacted with the content before it was removed;
   - the amount of content flagged to human moderators by AI or other tools developed by the platform;
   - the number of users to whom the TVE content was disseminated before it was removed;
   - the type of content that was removed (manifestos, manuals, in-action videos, propaganda images, etc.) in percentages;
   - the type of users who share TVE content (bots, fake accounts, verified accounts, etc.) in percentages;
   - the number of public messages via social media from national and international government officials pertaining to the incident;
   - the number of instances of false positives in takedown processes (legitimate content misidentified as TVE content);
   - the number of new companies/online service providers that react to TVE content;
   - the number of stakeholders (e.g., technology companies, national and international governments, academics) involved in responding to the incident;
   - the number of online trend reports during the TVE incident;
   - number of hashes shared with other stakeholders; and
   - the distribution of frequency of exposures for users (how many users had 0, 1, 2, 3, 4, 5+ harmful exposures).

c. Transparency quantitative indicators that measure the impact of the IR team's activities *after the incident*:
   - the percent of TVE content reported vs. removed;
   - the amount of manipulated content that was not detected by hashes of the original content;
   - the number of requests for takedowns of TVE content (hashes, URLs, etc.) from governments

and law enforcement agencies (including the number of these requests that were granted and the amount of content that was removed as a result);

- the percentage of requests/content providers that were denied or challenged by technology companies; and
- the amount of content preserved for investigative purposes.

Conclusions

1. Measuring the impact of IR is critical to improving response effectiveness. Transparency is a key component of IR, and measuring its impact is critical to improving the overall response.
2. Accuracy is more important than speed in providing incident-related information/statistics.
3. Some IR mechanisms are based on the relationships among GIFCT members rather than clear and predefined communication channels with stakeholders (including government and law enforcement agencies).
4. Technology companies' efforts to surface information and alerts about emerging offline terrorist and mass violent attacks that may have an online dimension are still quite limited (due to the number of GIFCT technology company partners, online content moderators, etc.).

Recommendations

1. Encourage cooperation and collaboration among technology companies, civil society organizations, national and international government bodies, and academic experts to prevent terrorists and violent extremists from exploiting digital platforms by highlighting the importance of transparency in this respect.
    a. Participate in joint activities (e.g., TTX, meetings, briefs) to encourage GIFCT and its member companies to communicate transparently during incidents and to ensure transparency in their takedown processes.
    b. Adopt a three-phase approach when measuring the impact of IR work during a TVE event.
2. Develop/improve the methodologies that ensure increased transparency of actions taken by the IR team.
    a. For technology companies, comprehensive algorithmic transparency, contact person in authority, and comprehensive transparency in company processes.
    b. For government/ international agencies, periodic press releases about the incident, liaison officer designation, and establishing communication channels with social media platforms and other IR team members.
3. Developing AI software and supplementing human moderators to identify online TVE content.

Human Rights Due Diligence Indicators During an Incident Response

Ottavia Galuzzi, Christchurch Call Advisory Network [32]

Executive Summary

This section of the Handbook aims to explore the potential human rights impacts during various stages of incident response and how these impacts can be measured. In particular, it builds upon the insights of last year's GIFCT Incident Response Working Group (IRWG) and aims to contribute to the Human Rights Matrix[33] by identifying quantitative metrics, key performance indicators (KPIs), and proposing a measurement process of human rights impacts intended for incident response protocol operators. In addition, this output focuses on tracking and communication principles of the United Nations Guiding Principles (UNGPs) to analyze potential challenges that tech companies and governments may face in applying these principles concretely while responding to a terrorist incident online. Although this output suggests metrics and KPIs, the outlined lists and processes may not be exhaustive and require clear definitions of stakeholders' achievements to refine indicators.

This output aims to provide concrete tools and actionable recommendations for incident protocol operators and specific actors involved in incident response and human rights (such as tech companies, governments, and civil society organizations). The main finding is the necessity to embed human rights impact assessment and due diligence within existing processes instead of creating new procedures that would only require more time and resources. While it may be a more challenging for smaller or less-resourced companies that would benefit from external support and mentorship programs,[34] adopting and implementing a human rights impact assessment is attainable for larger companies within existing transparency reporting efforts and other internal procedural rhythms. Through a few recommendations, this output seeks to provide stakeholders with practical ideas for improving current practices and strengthening their collaboration to ensure human rights protection.

The information and content presented in this section of the Handbook are a result of a presentation by Business Social Responsibility (BSR) on "Measuring Human Rights Impacts of Company Actions in Incident Response" and from the subsequent discussion among GIFCT IRWG participants on the topic of "Measuring the Impact of Incident Response on Human Rights." The IRWG participant insights are consolidated into this output in order to present a comprehensive view of the stakeholders involved[35]

32 Opinions presented are that of the author and not of the organization the author is affiliated with.

33 The Human Rights Matrix was developed by Dr. Farzaneh Badii as part of the Human Rights Lifecycle of a Terrorist Incident Online (output of GIFCT IRWG 2021-2022). This human rights matrix maps out and evaluates the impact of the incident protocol at each stage on human rights.

34 For example, the Tech Against Terrorism Mentorship Program.

35 GIFCT IRWG participants are composed of government representatives, technology company representatives, national and international protocol holders, and civil society representatives.

and summarize the main takeaways with suggested recommendations for a way forward. The background and additional information are gathered using desktop research, open-source intelligence research, and a review of existing literature and resources.

Introduction

In its work, GIFCT aims to be transparent, inclusive, and respectful of the fundamental and universal human rights that terrorists and violent extremists seek to undermine.[36] As stated in GIFCT's Human Rights Impact Assessment (HRIA)[37] and Human Rights Policy,[38] it is crucial to pursue a human-centric approach and embed human rights into strategies rather than considering human rights as extra aspects that need to be taken into consideration. This is particularly relevant in preventing and countering terrorist and violent extremist content (TVEC) online, where efforts to counter terrorism and violent extremism and protect human rights must be complementary and mutually reinforcing. Since its inception, GIFCT has worked to embed human rights considerations in its multi-stakeholder and operational processes, including in its Incidence Response Framework (IRF)[39] and Content Incident Protocol (CIP).[40]

Within this work, stakeholders must collaborate to continually understand the human dimension and assess what impact digital processes and technologies may have on human rights. Thanks to this approach, experts have put together insightful resources aimed at defining and addressing the impact of incident response stages on human rights. Through the IRWG, GIFCT has gathered groups of stakeholders eager to address these issues and has been working to create practical guidelines and outputs for all incident response operators. Although far from an exhaustive resource, this output strives to provide stakeholders with constructive content and actionable recommendations on measuring the human rights impact of incident response.

Existing Knowledge and Lessons Learned

The insights in this output rely on existing knowledge and research carried out to define international and domestic incident response protocols and address the measurement of their effectiveness and impact with regard to several aspects, with human rights being a priority. Last year's GIFCT IRWG consolidated information about existing mechanisms in the Crisis Response Protocols' "Mapping & Gap Analysis," which states that "all the protocols are voluntary in nature but grounded in robust legal frameworks that

36 See GIFCT's website.

37 Dunstan Allison-Hope, Lindsey Andersen, and Susan Morgan, "Human Rights Assessment: Global Internet Forum to Counter Terrorism," Business for Social Responsibility (BSR) (2021).

38 See GIFCT Human Rights Policy.

39 From GIFCT's website: "GIFCT's Incident Response Framework (IRF) guides how GIFCT and members respond to a mass violent incident, streamlining how members can communicate and share situational awareness as an incident unfolds in order to identify any online dimension to the offline attack."

40 From GIFCT's website: "The Content Incident Protocol (CIP) is a process by which GIFCT member companies quickly become aware of, assess, and address potential content circulating online resulting from an offline terrorist or violent extremist event."

ensure due process and protection or respect for human rights."[41] Additionally, the Human Rights Lifecycle of a Terrorist Incident Online identified what human rights could potentially be impacted at each stage of incident response, whose human rights, and what qualitative indicators stakeholders could use to measure the impact on human rights.[42] From this resource, the Human Rights Matrix is an essential tool to "map out and evaluate the impact of the incident protocol at each stage on human rights,"[43] as well as to understand the nature of such impact. In fact, different actions taken during stages of an incident response can lead to either a positive impact (i.e., human rights opportunity) or an adverse impact (i.e., human rights risk).

The UNGPs are an important starting point for understanding how tech companies and governments can evaluate the compliance of their services and products aimed at preventing and countering TVEC online, including incident response protocols. The UNGPs embody a key set of principles aimed at guiding governments and businesses like tech companies in having their products and services comply with human rights due diligence in an operational way.[44] Mechanisms like HRIA and human rights due diligence (HRDD) are essential for assessing a company's services, such as the response protocol to a terrorist incident online. GIFCT's HRIA is an example of commitment to human rights. It paves the way for accountability on the part of other actors in a sector where government and tech companies operate at the intersection of TVEC and human rights. While the community is learning how to incorporate HRDD in existing processes, much more can be done by actors in undertaking an HRIA and ensuring transparency.

