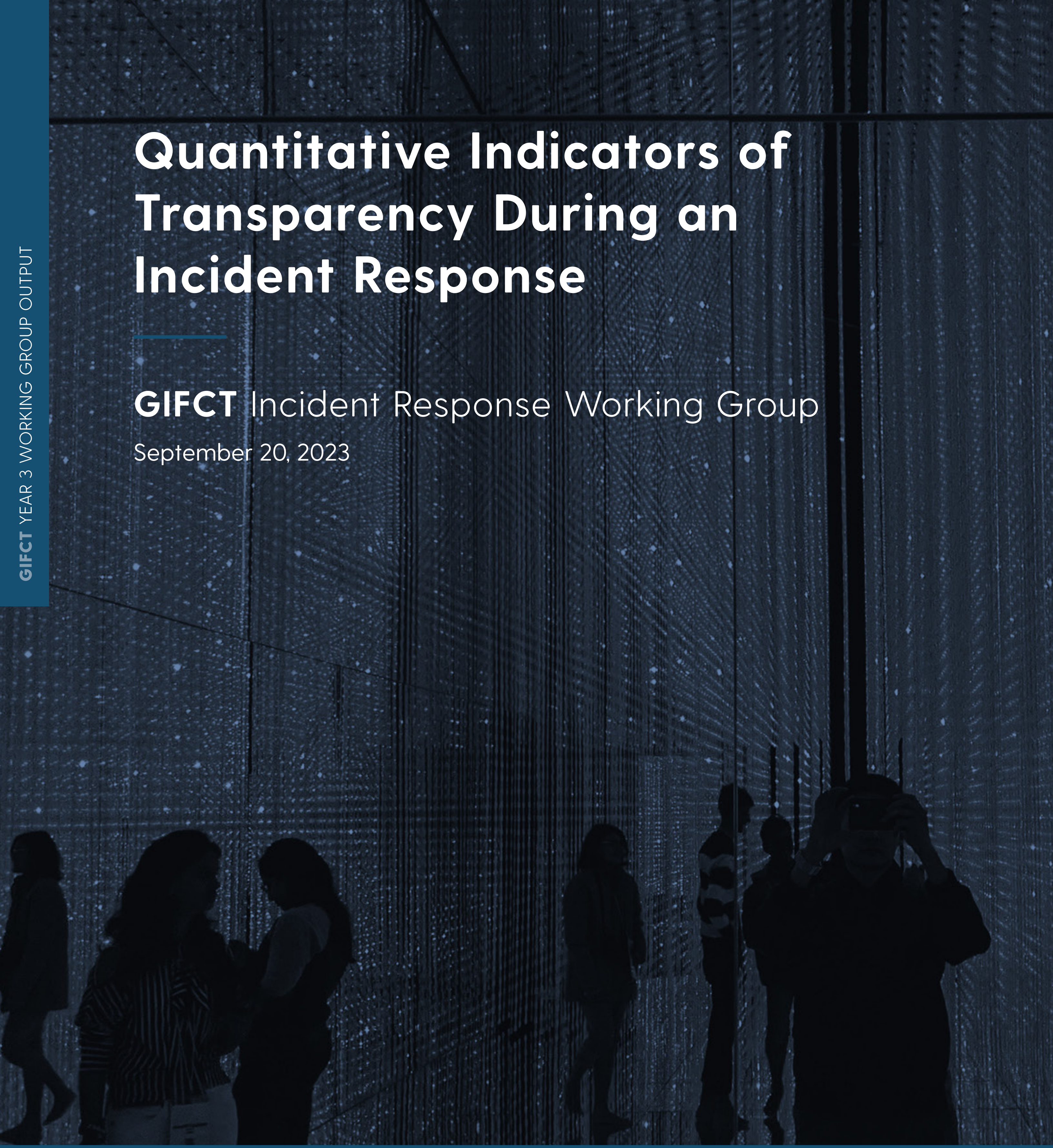
GIFCT YEAR 3 WORKING GROUP OUTPUT



Quantitative Indicators of Transparency During an Incident Response

GIFCT Incident Response Working Group

September 20, 2023

GIFCT
Global Internet Forum
to Counter Terrorism

Emil Gîrdan
Romanian Ministry of Internal Affairs

About GIFCT Year 3 Working Group Outputs

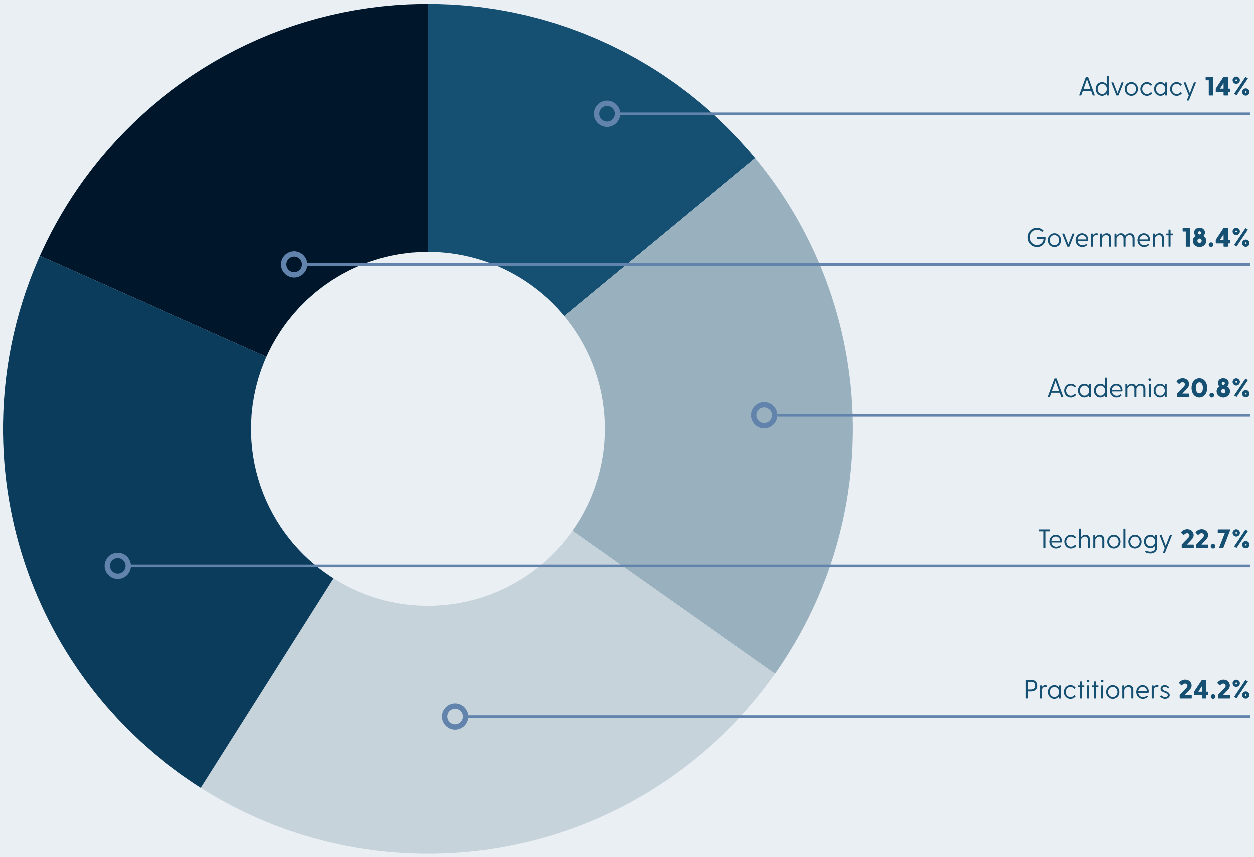
By Dr. Nagham El Karhili, Programming and Partnerships Lead, GIFCT

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. Started in 2020, GIFCT Working Groups contribute to growing our organizational capacity to deliver guidance and solutions to technology companies and practitioners working to counter terrorism and violent extremism.

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**.

Quantitative Indicators of Transparency During an Incident Response

The GIFCT Incident Response Working Group explored incident response processes and protocols of tech companies and GIFCT resulting in a Handbook on Measuring the Impact of Incident Response. The handbook provides guidance on how to better measure and evaluate incident response around questions of (1) communication, (2) qualitative and (3) quantitative transparency metrics, (4) human rights evaluation frameworks, (5) potential inclusions on measuring bystander footage, (6) and how to assess virality. This represents one section of the wider Handbook. All Working Group outputs are made available on the GIFCT Working Groups page.

Introduction

This output is a contribution to the wider *Handbook on Measuring the Impact of Incident Response* developed by the GIFCT Incident Response Working Group (2023). This section 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/violent extremism (TVE) content, the number of public messages about the incident, and the number of instances of false positives.

This contribution to 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 (AI) 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).[1]

The main approach used in the research was comparative legal analysis of the national and international crisis response protocols.[2] 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.[3]

Quantitative Transparency Metrics in Incident Response

GIFCT recognizes the importance of transparency in its mission to prevent terrorists and violent

[1] 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.

[2] 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

[3] 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/.

extremists from exploiting digital platforms.[4] 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.[5] 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[6] is activated after an incident has occurred with specific online criteria, incident management begins even before the live-streaming of the TVE event, 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).

