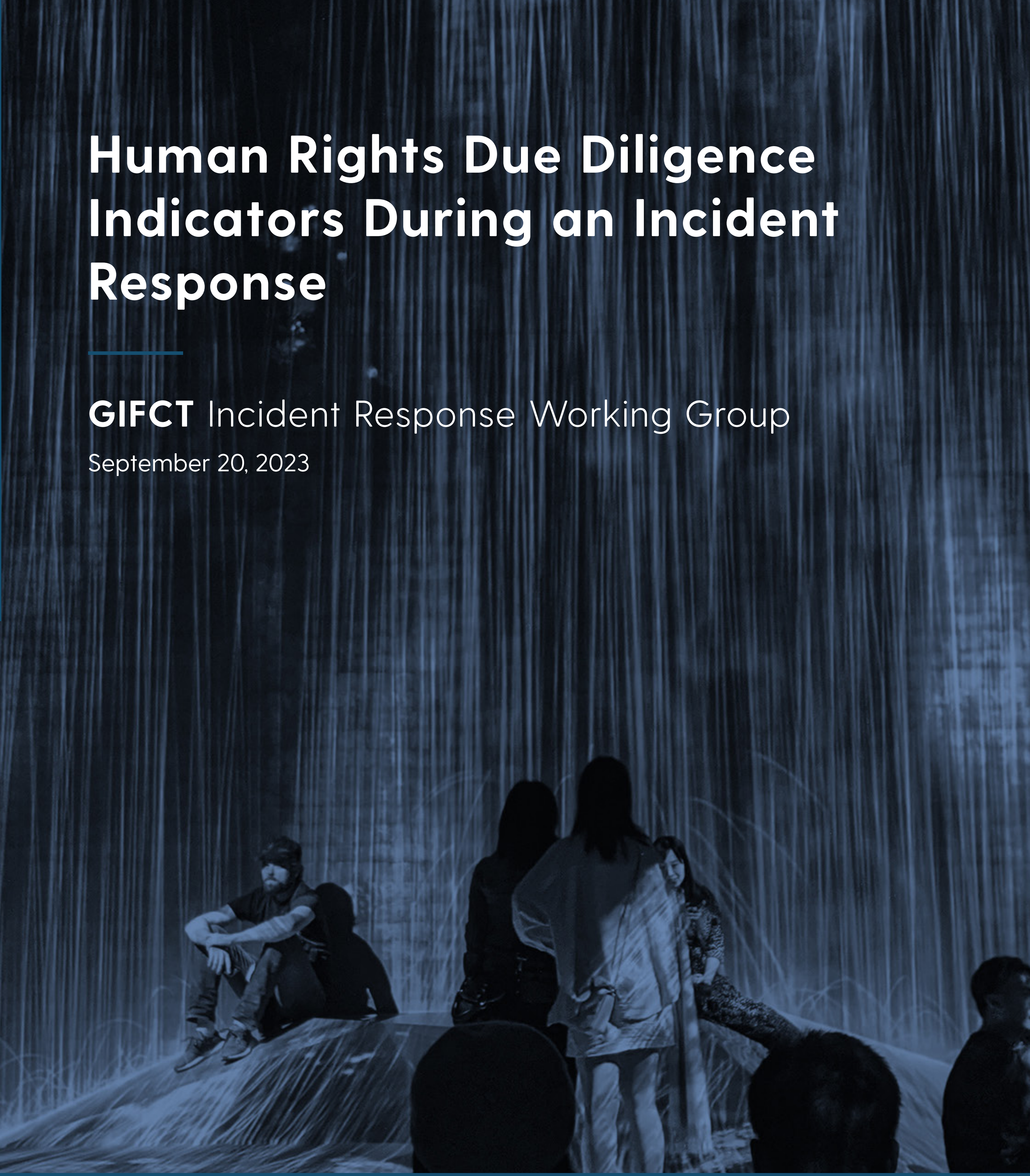
GIFCT YEAR 3 WORKING GROUP OUTPUT

# Human Rights Due Diligence Indicators During an Incident Response

## GIFCT Incident Response Working Group

September 20, 2023



GIFCT
Global Internet Forum
to Counter Terrorism

Ottavia Galuzzi
Christchurch Call Advisory Network

# About GIFCT Year 3 Working Group Outputs

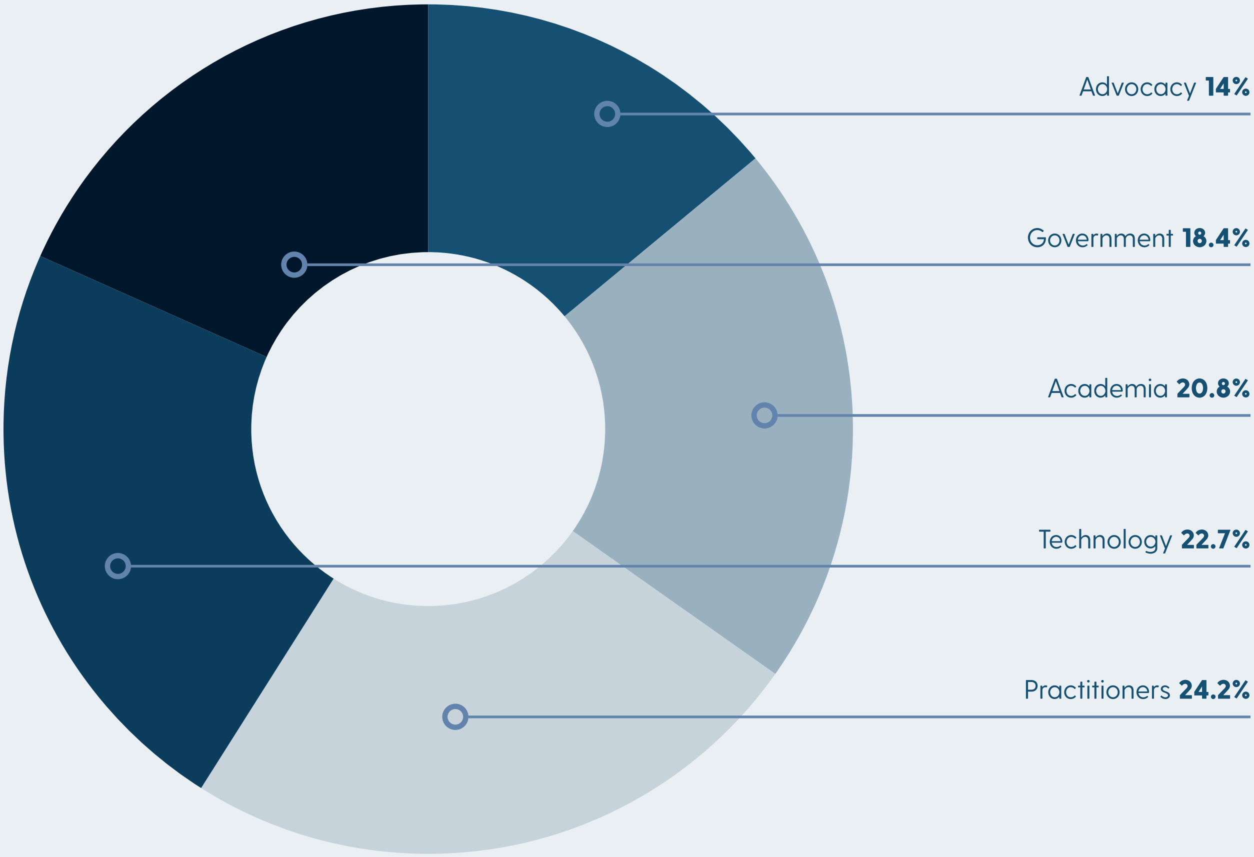
**By Dr. Nagham El Karhili,** Programming and Partnerships Lead, GIFCT

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**.

# Human Rights Due Diligence Indicators During an Incident Response[1]

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.

## Executive Summary

This section of the GIFCT Incident Response handbook aims to explore the potential human rights impacts during various stages of incident response and how these impacts can be measured. In particular, this section builds upon the insights of last year's GIFCT Incident Response Working Group (IRWG) and aims to contribute to the Human Rights Matrix[2] 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.

Overall, 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 undertaking for smaller or less-resourced companies that would benefit from external support and mentorship programs,[3] adopting and implementing a human rights impact assessment is attainable for

[1] Opinions presented are that of the author's not of the organization the author is affiliated with.

[2] 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.

[3] For example, the Tech Against Terrorism Mentorship Program.

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 section in order to present a comprehensive view of the stakeholders involved[4] 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.[5] As stated in GIFCT's Human Rights Impact Assessment (HRIA)[6] and Human Rights Policy,[7] It is crucial to pursue a human-centric approach and embed human rights into strategies rather than considering human rights as extra aspects that additionally 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 protecting 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)[8] and Content Incident Protocol (CIP).[9]

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

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

[5] See GIFCT's website.

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

[7] See GIFCT Human Rights Policy.

[8] 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."

[9] 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."

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 & 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 ensure due process and protection or respect for human rights."[10] 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.[11] 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,"[12] 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.[13] Mechanisms like human rights impact assessments (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

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

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

[12] Badii, "Human Rights Lifecycle of a Terrorist Incident Online."

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

in such processes or those that reveal their HRDD rarely disclose their full outcomes, like publishing a HRIA.[14] 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.[15] 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.

## 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, it becomes complicated to measure only the incident response impact on human rights. 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.[16] 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[17] that may have different names depending on companies and may not be applicable to all kinds of services, and 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.[18]

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