Through an independent ongoing project being carried out by the Christchurch Call Advisory Network (CCAN) aimed at evaluating the impact of government and company commitments under the Christchurch Call to Action, one of the main preliminary findings is that evidence and outcomes of HRDD processes are challenging to find, either because these actors do not regularly disclose if they engage in such processes or those that reveal their HRDD rarely disclose their full outcomes, like publishing a HRIA.[45] Although the number of governments and tech companies evaluated in this project is limited, it still gives a glimpse into the issues at stake. Thanks to the work of organizations like BSR, some tech companies sought support for undertaking a HRIA of their services and products.[46] In order to drive effective change with respect to human rights, similar work should be pursued by the growing number of governments and tech companies involved in preventing and countering TVEC online. For those stakeholders that have already undertaken a HRIA, they should be committed to conducting HRDD of their businesses and operations over time.

41 GIFCT Working Group Output 2022, "Crisis Response Protocols: Mapping & Gap Analysis," GIFCT (2022).

42 Farzaneh Badii, "Human Rights Lifecycle of a Terrorist Incident Online," GIFCT Working Group Output 2022, GIFCT (2022).

43 Badii, "Human Rights Lifecycle of a Terrorist Incident Online."

44 United Nations High Commissioner for Human Rights Office (UNHCHR), "Guiding Principles on Business and Human Rights," (2011).

45 Christchurch Call Advisory Network, "Evaluating the Impact of Government and Company Commitments Under the Christchurch Call to Action: A Pilot Project of the Christchurch Call Advisory Network" (2022).

46 Examples of BSR's work: HRIA of Meta's Expansion of End-To-End Encryption (2022); Twitch HRIA (2023).

Quantitative Indicators

The measurement of human rights impact is a changing field, and it is tied up in broader discussions around transparency, security, and related metrics. As human rights and transparency metrics often overlap, measuring only the incident response impact on human rights becomes complicated. This highlights the importance of pursuing a comprehensive approach aimed at measuring the impacts of incident response on several elements while embedding human rights protection. This approach must be at the core of incident response protocols, where content moderation measures are used to limit the spread of harmful content online. However, activities inherent to content moderation may impact human rights positively or negatively. For this reason, several existing trust and safety metrics (like those around harmful exposure) are relevant for measuring human rights impacts during incident response stages.

In the Human Rights Lifecycle of a Terrorist Incident Online, a set of qualitative indicators are identified as metrics that might be present at each stage of an incident response protocol and impact human rights.[47] To build on this, the IRWG discussed quantitative indicators associated with the defined qualitative indicators that could provide a measurement approach of the impacts that incident response stages may have on human rights. The quantitative indicators are selected from established trust and safety metrics[48] that may have different names depending on companies and may not be applicable to all kinds of services. In this context, they indicate the impacts on human rights of actions taken during incident response stages. These indicators also rely on the findings and recommendations of the Integrity Institute on transparency metrics.[49]

With regards to protocol owners, the listed indicators are more intended for company reporting and multi-party protocol operators:

- Number of users exposed to harmful content related to an incident
- View rates: the number of times a piece of content was shown
- Sharing rates: the number of times a piece of content was shared
- Number of accounts and pieces of content flagged
- Number of accounts suspended or closed
- Number of platforms where harmful content was shared (as a cross-platform metric it should be compiled externally to any single platform)
- Number of stakeholders involved in consultations and processes
- Percentage of false positives: the percentage of content identified falsely as terrorist and violent extremist

47 Badii, "Human Rights Lifecycle of a Terrorist Incident Online."

48 Harsha Bhatlapenumarthy, James Gresham, "Metrics for Content Moderation," Trust & Safety Professional Association.

49 Integrity Institute, "Metrics and Transparency: Data and Datasets to Track Harms, Design, and Process on Social Media Platforms," (August 22, 2019).

- Percentage of accuracy in identifying perpetrator content only
- Number of hashes created and shared
- Number of users flagging harmful content
- Number of government requests and/or law enforcement authorities' orders to remove content
- Number of false reports by users (e.g., bad faith reporting and concerned content ultimately determined not violative)
- Number of appeals
- Number of successful appeals or overturns
- Time to resolution: the time taken for an appeal to be determined
- Response time from posting to removal of violating content
- Response time from detection to removal of violating content

Even though this selection is limited, it summarizes quantitative indicators that are relevant in certain stages of incident response and may have an impact, from physical to psychological, on human rights. These indicators can be combined with the qualitative indicators identified in the Human Rights Matrix[50] to provide a comprehensive picture of what and whose human rights are impacted, how stages of incident response impact human rights, and what indicators can measure the impacts.[51]

Measurement

The identification of metrics is important for measuring impact, but it is not the only necessary step. A measurement process requires several activities and multi-stakeholder collaboration involving different teams, which means resources and time. However, metrics and reporting processes should be considered mutually reinforcing, and a measuring strategy should be embedded in a company's existing processes instead of requiring the creation of new procedures.

This output proposes a practical measurement process based on the expertise of BSR and IRWG experts. This measurement process is recommended for protocol operators and participants, including tech companies, governments, non-governmental organizations and multi-party protocol operators involved in incident response protocol-related activities (collectively referred to below as stakeholders). In light of the hurdles in implementing an entirely new process, stakeholders are invited to implement the discussed steps within existing transparency reporting efforts and other internal procedures.

50 The Human Rights Matrix can be found in the Human Rights Lifecycle of a Terrorist Incident Online.

51 A focus on qualitative and quantitative indicators more relevant to governments should be considered in further work. This could be achieved by assessing the human rights impact of designation lists' invocation, unilateral government action in crisis context instead of participation in multi-stakeholder forums, and after-action review and assessment.

Define Activities and Goals

- Clearly define what activities should occur at each stage of the incident response protocol and by whom. This would help identify where human rights impact may be coming from and who should be involved to mitigate such impacts.
- Lay out the stakeholder's goals for each stage of the incident response and understand what the stakeholder attempts to accomplish.

Identify Potential Errors

- Determine the risks and errors that may occur at each stage, assess how they can impact human rights, and identify avoidance and mitigation strategies.

Consider Domino Impacts

- Consider how errors made by others (media or other protocol operators and participants) could affect the stakeholder's objectives and services and lead to adverse human rights impacts (e.g., the misidentification of a suspect).

Identify Relevant Existing Indicators and Narratives

- Combine relevant quantitative and qualitative indicators and analyze how each indicator can be relevant to the actions taken in each stage of incident response to measure the impacts on human rights effectively.
- Provide the background and context to the indicators identified by offering a narrative of the stakeholder's performance and efforts towards tackling terrorist incidents online.

Track the Performance

- Track the stakeholder's capacity to meet the goals for each stage of the incident response while measuring its human rights performance and other aspects like transparency based on identified metrics and narrative.
- Consider using KPIs and key performance narratives (KPNs) to articulate a clear vision of the stakeholder's achievements. KPIs are directly measurable values that demonstrate performance against a goal.[52] KPNs are narratives that explain how KPIs should be interpreted, describe the performance, the reason for the performance (e.g., users' appeals, errors committed, coordination with members, etc.), and future expectations (human rights protection, errors avoided, diversity of

52 Nina Hatch, Adam Fishman, Dunstan Allison-Hope, "Five Steps to Good Sustainability Reporting: A Practical Guide for Companies," BSR (2020).

stakeholders consulted, etc.).[53]

Consultation with Other Actors

- Engage with other actors involved in each stage of the incident response to assess their perspectives on the potential impact on human rights. Equally important is to engage with communities directly affected by the terrorist incident (both offline and online) to understand their needs and listen to their perspectives.

Assemble the Whole Picture and Communicate About It Externally

- Carry out a complete assessment of the impacts that each stage of incident response has had (from human rights to transparency). Include a summary narrative that explains the impact assessment and discuss any potential long-term implications for the stakeholders, the affected communities, and other actors involved.
- Communicate externally about the stakeholder's actions to tackle the terrorist incident online and share a copy of the impact assessment. Include external communication during different stages of the incident response as well as the measurement process to transparently inform the different audiences involved.

This strategy can be further tailored to different stakeholders' needs in responding to terrorist incidents online. Beyond offering this practical approach, this section of the Handbook represents a go-to resource that incident response protocol operators and participants can use to assess the different impacts caused by their processes.

