The Tech Against Terrorism Guidelines

–

Government Transparency Reporting on Online Counterterrorism Efforts

Tech Against Terrorism – An initiative supported by UN CTED
contact@techagainstterrorism.org

Tech Against Terrorism Guidelines
Government transparency reporting on online counterterrorism efforts

INTRODUCTION

Transparency is vital to ensure accountability towards the public and internet users. Transparency reporting provides insight on to what extent human rights and fundamental freedoms such as freedom of expression and the right to privacy are respected across the internet. It can also encourage and recognise meaningful action in tackling terrorist use of the internet and provide crucial insight on this threat. Transparency should therefore be considered a key aspect of counterterrorism online, and transparency reporting has been a core part of Tech Against Terrorism's support for the tech sector since 2017, including in the Mentorship Programme.

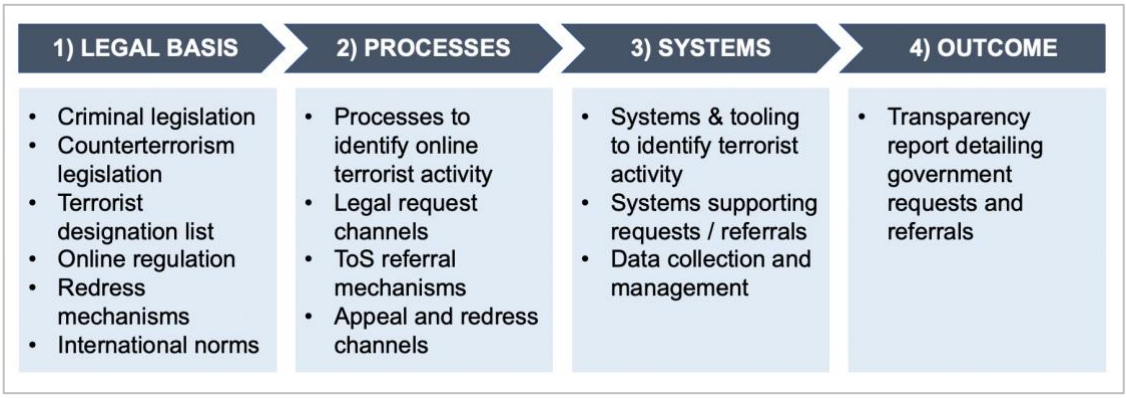
Whilst efforts to tackle terrorist use of the internet have largely been led by tech companies, governments are increasingly active in shaping counterterrorism online. Some governments submit legal orders requesting that companies remove material based on local laws. Other governments have established law enforcement teams, often dubbed Internet Referral Units (IRUs), that monitor the web for terrorist content and refer it to platforms for examination against tech companies' own policies. Further, several governments and intergovernmental organisations have set up voluntary collaboration frameworks aimed at tackling violent extremism and terrorist activity online.

Regrettably, meaningful government and law enforcement transparency on their online counterterrorism efforts is limited. Whilst some IRUs publish transparency reports,[1] most governments do not, and most data on government legal requests can only be found in tech company transparency reports.[2] To be accountable to their citizens and internet users globally, governments should be transparent about their efforts to counter terrorist use of the internet. Such reports could also prove valuable in providing insight on the evolution of terrorist use of the internet.

To that end we are **introducing the Tech Against Terrorism Guidelines for government transparency reporting on online counterterrorism efforts**. These Guidelines are – just like our Guidelines for tech companies – designed to drive increased transparency around a small set of core principles and key metrics to improve overall transparency from governments. The aim is to encourage governments to be more transparent and accountable towards their citizens and to allow for civil society oversight of government online counterterrorism activities.

[1] However, IRUs transparency towards users is often lacking, as it is often not clear whether content is removed as a result of IRU referrals.

[2] Company transparency has increased over the past decade, with a large number of companies reporting on their moderation efforts and the larger platforms are producing detailed reports.



Transparency is a process in which transparency reporting is an outcome. For governments, this process starts with establishing a clear legal basis for government activity in encouraging or compelling tech sector action against terrorist use of the internet, and continues with building engagement mechanisms and channels with tech companies.