[14] 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).

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

[16] Badii, "Human Rights Lifecycle of a Terrorist Incident Online."

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

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

- 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
- 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[19] 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.[20]

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

[20] 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.

## 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.

## 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 relevant narrative

- 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.[21] 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 stakeholders consulted, etc.).[22]

## 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 people and 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

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

[22] Hatch et al., "Five Steps to Good Sustainability Reporting."

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.[23]

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 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.

[23] Hatch et al., "Five Steps to Good Sustainability Reporting."

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

| Stage | Goal | KPIs / KPNs | Qualitative | Quantitative (# of) |
|---|---|---|---|---|
| Horizon[24] | 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;<br>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. |
| 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;<br>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;<br>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. |

[24] 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.

## 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.[25]

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.[26] 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.[27]

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[28] and pursue a defined set of principles of good reporting,[29] 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.

[25] UNHCHR, "Guiding Principles on Business and Human Rights."."

[26] UNHCHR, "Guiding Principles on Business and Human Rights."."

[27] UNHCHR, "Guiding Principles on Business and Human Rights"

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

[29] 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.[30] 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 through informal channels 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.

[30] Different governments present differences in their judicial processes.

## Recommendations

- 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 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.
- 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.
- Ensure the inclusion of a HRIA covering each stage of the incident response, including the debriefing process.
    - » 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.
- 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.
- 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.
- Actively listen to the needs of communities whose human rights may be adversely impacted by incident response stages.
    - » 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).

Recommended citation: Ottavia Galuzzi, Human Rights Due Diligence Indicators 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**