With reference to the above-mentioned measurement method and the Human Rights Matrix, this output identifies quantitative indicators that may be present at different stages of the incident response and may impact human rights. Depending on the nature of the indicator, their higher or lower levels may cause more or less severe positive or negative impacts on human rights. By combining these indicators with the identified qualitative indicators, this output suggests KPIs and KPNs that protocol owners and participants can set for each stage to track performance and signal the directions of future achievements.[54]

Even though it is not always easy to identify directly measurable KPIs for the human rights impact of incident responses, crisis protocol operators and participants can set KPIs on numbers and percentages registered for terrorist incidents online and related responses that occurred in the past. The KPIs and KPNs presented below rely on the findings of the Human Rights Matrix, particularly with respect to

53 Hatch et al., "Five Steps to Good Sustainability Reporting."

54 Hatch et al., "Five Steps to Good Sustainability Reporting."

what and whose human rights may be impacted during stages of incident response. The indicators and KPIs listed may not be exhaustive, do not apply to all services, and differ depending on the role that a stakeholder has within a protocol. The table represents useful metrics and guidelines for protocol owners and participants and offers an initial suggestion that they can adapt and refine for their purposes.

KPIs, qualitative, and quantitative indicators of human rights impact during stages of incident response

Stage	Goal	KPIs / KPNs	Qualitative	Quantitative (# of)
Horizon[55]	Understand and identify the threat posed by the individual/group live-streaming before the attack is undertaken.	Monitor platforms used by the individual/group; Gather numbers of accounts/ content monitored or flagged to assess potential virality; Ongoing consultation with involved stakeholders.	Monitoring; Virality; Cross platforms; Broadening GIFCT's scope; Diversity of stakeholders' consultation.	Accounts/pieces of content flagged; stakeholders involved in consultations and processes.
Identify and validate	Seek information to understand what has happened/ is happening and ensure that understanding is valid.	Gather reliable information about what happened in the real world; Identify if there are online impacts and what these impacts are.	Monitoring; Use of OSINT; Probability of false positive; Verification of information.	Accounts /pieces of content flagged; users flagging harmful content.
Assess	Determine the scope of the incident and its online presence to assess next steps.	Evaluate the nature of the incident against the defined scope of action (e.g., mass violence, violent or extremist content shared by perpetrator); Identify if/how spread content is online by gathering no. of users exposed to harmful content, view rates and sharing rates.	Accuracy in identifying perpetrator-only content; Criteria to assess significance of online presence; Assessment of violent or extremist content; Probability of false positives.	Users exposed to harmful content; View rates; Sharing rates; accounts/ pieces of content flagged; % of false positives; % of accuracy in identifying perpetrator content only.

55 It is important to acknowledge the difficulty of measuring the horizon stage due to its nature, as to act on violative content a degree of foreknowledge is needed, and this is often unlikely to occur.

Activate and notify	Activate the protocol, notify the members, and inform them of the level of action needed.	Ensure that the incident falls within the scope to activate protocol; Provide accurate information to members about violating content and its spread online; Provide correct guidance to members about actions and monitor potential false positives.	Accuracy in identifying perpetrator-only content; Criteria to assess significance of online presence; Assessment of violent or extremist content; Probability of false positives.	Platforms where harmful content was shared; View rates; Sharing rates.
Prepare and Act	Share information about the incident online and the content's location with members to limit spread and take down harmful content.	Ensure accurate and complete information sharing with members; Create and share hashes in scope with the incident; Take down violating content/ accounts with preservation of evidence within appropriate timing and with human rights safeguards in place.	Expansion of hash database; Accuracy and completeness of information sharing; Take down of content; Actions taken other than content take down; Sharing hashes with third parties.	Hashes created and shared; Accounts suspended/closed; Gov't requests and/ or LEAs' orders to remove content; appeals; false reports; Response time from posting to removal of violating content; Response time from detection to removal; Response time from receipt to response to appeals.
Conclude	Summarize incident response actions and assess them against threshold.	Resolve potential false positives; Convene constructive multi-stakeholder debriefs; Gather lessons learned and explore how to implement them in future mitigation plans; Assess future human rights implications.	Diversity of stakeholder consultation; Mitigation plan with a human rights analysis for future events.	Successful appeals; Time to resolution.

The Challenges of Tracking and Communication

As seen in the proposed measurement process of human rights impact, tracking and communication are important steps to ensure a thorough assessment of governments' and companies' HRDD. On this matter, a go-to resource for tech companies and governments is the UNGPs, where HRDD is described as an ongoing 4-step process consisting of:

1. Assessing human rights impacts;
2. Integrating insights from the assessment into existing processes;
3. Tracking the effectiveness of the response to human rights impact; and
4. Communicating about this publicly.[56]

For the purpose of this output, the IRWG explored how the guidelines inherent to the steps of tracking (principle 20) and communication (principle 21) could be applied to incident response protocols. In particular, tracking should be based on qualitative and quantitative indicators, draw on feedback from internal and external sources (including affected stakeholders), and integrated into internal processes.[57] Communication should be of a form and frequency that reflect human rights impacts, accessible to the intended audience, provide sufficient information to evaluate the adequacy of the company's response, and not pose risks to affected stakeholders, staff, or legitimate commercial confidentiality requirements.[58]

Although these principles offer insightful guidance, they tend to be high-level and difficult to apply concretely. In addition, these principles are valid for businesses of any size and operating in different sectors, which may require an additional adjustment for their implementation. What is effective for companies and governments is to embed the reporting of human rights impacts within their business contexts[59] and pursue a defined set of principles of good reporting,[60] which are interconnected with the elements of tracking and communication and mutually reinforced by a set of metrics. For the purpose of measuring the human rights impact of incident response stages, the IRWG discussed how it might be fruitful to combine the following principles of good reporting with the tracking and communication guidelines:

- **Context**: Information is presented in its wider social, economic, human rights, and environmental context.

56 UNHCHR, "Guiding Principles on Business and Human Rights."

57 UNHCHR, "Guiding Principles on Business and Human Rights."

58 UNHCHR, "Guiding Principles on Business and Human Rights."

59 Shift & Mazars LLP, "UN Guiding Principles Reporting Framework with implementation guidance," (2017).

60 Hatch et al., "Five Steps to Good Sustainability Reporting."

- **Numbers and Narrative**: Key metrics, indicators, and targets are supported by an accompanying narrative that explains past trends and future expectations.
- **Connectivity**: Information enables the audience to assess the connections between risks and opportunities.
- **Clarity and Understanding**: The report is clear, minimizes legal or technical jargon, and enables all target audiences to readily comprehend the information being communicated.
- **Consistency and Comparability**: Reports are issued on an annual basis in formats that allow comparability between years so that readers can ascertain progress over time. The information presented (e.g., metrics) adheres to industry and global best practice to allow comparability across entities.
- **Stakeholder engagement**: Key internal and external stakeholders are identified and engaged on a regular and structured basis. The results of these engagements are transparently communicated and company responses to feedback are clear.

Looking at the tracking principle, the IRWG discussed what is feasible for tech companies and governments and how they can concretely measure the potential impact of incident response stages on human rights. Three main challenges were identified:

1. The lack of standardization across the sector with regard to tracking processes and reporting principles makes it hard for incident protocol operators to compare reporting procedures and achievements and identify a common measurement process for the impact of incident response stages on transparency, human rights, and other elements.
2. From a pragmatic point of view, it is challenging for incident protocol operators to track activities against indicators and communicate about it externally in the short and reactive timeline that a terrorist incident online may require due to its sudden and unforeseen nature.
3. External actors (often the media) have been very prompt in finding instances of "failure" or mistakes in the responses of incident protocol operators while failing to acknowledge the many "successes" in removing harmful content or other actions. This leads to the risks that actors may cause in reporting too quickly on a terrorist incident online, considering how much information can remain unknown about an event or a perpetrator in the aftermath. The potential adverse human rights impact of this requires a fine balance in the reporting process of every stakeholder involved (including external ones).

With regard to the communication principle, the IRWG discussed how mistakes are communicated and who has the authority and responsibility to think about affected communities. The discussion focused on the general scenario of a violation of human rights during a terrorist incident online and the IRWG agreed on the necessity of different layers and communication times.

From a government's perspective, specific judicial processes are in place to deal with such scenarios,

and they often consist of contacting the prosecutor who will take the next steps.[61] From a tech company's perspective, there are different communication channels established to ensure clear communication with users about content removed or actions taken on their accounts. In addition, it is important for tech companies to consider to what extent the information shared empowers the users to appeal if they think the actions taken are wrong. If in this case not enough information is shared with users to know what is happening, there may be errors and adverse human rights impact.

The IRWG outlined the role of civil society organizations (CSOs) in building trusted relationships with vulnerable communities and being in the best position to understand their needs and assess how they can be affected by adverse human rights impacts. The IRWG stressed the importance for tech companies to have a triage process that enables them to intake information from CSOs, resolve potential issues, and restore a balance between the external perception and the actual internal situation. An example of this is the information that CSOs communicate to tech companies about users' complaints of their accounts being deactivated for unknown reasons.