We acknowledge the legitimate security concerns that governments might have with being transparent about their online counterterrorism efforts. Whilst an adversarial shift from terrorist actors is a legitimate concern, the Guidelines do not, in our view, ask for any granular details that might cause this. Furthermore, whilst we ask companies to also report on government requests, we believe that it is important that governments recognise the value of proactive transparency.

We ask governments to report on the legal basis and processes that underpin their legal orders, requests, referrals, and other engagement with the tech sector in addition to quantitative metrics. We also ask governments to detail their commitments to international law. Given the transparency demands some governments have placed on tech companies, we hope that governments will be able to live up to similar standards and produce reports in line with our Guidelines.

OUR TRANSPARENCY GUIDELINES FOR GOVERNMENTS

We ask governments to report on a small number of key metrics that reflect the transparency process. In Part A we ask governments to clarify on which legal basis they request removal of content to platforms, including their commitment to international law. In Part B we ask governments to detail any processes and systems introduced to discover and refer content and activity to tech companies for actioning, and what channels they provide for redress. In Part C, we ask governments to report quantitatively on their activities with regard to requesting and encouraging actioning of terrorist activity online, and the legal basis (if any) on which they ground such requests.

Part A: Legal basis

Explain the legal basis for the activities undertaken by your government and its law enforcement agencies to discover, report and/or refer terrorist activity to tech companies, by detailing:

Definitions

1. Your country's definition of terrorism as defined in legislative frameworks
2. Your country's definition of terrorist content (if any) in legislative frameworks
3. Your country's terrorist designation list(s) (if any) and/or the inter-governmental designation lists to which your country adheres

Legal framework

4. Your country's legislative framework's provisions with regard to online terrorist activity
5. The legislative framework that enables government entities and law enforcement in your country to send requests ordering action against online terrorist activity

International law

6. The international treaties your country has signed and ratified

Part B: Process & Systems

Explain the processes and systems supporting your government's online counterterrorism efforts by detailing:

Discovery

7. The processes through which state actors discover terrorist activity online
8. The systems used by state actors to discover terrorist activity online (including automated tooling and software)

Orders and requests

9. The processes and systems state actors use to submit orders and requests to tech companies to action terrorist content and/or activity, or to demand user information, in accordance with specific legislation

ToS Referrals

10. The processes and systems through which state actors refer terrorist content and activity to tech companies for examination against tech company Terms of Service, including via Internet Referral Units (if applicable)

Due diligence

11. The review your government or state actors carry out before sharing orders, requests, and ToS referrals to tech companies

Record-keeping

12. The type of content and data your government and/or state actors store and record following discovery of terrorist activity online
13. The type of content and data your government and/or state actors store and record following requests and referrals to tech companies

Redress

14. The processes through which your government supports companies in facilitating redress for content or activity that was wrongfully removed as a result of a government request

Part C: Report

Provide data on your engagement with tech companies around online terrorist activity by detailing your government's:

Source of discovery

15. Proactive discovery via the processes outlined in Part B
16. Proactive discovery via the systems outlined in Part B
17. Reports from the public (if relevant)

Information on content and activity discovered

18. Total amount of content or activity discovered to violate the legislative framework mentioned in Part A, broken down by:
    - Company
    - Source of discovery
    - Type of violation
    - Terrorist group or actor (and designation status)

Orders and requests made to tech companies (in numbers)

19. Removal requests, broken down by:
    - Company
    - Type of violation
    - Terrorist group or actor
20. User information requests
    - Broken down by company

Referrals made outside legal channels (in numbers)

21. ToS Referrals
    - Broken down by company
    - Broken down by terrorist group or actor

Contestations

22. Appeals or other contestations made against government reports, segmented by
    - Company
    - Type of report (legal order or ToS referral)
    - Success rate for each of the above

Example report

The below is an example report based on what reporting in accordance with the Guidelines could look like. The Guidelines should act as a baseline, and it is up to each government entity to decide how much detail they wish to provide on each given data point.

