Schedule B: Manner and form of the required report.

This document relates to the non-periodic reporting notice given to **Telegram FZ LLC** under section 56(2) of the *Online Safety Act 2021 (Cth)* (the **Act**) (the **Notice)**

Instructions

Respond to the questions below, which relate to your implementation of the Basic Online Safety Expectations. The relevant expectations are stated alongside each question.

Each of the following is a **requirement** under the Notice:

- Answer each question **truthfully and accurately**. Other information provided to eSafety in relation to the Notice, including statements about a provider's capability to respond to the Notice and submissions related to confidentiality of information, must also be truthful and accurate.
- Use this template for your response. Responding in an alternative way without prior agreement with eSafety, including by altering this template, is likely to amount to non-compliance with the Notice.
- Provide a response to each question, and its parts, in Schedule B of the Notice. Leaving a response blank without first providing eSafety with a reasonable explanation is likely to be non-compliance with the Notice.
- Provide information for the period from **1 April 2023 to 29 February 2024** (the **Report Period)**, unless a question requires information on a specific time period within the Report Period. Some questions may require information 'at 29 February 2024'. The use of this date within the Report Period ensures that eSafety can obtain up to date information on providers' current practices, and aims to reduce burden on providers.
- You **must contact eSafety prior to the Notice deadline** if any of the following apply, and take the steps indicated:
    - You consider a question does not apply to your service(s). Provide reasons and an explanation of why the question does not apply.
    - You consider that you are not capable of answering a question. Identify the question(s) and provide reasons and an explanation of why you are not capable of answering the question(s). eSafety will consider your reasons and explanation and provide advice on an appropriate course of action.
    - You require additional time to comply with the Notice and to provide the required information. You must request an extension of time in which to provide the information, and provide specific reasons why you are not capable of meeting the existing deadline.
    - You would like to request an alternative means of returning Schedule B to the Notice other than via email. Provide reasons and suggest an alternative method of providing the information. eSafety will arrange appropriate alternate means where appropriate.
- **eSafety will engage with all providers who contact eSafety. Any instructions given by eSafety in response to any of these requests should be taken as a requirement under the Notice.**

Please also note the following:

- The size of text boxes should not be taken as an indication of any text limit. Include as much information as is necessary to provide a fulsome response to the question. The boxes should resize automatically as necessary.
- You can raise concerns about confidentiality and/or the sensitivity of information. eSafety will carefully consider your objections when deciding what information to publish. Attachment A has been provided to you to outline the criteria that eSafety will have regard to in considering confidentiality submissions. However, concerns about the confidentiality and sensitivity of certain information are not reasons to not provide information required by the Notice. eSafety will prepare a summary of information for publication. Submissions about confidentiality and sensitivity of information will be considered by eSafety when preparing that summary. Before eSafety publishes or publicly discloses that summary, you will be given an additional opportunity to make submissions about the information that eSafety proposes to publish.
- Unless otherwise specified, '**TVE**' refers to **terrorist and violent extremist material and activity**[1]. Some questions will ask about material only, or specific kinds of material (such as images). Questions should be answered in relation to TVE using Telegram's closest equivalent definitions in its terms of service, guidelines and policies.
- Where not specified, "**CSEA**" should be taken to refer to **child sexual exploitation and abuse**[2] **material and activity**. Some questions will ask about material only, or specific kinds of material (such as images). Questions about CSEA should be answered with reference to the definition in footnote 2.

[1] This may include but is not limited to material or activity that:

a. depicts or includes a 'terrorist act' as defined in section 100.1 of the Criminal Code (Cth) no matter where the action occurs, the threat of action is made, or where the action would occur if carried out;
b. depicts or includes advocating the doing of a 'terrorist act', e.g. 'pro-terror material', as defined in the Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material) Head Terms – Annexure A ;
c. depicts or includes promoting, inciting or instructing in matters of crime or violence with the intention of advancing a political, religious or ideological cause;
d. has the effect of – whether intentionally or unintentionally – promoting or glorifying material or activity that is underpinned by violent extremist or terrorist ideologies; or
e. promotes or celebrates terrorist leaders, organisations and groups, their actions or ideologies.

Not all material or activity that falls within these, or other, categories will constitute TVE. For example, see the defences that apply to the access of abhorrent violent material at section 474.37 of the Criminal Code, which includes defences for news reports, and scientific, medical, academic or historical research, amongst others.

[2] Child sexual exploitation and abuse material is Class 1 Material, as defined under the Act by reference to the National Classification Scheme and includes material that is both sexually exploitative and that depicts or describes child sexual abuse. CSEA material encompasses both 'child sexual exploitation material' (a broad category of content that encompasses material that sexualises and is exploitative to the child, but that does not necessarily show the child's sexual abuse) and 'child sexual abuse material' (which shows a sexual assault against a child, is a narrower category and can be considered a sub-set of child sexual exploitation material).

- Further information can also be found in the Basic Online Safety Expectations Regulatory Guidance[3] and on the relevant page on eSafety's website.[4]
- You can contact s 47E(d) @safety.gov.au regarding any of the above points, or with any other questions regarding your response.

Questions

1. s 47E(d)

s 47G(1)(a)

2. s 47E(d)

s 47G(1)(a)

3. s 47E(d)

[3] Office of the eSafety Commissioner, 'Basic Online Safety Expectation Regulatory Guidance', September 2023, accessed 8 February 2024, URL: https://www.esafety.gov.au/about-us/who-we-are/regulatory-schemes

[4] Office of the eSafety Commissioner, 'Basic Online Safety Expectations', accessed 8 February 2024, URL: https://www.esafety.gov.au/industry/basic-online-safety-expectations

s 4

s 47

s 47E(d)

s 47G(1)(a)

s 47E(d)

s 47G(1)(a)

s 47E(d)

s 47G(1)(a)

s 47G(1)(a)

4. s 47E(d)

(Reply to Question 4 follows.)

s
4

s 47G(1)(a)

5. s 47E(d)

s 47E(d)

s 47E(d)

s 47G(1)(a)

s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47E(d) s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

6. s 47E(d)

a. s 47E(d)

s 47E(d) s 47G(1)(a)

s 47E(d)

s 47G(1)(a)

s 47G(1)(b)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

7. s 47E(d)

a. s 47E(d)

s 47G(1)(a)

s 47E(d)

s 47G(1)(a)

s

4
7

s 47G(1)(a)

8. s 47E(d)

a. s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

9. s 47E(d)

a. s 47E(d)

s 4

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

10. s 47E(d)

a. s 47E(d)

s 47G(1)(a)

s 47G(1)(b)

s 47E(d)

b. s 47E(d)

s 47G(1)(a)

s 47G(1)(a)

s 47E(d)

c. s 47E(d)

s 47G(1)(a)

11. s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

12. s 47E(d) [redacted]

a. s 47E(d) [redacted]

s 47E(d) [redacted] s 47G(1)(a) [redacted]

b. s 47E(d) [redacted]

s 47E. [redacted]

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

s 47G(1)(a)

g. s 47E(d)

s 47G(1)(a)

h. s 47E(d)

s 47G(1)(a)

13. s 47E(d)

(Reply to Question 13 follows.)

s 47

s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

g. s 47E(d)

s 47G(1)(a)

h. s 47E(d)

s 47G(1)(a)

14. s 47E(d)

a. s 47E(d)

s 47E(d)	s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47

s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

g. s 47E(d)

s 47G(1)(a)

h. s 47E(d)

s 47G(1)(a)

15. s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

s 47

g. s 47E(d) [redacted]

s 47G(1)(a) [redacted]

16. s 47E(d) [redacted]

(Reply to Question 16 follows.)

s 4 [redacted]

s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

17. s 47E(d)

a. s 47E(d)

s 47E(d)	s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

s 47

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

18. s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

s 47

b. s 47E(d) [redacted]

s 47G(1)(a) [redacted]

c. s 47E(d) [redacted]

s 47G(1)(a) [redacted]

d. s 47E(d) [redacted]

s 47G(1)(a) [redacted]

e. s 47E(d) [redacted]

s 47G(1)(a) [redacted]

19. s 47E(d) [redacted]

a. s 47E(d) [redacted]

s 47E(d) [redacted]

s 47G(1)(a) [redacted]

s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

20. s 47E(d)

s 4

s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

21. s 47E(d)

a. s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

22. s 47E(d)

s 47E(d)

s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d s 47E(d)

s 47G(1)(a)

s 47E(d)

23. s 47E(d)

a. s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

24. s 47E(d)

s 47G(1)(a)

s 47G(1)(a)

25. s 47E(d)

s 47E(d)	s 47G(1)(a)

s 47E(d)

s 47G(1)(a)

26. s 47E(d)

a. s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

s 47G(1)(a)

d. s 47E(d) s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

27. s 47E(d)

a. s 47E(d) s 47G(1)(a)

b. s 47E(d)
s 47G(1)(a)

c. s 47E(d)
s 47G(1)(a)

d. s 47E(d)
s 47G(1)(a)

28. s 47E(d)

s 47G(1)(a)

29. s 47E(d)

a. s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c s 47E(d)

s 47G(1)(a)

30. s 47E(d)

a. s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

s 4

c. s 47E(d)

s 47G(1)(a)

31. s 47E(d)

a. s 47E(d) s 47G(1)(b)

b. s 47E(d)

s 47G(1)(a)

32. s 47E(d)

a. s 47E(d) s 47G(1)(a)

b. s 47E(d) s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

33. s 47E(d)

a. s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

s 4

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

34. s 47E(d) [redacted]

a. s 47E(d) [redacted] s 47G(1)(a) [redacted]

b. s 47E(d) [redacted]

s 47E(d) [redacted] s 47G(1)(a) [redacted]

35. s 47E(d) [redacted]

s 47E(d)

s 47G(1)(a)

36. s 47E(d)

a. s 47E(d) s 47G(1)(a)

s 47E(d)

b. s 47E(d) [redacted]

s 47G(1)(a) [redacted]

c. s 47E(d) [redacted]

s 47G(1)(a) [redacted]

d. s 47E(d) [redacted]

s 47G(1)(a) [redacted]

37. s 47E(d) [redacted]

s 47E(d)

s 47E(d) s 47G(1)(a)

s 47E(d)

38. s 47E(d)

a. s 47E(d)

s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

39. s 47E(d)

a. s 47E(d)

s 47G(1)(a)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

40. s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

s 47E(d)

c. s 47E(d)

s 47G(1)(a)

41. s 47E(d)

a. s 47E(d)

s 47E(d)	s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

42. s 47E(d)

s 47

s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

g. s 47E(d)

s 47G(1)(a)

h. s 47E(d)

s 47G(1)(a)

43. s 47E(d)

s 47

s 47E(d)

a. s 47E(d)

s 47E(d)	s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

g. s 47E(d)

s 47G(1)(a)

h. s 47E(d)

s 47G(1)(a)

44. s 47E(d)

(Reply to Question 44 follows.)

s 47E

s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

f. s 47E(d)

s 47G(1)(a)

45. s 47E(d)

(Reply to Question 45 follows.)

s 4

s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

46. s 47E(d)

a. s 47E(d)

s 47E(d)	s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

s 47

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

47. s 47E(d)

a. s 47E(d)

s 47E(d)

s 47G(1)(a)

b. s 47E(d)

s 47G(1)(b)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

s 47G(1)(a)

48. s 47E(d)

s 47E(d)

a. s 47E(d) s 47G(1)(a)

b. s 47E(d)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(a)

d. s 47E(d)

s 47G(1)(a)

e. s 47E(d)

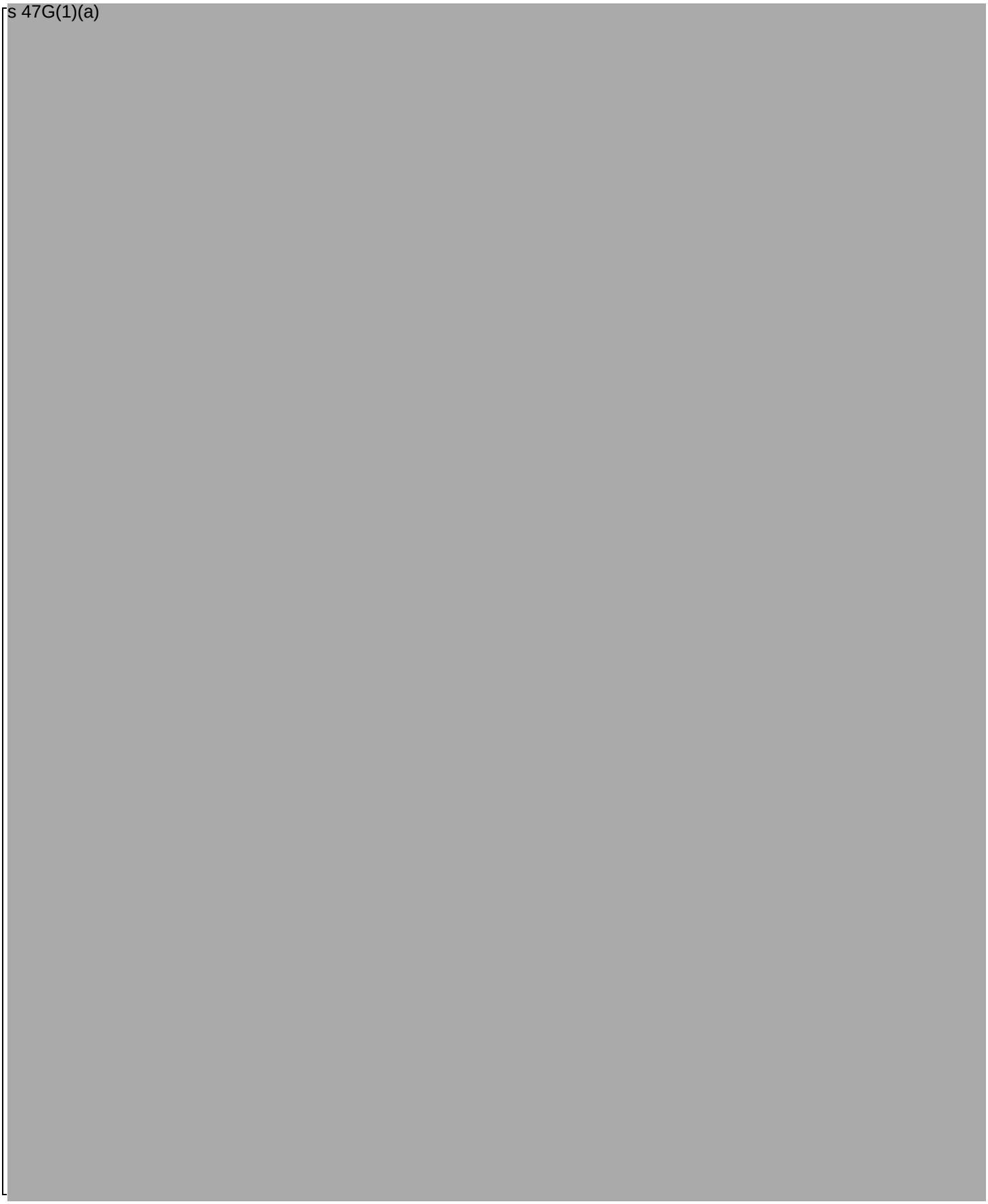
s 47G(1)(a)

f. s 47E(d)