While errors and adverse impacts are often discussed in the multi-stakeholder debrief held after an incident occurred, the IRWG struggled to identify clear processes of when and where to communicate about a failure – like a clear violation of human rights – during the unfolding of a terrorist incident online and the consequent incident response. Although the perspectives and examples mentioned above represent effective communication methods, it is clear that there is still room for improvement for incident protocol operators to embed tracking and communication of human rights impacts in existing reporting processes.

Conclusion

Even though measuring human rights impact is an evolving field, the multi-stakeholder community agrees on the complementarity and mutually reinforcing nature of counter terrorism strategies and human rights. Now that it is time to move from theory to practice, the community needs practical guidance and actionable tools to measure the adverse human rights impacts of incident response and identify potential mitigation and avoidance strategies for such negative impacts. Based on insights from previous IRWG resources, this output offers a practical starting point for identifying metrics and defining measurement processes of human rights impacts. While it is clear that there are still many unknowns, it is also clear that improvements can be made only through a trusted multi-stakeholder collaboration.

Recommendations

1. Work towards embedding a measurement process of human rights impacts in existing reporting processes.
    - Tech companies, governments and protocol operators can test the proposed measurement

61 Different governments present differences in their judicial processes.

steps by tailoring them to their processes, particularly regarding identifying goals, metrics, and KPIs. The Human Right Matrix can be used as a starting point for discussion and implementation.

2. Consider possible scenarios unfolding during the response to terrorist incidents online (e.g., hashes removed in case of homicide as out of GIFCT scope, accounts takedowns, etc.) and assess the potential human rights impacts related to these scenarios.
    - Domestic and multi-party protocol operators can identify a set number of scenarios to go through and measure the potential human rights impacts in internal and multi-stakeholder training or tabletop exercises.
3. Ensure the inclusion of a HRIA covering each stage of the incident response.
    - CSOs can collaborate with tech companies and governments to provide guidance and standardize the HRIA process during a debrief by drafting a tailored paragraph about human rights protection to be added to debrief documentation and transparency reports.
4. Determine who is responsible for the decisions and actions taken during the stages of an incident response and clearly define accountability mechanisms tracking successes and errors.
    - Domestic and multi-party protocol operators can disclose what team/department owns responsibility and what accountability mechanisms are in place through their transparency reporting efforts.
5. Identify mitigation actions and strategies to avoid or limit the adverse impacts that incident response stages and potential errors may have on human rights.
    - Tech companies, governments, and protocol operators can assess if and how their existing mitigation actions can be implemented in the Incident Response Framework. CSOs can be consulted to ensure that such actions are compliant with the respect of human rights.
6. Actively listen to the needs of communities whose human rights may be adversely impacted.
    - CSOs represent the voices of these communities and can collaborate with tech companies and governments to strengthen their involvement in incident response stages to reduce the potential harm to human rights (such as informing vulnerable communities about terrorist incidents online while they are happening to protect their right to life, liberty, and security).

Metrics to Measure Bystander Footage During an Incident Response

Friederike Wegener, Directorate-General for Migration and Home Affairs, European Commission

Executive Summary

This section of the Handbook focuses on bystander footage of violent extremist and terrorist attacks circulating online. Bystander footage can broadly be defined as audio and/or visual content produced by people (or devices belonging to people) who are not perpetrators or accomplices in an attack and displays aspects of an attack (such as the perpetrator and/or victims). Examples of bystander footage include material from bystanders, CCTV, or police body cam recordings. By addressing this topic, the IR Working Group builds on the outcomes of the CRWG's tabletop exercise in April 2022, which highlighted the challenges of differentiating content captured from different vantage points (perpetrator, CCTV, bystander) and/or shared for different purposes, such as in condemnation or support of the attack, as an eyewitness account, as a safety message, or as part of journalist reporting. The purpose of this output is to explain why a common approach among crisis response protocols (and as far as possible among platforms) is needed, shed light on the complexity of bystander footage, propose a joint approach to address this content, and recommend measurements to assess crisis response mechanisms on bystander footage.

This chapter is based on the 2023 GIFCT Incident Response Working Group (IRWG) mapping exercise of existing crisis response mechanisms, exchanges between governments and law enforcement agencies, a multi-stakeholder discussion in the IRWG (including civil society and governments regarding a set of questions),[62] and written feedback from tech companies on the same set of questions. Responses from stakeholders are based on their professional experience and research, and information gathered in this process will be referred to as "IRWG discussion."

This output finds that existing incident response frameworks account for bystander footage to varying degrees. They range from explicit exclusion to broad mentioning and even specific inclusion of bystander footage to trigger the response. However, finding a common approach to bystander footage is crucial to prevent potential exploitation for terrorist and violent extremist (TVE) purposes, such as incitement of hate or further terrorization of victims, propagation of the ideology, radicalization, and recruitment.

62 The questions are as follows:

1. What role does/can bystander footage play in radicalization? To what extent is bystander footage manipulated and used to glorify the attack, perpetrator, or cause and incite others to commit violent and terrorist acts?
2. What are the challenges in actioning bystander footage and which actions are possible?
3. What aspects are important to support regarding investigations?
4. What impact does/can bystander footage have on affected communities/victims? How can journalistic and fundamental freedoms be protected?

Bystander footage can take various forms, and any content moderation action has to account for numerous considerations, ranging from the intent of the uploader, the role of news reporting, and fundamental rights concerns about the potential risk of exploitation. This output recommends a human rights impact assessment (HRIA) including bystander footage to see how it can be integrated into the GIFCT crisis response mechanisms found in the Incident Response (IR) framework while accounting for fundamental freedoms. The IRWG proposes a broad range of possible moderation actions to be taken based on assessing different criteria for companies, considering the needs of governments and media, and preserving the content for investigative purposes and potential reinstatement. The recommendations, proposed metrics, and measurements included in this document should be understood as initial ideas for further multi-stakeholder discussion to develop a proportionate tiered response to various types of bystander footage.

Why Do We Need a Common Approach to Bystander Footage?

In recent years, we have increasingly seen violent extremist and terrorist attacks with considerable online dimensions. This online content, usually produced by the perpetrators, is disseminated quickly across different platforms by terrorists and violent extremists (as well as their sympathizers and supporters) to foster their aims. Existing IR frameworks were set up with the goal of ensuring a coordinated, rapid, effective, and cross-border response to prevent the dissemination of terrorist or violent extremist content (TVEC) during and following a terrorist event while protecting human rights. The protocols thereby aim to prevent radicalization, safeguard the dignity of victims, and protect targeted communities.

In our full-scale digital world, with smartphones, CCTV, and body cams being widely used in daily life, the community—aiming to prevent the spread of violent extremist and terrorist content—is increasingly faced with non-perpetrator-produced content depicting attacks. Governments and civil society observe more and more cases where this content is increasingly used to terrorize communities, radicalize, recruit, and inspire individuals to commit terrorist attacks and glorify perpetrators.[63] In some instances, the content has also been used to incite hatred against the marginalized community to which the perpetrator belongs. In addition, bystander footage is not only used for terrorist purposes during the attack, but the community has also witnessed bystander footage of past attacks being reappropriated. In several instances, bystander footage has been shared to discredit the attack. Despite the good intentions behind such actions, this can negatively impact vulnerable communities.

A mapping of existing crisis response protocols[64] has found varying approaches to bystander footage, ranging from the inclusion of such content under certain conditions (EU Crisis Protocol) to the exclusion of this content (GIFCT Incident Response Framework).[65] In addition, online service providers do not

63 See for example the shooting in Nashville, U.S., March 27, 2023; the Ellen Premium Outlets shooting in Allen, Texas, U.S., May 6, 2023; and the attack in Annecy, France, June 9, 2023.

64 See "GIFCT IRWG Jan 2023 Mapping Protocols," GIFCT.com, 2022, https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-CRP-MapGap-1.1.pdf.