A: Legal basis	B: Processes & Systems
Definitions: "In Country, terrorism is defined as a method – justified by political ideology – that includes violent acts aiming to kill, harm, or spread fear amongst civilians."	"The online investigations team in the Anti-Terrorism Unit monitors online terrorist activity and submit legal requests and ToS referrals to tech companies. The Unit has a dedicated review process in place to ensure that it does not encourage undue takedown of legal content, or that there is no risk

"In Country, terrorist content is defined as any content produced by designated terrorist groups." "There are 23 groups designated as terrorist in Country. You can find the full list of groups here [hyperlink]." "Country is signatory to the International Covenant on Civil and Political Rights (ICCPR)" **Legal framework:** "Use of the internet for terrorist purposes is criminalised in the Counterterrorism Act of 2005." "The 2016 Counterterrorism Statute authorises the Anti-Terrorism Unit to submit legal orders on behalf of Country's government"	associated with human rights before sharing with a tech company." "This team uses web scraping methods to collect content for analysis from online platforms. It also relies on public reports made via a dedicated reporting platform. A record is kept of all content discovered and reported to platforms." "The Anti-Terrorism Unit has a dedicated staff devoted to managing appeals, in case we erroneously request or refer content, or if tech companies disagree about a referral violating their ToS."

C: Moderation statistics					
			2019	2020	2021
Terrorist content* or activity discovery					
	Cases of terrorist content or activity discovered		350	450	550
		Breakdown by company			
		Breakdown by source of discovery			
		Breakdown by type of violation			
		Breakdown by terrorist group or actor			
Orders and requests					
	Removal requests		250	350	450
		Breakdown by company			
		Breakdown by type of violation			
		Breakdown by terrorist group or actor			
	User information requests		150	250	350
		Breakdown by company			
ToS referrals					
	Referrals		100	200	300
		Breakdown by company			

		Breakdown by terrorist group or actor			
Contestations					
	Total appeal notices received		10	20	30
		Breakdown by company concerned			
		Breakdown by type (legal requests or ToS referral)			
	% successful		90%	90%	90%
		Breakdown by company concerned			
		Breakdown by type (legal requests v ToS referral)			

Glossary of key terms	
"Processes"	Any process or workflow that your government has introduced to support the discovery, moderation, and/or statistics collection of terrorist content and/or activity. This includes but is not limited to specialised teams or capability introduced to support with the above aims.
"Systems"	Any systems or tooling, such as automated data-driven tools, that your government uses that supports discovery, reports and/or statistics collection of terrorist content and/or activity.
"Removal orders and/or requests"	Request to moderate content activity submitted by a government-affiliated body or law enforcement agency via appropriate legal channels, including court orders or other clearly legally defined channels, that references content illegality under specified legal framework.
"Government and law enforcement ToS referral"	Flagging of content or activity made by a government-affiliated body or law enforcement agency, sometimes via extra-legal channels, for companies to examine against their own policies and content standards.
"Proactive discovery"	Activities you undertake on your own initiative to discover and surface terrorist content or activity on their platforms. This includes but is not limited to the use automated tooling.

ANNEX

Annex 1. Existing government and law enforcement transparency reporting

The current existing transparency initiatives from governments and law enforcement agencies include reports from Australia's e-Safety Commissioner and the EU Internet Referral Unit (EU IRU). The United Kingdom has stated in its Online Harms White Paper that it plans to start producing transparency reports, however the first report has not yet been published and there is currently limited information available on progress of this initiative.

Both Australia's e-Safety Commissioner and the EU IRU have been publishing annual reports since the year covering for 2015-2016 and include information on their major activities and projects, as well as on performance reports related to assistance and investigations. These two reports, though varying in the content and metrics included, are extremely limited compared with most tech company transparency reports.

EU IRU 2019 Report

> "This report gives an account of the EU IRU's major activities in 2019, and more specifically, it sheds light on both the prevention activities and the investigative support the EU IRU provided upon request of EU Member States."