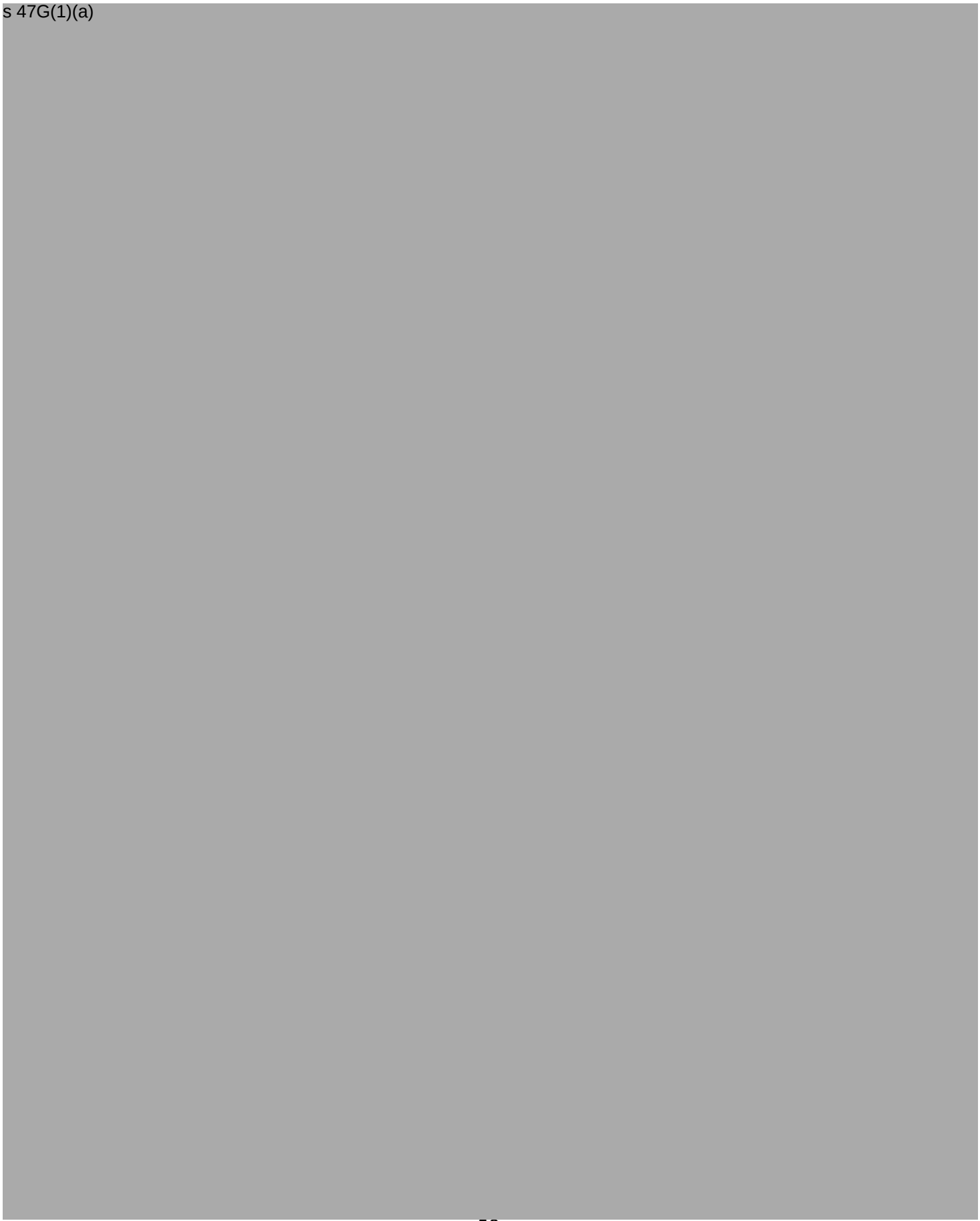
s 47G(1)(a)

49. s 47E(d)

s 47G(1)(a)



s 47G(1)(a)



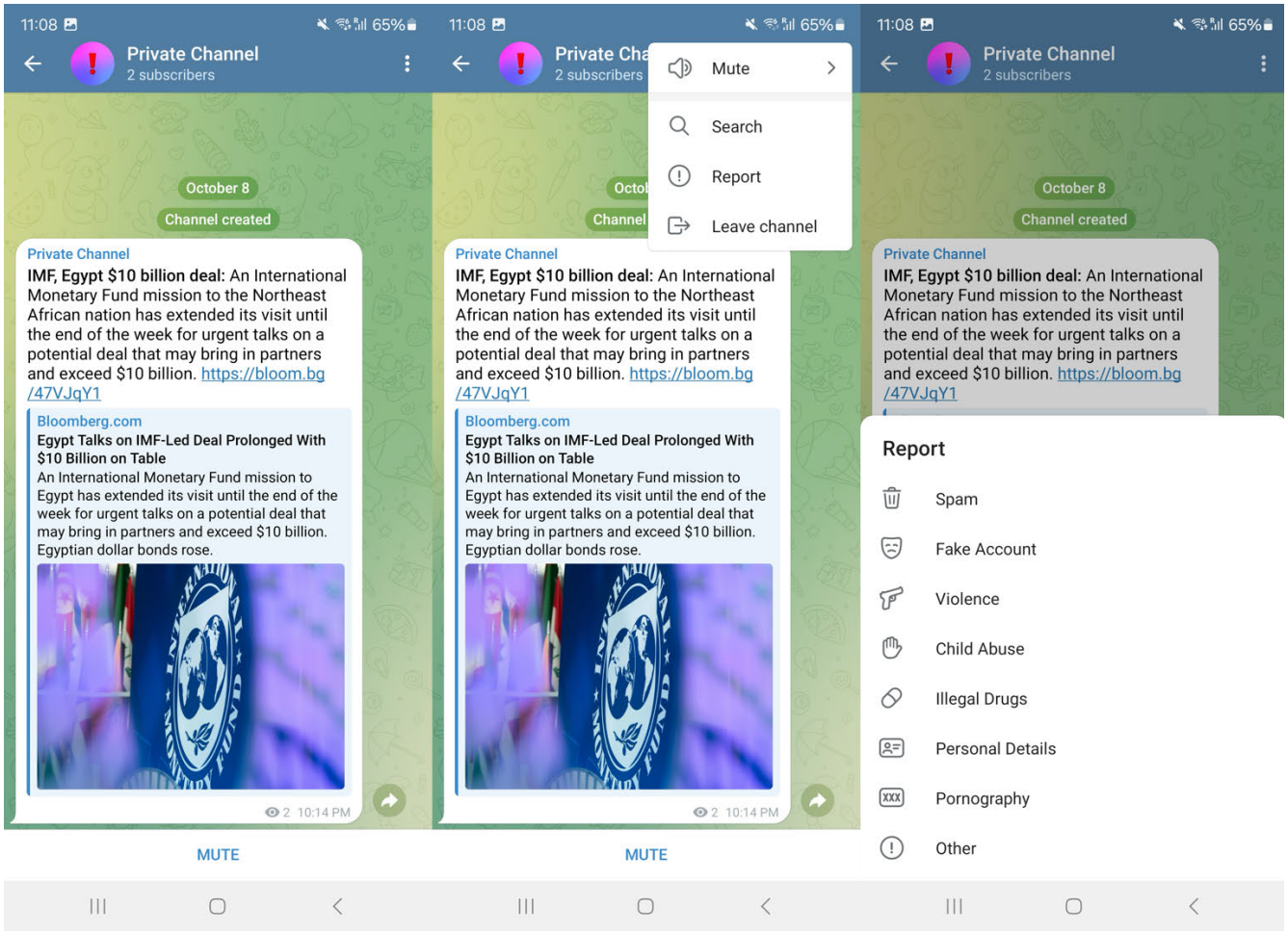
s 47G(1)(a)

[redacted]

Annex 01

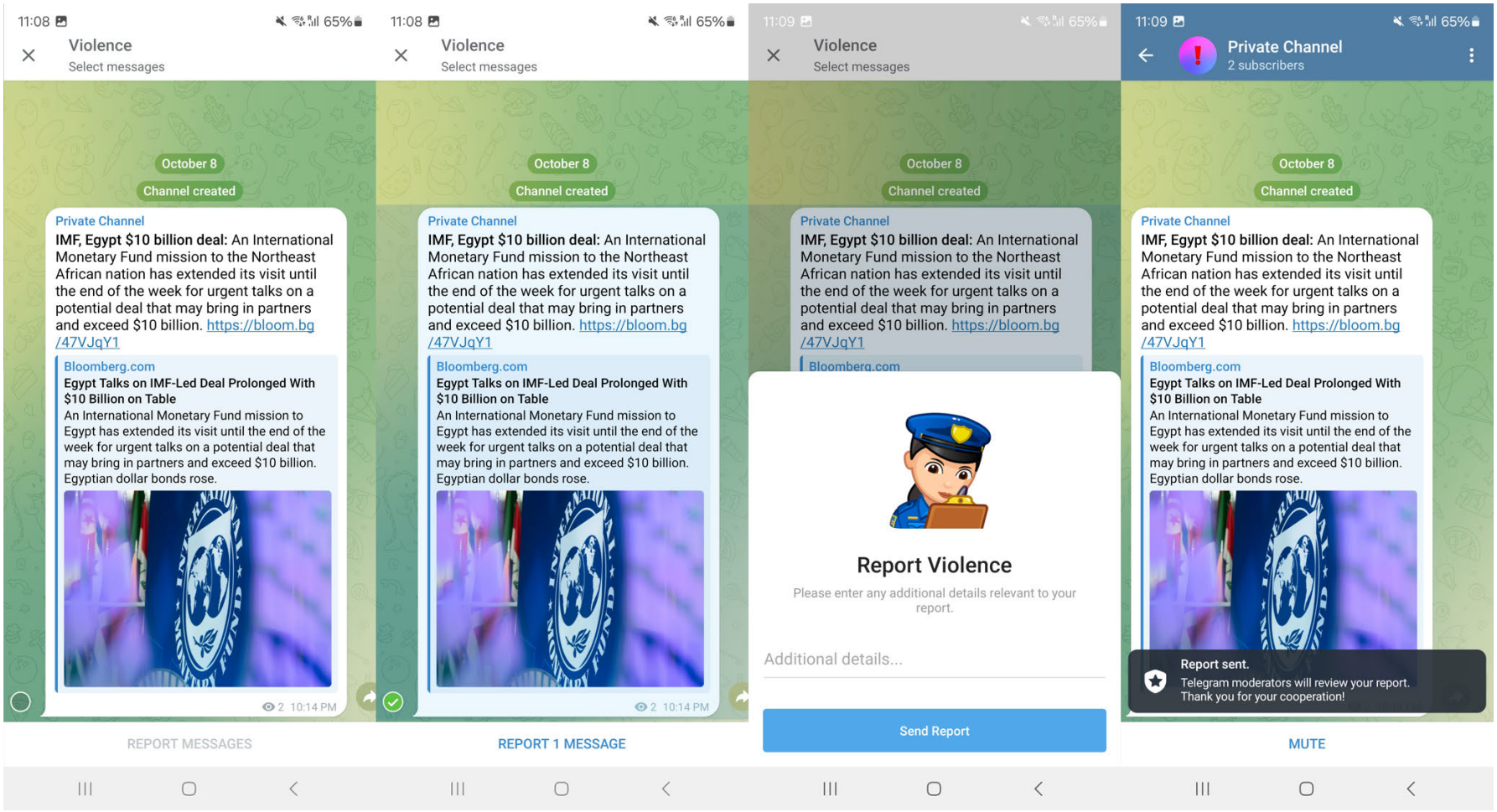
s 47E(d)