65 See "GIFCT Incident Response Framework," GIFCT.com, n.d., https://gifct.org/incident-response/.

necessarily have a dedicated category for bystander footage in their policies, resulting in no dedicated actioning capabilities or processes being in place. However, depending on the type of bystander footage, many online service providers treat the content under TVEC-adjacent categories, such as graphic violence and gore. Agreeing to a common approach is key to account for the following risks:

- **Immediate and easy manipulation of content for terrorist purposes**: Due to the nature of social media and online sharing platforms, bystander and other grey-zone footage can easily be multiplied and disseminated widely. It allows sympathizers and extremists to appropriate the content for malign purposes. This easily leads to an inundation of derivative content, exacerbated by malicious actors.
- **Easy dissemination and out linking to platforms with less content moderation:** The majority of TVEC is shifting from mainstream platforms to fringe/alt-tech platforms, where it is harder to moderate, and such content runs the risk of becoming eternal on the internet.
- **Protection of victims' dignity:** Victims can be shown in bystander footage, and their fundamental rights should be safeguarded. Content showing victims should be prevented from circulating online.
- **Avoiding negative impact on victims and survivors of previous attacks**: This includes their families, friends, and vulnerable communities.
- **Ambiguity of intent:** In a crisis scenario, it is almost impossible to individually assess the 'intention' of each uploader to gauge whether they have a valid exemption for posting footage of an attack. Without contextual explanatory information, bystander footage can be indistinguishable from perpetrator-produced footage and therefore can carry the same risks of real-world harm.

A common approach would ensure a holistic response to bystander footage that would account for the exploitation of this content and mitigate risks while protecting various rights and freedoms—most notably freedom of expression and protection of victims' dignity. It would also raise awareness of human rights violations and ensure accountability mechanisms (for instance when releasing body cam footage). Finding a common approach can also allow GIFCT to underline its ability to deliver meaningful multi-stakeholder outcomes.

Complexity of Bystander Footage

This output started with a broad definition of bystander footage for a common understanding of the issue. The IRWG discussion, however, showed that a more granular understanding is needed for it to have practical use and ensure protection of fundamental freedoms. The complexity of defining bystander footage centers around four key aspects and respective considerations:

- **Type of bystander footage**: Is the material produced by 1) an individual bystander, 2) from a government source (e.g., police body cam), or 3) private party (e.g., CCTV)? Factors to consider in relation to these include:

  - How and by whom is the content released;
  - During unfolding incidents, the content can be;
  - The importance of content of unfolding incidents for investigations, awareness-raising, or public safety; and
  - The viewpoints and quality of the content (which can differ greatly, posing a challenge for automated detection systems).
- **Intent**: What is the intent of the uploader? Is the content uploaded for legitimate reasons or to propagate a violent extremist ideology? Can the content provider be clearly identified as not cooperating, associating, or sympathizing with the perpetrator?
- **Scenery depicted**: What is shown can vary greatly, ranging from the perpetrator during the attack, the perpetrator pre/post-attack (i.e., not engaged in violence), and victims in various states (panic, injured, dead). (Depending on the scenery, the original uploaded content might not violate terms of service, but the risk persists that the content is misappropriated later and exploited.)
- **Degree of exploitation**: To what extent is or can the content be exploited for violent extremist and terrorist purposes. Factors to consider include:
  - Whether the bystander footage can be disseminated as part of journalistic work or to discredit the attack (recognizing that there is the possibility that journalistic content can be decontextualized and misused for terrorist purposes);
  - Irrespective of the intent (terrorism, information), whether the content can negatively impact victims; and
  - The repurposing of content over time (i.e., can the content be exploited after the initial incident response reaction?)

Recommended Approach

Given the considerable risk for exploitation and potential adverse effect on victims and vulnerable individuals/communities that bystander footage can have, it is crucial to agree on a common approach in dealing with this type of content. Given the different nature of the footage and intentions behind the upload, it is key that any response minimally interferes with freedom of expression and fundamental rights. It is therefore strongly recommended that companies and GIFCT engage in multi-stakeholder consultations, develop a definition and indicators relating to the complexity of bystander footage, and build a tiered system guiding proportionate actioning in response to various types of bystander footage.

In addition, based on the IRWG discussion and to account for the complexity of bystander footage, this output contains initial recommendations for general/foundational actions, actions during the attack, actions after the attack, and long-term considerations.

General/Foundational Actions

- Civil Society recommends a **HRIA** of including bystander footage in the GIFCT Incident Response Frameworks to assess if and how it can be included;
- **Governments recommend including bystander footage in the GIFCTs Incident Response Framework** to allow for activation of the protocols based on its exploitation for TVE purposes which has the potential for high virality;
- GIFCT to amend its criteria for activation and potentially develop an objective test to support crisis decision-making when facing significant uncertainty;
- Following a HRIA and developed criteria for a tiered approach, GIFCT includes harmful bystander footage in the hash-sharing database;
- **Companies to include bystander footage in their policies (based on the definition and indicators mentioned above) and provide transparency** on potential actioning tools and processes **or** provide transparency on how bystander footage is included and treated under TVEC-adjacent categories. Complaint and redress mechanisms should be accessible for users in relation to actioning of bystander footage;
- **Companies** are encouraged to **work with victims and civil society organizations** to further their understanding of adverse impacts on specific vulnerable communities, specifically which/how content can negatively impact victims and vulnerable users to inform their responses accordingly; and
- **Exclude from sanctions content** produced for reasonable and legitimate journalistic, artistic, or educational purposes.

Actions During the Attack

- **Industry, in a multi-stakeholder process, to explore technical capacity and human rights considerations to pause attempts to post derivative content** related to the attack for a limited amount of time in a crisis situation by default until a moderator can contextualize the content and assess the intent of the uploader.

Actions After the Attack

- **Industry to assess the content against provisions in their terms of service** other than terrorism, such as graphic violence (or during an attack if uploaded content cannot be paused by default);
- This assessment can include the type of bystander footage, the context of the released bystander footage, the possibility of repurposing the context, the scenery depicted, and the current and future degree and use of exploitation;
- Where the content is assessed not to violate terms of services but is assessed as having a

risk/potential for exploitation for terrorist purposes, adversely impact victims, or be used to incite hate against marginalized communities, **companies should use other actions to limit the dissemination of the content** (e.g., by reducing visibility (based on age or user settings), labeling, temporary suspending, demonetization, or downranking the content);

- Any actions should be taken in a transparent, consistent, and fair way. Platforms are encouraged to communicate their policies and practices to users and provide opportunities for feedback;
- **Preserve any bystander footage** related to an attack for an adequate period of time that would allow **law enforcement** to lawfully request and assess it to:
    - » Facilitate investigations and judicial proceedings. This may allow law enforcement to identify the offender(s)/ accomplices and avert potential escalation of the event or imminent threat to life, and may be of evidential value for the attribution of crimes to the perpetrator(s);
    - » Allow for reinstatement after the closure of the incidence response mechanism and a thorough assessment of the exploitability of the content, weighing fundamental freedoms; and
- **Law enforcement and legal authorities** are encouraged to use **court orders** to facilitate online service providers to supply the removed content.

Long-Term Considerations After an Attack

- Beyond the initial crisis response, **collaboration between law enforcement and tech companies** is recommended to **inform about the potential repurposing** of content over time. GIFCT is encouraged to ensure a comprehensive industry response in this case. Responses to this content could be in line with actions taken on so-called borderline content to reduce visibility and limit vitality.

Additional Considerations

- **Industry** is strongly encouraged to develop/use appropriate mechanisms and communication channels to **prevent out links** to other platforms where derivative content is hosted;
- **Industry** is encouraged to develop policies that **allow users to edit their uploads** with **explanatory information**, when signals of intent to document (and not to propagate) are present. This would allow the clarification of uploader intent, contextualize the content for viewers to reduce harm and ensure freedom of expression, and preventing decontextualization and misuse of the content;
- **Governments and law enforcement** are encouraged to **assess how (context, format) bystander footage** (for example police body cams) **is released** to avoid exploitation of this content. This can include:

  - The timeframe for releasing information;
  - What information is necessary to be released under accountability, transparency, and freedom of information provisions;
  - Where information is released;
  - How information is released, the format and context;
- **Traditional media outlets** are encouraged to:
  - apply ethical standards when depicting terrorist events online;
  - contextualize the content shown;
  - assess the potential negative impact of their coverage on victims and raise awareness about the adverse impact of sharing bystander content; and
  - apply measures to limit the risk of decontextualization and misappropriation of the content used in the coverage to avoid amplifying TVE content or instrumentalization for the propagation of hate and violence.

Metrics and Measurement

Based on the recommendations above, the following (mainly qualitative) metrics can be used to measure the success of the incidence response. These should be understood as initial metrics open for further development and assessment:

- Has bystander footage been integrated into the incident response mechanism?
- Do/how do GIFCT and companies provide transparency on tools and processes used to classify bystander footage, including the intent of the uploader and actions taken on this content?
- How has bystander footage been identified – user reports or automated detection? What were the follow-up actions to the identification (i.e., how has the content been actioned)?
- How many pieces of bystander footage have been identified?
- Have/how many pieces of misappropriate bystander footage have been identified?
- If applicable, how many pieces of content have been uploaded despite a default stop on uploads of derivative content?
- If applicable, how long did it take human verifiers to assess the content included in the default stop?
- If applicable, how many and which pieces of content have been reinstated and for which reasons?
- Has (and how much) content been preserved and made accessible to law enforcement?
- Were the follow-up actions successful in preventing the exploitation and virality of the content?