Australia e-Safety Commissioner 2019-20 Report

> "This report provides information about the Australian Communications and Media Authority (ACMA)'s and eSafety's performance for 2019–20, key corporate information, and details against the mandatory reporting requirements. As a primary mechanism of accountability to the Parliament of Australia, this report has been prepared in line with the requirements for annual reports for entities under the Public Governance, Performance and Accountability Act 2013."

Report Characteristics

Reports	Gov / LEA	
	EU IRU	Australia's e-Safety Commissioner
Annual or bi-annual report	annual	annual
Latest report published	2019	2019 - 2020
Date first published	2015 - 2016	2015 - 2016

Report Metrics

The latest EU IRU report includes information on context such as the mandate of the EU IRU and its legal framework, as well as on referrals, terrorist propaganda and monitoring analysis, support to Member States' investigations, and ongoing projects and outreach activities.

In terms of metrics included in the report, the EU IRU specifically reports on referrals, categorised by terrorist/violent extremist content as well as content promoting illegal immigration services. It additionally provides information on the amount of Referral Action Days (RADs)[3] organised, and how many items were referred to Online Service Providers[4] (OSP) during the RADs. Finally, the EU IRU reports on its operational support to Member States investigations. These reporting methods and metrics are outlined below:

- "Volume of assessed content in 2019", divided by:
  - Terrorist/Violent Extremist content
  - Content promoting illegal immigration services

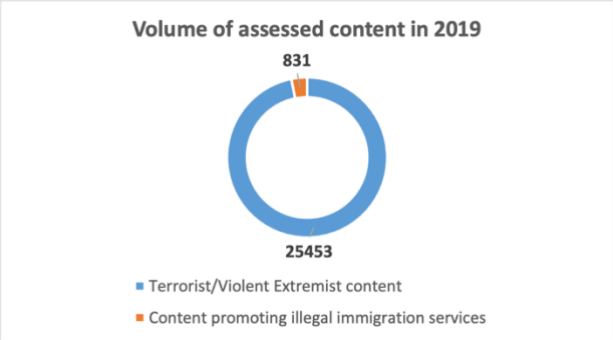


Figure 1: The EU IRU's "volume of assessed content in 2019"

- Support to Member State Investigations, the number of:
  - Intelligence notifications
  - Cross-match reports
  - Intelligence packages
  - Preliminary forensic reports
  - Provision of expertise

2019 Operational Support	
Intelligence notifications	6
Cross-match Reports	1
Intelligence Packages	314
Preliminary Forensic Reports	9
Provision of Expertise	94
TOTAL	**424**

Figure 2: EU IRU's Operational Support to Member States

[3] According to the report, during the targeted RADs there is a swift exchange of best practices between MS IRUs and OSPs with the aim of enhancing the referral process and improving critical elements such as feedback and response times.
[4] The working definition of "Online Service Provider" (OSP) used in the report is any company providing online services to EU citizens.

Tech Against Terrorism Guidelines
Government transparency reporting on online counterterrorism efforts

The Australian e-Safety Commissioner releases annual reports including evidence on performance, key corporate information, and details against the mandatory reporting requirements. The latest report includes the number of investigations conducted into "potentially prohibited online content", the number of URLs hosting material identified as likely to be prohibited, and the amount of notices to overseas services conducted in relation to abhorrent violent material.

A large percentage of the metrics found in the report focus on the demographics of the users reporting the material, as well as categorising the material within the given prohibition, and the location within Australia that the content has occurred. However, the metrics in the report that highlight the e-safety commissioner's actions, include the following:

Metrics related to (potentially) prohibited content or "sufficiently serious internet content"

- number of reports received about potentially prohibited online content
- number of URLS identified hosting material that was "sufficiently serious" to warrant referral to law enforcement, and the percentage of which these provided access to child sexual abuse material
- number of items of prohibited and potentially prohibited content that the e-safety commissioner finalised investigations into
- number of complaints about serious cyberbullying targeting Australian children, and the number of which was referred to the Kids Helpline website
- items actioned for "sufficiently serious internet content", (divided by content hosted in or provided from Australia or hosted from overseas)
    - "refused classification content for offensive depictions/descriptions of children"
    - "refused classification content for instruction, incitement or promotion of crime or violence"
    - "refused classification for films that advocate terrorist acts"
- number of abhorrent violent material notices issued
- referrals made to key support services, including Kids Helpline
- number of days it took for a percentage of investigations into CSAM to be investigated, actioned, or notified to the International Association of Internet Hotlines (INHOPE) [5] network or the Australian Federal Police (AFP)[6]
- number of reports contributed through the INHOPE network