Private channel

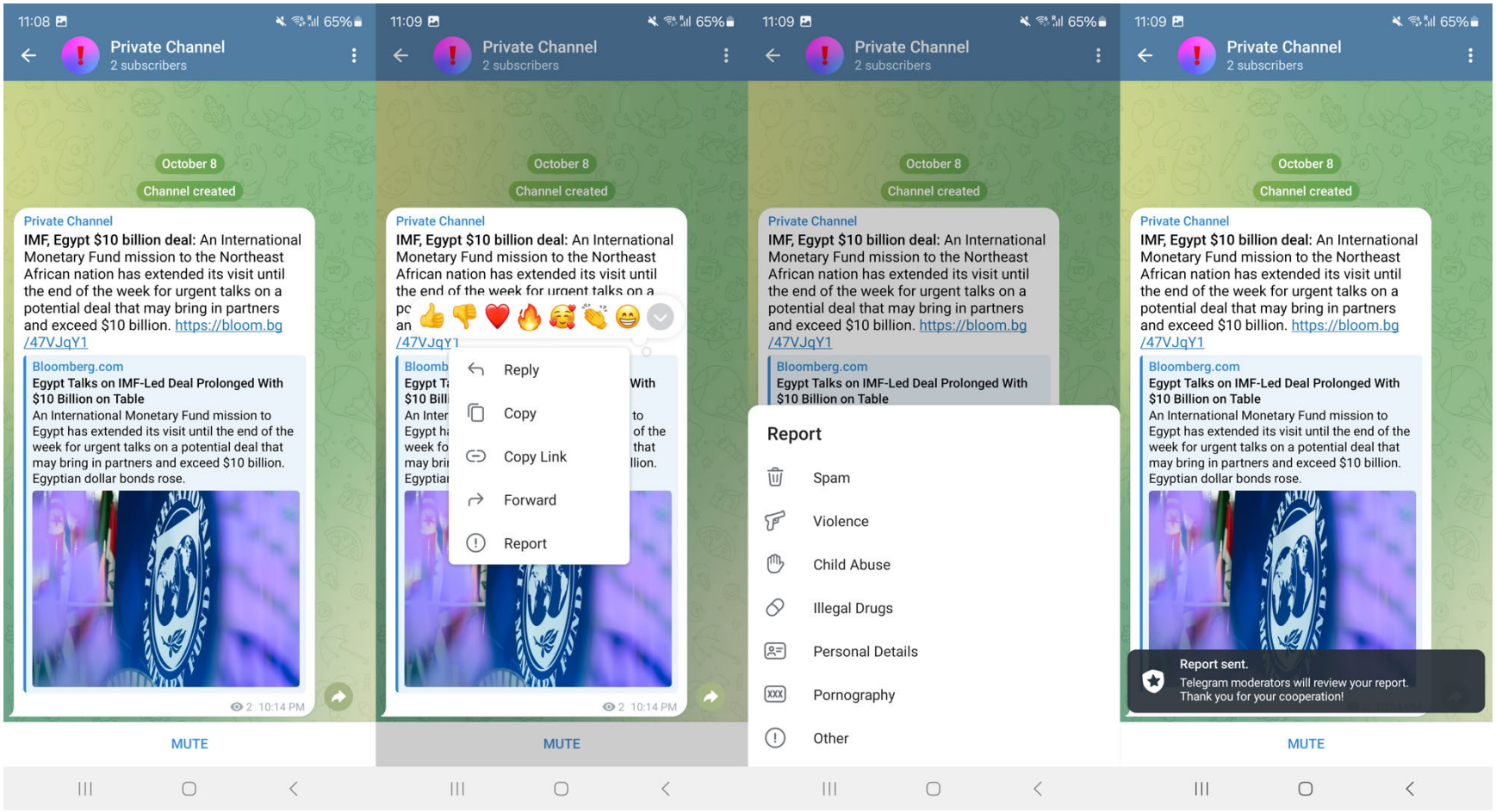


Private channel

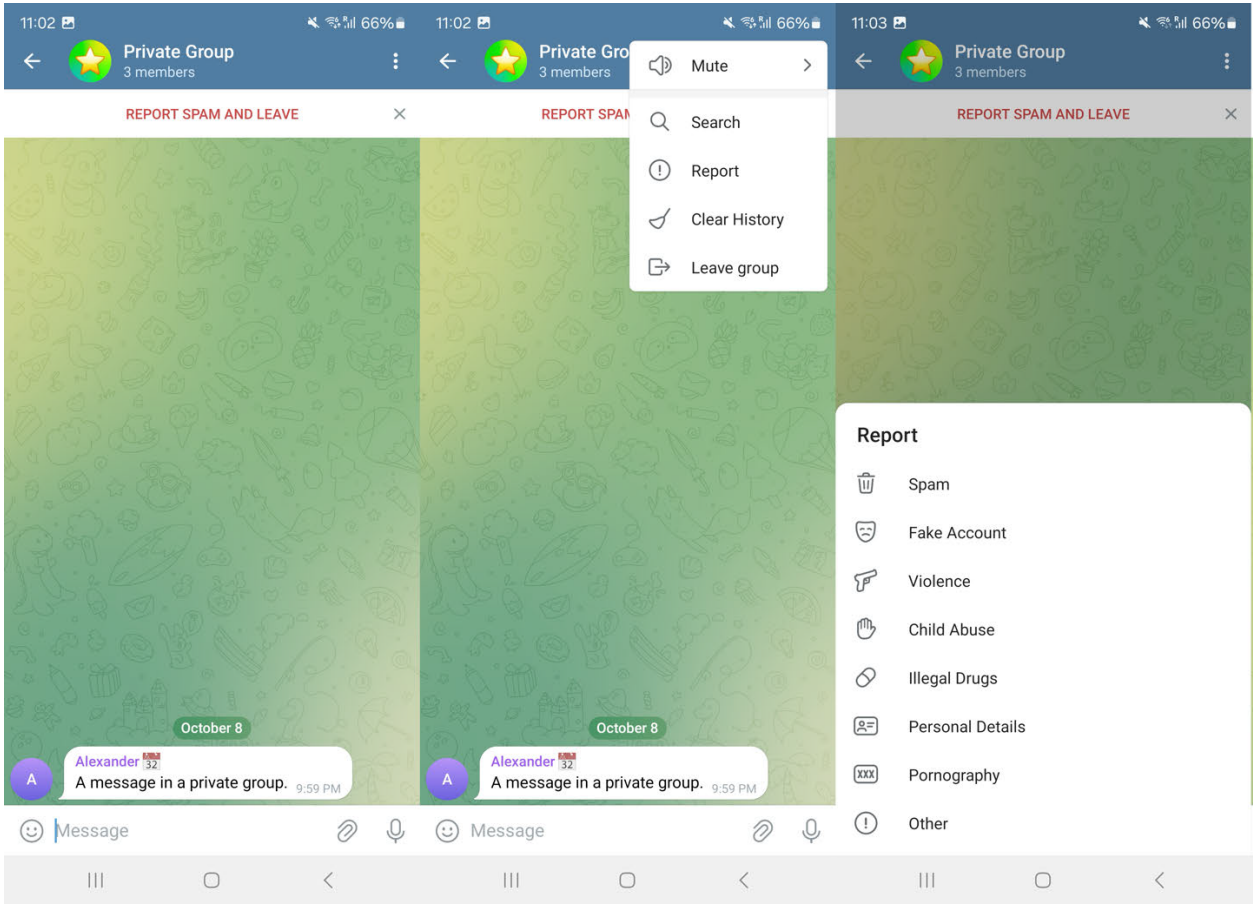
(continued)



Message in a private channel

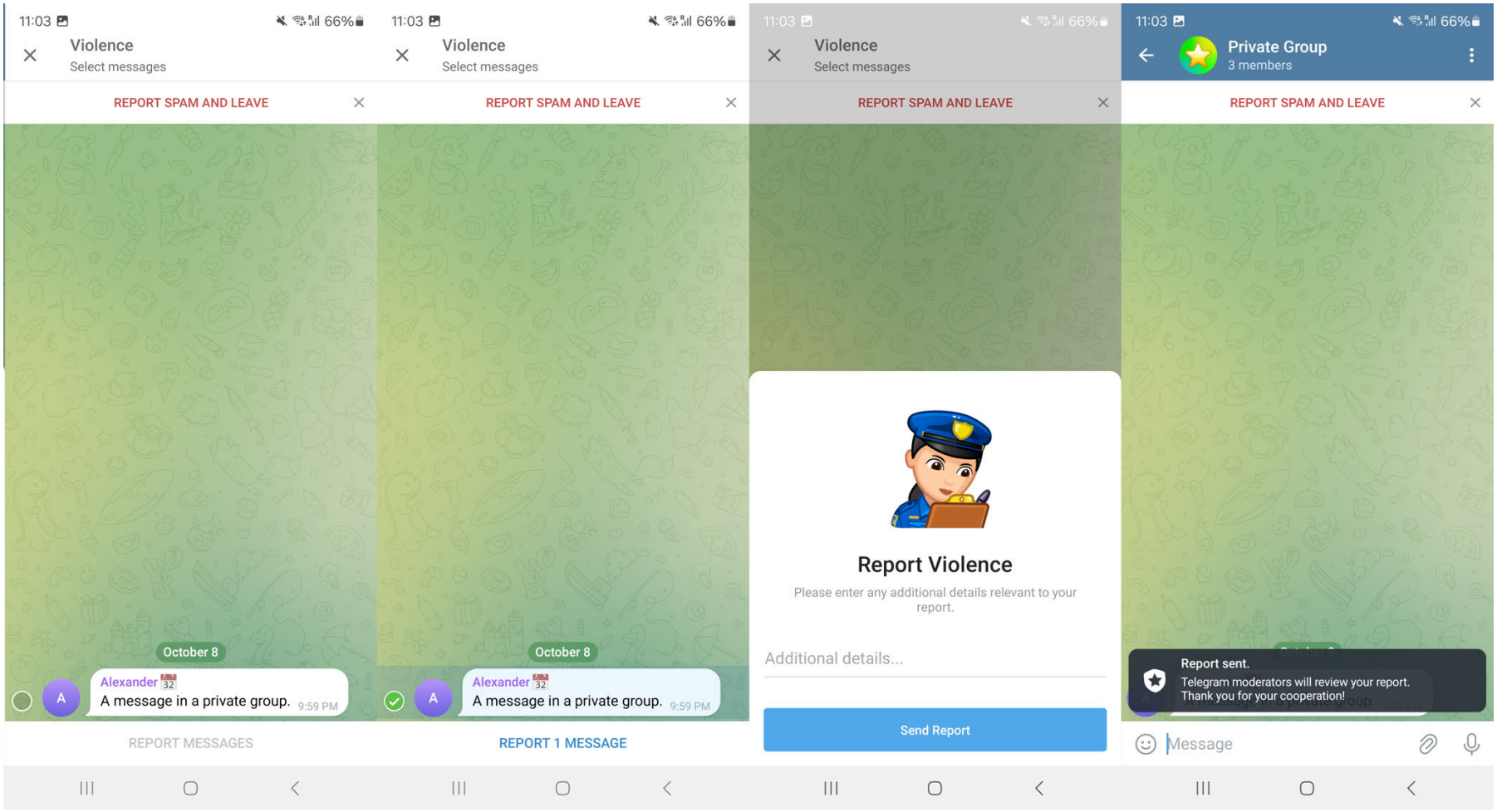


Private group

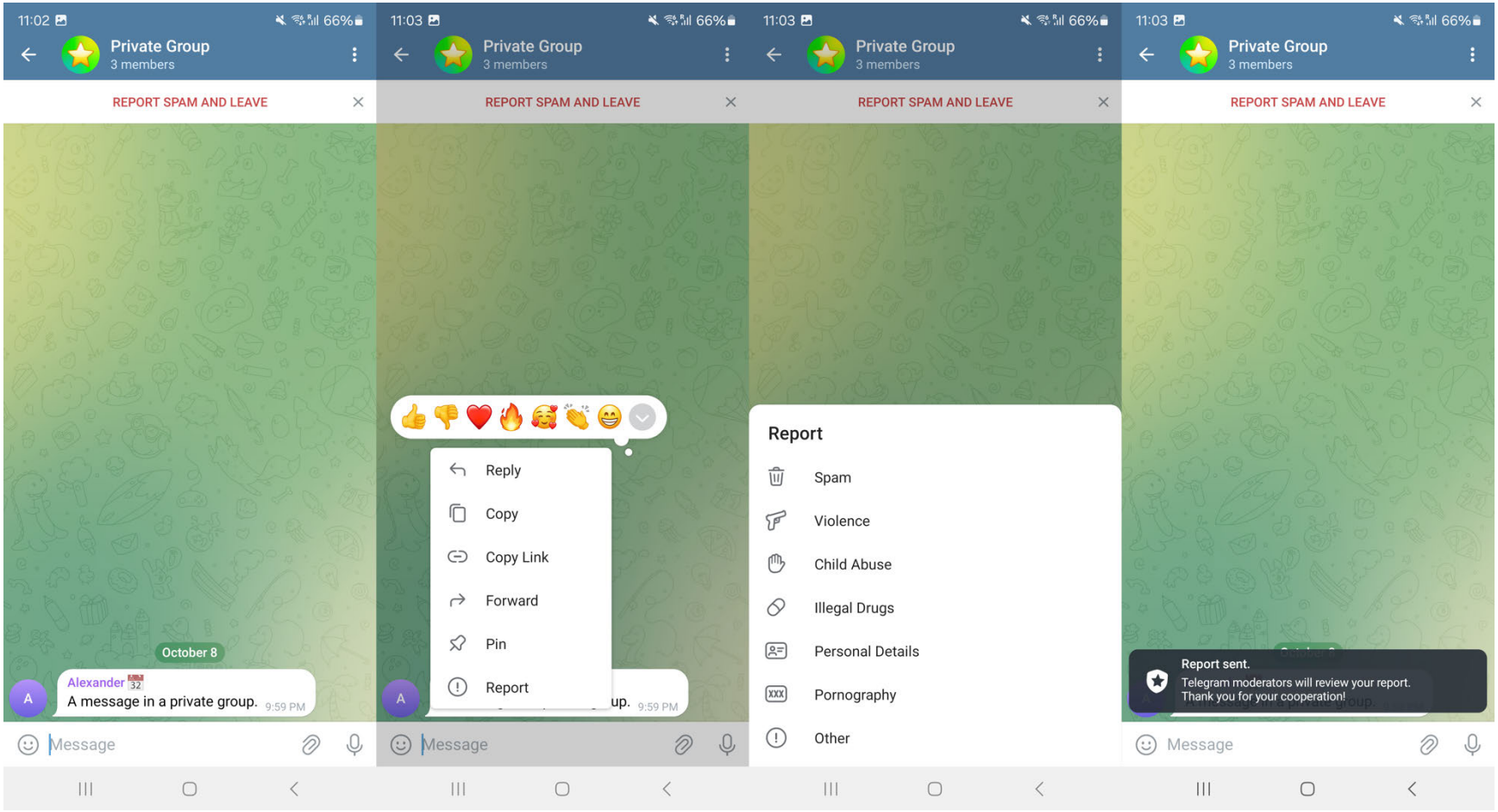


Private group

(continued)



Message in a private group



Attachment A
Transparency: Opportunity to Make Submissions

As set out in eSafety's regulatory guidance, eSafety considers that the transparency and accountability objectives of the Act in relation to the Basic Online Safety Expectations, and eSafety's broader statutory functions, will be met most effectively by making public the information received from industry in response to a reporting notice, where appropriate. eSafety therefore intends to publish on its website a summary of the information you provide in response to the notice given under s 56(2) of the *Online Safety Act 2021* (Cth) (**the Act**) (**the Notice**)**,** pursuant to the Commissioner's powers under the Act, including section 217. **Part of the purpose of obtaining the information through the Notice is disclosure.**

eSafety recognises however that some information may not be suitable for publication and invites you to make any submissions about the publication of the information provided in response to the Notice.

eSafety does not intend to publish information where it is satisfied that:

- the information falls into one of the categories in the table below; **and**
- the reasons provided establish that the harm that is identified outweighs the public interest in transparency and promoting the objectives of the Act and the functions of the Commissioner.