- How was the risk/potential for exploitation for terrorist purposes assessed? Was content that was assessed as having a low risk exploited or misused for violent extremist/terrorist purposes?
- Did the company engage with victim organizations to understand the impact and potential improvement of the content moderation action taken on the content? What was the feedback of the organization?
- Which human rights were impacted by the actions taken on bystander footage? Did mitigation actions exist?
- If applicable, for governments/law enforcement: was the contextualization of officially released bystander footage (such as body cams) successful in preventing misuse?

Conclusion

To sum up, bystander footage is a highly complex part of crisis response that may require different responses based on the case. It is recommended to find a common approach to bystander footage, taking into account fundamental rights concerns, to prevent its potential negative impact and allow for a holistic, consistent, and tailored response. Responses need to be flexible to account for the complexity of bystander footage, including the type of bystander footage, the scenery depicted, the intent of the uploader, and the (potential) use and degree of exploitation. Based on this, initial recommendations have been made which can serve as the basis for further multi-stakeholder discussion. These include an assessment of the content against provisions of terms of service that are not terrorism, exploration of technical capacities, and ethical considerations (e.g., to pause and verify the content before upload during a crisis situation). After the crisis, it is recommended to assess the risk of exploitation and decide on removal or upload, for companies to preserve content, and any additional measures to reduce virality of the content while respecting fundamental freedoms. In addition, industry is encouraged to further develop tools to detect altered content and out links pointing to related content on other platforms. Government and media are encouraged to assess context and format of sharing bystander footage, such as body cam recordings, to prevent exploitation and negative impacts on communities.

Recommendations

Governments recommend GIFCT include a reference to bystander footage in its protocol to allow for activation of the IRF in exceptional cases. In addition, GIFCT should conduct a HRIA of including bystander footage in its IR framework.[66] Moreover, it is important that incident response teams map different types of bystander footage to enable the development of criteria to assess the risk of exploitation. Building on this, it is recommended that the IRWG develop a tiered response to different types of bystander footage to guide decisions on how to moderate bystander footage (removal, reducing visibility, warning, etc.) in crisis situations. In this process, the IRWG is also encouraged to engage with human rights organizations and news media to ensure a more holistic response to

66 "GIFCT Incident Response Framework," https://gifct.org/incident-response/.

bystander footage while protecting fundamental freedoms. This work could provide the basis for including bystander footage as a dedicated category in companies' terms of service to ensure a proportionate response and prevent negative consequences of bystander footage circulating online while protecting fundamental freedoms.

Implications of Virality During Incident Response

Farzaneh Badii and Angie Orejuela, Digital Medusa

Executive Summary

This section of the Handbook discusses success measures for holistically mitigating viral terrorist incidents online while upholding values such as human rights and access to the internet. The nature of virality is important to consider as it is difficult to mitigate the impact of viral terrorist incidents given their potential to involve multiple jurisdictions and scale globally. A viral piece of content online can have ramifications for victims, bystanders, marginalized communities, internet users and others before, during, and after a terrorist incident. Mitigation techniques might also have implications for human rights and the global internet community. To achieve these goals, we have focused on techniques and processes such as the use of upload filters, detection algorithms, user suspension, disabling algorithmic amplification, nudges for third parties, and disabling certain search terms. While employing tried and tested techniques is essential, our goal was to take strides toward providing a more holistic approach to success measures, going beyond techniques to consider the broader impact on the community and human rights.

Why is Consideration of Virality During Incident Response Important?

The goal of this section of the Handbook is to understand the impact of virality during online terrorist crisis management and how to manage it successfully. The first step is to recognize the factors involved in identifying viral terrorist content.

The GIFCT Incident Response Working Group[67] identified the following reasons why it is important to consider virality during a crisis:

- Virality across platforms means that the speed and volume of content spread is enormous.
- Research findings from psychology and information science reveal that the speed of viral spread hinges on several specific factors, such as emotional content (e.g., intensity and valence of emotions), the importance of information for survival, cognitive coherence, plausibility, people's personal implication in the content, and the nature of images.[68]
- On video- and image-based platforms, viral content it might undergo various modifications to circumvent content policy.
- Viral content can be modified and used to violate the platform policy more easily.

67 See GIFCT's Working Group page: https://gifct.org/year-three-working-groups/.

68 See Soroush Vosoughi, Deb Roy, and Sinan Aral, "The spread of true and false news online," Science 359, no. 6380 (2018): 1146–1151, https://doi.org/10.1126/science.aap9559.

- The virality of content can lead to more "innocent user" account suspensions, as content may go viral due to sharing by bystanders who aim to raise awareness rather than support the attack.
- In some situations, virality might contribute to raising global awareness about an issue or incident, leading to collective action; hence, responses should be more nuanced and not merely aimed at preventing virality.
- It is essential to have a team that understands the local context.
- Virality can have a profound impact on victims and spur radicalization, especially if an attack targets marginalized communities.
- Terrorist groups may reuse viral content to demonstrate success.
- Viral content might achieve a global reach, involving many jurisdictions and making it difficult to contain and contextualize.
- Exposure to viral content can normalize terrorist content for average users.
- In a crisis's initial stages, content virality can be particularly harmful because companies are still researching the incident.
- The importance of virality increases when content is shared on platforms that do not cooperate with networks such as GIFCT.
- The need for rapid action in the face of virality must be balanced against the importance of allowing user expression.

Definition of Virality

Virality refers to the rapid spread and circulation of information, ideas, trends, or content through social networks, online platforms, or other communication channels.[69] When something becomes viral, it is shared and disseminated widely and rapidly, often reaching a large audience in a short period.

Defining virality is challenging as there are many technical but also social factors involved[70] and it is used in various social, economic, political, and legal contexts. Because virality is not solely a technical phenomenon, it is difficult to isolate technical vulnerabilities. Some scholars alternatively view virality as strictly connected to the nature of the content and not necessarily to the influencer or poster.[71]

69 Anastasia Denisova, "Viral journalism. Strategy, tactics and limitations of the fast spread of content on social media: Case study of the United Kingdom quality publications," Journalism, 2022, https://doi.org/10.1177/14648849221077749.

70 Karine Nahon and Jeff Hemsley, Going Viral (New York: Polity, 2013).

71 Marco Guerini, Carlo Strapparava, and Gozde Ozbal, "Exploring Text Virality in Social Networks," Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), July 17-21, 2011, https://doi.org/10.48550/arXiv.1203.5502.

Spectrum of Virality

The table below illustrates the spectrum of virality. At each level of virality, specific actions can be taken to mitigate harmful effects. Depending on the level of virality, the impact on human rights and the internet community can vary. Protocol operators need to consider success measures based on the level of virality. High-level virality will naturally involve more diverse actors (for example media might be more involved) and may have higher levels of impact on the internet and human rights that should be considered. The impact can also be greater if the content is shared on mainstream, large, and popular platforms.

Virality Level	Description
Low-Level Virality	Content or information that spreads to a limited extent within a small circle or niche audience. It generates some engagement and sharing, but the impact and reach remain relatively contained.
Moderate-Level Virality	Content or information that gains traction and spreads across a broader audience. It captures the attention of a larger number of people, resulting in increased sharing, comments, and engagement. The content may become popular within specific communities or interest groups but has not reached widespread or mainstream attention.
High-Level Virality	Content or information that experiences explosive growth in popularity and widespread sharing. It reaches a vast audience and gains significant attention from the general public. This level of virality often involves viral trends, memes, or news stories that dominate social media platforms and generate extensive media coverage.
Extreme-Level Virality	The most intense and widespread form of virality. Content or information becomes a cultural phenomenon, spreading rapidly and extensively across various platforms and media outlets. It captures the attention of millions, goes viral globally, and permeates popular culture.