Metrics related to image-based abuse material

- the number of image-based abuse reports handled
- number of enquiries responded to about image-based abuse
- number of removal notices to website and hosting services providers given
- number of formal warnings issued to the person responsible for image-based abuse
- number of informal warnings issued to the person responsible for image-based abuse
- percentage of successful removal of image-based abuse material on request
- number of alerted social media services to accounts that were being misused to share or threaten to share intimate content, or to elicit intimate content from minors

[5] INHOPE is a network of 46 hotlines around the world dedicated to rapidly removing child sexual abuse material from the internet. According to the report, content referred to an INHOPE hotline is passed on to local police or service providers for follow-up action.
[6] The Australian Federal Police (AFP) is the national and principal federal law enforcement agency of the Australian Government.

Figure 3: Australia's e-Safety Commissioner "Our year at a glance" on fighting illegal content and abhorrent violent material

Actual or likely classification and description of online content	Online content hosted in or provided from Australia	Internet content items hosted overseas
RC 1(b) (Refused Classification content for offensive depictions/descriptions of children)	0	13,359
RC 1(c) (Refused Classification content for instruction, incitement or promotion of crime or violence)	0	16
RC9A (Refused Classification for films that advocate terrorist acts)	0	18
Total	**0**	**13,393**

Figure 4: items actioned by Australia's e-Safety Commissioner for "sufficiently serious internet content" in 2019-20

Proposed government transparency

<u>EU regulation on preventing the dissemination of terrorist content online</u>

In the EU's regulation on preventing the dissemination of terrorist content online, so-called competent authorities – bodies which will be empowered to request content removal from tech platforms in line with the regulation and to issue penalties for lack of compliance – will have to report on the following metrics:

- The number of removal orders issued, specifying
    - the number of removal orders subject to Article 4(1)
    - the number of removal orders scrutinised under Article 4
    - information on the implementation of those removal orders by the hosting service providers concerned, including the number of cases in which terrorist content was removed or access thereto was disabled and the number of cases in which terrorist content was not removed or access thereto was not disabled
- The number of decisions taken in accordance with (and tech company implementation of):
    - Article 5 of the regulation, which empowers competent authorities to instruct platforms to introduce "specific measures' to remove terrorist content
    - Article 6, which instructs platforms to preserve content that has been removed as a result of a removal order or of Article 5

- The number of cases in which removal orders and decisions taken in accordance with Article 5 were subject to administrative or judicial review proceedings and information on the outcome of the relevant proceedings
- The number of decisions imposing penalties on platforms, and a description of the type of penalty imposed

Annex 2. Tech Against Terrorism Pledge

Introduction

The increased exploitation of information and communication technologies for terrorist and violent extremist purposes raises new challenges related to countering terrorism whilst respecting human rights, in particular with regards to freedom of expression and privacy. In the context of preventing and countering terrorism and violent extremism, effective counter-measures and the protection of human rights are not conflicting goals, but complementary and mutually reinforcing.

Tech Against Terrorism has developed six guiding principles (the Tech Against Terrorism Pledge) which inform our approach and underpin our framework for engaging with the very smallest technology companies.[1] These reinforce the importance of addressing challenging content and will support small tech companies in articulating their commitment to human rights and diversity in transparent, accountable, and collaborative ways. Our pledge complements the Global Network Initiative (GNI)[2] Principles as it is specifically designed for smaller tech platforms.