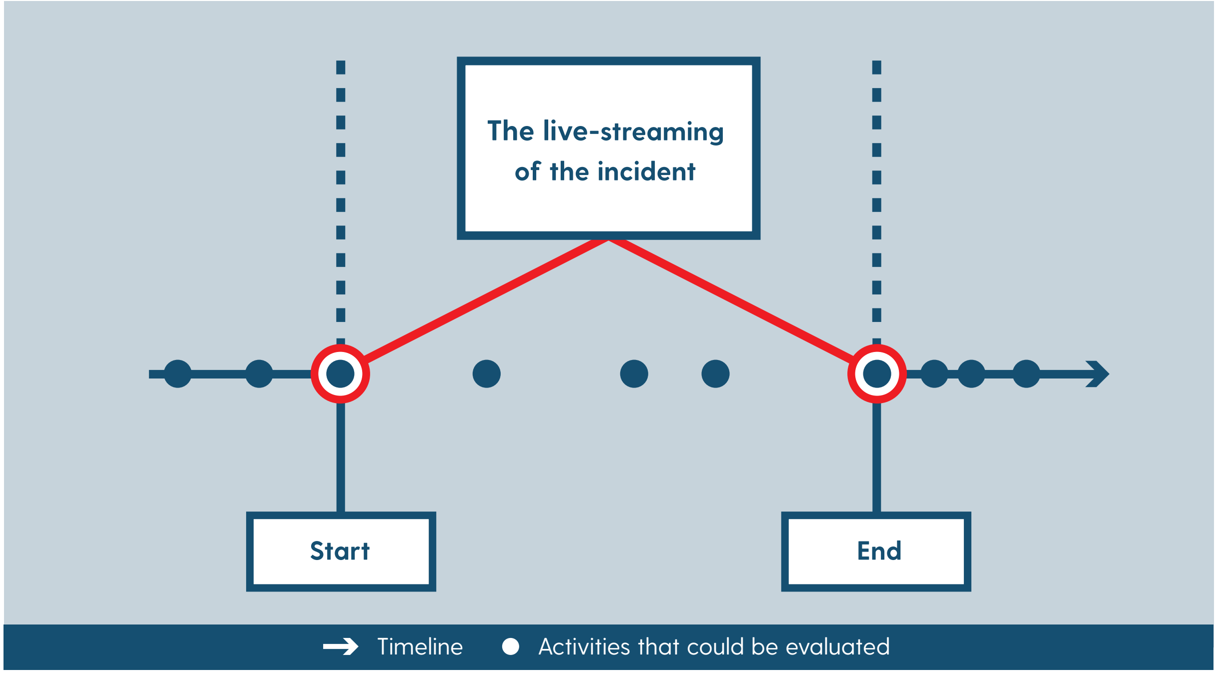

The live-streaming of the incident
Start
End
→ Timeline
● Activities that could be evaluated

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

[4] See GIFCT.com, "Transparency," https://gifct.org/transparency/.

[5] 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/).

[6] 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.

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.[7]

In addition, the IRWG considered the results of the tabletop exercises (TTX)[8] 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 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);[9]
- 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

[7] 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/).

[8] See "Introducing 2022 GIFCT Working Group Outputs," GIFCT.com, 2022, https://gifct.org/wp-content/uploads/2022/07/GIFCT-22WG-CR-TableTop-1.1.pdf.

[9] "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.

- 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:

1. Transparency indicators that measure the impact of the IR team's *pre-incident* activities:
    a. the number of suspended accounts for TVE activity;
    b. the average time taken to suspend social media accounts sharing TVE online content;
    c. the type of content that was removed (manifestos, manuals, in-action videos, propaganda images, etc.) in percentages; and
    d. the number of instances of false positives in takedown processes (legitimate content misidentified as TVE content).
2. Transparency quantitative indicators that measure the impact of the IR team's activities *during the incident*:
    a. time to activation and activation level;
    b. the amount of TVE content removed;
    c. the time needed by social media administrators to remove the flagged content;
    d. the number of users that saw or interacted with the content before it was removed;
    e. the amount of content flagged to human moderators by AI or other tools developed by the platform;
    f. the number of users to whom the TVE content was disseminated before it was removed;
    g. the type of content that was removed (manifestos, manuals, in-action videos, propaganda images, etc.) in percentages;
    h. the type of users who share TVE content (bots, fake accounts, verified accounts, etc.) in percentages;
    i. the number of public messages via social media from national and international government officials pertaining to the incident;
    j. the number of instances of false positives in takedown processes (legitimate content misidentified as TVE content);
    k. the number of new companies/online service providers that react to TVE content;

l. the number of stakeholders (e.g., technology companies, national and international governments, academics) involved in responding to the incident;

m. the number of online trend reports during the TVE incident;

n. number of hashes shared with other stakeholders; and

o. the distribution of frequency of exposures for users (how many users had 0, 1, 2, 3, 4, 5+ harmful exposures).

3. Transparency quantitative indicators that measure the impact of the IR team's activities after the *incident*:

   a. the percent of TVE content reported vs. removed;

   b. the amount of manipulated content that was not detected by hashes of the original content;

   c. 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);

   d. the percentage of requests/content providers that were denied or challenged by technology companies; and

   e. the amount of content preserved for investigative purposes.

Recommended citation: Emil Gîrdan, Quantitative Indicators of Transparency During an Incident Response (Washington, D.C.: 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