Government Approaches to Virality

Some governments prefer not to define virality, allowing them to combat content without having to meet a specific threshold. However, it is essential to understand what constitutes viral content, as exceptional circumstances involving such content would require initiating the crisis protocol. However, this does not mean that evidence of virality must be present to take action and trigger the crisis protocol;

the protocol can be triggered to prevent virality.[72]

Some governments consider virality in terms of 1) the amount of disseminated content, 2) the number of platforms affected, and 3) how many users are exposed to or interact with the content (thereby contributing to virality). However, it is also important to focus on the speed of dissemination and context. Typically, in incidents involving the spread of terrorist content, governments and legislators do not regard virality as a factor for applying legislation. But certain policies and laws can indirectly prevent viral content from spreading. For example, the EU Terrorist Content Online Regulation obligates online platforms to take down terrorist content within one hour of receiving a removal order.[73]

Company Approaches to Virality

Companies and online platforms have recognized the need to prevent and mitigate the negative consequences of virality, including the spread of violent and terrorist content. Here are some steps that companies typically take to address these issues:

- Community Guidelines and Content Policies: Companies establish clear community guidelines and content policies that explicitly prohibit the dissemination of violent, terrorist, or extremist content. These guidelines outline acceptable behavior on the platform and provide a framework for content moderation and enforcement.
- Content Moderation: Companies invest in content moderation systems and teams to review and monitor user-generated content. Moderators are trained to identify and remove content that violates community guidelines, including violent and terrorist material. Advanced technologies like artificial intelligence and machine learning algorithms are often employed to assist in content analysis and identification. (This function can be outsourced and several companies provide this as a service.)
- Reporting and Flagging Mechanisms: Platforms implement reporting and flagging mechanisms that allow users to report potentially harmful or inappropriate content. This empowers the user community to be actively involved in identifying and flagging content that violates community guidelines. These reports are then reviewed and acted upon by the content moderation teams.
- Hashing and Digital Fingerprinting: Companies use hashing and digital fingerprinting technologies to identify and track the spread of known violent or terrorist content (for example, Meta offers this service to other companies). Hashing involves creating a unique digital fingerprint or signature for specific content, enabling companies to detect and remove duplicates quickly or reuploads of the same material.

72 Because of the complexity of virality, having to establish if a threshold is met (by counting content and platforms) is nearly impossible in a crisis moment. The government protocols such as the EU Crisis Protocol therefore leaves the judgment to experienced professionals to assess in the crisis situation if the virality is expected to be exceptional so that the protocol can be triggered or not.

73 See REGULATION (EU) 2021/784 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 29 April 2021 on addressing the dissemination of terrorist content online.

- Machine Learning and Automation: Companies leverage machine learning algorithms to improve content moderation processes. These algorithms can learn from vast amounts of data and identify patterns associated with violent or extremist content. They can aid in the automated detection and removal of such content, thereby enhancing the efficiency and effectiveness of moderation efforts.
- Collaborative Partnerships: Companies often collaborate with external organizations, such as law enforcement agencies, non-governmental organizations, and civil society groups, to share information, best practices, and intelligence related to the identification and prevention of violent and terrorist content. These partnerships facilitate a collective approach to addressing the issue across multiple platforms and jurisdictions.
- Transparency Reports and Accountability: Companies may release regular transparency reports that provide insights into their content moderation efforts. These reports share statistics, trends, and data regarding removing violent and terrorist content, showcasing the company's commitment to transparency, accountability, and continuous improvement.

It is important to note that while companies play a vital role in mitigating the spread of violent and terrorist content, the responsibility to address these issues extends to governments, policy makers, and users themselves. Collaboration among all stakeholders is essential to effectively tackle the challenges posed by virality and extremist content on digital platforms.

Analysis of Several of the Mitigation Techniques

Following are some of the mitigation techniques that tech companies use to remove viral terrorist or violent extremist content from their platforms:

Content removal: There is a general belief that to successfully mitigate harmful viral content, platforms, and governments should intervene to remove it. There are arguments that measures should be in place to stop content from going viral even before it reaches that point. According to Bak-Coleman et al., the removal of sensational content that could go viral within one to four hours can be very effective in stopping its spread.[74]

However, as some civil society positions assert, "removal" is not always the success measure, even in a terrorist context.[75] Viral terrorist content can raise awareness about an event, foster solidarity, make evidence more accessible, and even encourage civil society, governments, and tech companies to address issues from a global perspective to prevent future incidents. Viral content can also create a user journey and shed light on how and why the incident had a significant online angle.

74 Bak-Coleman et al., "Combining interventions to reduce the spread of viral misinformation," Nature Human Behaviour 6, no. 10 (2022): 1372-1380.

75 Civil Society Positions on Christchurch Call, 2019, https://www.eff.org/files/2019/05/16/community_input_on_christchurch_call.pdf.

Detection algorithms: In the case of the Buffalo shooting, Twitch's algorithm detected the incident. "Approximately 22 minutes into the livestream, the perpetrator stepped out of his car and began shooting. Twitch stopped the livestream approximately two minutes after the first person was shot."[76] Detection algorithms might be preferable to using upload filters due to the potential impact of upload filters on human rights and the internet architecture of upload filters.[77] Though the use of upload filters is now prevalent among governments and tech companies,[78] it is necessary to conduct an internet impact analysis to see how they affect the open and secure internet.

User suspension: User suspension is another technique that is used to prevent viral content. However, it is important to note that user suspension, especially of bystanders or users who want to raise awareness about the incident, can lead to serious implications for the user. It is important to conduct a human rights impact assessment and also consider the vital role of messaging before, during, and after the crisis.

Upload filters: Upload filters can filter out terrorist and extremist content before it is uploaded, and the number of accurate upload filters is often considered a quantitative measure for stopping terrorist content from going viral. However, upload filters can impact global access to the internet and inaccuracies in these can violate human rights.

Stopping highly influential spreaders: Another method to contain virality is to identify influential repeat spreaders at an early stage. These repeat spreaders are not necessarily human accounts; they can be bots developed by networks. Stopping highly influential spreaders and disrupting the spreader network that intentionally disseminates the content usually poses lower risks to human rights and can effectively contain virality.

Hide or remove content matching search terms: In extreme cases, this method is used to minimize access to exposure to content and prevent virality.

Disabling algorithmic amplification: Called the "virality circuit breaker," this approach might be easier and more efficacious than total removal, which requires very timely and sophisticated detection mechanisms. According to Bak-Coleman et al., virality circuit breakers can have efficacy similar to removal if the amount of content that needs to be removed is small.[79]

Nudges for third party or accidental share: Nudging users before they share a piece of terrorist

76 Office of the New York State Attorney General Letitia James, Attorney General Investigative Report on the role of online platforms in the tragic mass shooting in Buffalo on May 14, 2022, https://ag.ny.gov/sites/default/files/buffaloshooting-onlineplatformsreport.pdf.

77 Konstantinos Komaitas and Farzaneh Badiei, "Upload Filters And The Internet Architecture: What's There To Like?," Techdirt.com, November 16, 2020, https://www.techdirt.com/2020/11/16/upload-filters-internet-architecture-whats-there-to-like/.

78 "Upload filters: A danger to free internet content?," Ionos.com, March 28, 2019, https://www.ionos.com/digitalguide/websites/digital-law/upload-filters/; Eva Simon, "Upload Filters Are Back, and We Are Still Strongly Against Them," *Liberties*, November 3, 2020, https://www.liberties.eu/en/stories/uploa-filter-back-eu-2020/18938.

79 Bak-Coleman et al., "Combining interventions."

content can also reduce or prevent virality. Nudges can reduce sharing[80] and typically have lower implications for human rights and the global internet, as they do not usually lead to the suspension of accounts or removal.

Success measures: Success measures in this output include both qualitative and quantitative indicators that take human rights into account. The measures are not exhaustive but offer a starting point for discussion.

Protecting the privacy and security of the victims: Because terrorist content is sensational, it often leads to numerous instances of sharing. When content reaches virality, the human rights implications might increase. For example, when bystander footage goes viral, the identity of victims can be revealed, thereby violating their privacy. The images might also be used to target marginalized communities, leading to continued abuse and hate offline. Protecting the privacy and security of victims during the operation of the crisis protocol is essential.

Protecting the security of bystanders and other community members: Anecdotally, viral content can provide community members with vital information about where an attack is taking place and where they should seek shelter. While the speedy take down of content may protect the privacy and security of victims, restricting access to content could hamper efforts to raise awareness about the crisis. Civil society, companies, and other stakeholders should provide alternatives for raising awareness so that communities do not have to rely on viral terrorist content to protect themselves.

Lowering the number of vulnerable community interactions with the content:
In addition to the number of take downs, a quantitative success factor is to reduce the number of interactions with content by working on algorithms to prevent amplification. This measure, however, should be used cautiously, and the trade-off between this approach and access to digital services should be carefully considered.

Lowering the impact on account access and access to digital services: One method to contain viral content is to restrict access to accounts. However, social media accounts are potentially used for other purposes (for example to log in and use other digital services). Some companies may temporarily restrict access to accounts and then restore them, ensuring that restrictions do not lead to permanent removal (this might not be available on all platforms). When using account removal as a measure to limit the spread of content, companies and governments should consider strategies to mitigate the impact on access.

Reducing the quantity of terrorist content and the degree of interaction: Efforts should be made to reduce the quantity of terrorist content that circulates online and the degree of interaction with this content, including how widespread the content circulates across platforms, accounts, and exposed users.