The Tech Against Terrorism Pledge provides simple and accessible guidelines to help the very smallest companies understand the importance of tackling terrorist exploitation in a manner that respects human rights and freedom of speech. With our pledge, we want to ensure that small companies – who often do not have enough resources to familiarise themselves with the myriad of legal regimes and social contexts which may apply to their services – can contribute to a free internet. The pledge is a starting point from which companies can build their own appropriate systems and policies. Company commitments to the pledge should be understood as aspirations to be achieved as quickly and thoroughly as possible, consistent with available resources and scale.

Our pledge is based on the GNI Principles and internationally recognised norms as articulated in the Universal Declaration of Human Rights ("UDHR"), the International Covenant on Civil and Political Rights ("ICCPR"), the International Covenant on Economic, Social and Cultural Rights ("ICESCR"), UN Security Council resolutions and documents S/RES/1624 (2005), S/RES/2129 (2013), S/RES/2322 (2016), S/RES/2354 (2017) and S/2017/375, and the UN Guiding Principles on Business and Human Rights ("UN Guiding Principles"). These constitute crucial normative precepts to help technology companies tackle exploitation of their services in a manner that promotes and protects human rights.[3]

The Tech Against Terrorism Pledge

1. **Freedom of Expression**

"We respect the right to freedom of expression that should be enjoyed by our users and will take actions consistent with applicable law to protect it from unlawful or unnecessary restrictions."

Article 19 of the ICCPR provides that "*1. Everyone shall have the right to hold opinions without interference. 2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.*

The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals."

2. **Non-Discrimination and Diversity**

"We respect the right of our users to express diverse views and opinions, and commit to educating users regarding what content and expression is not permitted on our platforms through clear terms of service and their transparent and consistent application."

Article 24 of the ICCPR states that *"All persons are equal before the law and are entitled without any discrimination to the equal protection of the law."* Article 15 of the ICESCR recognises the rights of everyone to take part in cultural life.

3. **Privacy**

"We respect the privacy of all our users and will take actions consistent with applicable law to protect it from arbitrary or unlawful interference."

UNDHR Article 12 and ICCPR Article 17 states *"No one shall be subjected to arbitrary or unlawful interference with his privacy, family, home or correspondence, nor to unlawful attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks."*

4. **Transparency and Accountability**

"We appreciate the need to account for what content we deem impermissible on our platforms, how we address government requests related to content on our platforms, and how we make determinations about content. To this end, we value and strive for transparency regarding those policies and practices, especially with regard to how they may impact the above-mentioned human rights-principles."

Guiding Principle 21 articulates an expectation that companies will account for how they address human rights and the commentary further explains that this "*requires that business enterprises have in place policies and processes through which they can both know and show*

that they respect human rights in practice. Showing involves communication, providing a measure of transparency and accountability to individuals or groups who may be impacted and to other relevant stakeholders, including investors."

5. **Remedy**

"While we strive to apply content policies fairly and consistently, we recognise that resource limitations, cultural contexts, and other factors may result in decisions that unintentionally cause negative impacts. To address this eventuality, we commit to devising appropriate mechanisms to allow individuals impacted by our policies and practices to bring information to our attention."

Guiding Principle 20 states: *"To make it possible for grievances to be addressed early and remediated directly, business enterprises should establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted."*

6. **Collaboration**

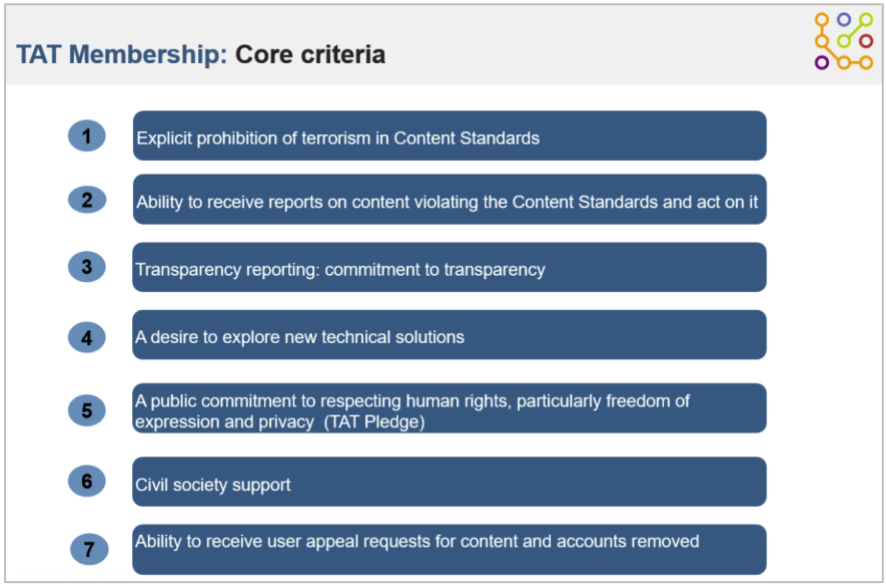
"We commit to work with partner organisations and enterprises to collaboratively develop strategies to keep our platforms and products safe from abuse by terrorist organisations and their supporters, and to promote tolerance, coexistence and diversity."