Any objections you make should be accompanied by specific reasons that explain any real, significant or material possibility of substantial and adverse consequences likely to follow publication, with reference to specific material in your response to the Notice. Please provide detailed reasons to identify any issues of concern. It is not sufficient to simply assert that the publication would fall into a category below.

eSafety will carefully consider your submissions in line with the guidance in Table 1 below. In determining what is appropriate to publish, eSafety will take into account the objectives of the Act and relevant functions of the eSafety Commissioner pursuant to section 27 of the Act:

- promoting online safety for Australians
- supporting and encouraging the implementation of measures to improve online safety for Australians
- the collection, analysis, interpretation and dissemination of information relating to online safety for Australians
- supporting, encouraging, conducting and evaluating research about online safety for Australians
- publishing reports and papers relating to online safety for Australians; and
- promoting compliance with the Act.

Please note that any submissions are voluntary and do not form part of your obligation to response to the Notice. Any submissions will not be published, unless required by another legal process. These submissions are solely for the purpose of eSafety's consideration of information it will publish.

Please provide any submissions to the same deadline as your response to the Notice.

Table 1: Categories of information that eSafety will consider not publishing.

Category of information	**Includes**	**Relevant Factors**
Commercial in confidence	Trade secrets. Information with commercial value, where that value would be diminished if the information were published.	Matters eSafety will consider include: - the extent to which information is already publicly known - measures taken to guard secrecy - the value of the information to its owner and competitors - the effort and money spent by the owner in developing the information - the ease or difficulty with which others might acquire or duplicate the information - the commercial harm that could occur from publication - other relevant information or submissions raised.
Other business information that would be unreasonable to publish	Information about an individual's business or professional affairs, or information about the business, commercial or financial affairs of an organisation or undertaking (business information) that would unreasonably affect that person adversely in respect of their lawful business or professional affairs or that organisation or undertaking in respect of its lawful business, commercial or financial affairs.	Matters eSafety will consider include: - whether the information is business information - how the publication of the information could have an unreasonable adverse impact on the individual or business.

Law enforcement and public safety	Information that could affect law enforcement or public safety, including disclosing methods or procedures for preventing, detecting, investigating, or dealing with matters arising out of, breaches or evasions of the law.
	
	Information that could assist individuals and groups from deliberately contravening or circumventing safety and security measures.
		Matters eSafety will consider include:
		- whether the methods or procedures are publicly known or prevalent across industry.
		- the level of detriment that is likely to occur from the disclosure of any lawful methods or procedures for investigating, preventing, detecting or dealing with breaches of the law.
		- how information could assist individuals in contravening company safety policies and interventions and the level of detriment that is likely to occur. Information that eSafety will not normally publish:
		  - Specific indicators, for example behavioural indicators (new account contacting multiple children) that might companies use, or technical indicators (e.g. device and IP addresses);
		  - Language/terms searched for;
		  - Detailed explanations or information on how technologies work and their weaknesses/vulnerabilities;
		  - "New technology" that is not currently in the public domain.

Personal information	Information about a natural person.	Matters eSafety will consider include: - whether the information is about an identified individual or an individual who is reasonably identifiable from the summary or other sources - whether the information in the summary could be de-identified so that is no longer about an identifiable individual or individual who is reasonably identifiable.

Table 2: Submission by provider

You are invited to provide submissions in the table below, via separate correspondence, or within your Schedule B response. Whichever method you choose, it will assist eSafety's consideration if you provide submissions in the following format:

1) Question # in Schedule B of the Notice;
2) Specific content in provider's response that submission relates to;
3) Which '**Category of information'** in Table 1 the submission relates to (you may identify more than one where relevant);

4) Specific reasons explaining any real, significant or material possibility of substantial and adverse consequences likely to follow publication (carefully consider '**relevant factors'** in Table 1).

Question number in Schedule B Notice	s 47G(1)(b)
3b	
8b	

9a	s 47G(1)(b)
10a, 10b	
12, 13, 14	
15, 16, 17	

	s 47G(1)(b)
18, 19, 20	
22, 23	

	s 47G(1)(b)
24, 25, 26c, 26e, 26f	
33c, 33f	
34b	
35	
36	

39a	s 47G(1)(b)
40a, 40b	
42, 43	
44, 45, 46	
47	
48b, 48f	
49	

s 47G(1)(b)

Follow- Up Questions to Telegram related to its non-periodic reporting notice

Background

On 18 March 2024, eSafety gave Telegram a non-periodic reporting notice under section 56(2) of the *Online Safety Act* (2021) (**the Notice**). Telegram was required to respond to the Notice with a report by 6 May 2024, but did not. As a result, Telegram did not comply with the Notice.

Since this time, Telegram has engaged with eSafety and, on 13 October 2024, provided a response to the questions originally asked in the Notice (**Telegram's Response**).

Follow-up Questions

The questions contained in this document relate to Telegram's Response.

These questions:

1) identify where certain information may not have been provided in answer to a question in the Notice;
2) seek to obtain or confirm information required by the Notice; and/or
3) seek to understand the reasons why Telegram is not capable of providing the information.

Please provide Telegram's response to the questions below by **17:00 Australian Eastern Daylight savings Time on 27 October 2024.**

eSafety is also available to discuss via a virtual meeting or call.

Question number in Schedule B of Notice	eSafety question for Telegram to clarify	Telegram response Provide the information sought by the Notice or explain why Telegram is not capable of providing the information. **Note**: Telegram's responses have been enumerated [TR 1..N] for your convenience.
All questions	s 47E(d)	[TR 1] s 47G(1)(a)

12, 13, 14, 16, 17, 42, 43, 45, 46	s 47E(d)	[TR 2] s 47G(1)(a)
12, 13, 15, 16, 17, 33, 42, 43, 44, 45, 46, 48	s 47E(d)	[TR 3] s 47G(1)(a)
5b	s 47E(d)	[TR 4] s 47G(1)(a)


s 47E(d)
s 47G(1)(a)

		s 47G(1)(a)
11	s 47E(d)	[TR 5] s 47G(1)(a)
12, 13, 14, 15, 16, 17, 42, 43, 44, 45, 46	s 47E(d)	[TR 6] s 47G(1)(a)
12a	s 47E(d)	[TR 7] s 47G(1)(a)


s 47G(1)(a)
13a
s 47E(d)
[TR 8]
s 47G(1)(a)
14a
s 47E(d)
[TR 9]
s 47G(1)(a)
15a
s 47E(d)
[TR 10]
s 47G(1)(a)

15b	s 47E(d)	[TR 11] s 47G(1)(a)
16a	s 47E(d)	[TR 12] s 47G(1)(a)


16c
s 47E(d)
[TR 13]
s 47G(1)(a)
17a
s 47E(d)
[TR 14]
s 47G(1)(a)


s 47E(d)
17c
s 47E(d)
[TR 15]
s 47G(1)(a)
17d
s 47E(d)
[TR 16]
s 47G(1)(a)

	s 47E(d)	
25	s 47E(d)	[TR 17] s 47G(1)(a)

26c	s 47E(d) [redacted]	[TR 18] s 47G(1)(a) [redacted]

26f	s 47E(d)	[TR 19] s 47G(1)(a)
33f	s 47E(d)	[TR 20] s 47G(1)(a)


s 47E(d)
s 47G(1)(a)
33b and 33f
s 47E(d)
[TR 21]
s 47G(1)(a)


s 47E(d)
38b
s 47E(d)
[TR 22]
s 47G(1)(a)
41a
s 47E(d)
[TR 23]
s 47G(1)(a)
42a
s 47E(d)
[TR 24]
s 47G(1)(a)


43a
s 47E(d)
[TR 25]
s 47G(1)(a)
44a
s 47E(d)
[TR 26]
s 47G(1)(a)
45a
s 47E(d)
[TR 27]
s 47G(1)(a)
45
s 47E(d)
[TR 28]
s 47G(1)(a)


s 47E(d)
46a
s 47E(d)
[TR 29]
s 47G(1)(a)
46d
s 47E(d)
[TR 30]
s 47G(1)(a)

	s 47E(d) 	
48	s 47E(d) 	[TR 31] s 47G(1)(a) 

	s 47E(d) 	

Attachment A
Transparency: Opportunity to Make Submissions

As set out in eSafety's regulatory guidance, eSafety considers that the transparency and accountability objectives of the Act in relation to the Basic Online Safety Expectations, and eSafety's broader statutory functions, will be met most effectively by making public the information received from industry in response to a reporting notice, where appropriate. eSafety therefore intends to publish on its website a summary of the information you provide in response to the notice given under s 56(2) of the *Online Safety Act 2021* (Cth) (**the Act**) (**the Notice**)**,** pursuant to the Commissioner's powers under the Act, including section 217. **Part of the purpose of obtaining the information through the Notice is disclosure.**

eSafety recognises however that some information may not be suitable for publication and invites you to make any submissions about the publication of the information provided in response to the Notice.

eSafety does not intend to publish information where it is satisfied that:

- the information falls into one of the categories in the table below; **and**
- the reasons provided establish that the harm that is identified outweighs the public interest in transparency and promoting the objectives of the Act and the functions of the Commissioner.

Any objections you make should be accompanied by specific reasons that explain any real, significant or material possibility of substantial and adverse consequences likely to follow publication, with reference to specific material in your response to the Notice. Please provide detailed reasons to identify any issues of concern. It is not sufficient to simply assert that the publication would fall into a category below.

eSafety will carefully consider your submissions in line with the guidance in Table 1 below. In determining what is appropriate to publish, eSafety will take into account the objectives of the Act and relevant functions of the eSafety Commissioner pursuant to section 27 of the Act:

- promoting online safety for Australians
- supporting and encouraging the implementation of measures to improve online safety for Australians
- the collection, analysis, interpretation and dissemination of information relating to online safety for Australians
- supporting, encouraging, conducting and evaluating research about online safety for Australians
- publishing reports and papers relating to online safety for Australians; and
- promoting compliance with the Act.

Please note that any submissions are voluntary and do not form part of your obligation to response to the Notice. Any submissions will not be published, unless required by another legal process. These submissions are solely for the purpose of eSafety's consideration of information it will publish.

Please provide any submissions to the same deadline as your response to the Notice.

Table 1: Categories of information that eSafety will consider not publishing.

Category of information	**Includes**	**Relevant Factors**
Commercial in confidence	Trade secrets. Information with commercial value, where that value would be diminished if the information were published.	Matters eSafety will consider include: - the extent to which information is already publicly known - measures taken to guard secrecy - the value of the information to its owner and competitors - the effort and money spent by the owner in developing the information - the ease or difficulty with which others might acquire or duplicate the information - the commercial harm that could occur from publication - other relevant information or submissions raised.
Other business information that would be unreasonable to publish	Information about an individual's business or professional affairs, or information about the business, commercial or financial affairs of an organisation or undertaking (business information) that would unreasonably affect that person adversely in respect of their lawful business or professional affairs or that organisation or undertaking in respect of its lawful business, commercial or financial affairs.	Matters eSafety will consider include: - whether the information is business information - how the publication of the information could have an unreasonable adverse impact on the individual or business.

Law enforcement and public safety	Information that could affect law enforcement or public safety, including disclosing methods or procedures for preventing, detecting, investigating, or dealing with matters arising out of, breaches or evasions of the law.
	
	Information that could assist individuals and groups from deliberately contravening or circumventing safety and security measures.
		Matters eSafety will consider include:
		- whether the methods or procedures are publicly known or prevalent across industry.
		- the level of detriment that is likely to occur from the disclosure of any lawful methods or procedures for investigating, preventing, detecting or dealing with breaches of the law.
		- how information could assist individuals in contravening company safety policies and interventions and the level of detriment that is likely to occur. Information that eSafety will not normally publish:
		  - Specific indicators, for example behavioural indicators (new account contacting multiple children) that might companies use, or technical indicators (e.g. device and IP addresses);
		  - Language/terms searched for;
		  - Detailed explanations or information on how technologies work and their weaknesses/vulnerabilities;
		  - "New technology" that is not currently in the public domain.

Personal information	Information about a natural person.	Matters eSafety will consider include: - whether the information is about an identified individual or an individual who is reasonably identifiable from the summary or other sources - whether the information in the summary could be de-identified so that is no longer about an identifiable individual or individual who is reasonably identifiable.

Table 2: Submission by provider

You are invited to provide submissions in the table below, via separate correspondence, or within your Schedule B response. Whichever method you choose, it will assist eSafety's consideration if you provide submissions in the following format:

1) Question # in Schedule B of the Notice;
2) Specific content in provider's response that submission relates to;
3) Which '**Category of information'** in Table 1 the submission relates to (you may identify more than one where relevant);

4) Specific reasons explaining any real, significant or material possibility of substantial and adverse consequences likely to follow publication (carefully consider '**relevant factors'** in Table 1).

Note: Telegram issued confidentiality submissions with respect to its initial answers in Schedule B. The below outlines which of its replies to the follow up request from eSafety fall within the scope of Telegram's previous confidentiality request. SR1..N refer to Telegram's Specific Reasons as defined in its original confidentiality submissions. TR1..N refer to Telegram's replies to eSafety's follow up request.			
Question number in Schedule B Notice	**Content in provider's response that submission relates to**	**Submission Category**	**Specific Reason**

s 47G(1)(a)

[redacted]


s 47G(1)(a)

Telegram summary

Overview

Telegram FZ LLC was asked about its Telegram service.

Part 1. Questions in relation to Terrorism and Violent Extremism (TVE)

1. Questions about Telegram's definitions of 'terrorist material and activity' and 'violent extremist material and activity'

A. Terrorist material and activity

In response to a question about how Telegram defines 'terrorist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, Telegram stated:

> For the purposes of its moderation procedures, Telegram is guided by the notion of "terrorist content" that comprises texts, imagery, recordings, and footage promoting and glorifying violence and terrorist ideology, soliciting funds for terrorist causes, instructing or advising on planning or carrying out of terror attacks.

B. Violent extremist material and activity

In response to a question about how Telegram defines 'violent extremist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, Telegram stated that, in its moderation procedures, it defines 'violent extremist content' as:

> texts, imagery, recordings, and footage advocating for violence against a person or a group (i.e., specific threats of physical harm, etc.), as well [sic] instructions for creating and obtaining weapons, explosives and other means of carrying out violent attacks.

Telegram also stated that in its moderation procedures, it defined 'violent content' as:

> [content which] also covers graphic, gruesome or shocking materials, like graphic details of torture—unless such content is clearly communicated for the purposes of news reporting or raising awareness of human rights violations.