80 Bak-Coleman et al., "Combining interventions."

Crisis Protocols and Virality

Protocol stage	Description	Who and what virality affects	Level of virality	Success measures and obstacles
Horizon	Horizon stage is when the attack is about to take place. Usually upload filters can be used at this stage.	Victims of terrorism, vulnerable groups, users of the internet	The level of virality at this stage is unknown.	Security and privacy of the victims might be at stake if the content goes viral. However, the perpetrator might become more violent if unable to stream. The impact of upload filters on other users freedom of expression and assembly should be assessed. Success measures include containing virality with minimum removal of evidence (removal does not mean destroying evidence). Raise awareness about the incident.
Identify and validate, incident detection, assess if the incident meets the criteria of mass violence and protocol threshold	At this stage the operator has identified the incident and is validating the identification.	Victims of terrorism, victims of counter terrorism efforts, vulnerable groups users of the internet	The level of virality might come to light at this stage. Low virality might result in not activating the protocol (in some circumstances). In the EU Crisis Protocol, for instance, anticipated virality is a factor to be considered when triggering the activation. Higher levels of virality might make it easier at this stage to identify the attack and assess it.	The detection algorithm might work effectively during this stage (see for example the Buffalo mass shooting). However, the detection algorithm should also be analyzed and assessed in light of its accuracy, as it can have an impact on the internet and human rights. At this stage having some criteria for identifying virality might be useful, especially if the protocol is triggered if meeting certain thresholds.
Activate and Notify	At this stage, the operators activate the protocol, notify members, inform them the level of action needed (only monitoring or doing more), and maybe also notify civil society and the public.	Victims of terrorism, victims of counter terrorism efforts, vulnerable groups, bystanders, users of the internet	Usually at this stage level of virality might be higher.	Notifying other operators and tech companies accurately about the attack and evaluating the right threshold for action. The public and civil society should also be informed about the incident effectively.

Prepare and Act	Look at OSINT materials, share hashes, share awareness about where the content is, take action to find/ moderate/ remove content, preserve data, share actions and outcomes, engage in ongoing strategic communications. May involve contact/ cooperation with third-party governments, OSPs and industry bodies like TAT.	Victims of terrorism, victims of counter terrorism efforts, vulnerable groups, bystanders, users of the internet	This stage contains a high level of virality as the protocol is activated.	During this stage, the algorithm amplifier should be deactivated and the spreader networks be identified and disabled. Because there is a high level of various content being distributed, the hash database might be expanded in ways that could have human rights and internet implications. It is important to protect the security and privacy of the victims at this stage, inform the public about the incident, increase the levels of accuracy in detection of spreader networks, and use nudging mechanisms to stop the spread further.
Recover and conclude	Assessment against threshold, stand down response, notify members/ stakeholders/ public of stand down, continue to monitor through documenting decisions/actions, organizing debrief/ multi-stakeholder review, sharing findings with stakeholders and public.	Victims of terrorism, victims of counter terrorism efforts, vulnerable groups, bystanders, users of the internet	Higher levels of virality.	Success measures at this stage could involve the protection, security, and privacy of victims, lowering the quantity of terrorist content that circulated online, limiting the degree of interaction with this content, and restricting how widespread the content circulates (platforms, accounts, exposed users). Provide enough information to bystanders and members of vulnerable communities, identify how the internet and human rights were impacted as a result of the protocol activation, and measure the accuracy of detection algorithm, upload filters, and hashes.

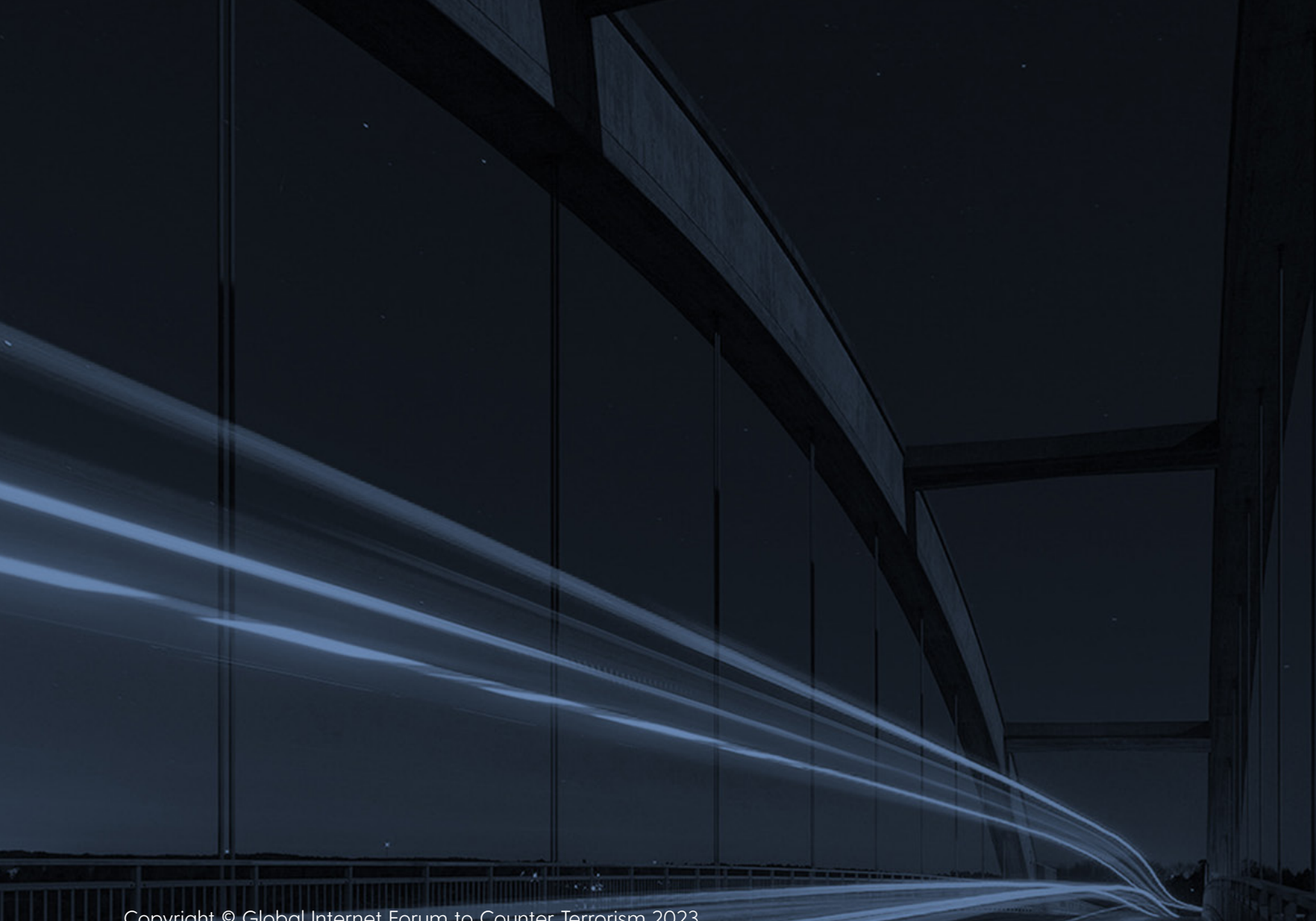
Conclusion

The virality of content during a crisis can make mitigating the harm more difficult right before, during, and after the crisis protocol is triggered. In order for companies and governments to mitigate the harm of viral content, they need to consider the extent of virality, the appropriate and proportionate response, and the method during each stage of the protocol, as well as the success factors.

Recommendations

This section of the Handbook does not provide a detailed analysis of how the recommendations can be applied during each stage of the crisis protocol. The main takeaways are that governments and companies should (1) come up with success measures that are qualitative and quantitative, (2) undertake regular impact assessments of methods on human rights values and internet access, and (3) have a holistic approach when considering the impact of mitigation measures and success factors, considering context, the level of virality, the methods used, and the impacted communities (this includes those communities affected by the terrorist incident and those affected by the mitigation methods).

By considering these aspects, a balanced and effective response to viral content can be crafted that mitigates harm while respecting the complexities of the online environment and the rights and needs of various stakeholders.



Recommended citation: Nusrat Farooq, Laura DeBenedetto, Emil Girden, Ottavia Galluzi, Friederike Wegener, and Farzaneh Badii, Handbook on Measuring and Evaluating Incident Response Online (Washington, DC: Global Internet Forum to Counter Terrorism, 2023), Year 3 Working Groups.

GIFCT is a 501(c)(3) non-profit organization and tech-led initiative with over 20 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 terrorists and violent extremists from exploiting digital platforms through our vision of a world in which the technology sector marshals its collective creativity and capacity to render terrorists and violent extremists ineffective online. In every aspect of our work, we aim to be transparent, inclusive, and respectful of the fundamental and universal human rights that terrorists and violent extremists seek to undermine.

www.gifct.org **outreach@gifct.org**