Article 19 of the ICCPR states that the exercise of freedom of expression carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order of public health or morals.

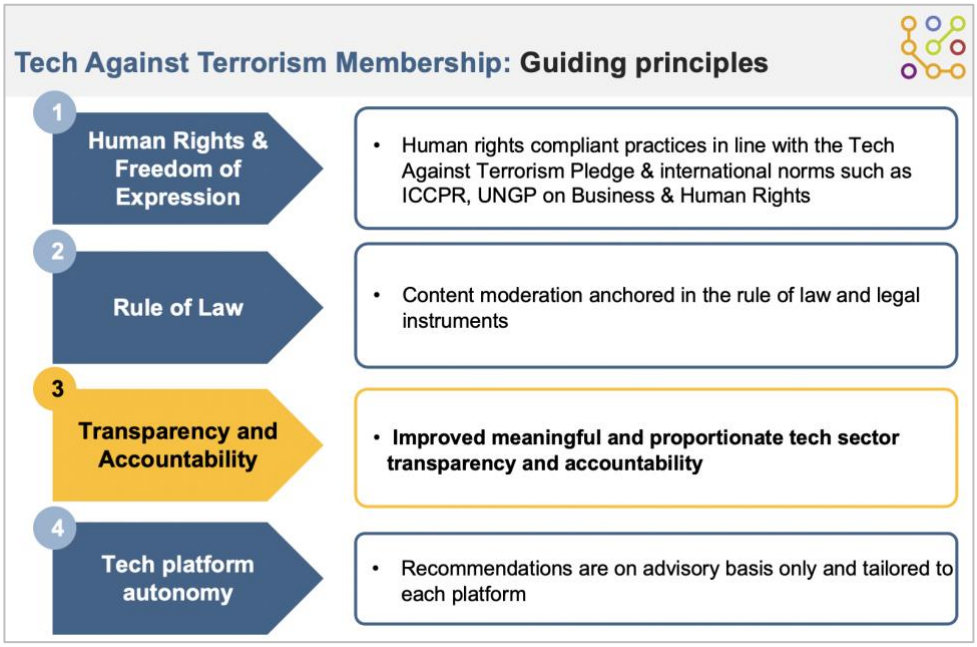
S/RES/1624 (2005) calls upon States to prohibit by law incitement to commit a terrorist act and S/RES/2354 (2017) condemns "*in the strongest terms the incitement of terrorist acts" and repudiates "attempts at the justification or glorification of terrorist acts that may incite further terrorist acts."*

S/RES/2354 (2017) further stresses the importance of the role of the business community *"in efforts to enhance dialogue and broaden understanding, and in promoting tolerance and coexistence, and in fostering an environment which is not conducive to incitement of terrorism, as well as in countering terrorist narratives."* It urges further development of initiatives to strengthen public-private partnerships in this area, and notes the benefits of engagement with a wide range of actors, including youth, families, women, community leaders, and other concerned groups of civil society.

Annex 3. Tech Against Terrorism Membership Criteria



Annex 4. Tech Against Terrorism Membership: Core Principles



Annex 5. Summary of Tech Against Terrorism Mentorship Programme