2. Prohibiting 'illicit content' on private parts of Telegram

On 18 March 2024, when Telegram was issued the Notice, the Telegram Terms of Service stated:

> by signing up for Telegram, you accept our Privacy Policy and agree not to:
>
> - Use our service to send spam or scam users
> - Promote violence on ***publicly viewable*** [emphasis added] Telegram channels, bots, etc.
> - Post illegal pornographic content on ***publicly viewable*** [emphasis added] Telegram channels, bots, etc.[1]

eSafety highlighted this in the Notice, and asked Telegram to specify whether these rules permitted end-users to promote violence or post illegal pornographic content on private parts of the service – namely, Private Channels, Groups, and Secret Chats.

Telegram stated that this was not the case, and that the 'Telegram Terms of Service … apply throughout the app, regardless of chat type.' Telegram stated that it 'doesn't tolerate illicit content' in these parts of the service, and that it will 'take appropriate action within its technical capabilities whenever it becomes aware of such content'.

Telegram stated that the references quoted by eSafety in Telegram's terms of service to publicly viewable parts of the service refer to the areas of its service that Telegram moderators proactively monitored, but that is not to say that Telegram 'tolerate[d] illicit content in private channels, groups, or secret chats'.

Telegram provided the following information about how it detected violative material in the private parts of Telegram where moderators cannot proactively check messages:

---

[1] Telegram, 'Terms of Service', accessed 9 February 2024, URL: https://telegram.org/tos

- User reporting – Telegram stated that content in private Communities[2] can be reported by users. Telegram also stated that 'regular users, non-registered viewers, and organisations' can report material using in-service reporting options, or through dedicated email addresses.[3] Telegram stated that messages reported by end-users in private Communities (i.e., Channels and Groups) 'are forwarded to moderators', but messages reported in Secret Chats are not (see section 2B).

- Proactive detection measures – Telegram stated that it used 'algorithmic detection measures that prevent abuse' on its service. Telegram referred to its use of hash-matching to detect known TVE and CSEA material on public and private Communities, matched against previously identified TVE and CSEA on Telegram's public content. Telegram also referred to its use of 'automated content detection and manual search strategies tailored to locate Communities engaged in violations of Telegram Terms of Service'.

eSafety notes that responses captured in sections 5 and 9 highlight that there was inconsistent use of proactive detection tools across the 'private' parts of Telegram's service, s 47G(1)(b) Telegram also did not take any external hashes from external organisations which share hashes of terrorism and violent extremism.

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing TVE material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material has already been identified by other online service providers and hashed in extensive shared databases like those run by the GIFCT or Tech Against Terrorism.

A. Moderating 'private' Communities using 'invite links'

Telegram stated in circumstances where a 'private' Community is made accessible to the broader public via an 'invite link', it also changes Telegram's ability to monitor that Community. Telegram gave the example that if an 'invite link' to a

[2] Telegram used the term 'Communities' to refer generally to groups and channels on the service.
[3] Telegram provided the following e-mail addresses for reporting violative material on the service: abuse@telegram.org and StopCA@telegram.org.

private Community is shared on a public part of Telegram or another social media service, then

> the Community is considered to be public for content moderation purposes (thanks to the fact that moderators can follow the link and view the messages within before any user reports are made).

Telegram stated that it routinely detects such 'invite links' when the public channels or groups hosting them are taken down, and removes the associated Communities 'if appropriate based on the content they publish'.

B. Moderating E2EE 'Secret Chats'

Telegram added that it has 'no technical means of verifying the accuracy of user reports regarding content stored' inside Secret Chats because these messages are protected by E2EE. Telegram stated that messages in Secret Chats were not 'forwarded' to moderators when they were reported by an end-user.

Without access to the messages being reported, Telegram reported that it relies on alternative signals or indicators to determine if 'the reported user is not otherwise engaging in harmful or malicious behaviour'. eSafety has chosen not to publish these alternative signals to prevent them being misused.

eSafety notes that there are alternative measures that enable content moderators to review E2EE messages that have been reported by end-users as harmful or otherwise violative. For example, WhatsApp (which is E2EE) has processes in place that enable its moderators to receive the last 5 messages sent to an end-user from the account they are reporting.[4] eSafety considers that having measures in place that enable moderators to review the material being reported by end-users is key to ensuring that these reports can be responded to effectively.

C. Changes to Telegram's FAQs in September 2024

On 18 March 2024, when Telegram was issued the Notice, Telegram's 'frequently asked questions' web page stated:

> Q: There's illegal content on Telegram. How do I take it down?

[4] WhatsApp, 'About reporting and blocking someone on WhatsApp', accessed 15 October 2024, URL: https://faq.whatsapp.com/414631957536067/

> All Telegram chats and group chats are private amongst their participants. We do not process any requests related to them.
>
> But sticker sets, channels, and bots on Telegram are publicly available. If you find sticker sets or bots on Telegram that you think are illegal, please ping us at abuse@telegram.org.

eSafety pointed to this in the Notice and asked Telegram to specify the steps it was taking to comply with the Expectations in relation to the safe use of private chats and private group chats, including after a user report was made about illegal or harmful content, given this statement .Telegram responded by referring to its use of proactive moderation tools, user reporting tools, and specialised moderation teams to address abuse on its service. Telegram stated

> If a report is confirmed, Telegram acts with due regard to all known circumstances. Disseminating terrorist and violent content or CSAM on any parts of Telegram service leads to permanent removal of associated accounts and Communities.

Regarding the statement on its 'frequently asked questions' page that it would not process removal requests for 'illegal content' on certain parts of its service, Telegram stated this information was outdated and was the result of statements about Telegram's stance on copyright infringement having been mistakenly copied to a section dealing with Telegram's stance on illegal content. Specifically, Telegram stated

> As at February 29, 2024, certain portions of the Telegram FAQ may have featured outdated information, some of which was updated in September 2024 (with further updates planned in the coming months). This included the item quoted under "Context" in this question, which mistakenly included text from an earlier revision of the FAQ. Namely, the mention of "not processing requests regarding private chats" had been erroneously copied from the FAQ section related to copyright infringement, where it is present to this day.[5]...
>
> The quoted paragraph was aimed at law-abiding Telegram users who rely on Telegram for the privacy for their personal communication. It was meant to emphasize that, as Telegram moderators cannot proactively inspect the

[5] Telegram, 'Telegram FAQ - Q: A bot or channel is infringing on my copyright. What do I do?', URL supplied by Telegram on 13 September 2024, URL: https://telegram.org/faq#q-abot-or-channel-is-infringing-on-my-copyright-what-do-i-do

private messages of its users, they cannot act on unsupported requests which do not rely on reporting mechanisms.

The text was never meant to imply that Telegram's Terms of Service could be violated in private chats. This is evidenced by the fact that users could always report both incoming private and secret chats, and messages in private groups and channels to moderators.

eSafety notes that online media outlet, The Verge, first reported this change to Telegram's FAQs page as having occurred on 6 September 2024.[6]

eSafety also notes that the Internet Archive's Wayback Machine shows that the item Telegram states was 'mistakenly included text from an earlier version of the FAQ' was present on Telegram's FAQ page as far back as 15 March 2016, and that this item appears to pre-date any reference to copyright infringement in Telegram's FAQs.[7]

3. Thresholds/criteria to determine action on TVE breaches

Telegram was asked if it had criteria or thresholds in place to determine what action would be taken when TVE was identified on Telegram. Telegram provided the following information:

Table A

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported for Telegram
Permanent account ban	Telegram stated the following: • Disseminating material that calls for violence in the form of text, image, recordings, footage or otherwise. Telegram specified this means material 'like concrete and specified threats of physical harm'. • Disseminating material that is gruesome or shockingly graphic. Telegram gave such examples as 'graphic details of torture,

[6] The Verge, 'Telegram changes its tone on moderating private chats after CEO's arrest', 6 September 2024, URL: https://www.theverge.com/2024/9/5/24237254/telegram-pavel-durov-arrest-private-chats-moderation-policy-change

[7] Internet Archive Wayback Machine, 'Telegram FAQ – 15 March 2016', accessed 16 October 2024, URL: http://web.archive.org/web/20160315182715/https://telegram.org/faq

	accident photos' or material that 'glorif[ies] or promote[s] violent or terrorist ideologies'. • Soliciting funds for terrorist organisations or causes. • Owning or being an administrator of a Community involved in the above activities.
Account strikes	Telegram stated that if a Community, or an account belonging to a 'journalist' or 'researcher', reposts TVE with the intention of sharing 'legitimate scientific research, historical records, or news', then Telegram may either: • grant an exception; or • apply up to two warnings before terminating the Community or account. Telegram stated the decision on enforcement depends on the 'severity, purpose and relevance of the posted content under applicable law'.

Telegram also stated that 'where appropriate', it will remove publications in Communities and remove associated groups and channels.

s 47G(1)(b) [redacted]

4. Questions about reporting of TVE

A. In-service reporting of TVE on different parts of Telegram

In response to questions about whether users could report instances of TVE to Telegram within the service (as opposed to navigating to a separate webform or email address), Telegram responded:

Table B

Parts of the service	In-service reporting option?	Reporting category
Chats	Yes	'Block user > Report Spam'*
Secret Chats	Yes	
Group chats (public)	Yes	'Violence'
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Voice calls	No**	N/A
Video calls	No**	

*In response to a follow up question from eSafety, which highlighted that in eSafety's testing on the Telegram iOS app, for Chats and Secret Chats the option to 'Report spam' was not present in all cases. Telegram subsequently clarified that the 'Block + Report Spam' reporting flow is only available when the Chat or Secret Chat is 'initiated by non-contacts and strangers'. eSafety understands that when an end-user wishes to report a message from an account they have already added as contact, s 47G(1)(b) '.

Telegram provided the following reason for this discrepancy in reporting functionality

> In the extremely unlikely event that a user's friend or acquaintance began sending them TVE content, Telegram contends that it would be more reasonable and effective for said user to contact authorities directly, providing all relevant proof and contact information.

eSafety considers that limiting reporting tools to scenarios where the account sending harmful or violative material is not a contact of the end-user risks preventing Telegram from identifying and preventing bad actors from continuing to perpetrate harm on the platform even after they have been blocked by an end-user on the service.

Telegram stated that the single reporting option 'Block + Report Spam' for private and Secret Chats intended to simplify the user experience and minimise the length of time and number of interactions necessary for a user to end the chat. Telegram stated that 'once the report is processed by moderators, it is escalated as necessary – including via AI / ML if appropriate'.

eSafety notes that in response to other questions in the Notice, Telegram stated that it had no means of accessing messages reported by end-users from Secret Chats (see Section 2). Instead, Telegram stated it relies on alternative signals to assess and prioritise reports made about material in E2EE parts of the service.

eSafety notes that this may limit Telegram's ability to review, assess, prioritise, and respond to reports about harmful and illegal material or activity occurring in Telegram's Secret Chats.

**Telegram's original response to the Notice stated that end-users could make in-service reports about voice calls and video calls using a 'Violence (via the community info section)' reporting category. In response to a follow-up question from eSafety, Telegram subsequently stated that in-service reporting for voice and video calls was not available during the Report Period. Instead, Telegram stated that 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

s 47G(1)(b)

B. Reporting of TVE by third party services that use Telegram's API

eSafety asked whether Telegram had minimum safety requirements for third party services that use Telegram's APIs to access its service. Telegram responded that it did have minimum safety requirements and that this includes the requirement for user reporting functions on third party apps to notify Telegram of breaches of its terms of service.

Telegram provided links to its Telegram API Terms of Service[8] and Security Guidelines[9], and stated that the API Terms of Service expect that third-party services make available all basic functionalities of the Telegram service, including the reporting tools, and that third-party services are accountable for ensuring that these features function correctly.

Telegram stated that it would 'from time to time re-confirm that it correctly receives user reports from the most popular third-party clients, including reports of potentially terrorist and violent content'. Telegram also stated that third-party services that fail to comply with Telegram's Terms of Service are 'routinely flagged for removal to third-party app stores and blocked from accessing Telegram's core APIs'.

[8] Telegram, 'Telegram API Terms of Service', URL supplied by Telegram on 13 September 2024, URL: https://core.telegram.org/api/terms
[9] Telegram, 'Security Guidelines', URL supplied by Telegram on 13 September 2024, URL: https://core.telegram.org/mtproto/security_guidelines

C. Reporting mechanisms for other entities to report TVE

In answer to questions about having separate reporting mechanisms for certain entities to report TVE, Telegram stated that it did have dedicated reporting mechanisms for:

- Law enforcement;
- Trusted Flaggers;
- Regulatory and public authorities; and
- 'International organizations'.

Telegram stated that these reporting mechanisms enable '[f]aster processing times whenever possible; processing by a dedicated team member / task group; deeper review if needed'. Telegram pointed to its collaboration with the Global Center for Combating Extremist Ideology, or 'Etidal', as an example of a specialised reporting mechanism. Telegram stated that between February 2022 and June 2024, Telegram removed '93,99 million pieces of TVE content' through its collaboration with Etidal.

Telegram stated that during the Report Period, it did not have a separate reporting mechanism for civil society groups to report TVE to the service. Telegram reported

> While engagement with these groups regarding content reporting has been minimal as at 29 February 2024, Telegram recognizes the potential benefits of collaborating with more external experts. To this end, Telegram is actively considering the introduction of dedicated contact points and other initiatives to enhance its responsiveness to valid concerns raised by civil society groups and remains open to dialogue in furtherance of that goal.

D. Percentage of TVE sent for human review

Telegram was asked to provide the percentage of TVE reports it sent for human review and the criteria and thresholds used to determine when reports were sent for human review.

Table C

Percentage of user reports of TVE sent for human review	75%

Criteria and thresholds used to determine when a user report is sent for human review	Telegram stated some reports were not reviewed by humans because: • Reported content had already been removed by proactive measures. • Reported content had already been removed because the channel/group it was posted in was removed.
Percentage of TVE detected through automated tools sent for human review	65%
Criteria and thresholds used to determine when a report of TVE is detected through automated tools is sent for human review	Telegram provided a select number of scenarios where content would be sent for human review. eSafety has chosen not publish these specific scenarios to prevent this information being misused. Telegram stated that some of these criteria were intended to prevent Telegram's systems from automatically banning 'researchers, human rights activists, legitimate news sources etc' as well to prevent bad actors from attempting to 'silence' Telegram communities by deliberately posting violative material in them as 'abusive spam'. Telegram also stated that when its automated tools detect material that is a '100% hash match', in a private Community, human moderators are notified 'but do not receive a copy of the media item itself'. In response to other questions in the Notice, Telegram stated that detections of hash matched TVE material result in resulted in the automated removal 'of all users, Communities and publications involved' except for 'Communities or users that are likely to yield false positives' (see section 5Av).

E. Percentage of TVE detected proactively

Telegram was asked what percentage of TVE was detected proactively, compared to TVE reported by users, trusted flaggers or through other channels for the following parts of its service:

Table D

Parts of Telegram	Percentage of TVE detected proactively	Percentage of TVE reported by users, trusted flaggers or other
Chats	N/A	100%
Secret Chats	N/A	100%
Group chats (public)	67%	33%
Group chats (private)	82%	18%

Channels (public)	69%	31%
Channels (private)	79%	21%
Voice and video calls (public and private)	N/A*	N/A*
Group video calls (public and private)	'Included in group chats'**	
Stories	60%	40%

* In answer to a follow-up question from eSafety to clarify why its answer was 'N/A' for voice and video calls Telegram stated that voice and video calls could not be directly reported by end-users using in-service reporting tools. Instead, 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

**Telegram stated that its video group call data was included in the relevant group chat statistics because 'information on resulting bans is not stored separately'. In response to other questions in the Notice, Telegram stated that it did not use any proactive detection tools to detect livestreamed TVE in group video calls. eSafety therefore understands that 100% of any TVE detected in group video calls during the Report Period was reported by users.

F. Appeals against TVE-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for TVE, where Telegram was alerted by automated tools or user reports, and how many of those were successful, Telegram provided the following information:

Table E

How Telegram was alerted to TVE	Number of appeals made for accounts banned for TVE breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for TVE breach	Number of appeals that were successful for material removed
Automated tools	3,420	110	N/A*	
User reports	1,107	26		

*Telegram stated that because TVE-related content violations result in the users and Communities involved being removed from Telegram, 'It is not generally possible to appeal for reinstatement of removed TVE materials, so only account appeals are included'. Telegram stated that in some jurisdictions, such as the EU and the EU Terrorist Content Online Regulation, it may receive appeals against

content removals from legally mandated contact lines. However, Telegram reported that there were 'no actionable appeals connected to removal of terrorist, violent or extremist content' during the Report Period.

5. Questions about proactive detection

s 47G(1)(b)

In response to various questions in the Notice, Telegram stated that when TVE material was confirmed, the material was removed along with 'users, Communities and publications involved'.

s 47G(1)(b)

A. Detecting known material using hash-matching

i. Known TVE images

In response to questions about hash matching for known TVE images, Telegram provided the following information:

Table F

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	

Stories	Yes	Internal Telegram Hash Matching System
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why hash matching tools were not used on Chats or Secret Chats user reports, Telegram stated that Telegram was 'founded on the principle of defending user privacy and their right to private communication' and that 'this commitment prioritizes user privacy above all'. Telegram stated that because of this commitment to user privacy

> encrypted contents of private chats are always protected, ensuring that the confidentiality of private correspondence is never compromised.

eSafety notes that Telegram stated that it does use hash matching tools on other 'private' parts of the service – namely, private groups and private channels. eSafety further notes that Chats, Private Groups, and Private Channels all use the same form of encryption – which is not E2EE.

It is unclear to eSafety why tools capable of detecting known TVE, verified as harmful and/or violative by Telegram's own trust and safety staff, are not being used on Chats given Telegram stated that they are used on other private parts of Telegram's service, namely private groups and private channels. In relation to Secret Chats user reports, as noted at section 2B, alternative methods also exist which could enable hash-matching tools to review content reported in E2EE messages.

In response to a question about the alternative steps Telegram took to detect known TVE images on Chats and Secret Chats user reports, Telegram stated that it relied on users reporting messages via its 'Block + Report Spam' reporting tool (which as eSafety notes above, appears to only exist for messages from users that have not been added as contacts by the reporting users). Telegram stated these reports are 'processed by Telegram's tools and moderators…including via AI / ML if appropriate'. Telegram also stated that it

> employs extensive automated rate-limiting and spam-preventive measures, ensuring that no user or software is able to share content in bulk or to a significant number of users via any chat, irrespective of the nature of said content. In so doing, by design and without compromising user privacy, Telegram prevents malicious actors from effectively using its private messaging component to spread their messages.

ii. Known TVE video

In response to questions about hash matching for known TVE video, Telegram provided the following information:

Table G

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram Hash Matching System*
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Content in user reports	Yes	

*In response to a follow-up question, Telegram subsequently stated that it considered that 'there is no material difference in the way in which video and image hashing is performed on a technical level'.

In response to why hash matching tools are not used to detect known TVE videos in Chats and Secret Chats user reports, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known TVE videos on Chats and user reports about Secret Chats, Telegram referred to the alternative measures it took for known TVE images (see section 5Ai).

iii. Known TVE written material

In response to questions about hash matching for known TVE written material on Telegram, such as manifestos or text promoting, inciting, or instructing in TVE, Telegram provided the following information:

Table H

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	No	N/A

Group chats (private)	No	N/A
Channels (public)	No	N/A
Channels (private)	No	N/A
Stories	No	N/A
Content in user reports	Yes	Internal Telegram Hash Matching System*

*In response to a follow-up question, Telegram stated that it considered that 'there is no material difference in the way in which text and image hashing is performed on a technical level'.

In response to why hash matching tools are only used to detect TVE written material in content referred to Telegram in user reports, Telegram stated

> Relying on hash matching on all text content at scale is generally not advisable as any one message can be expressed in numerous ways across different languages and formats. Instead, Telegram relies on ML models finetuned on known TVE written material.... Conversely, hash matching on reported content provides a useful and efficient preliminary layer to the existing moderation pipeline.

In response to a question about the alternative steps Telegram took to detect known TVE written material on chats, secret chats, private group chats, and private group channels, Telegram referred to the measures it took for known TVE images on Chats and Secret Chats (section 5Ai). For 'Other chats', which eSafety assumes refers to Public Group Chats and Public Channels, Telegram stated that it used machine learning models 'finetuned on known TVE written material to check a subset of text messages sampled from all relevant chats according to reasonable criteria'. eSafety has chosen not to publish these criteria to prevent the information being misused.

iv. Sources of TVE hashes

Telegram reported that it sourced its hashes of known TVE images, videos, and text from **internal databases of hashes of TVE material that had previously been identified on Telegram and removed by its human moderators.** In answer to a question about how often it updated this database, Telegram stated that the database was updated every time a human moderator removed an item of new, or previously 'unknown', TVE material from the service.

Telegram noted that its '[e]xclusive reliance' on human moderators to compile its TVE hash database is designed to mitigate risks of 'circular data pollution' posed by automated systems being trained on decisions by previous automated systems.

Telegram stated that it 'purposefully avoids reintroduction of machine-labeled or synthetic data into datasets, because such measures can degrade the quality and reliability of the datasets'.

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing TVE material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material has already been identified by other online service providers and hashed in extensive shared databases like those run by the GIFCT or Tech Against Terrorism.

v. Action taken on known TVE

In response to questions about what action was taken when known TVE images and video were detected by its tools, Telegram stated that it resulted in the automated removal 'of all users, Communities and publications involved' except for 'Communities or users that are likely to yield false positives'. Telegram provided further information on the measures it took to address known TVE which eSafety has chosen not to publish to prevent the information being misused.

Telegram stated that when hashes of known TVE written material were detected in user reports, Telegram removed the material and the user who posted it, 'unless there is reason to believe the match may be a false positive … in which case a human moderator reviews the match and takes action accordingly.'

B. Detecting new TVE material

i. New or 'unknown' TVE images

In response to questions about the detection of new (or 'previously unknown') TVE images, Telegram provided the following information:

Table I

Parts of service	Used tools for images?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A

Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[10]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

When asked why it did not use technology to detect new TVE images on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using hash matching tools to detect known TVE images (see section 5Ai).

In a follow-up question to Telegram, eSafety noted that Telegram had stated that it used tools to detect *known* TVE on Private Group Chats and Private Channels. In light of this, eSafety asked Telegram to provide the reason it did not use tools to detect *new* TVE images on these parts of the service. Telegram stated that the 'technical architecture and access rules of private groups and channels' prevent anyone who is not a member of those groups from accessing them and seeing the content being shared inside. Telegram reported that moderators could only access such content when it was reported by an end-user, or the community became accessible via a public invite link. Telegram stated that it used hash matching tools to detect known TVE on these parts of the service because when its tools detected the material, moderators would receive a notification that a 100% match with violative content had occurred – rather than the actual image or video being disclosed to them for specific review. Telegram stated that it considered that this process should not be applied to detections of new TVE material because

> Telegram contends that notifying moderators of such matches would be insufficient, as matches for new content, while accurate, cannot be verified with absolute certainty without checking the content itself. Even assuming that such matches were taken at face value, moderators would then be unable to process potential appeals by the deleted accounts.

[10] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

eSafety considers that not using proactive detection tools to identify and review potential TVE material increases the likelihood that such material will remain undetected and continue to circulate on these parts of the service.

eSafety notes that Chats, private group chats, and private channels are not E2EE – This means technical options are available for content detection and review by human moderators. Telegram has stated it uses such tools on other parts of its service.

In response to a question about the alternative steps Telegram took to detect new TVE images on these parts of the service, Telegram referred to the measures it took to detect likely TVE in text (see section 5Biii).

ii. New or 'previously unknown' TVE videos

In response to questions about the detection of new (or 'previously unknown') TVE videos, Telegram provided the following information:

Table J

Parts of service	Used tools for videos?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[11]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
Content in user reports	Yes	

In response to why it did not use technology to detect new TVE videos on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect likely TVE in images (see section 5Bi).

[11] In response to a follow-up question seeking the names of the tools Telegram uses to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

In response to a question about the alternative steps Telegram took to detect new TVE videos on these parts of the service, Telegram referred to measures it took to detect likely TVE in text (see section 5Biii).

iii. Text analysis to detect TVE

In response to questions about technology to detect phrases, codes or hashtags, indicating likely TVE in text (for example manifestos or text promoting, inciting, instructing TVE), Telegram provided the following information:

Table K

Parts of service	Used text analysis tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[12]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
Profile username	Yes	
Profile description	Yes	
Group username	Yes	
Group description	Yes	
Channel username	Yes	
Channel description	Yes	
Content in user reports	Yes	

In response to why it did not use technology to scan Chats, Secret Chats, Private Group Chats, and Private Channels for indications of likely TVE in text, Telegram referred to the reasons it gave for not using hash matching tools to detect known TVE images (see section 5Ai).

In response to what alternative steps Telegram took to detect known phrases, codes, or hashtags indicating likely TVE on these parts of the service, Telegram

[12] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

stated that it provided reporting options for users to report such material on these parts of the service. Telegram noted that when users report such TVE material in 'private groups and channels', this results in the material being forwarded to Telegram moderators for their review. However, for what Telegram describes as 'private 1-on-1 chats' (i.e. chats and Secret Chats), Telegram reiterated that user reports in these parts of the service are 'processed by Telegram's tools and moderators...including via AI / ML if appropriate'.

Telegram also referred to the 'automated rate limiting and spam-preventative measures' it used to prevent bad actors from spreading known TVE material in private messages (see section 5Ai).

iv. Sources of phrases, codes, and hashtags

Telegram stated that it sourced phrases, codes, and hashtags indicating likely TVE from 'samples ... based on content items' that had been removed from Telegram by its human moderators.

v. Action taken when new TVE was detected

In response to questions about what action was taken when new TVE images, video, and likely TVE in text was detected by its tools, Telegram stated that the material was sent for human review and if the content was confirmed as TVE it resulted in the removal of 'users, Communities and publications involved'. Telegram reported that following removal, new TVE images and video were then added to Telegram's internal hash database so they could be actioned in the future as known TVE.

Telegram provided further information on the measures it took to address new TVE which eSafety has chosen not to publish to prevent the information being misused.

Telegram also noted that when a Community is removed, human moderators review the most common search terms that were used to find that Community in order for them to be 'possibly removed from Telegram's public search to limit future spread and reach of similar content'.

In answer to a question asking if the detection of phrases, codes, or hashtags indicating likely TVE in text resulted in Telegram blocking these words or phrases to users searching for them, Telegram responded that 'yes', it blocked 'certain keywords or text patterns' from its search results.

C. Languages covered by language analysis tools

i. Detecting TVE in text

In response to questions about the languages covered by Telegram's language analysis tools, Telegram did not provide a list of languages.

Telegram stated that its models and tools for detecting TVE text 'perform reasonably well in most languages', and that these technologies 'aim to abstract away the concept of language when deriving embeddings from text'.

In response to follow-up questions from eSafety, Telegram stated

> Telegram did not maintain a specific list of all languages included in the training sets of the underlying models, which is why it was unable to provide a list relevant for the Report Period. Telegram must note that a significant share of messages in its datasets contain text in multiple languages (i.e., multiple languages within the same message). As well, messages are often too short to automatically identify a specific language with absolute certainty.

ii. Detecting TVE in video

In response to a question about the languages covered by the tools Telegram used to detect new TVE in video, Telegram did not provide a list of languages.

Telegram stated the tools did not 'have particular regard for the specific language of the content in question' but focussed on detecting patterns the model had previously learned to associate with TVE from an existing dataset.

In response to follow-up questions from eSafety, Telegram referred to the answer it gave to eSafety's follow-up question regarding the languages covered by the tools Telegram used to detect likely TVE in text (see section 5Ci).

D. Livestreamed TVE

i. Detecting livestreamed TVE

The Notice specified that livestreaming includes one-on-one video calls and video calls where one or more multiple people stream material to a group of any size.

In response to questions about the measures Telegram had in place to detect the livestreaming of TVE on its service, Telegram provided the following information:

Table L

Parts of service	Measures in place to detect TVE in livestreams?	Interventions used	Name of tools used
Group video calls	No	N/A	N/A
Channel livestreams	No		

When asked why it did not have any measures in place to detect livestreamed TVE, Telegram stated

> While livestreaming functionalities are supported on Telegram, they represent a generally insignificant share of the service's overall usage, particularly so as it concerns the spread of harmful content. As such, Telegram finds that immediate user reports already provide reliable coverage to detect and address such incidents effectively.

ii. Reducing the likelihood of livestreamed TVE

In response to questions about the steps taken by Telegram to reduce the likelihood that TVE could occur in livestreams, Telegram stated that it used the following measures:

- Restrictions for those who have previously violated terms of service or community guidelines/standards.
- User reports – Telegram stated that it 'relies on immediate user reports' to detect livestreamed TVE **but did not indicate whether or how it prioritises reviews of reports of livestreamed content.** In response to other questions in the Notice, Telegram stated that it did not provide in-service reporting tools for video calls (see section 4A) during the Report Period.

E. Blocking links to TVE material

i. Detection and sources of URLs

Telegram was asked about its use of lists or databases to proactively detect and block URLs linking to TVE on other platforms. Specifically, Telegram was asked about:

- Known URLs linking to websites/services operated by individuals/organisations dedicated to the creation, promotion, or dissemination of TVE
- URLs linking to known TVE material on other services/websites (which may not be dedicated to TVE)
- Join-links to groups, Channels, communities, or forums on other services that were known to be associated with TVE.

Table M

Parts of service	Blocked URLs to websites/services dedicated to TVE?	Blocked URLs linking to known TVE material on other services/websites?	Blocked join-links to groups/channels on other services known to be associated with TVE?	URL sources
Chats	No	No	No	N/A
Secret chats (E2EE)	No	No	No	
Group chats (public)	No	No	No	
Group chats (private)	No	No	No	
Channels (public)	No	No	No	
Channels (private)	No	No	No	
Profile description	No	No	No	
Group description	No	No	No	
Channel description	No	No	No	

When asked why URLs to TVE material were not blocked and whether alternative steps were taken to block URLs, Telegram stated that 'focusing its efforts on ML-based classification tends to yield better results when compared to static link blacklists'. Telegram stated that links to harmful material tended to be 'routinely taken down by all hosting providers and tend to either rotate constantly or be hidden behind URL shorteners, proxies etc'. Telegram also stated that it used Internal Telegram AI and Machine Learning Models (see Section 12) that are trained on previous detections of TVE material, including material that may contain external links.

F. Off-platform monitoring

Telegram was asked if it used off-platform monitoring[13] either provided internally or by third-party services, to identify accounts, groups or channels on its service that were dedicated to TVE. Telegram stated it performs '[e]xtensive monitoring' of media sources as well as reviewing referrals sent by 'non-registered users and trusted organizations' to Telegram via email.

6. Questions about resources, expertise, and human moderation

A. Trust and Safety

i. Trust and Safety and other staff

Telegram was asked to provide the number of staff that were employed or contracted by Telegram to carry out certain functions at the end of the Report Period.

Table N

Category of staff	Number of staff*
Engineers employed by Telegram focussed on trust and safety	5
Content moderators employed by Telegram	0
Content moderators contracted by Telegram	150
Trust and safety staff employed by Telegram (other than engineers and content moderators)	4

* Telegram stated that these figures represented the number of staff who 'may from time to time be involved with decisions regarding content or reports from

[13] Monitoring of activity on other services.

Australia and do not reflect or approximate the total number of global content moderation and trust and safety personnel contracted by Telegram.' Telegram also stated that Australian end-users make up less than 0.2% of its monthly active users.

eSafety notes that Telegram did not provide its global resourcing. In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

eSafety notes that TVE is a global harm and the resources that a service has in place to respond to 'content or reports' internationally is highly relevant to the online safety of Australians, and implementation of the Expectations.

ii. Trust and Safety dedicated to minimising TVE

In response to a question asking if Telegram had a dedicated trust and safety team responsible for minimising TVE on the service, Telegram answered 'yes', reporting that '[t]he relevant team members are high-performing professionals trained in team management, threat mitigation and legal affairs'. Telegram stated that this team was responsible for 'overseeing content moderation and reviewing reports from external stakeholders, such as trusted flaggers and international organizations'.

Telegram provided the following information about the composition of its team:

Table O

Name of role/area of expertise	Number of staff	Number of contractors
Trust and Safety managers*	4**	0

* Telegram stated that it did not have specific titles for these positions because Telegram's 'hierarchy is informal and horizontal'.

** Telegram stated that this figure was specific to 'staff that may from time to time be involved in decisions regarding content or reports from Australia and do not reflect or approximate the total number of global trust and safety personnel contracted by Telegram'.

In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be

involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

iii. Surge teams to respond to a TVE crisis

Telegram was asked if it had a surge team(s) to respond to TVE crises, such as a livestreamed attack with content disseminated on the service. Telegram answered 'yes' and stated that it 'maintains crisis mitigation protocols in accordance with industry standards'. Telegram stated that on occasion it could establish task forces of '2-3 people tailored to resolving specific situations, e.g., appearance of terrorist or violent content in significant quantities'.

Telegram provided the following information about the composition of this team:

Table Q

Name of role/area of expertise	Number of staff	Number of contractors
Trust and Safety managers and contractors*	3**	13**

* Telegram stated that it did not have specific titles for these positions because Telegram's 'hierarchy is informal and horizontal'.

** Telegram stated that these figures were specific to 'staff that may from time to time be involved in decisions regarding content or reports from Australia and do not reflect or approximate the total number of global trust and safety personnel contracted by Telegram'.

In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

B. Languages human moderators operate across

In response to a question about the languages that its human moderators operated across (both employees and contractors), Telegram provided the following:

Table R

Languages covered by employees (all languages)	Languages covered by contractors (all languages)[14]	
N/A*	• English • Amharic • Arabic • Azerbaijani • Bulgarian • Chinese (traditional and simplified) • Croatian • Czech • Danish • Estonian • Farsi • Filipino • Finnish • French • Georgian • German • Greek • Hindi • Icelandic • Indonesian • Italian • Japanese	• Kazakh • Korean • Kyrgyz • Luganda • Lunyakore • Lusoga • Malay • Moldavian • Norwegian • Polish • Portuguese (Brazil) • Portuguese (Europe) • Romanian • Russian • Serbian • Shona • Spanish • Swahili • Swedish • Tajik • Turkish • Ukrainian • Urdu

[14] Telegram also advised that since the Report Period, it had expanded the languages covered by its contracted content moderators by adding Afrikaans, Bengali (Bangladesh), Chichewa (Zambia), Dhivehi (Maldives), Dutch, Gujarati, Kabyle (Algeria), Kinyarwanda, Lithuanian, Macedonian, Punjabi, Sinhalese (Sri Lanka), and Thai.

		• Uzbek • Yoruba

*Telegram stated that '[a]ll ordinary moderators' on Telegram are contractors.

eSafety notes that the top 5 languages, other than English, spoken in Australian homes are Arabic, Cantonese, Mandarin, Vietnamese and Punjabi.[15] Telegram's human moderators do not cover Vietnamese or Punjabi.

C. Median time to reach an outcome to a user report of TVE

Telegram was asked to provide the median time taken to reach an outcome after receiving a user report about TVE for the following parts of the service:

Table S

Parts of the service	Reports from users globally	Reports from users in Australia *
Chats	18 hours	18 hours
Secret Chats	18 hours	18 hours
Group chats (public)	15 hours	15 hours
Group chats (private)	15 hours	15 hours
Channels (public)	15 hours	15 hours
Channels (private)	15 hours	15 hours

In response to a question asking how median time was calculated Telegram stated that to calculate these figures it registered 'the net time frames between the submission of each individual report and the moderator's decision in respect of that report'.

* Telegram stated that it 'currently doesn't have the technical means to provide separate statistics by country'.

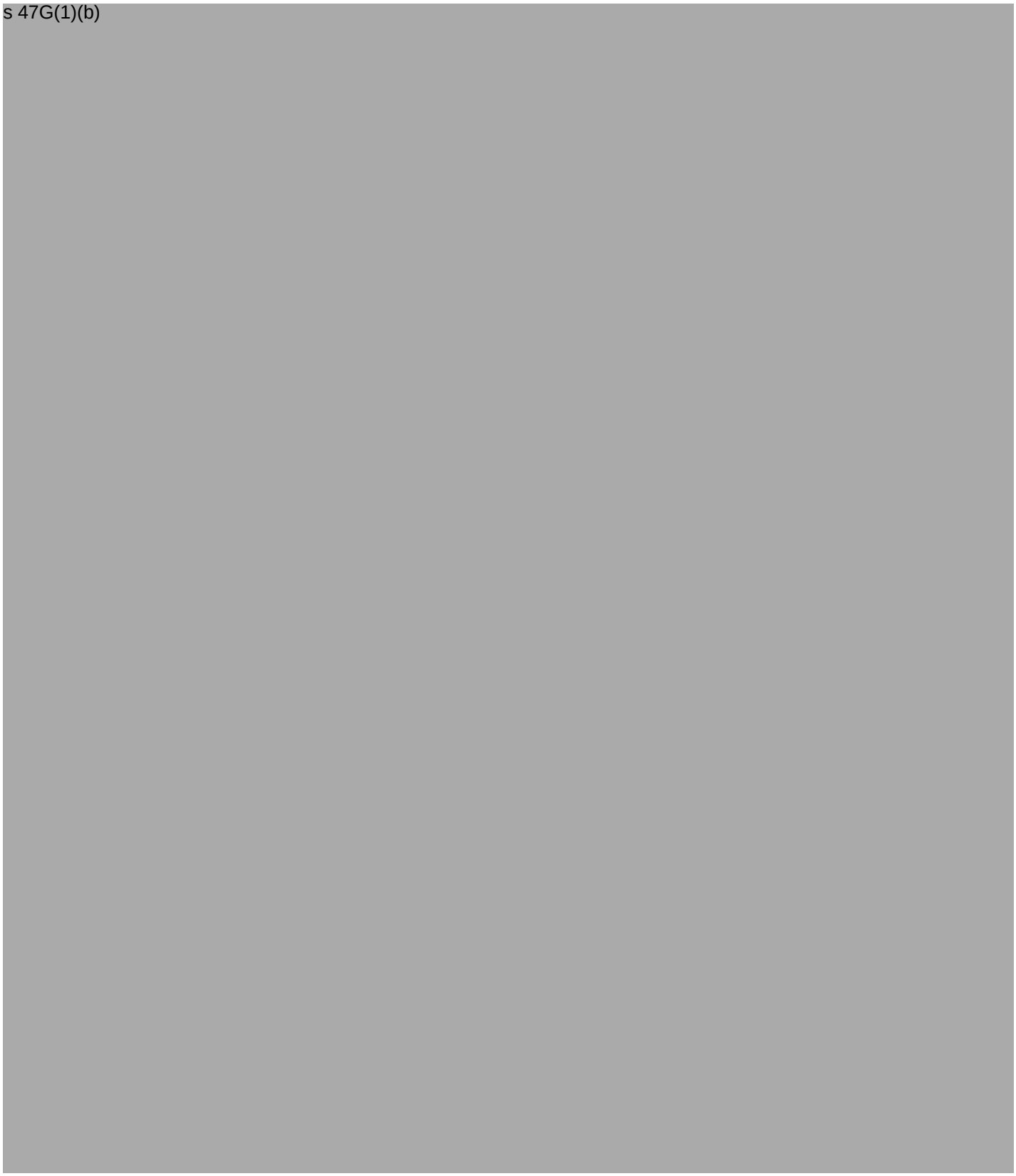
It is unclear to eSafety how Telegram determines that it has reached an outcome for Secret Chats when it has stated that it does not review the

[15] Australian Bureau of Statistics, 'Cultural diversity: Census', 28 June 2021, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Top%205%20languages%20used%20at,Punjabi%20(0.9%20per%20cent).

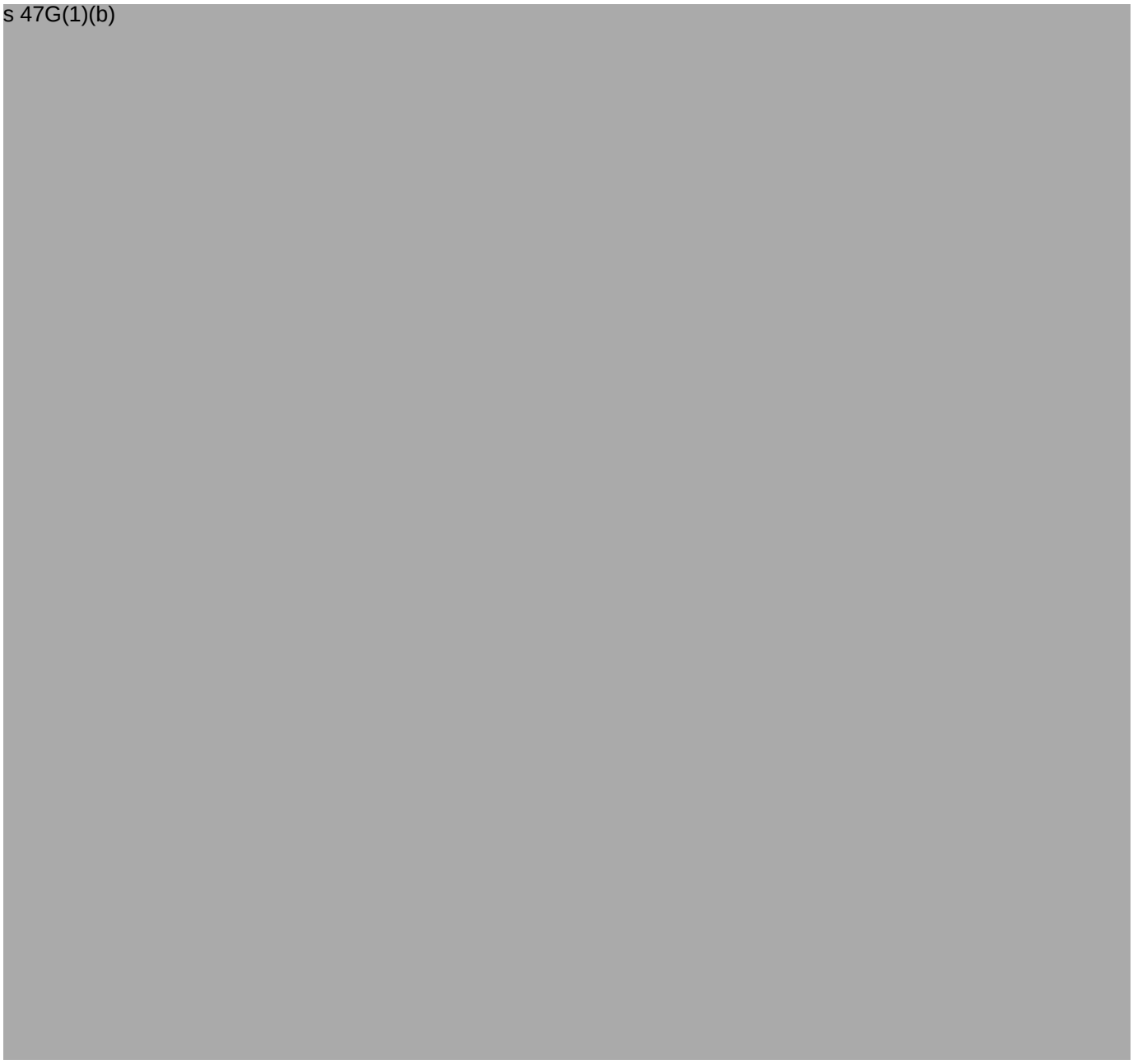
contents of the messages being reported on this part of the service (see section 2B).

D. Volunteer moderation

s 47G(1)(b)



s 47G(1)(b)



7. Questions about steps to prevent recidivism

A. Measures and indicators

In response to a question asking if Telegram had measures in place to prevent recidivism for TVE-related breaches on its service, Telegram responded 'yes' and

[16] Telegram, 'Channels FAQ', accessed 12 November 2024, URL: https://telegram.org/faq_channels?setln=en
[17] Telegram, 'Telegram FAQ', accessed 12 November 2024, URL: https://telegram.org/faq
[18] Telegram, 'Supergroups 10,000 admin tools and more', accessed 12 November 2024, URL: https://telegram.org/blog/admin-revolution?setln=tr

listed a minimal[19] number of indicators that it used to detect users that have previously been banned for TVE breaches. eSafety has chosen not to publish these indicators to prevent the information being misused.

B. Preventing banned TVE groups and channels from being recreated

In response to a question about the measures Telegram took to prevent banned TVE groups and channels from being recreated, Telegram stated that '[o]wners and administrators of infringing Communities may also face removal alongside the Communities themselves, preventing them from creating new Communities or accounts on Telegram'. Telegram also referred to a minimal number of signals it may use to 'routinely detect' Communities that bear similarities to previously banned Communities which eSafety has chosen not to publish to prevent the information being misused.

C. Applying TVE-related bans to associated accounts

Telegram was asked, when it took action against an account for a TVE-related breach, whether it applied bans to associated accounts. eSafety defined 'associated accounts' as 'other users who are associated with the banned user'. Telegram answered 'yes' and stated that when it identified a user disseminating TVE material, it reviewed 'further reports linked to this user, as well as to any Communities which the user owns or administrates'. Telegram stated that any Communities found to be involved in disseminating TVE would also be removed.

Telegram stated that Channel subscribers or group members 'who are neither managing nor directly spreading or promoting prohibited content, even if they are part of Communities that may contain such content, are not subject to automatic bans'. Telegram stated that this approach was adopted to avoid inadvertently disrupting law enforcement, journalists, activists and others who may be part of these groups for legitimate reasons.

Telegram further reported that while it 'does not compile data or operate tools conducive to internal investigative work on private user interactions', Telegram

[19] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:

- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

has developed tools to assist its moderators to identify 'interconnected management' between 'flagged Communities and their administrators'. eSafety has chosen not to publish further detail on the measures Telegram reported to prevent the information being misused.

D. Sharing of banned account details with other entities

Telegram was asked if it shared details of accounts banned for TVE with the following entities:

Table U

Entity	Shared details of accounts banned for TVE?
Other service providers	No
Law enforcement	Yes*
Regulatory or other public authorities	No
Global Internet Forum to Counter Terrorism	No
Civil society groups	No

* Telegram stated that it provided information to law enforcement in response to 'valid legal requests submitted by law enforcement agencies through designated channels'.

Telegram stated that as at 29 February 2024, Telegram had not received any legal requests from Australian law enforcement agencies 'via dedicated formal channels (e.g., mutual legal assistance requests to the governments in jurisdictions in which the relevant Telegram companies are located)'.

Part 2. Questions in relation to chid sexual exploitation and abuse (CSEA)

8. Questions about reporting of CSEA

A. In-service reporting of CSEA on different parts of Telegram

In response to questions about whether users could report instances of CSEA to Telegram within the service (as opposed to navigating to a separate webform or email address), Telegram responded:

Table V

Parts of the service	In-service reporting option?	Reporting category
Chats	Yes	'Block user > Report Spam'
Secret Chats	Yes	
Group chats (public)	Yes	'Child Abuse'
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Voice calls	No*	N/A
Video calls	No*	
Stories	Yes	'Child Abuse'

*Telegram's original response to the Notice stated that end-users could make in-service reports about voice calls and video calls using a 'Child Abuse (via the Community's info section)' reporting category. In response to a follow-up question from eSafety, Telegram subsequently stated that in-service reporting for voice and video calls was not available during the Report Period. Instead, Telegram stated that 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

s 47G(1)(b)

As noted at section 4Ai, Telegram subsequently clarified that the 'Block + Report Spam' reporting flow is only available when the Chat or Secret Chat is

'initiated by non-contacts and strangers'. eSafety understands that when an end-user wishes to report a message from an account they have already added as contact, s 47G(1)(b) .

With respect to TVE (which eSafety understands is applicable to CSEA), Telegram stated that this was because

> In the extremely unlikely event that a user's friend or acquaintance began sending them TVE content, Telegram contends that it would be more reasonable and effective for said user to contact authorities directly, providing all relevant proof and contact information.

As noted at section 4Ai, eSafety considers that limiting reporting tools to scenarios where the account sending harmful or violative material is not a contact of the end-user risks preventing Telegram from identifying and preventing bad actors from continuing to perpetrate harm on the platform even after they have been blocked by an end-user on the service.

Telegram stated that the single reporting option 'Block + Report Spam' for private and Secret Chats was intended to simplify the user experience and minimise the length of time and number of interactions necessary for a user to end the chat. Telegram stated that 'once the report is processed by moderators, it is escalated as necessary – including via AI / ML if appropriate'.

As noted at section 4Ai, eSafety notes that in response to other questions in the Notice, Telegram stated that it had no means of accessing messages reported by end-users from Secret Chats (see Section 2). Instead, Telegram stated it relies on alternative signals to assess and prioritise reports made about material in E2EE parts of the service.

eSafety notes that this may limit Telegram's ability to review, assess, prioritise, and respond to reports about harmful and illegal material or activity occurring in Telegram's Secret Chats.

9. Questions about proactive detection of CSEA

s 47G(1)(b)

In response to various questions in the Notice, Telegram stated that when CSEA material was confirmed, the material was removed along with 'users, Communities and publications involved'.

s 47G(1)(b)

A. Detecting known material using hash matching

i. Known CSEA images

In response to questions about hash matching for known CSEA images, Telegram provided the following information:

Table W

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram Hash Matching System
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why hash matching tools were not used on Chats or Secret Chats user reports, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known CSEA images on Chats and Secret Chats user reports, Telegram referred to the alternative measures it took for known TVE images (see section 5Ai). Telegram also noted that it maintains a dedicated hotline StopCA@telegram.org for reporting any content related to CSAM.

ii. Known CSEA video

In response to questions about hash matching for known CSEA video, Telegram provided the following information

Table X

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram Hash Matching System
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Content in user reports	Yes	

When asked why hash matching tools were not used to detect known CSEA videos in Chats and user reports about Secret Chats, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known TVE videos on Chats and user reports about Secret Chats, Telegram referred to the alternative measures it took for known TVE images and known CSEA images (see section 5Ai).

iii. Sources of CSEA hashes

Telegram reported that it sourced its hashes of known CSEA images and videos from internal databases of hashes of CSEA material that had previously been identified on Telegram and removed by its human moderators. In answer to a question about how often it updated this database, Telegram stated that the

database was updated every time a human moderator removed an item of new, or previously 'unknown', CSEA material from the service.

Telegram also referred again to the reasons it gave for its 'exclusive reliance' on human moderators to compile its TVE hash database (see section 5Aiv).

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing CSEA material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material has already been identified by other online service providers and hashed in extensive shared databases like those run by the IWF or NCMEC.

This means that to the extent that it used hash matching tools, Telegram did not have access to NCMEC's hash database, which contains more than 5 million hashes of verified CSEA material.[20] eSafety notes media reporting that NCMEC and the IWF both claimed to have made past efforts to contact Telegram that went ignored prior to the CEO of Telegram's arrest on 27 August 2024.[21]

iv. Action taken when CSEA hashes are detected

Telegram stated that detections of known CSEA images and videos through hash-matching resulted in the automated removal 'of all users, Communities and publications involved'. Telegram also referred to further information it had provided on the measures it took to address known TVE which eSafety has chosen not to publish to prevent the information being misused.

B. Detecting new CSEA material

i. Text analysis to detect CSEA

In response to questions about technology to detect terms, abbreviations, codes and hashtags indicating likely CSEA (for example grooming, sexual extortion, or the trading and sale of CSEA material), Telegram provided the following information:

Table Y

[20] Google Safety Center, 'NCMEC, Google and Image Hashing Technology', accessed 15 November 2024, URL: https://safety.google/stories/hash-matching-to-help-ncmec/

[21] NBC News, 'Telegram ignored outreach outreach from child safety watchdogs before CEO's arrest, groups say', 28 August 2024, URL: https://www.nbcnews.com/tech/security/telegram-ceo-pavel-durov-child-safety-rcna168266 .

Parts of service	Used text analysis tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[22]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
Profile username	Yes	
Profile description	Yes	
Group username	Yes	
Group description	Yes	
Channel username	Yes	
Channel description	Yes	
Content in user reports	Yes	

In response to why it did not use technology to scan Chats, Secret Chats, private group chats, and private channels for indications of likely CSEA, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to what alternative steps Telegram took to detect known terms, abbreviations, codes or hashtags indicating likely CSEA on these parts of the service, Telegram referred to the alternative steps it took to detect text indicating likely TVE (see section 5Bi). Telegram also referred to the dedicated hotline it maintains for receiving reports about CSAM (see section 9Ai).

ii. Sources of terms, abbreviations, codes, and hashtags

Telegram stated that it sourced phrases, codes, and hashtags indicating likely CSEA from samples of CSEA material that had been removed from Telegram by its human moderators.

[22] In response to a follow-up question seeking the names of the tools Telegram uses to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

iii. Languages covered by language analysis tools

When asked what languages were covered by technology used to detect terms, abbreviations, codes and hashtags indicating likely CSEA, Telegram did not provide a list of languages.

Telegram referred to the answer it gave in response to questions about the languages covered by the tools Telegram used to detect likely TVE in text (see section 5Ci).

iv. New or 'unknown' CSEA images

In response to questions about the detection of new (or 'previously unknown') CSEA images, Telegram provided the following information:

Table Z

Parts of service	Used tools for images?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[23]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why it did not use technology to detect new CSEA images on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect new TVE images (see section 5Bi).

In response to a question about the alternative steps Telegram took to detect new CSEA images on these parts of the service, Telegram referred to measures it

[23] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

took to detect known terms, abbreviations, codes and hashtags that indicate likely CSEA on the service and likely TVE in text (see sections 9Bi and 5Bi).

As noted at section 5Bi, eSafety considers that not using proactive detection tools to identify and review potential CSEA material increases the likelihood that such material will remain undetected and continue to circulate on these parts of the service.

eSafety notes that Chats, Private Group chats, and Private Channels are not E2EE – leaving alternative technical options available for content detection and review by human moderators.

v. New or 'previously unknown' CSEA videos

In response to questions about the detection of new (or 'previously unknown') CSEA videos, Telegram provided the following information:

Table AA

Parts of service	Used tools for videos?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[24]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
Content in user reports	Yes	

In response to why it did not use technology to detect new CSEA videos on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect new TVE images (see section 5Bi).

In response to a question about the alternative steps Telegram took to detect new CSEA videos on these parts of the service, Telegram referred to measures it

[24] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

took to detect known terms, abbreviations, codes and hashtags that indicate likely CSEA on the service and likely TVE in text (see sections 9Bi and 5Bi).

vi. Action taken when new CSEA is detected

Telegram stated that when likely new CSEA material was detected, it was either immediately processed by Telegram's automated tools or sent for human review 'depending on the degree of confidence to which the relevant model is able to issue a judgement, combined with other factors'. eSafety has chosen not to disclose these additional factors due to public safety reasons. If the detection was confirmed, it resulted in the removal of all 'users, Communities and publications involved'. Telegram reported that when new CSEA images and videos were removed, they were then added to Telegram's internal hash database.

C. Blocking links to CSEA material

i. URLs linking to known CSEA

In response to a question about whether Telegram blocked URLs linking to known CSEA, Telegram provided the following information:

Table BB

Parts of service	Blocked URLs linking to known TVE material on other services/websites?	URL sources
Chats	No	N/A
Secret chats (E2EE)	No	
Group chats (public)	No	
Group chats (private)	No	
Channels (public)	No	
Channels (private)	No	
Profile description	No	
Group description	No	
Channel description	No	

In response to why URLs to known CSEA material were not blocked and whether alternative steps were taken to block such URLs, Telegram reiterated that 'focusing its efforts on ML-based classification tends to yield better results when compared to static link blacklists'. Telegram stated that links to harmful material tend to be taken down routinely or hidden behind URL shorteners. Telegram stated that it used proactive detection tools that are trained on previous detections of CSEA material, including material that may contain external links.

Telegram also stated that, as at October 2024, it was 'in the process of joining the Internet Watch Foundation's safety programs involving *inter alia* access to URL lists containing links to known CSAM websites'.

As noted above, eSafety is aware of public statements made by the Internet Watch Foundation (IWF) asserting that prior to the arrest of Telegram's CEO in August 2024, the IWF had made repeated efforts to reach out to Telegram and that Telegram had refused to 'take any of its services to block, prevent, and disrupt the sharing of child sexual abuse imagery'.[25] It is unclear why Telegram did not take the opportunity to work with the IWF sooner.

D. Percentage of CSEA detected proactively

Telegram was asked what percentage of CSEA was detected proactively, compared to CSEA reported by users, trusted flaggers or through other channels for the following parts of its service:

Table CC

Parts of Telegram	Percentage of CSEA detected proactively	Percentage of CSEA reported by users, trusted flaggers or other
Chats	N/A	100%
Secret Chats	N/A	100%
Group chats (public)	71%	29%
Group chats (private)	85%	15%
Channels (public) s 47G(1)(b)	74%	26%
Channels (private)	80%	20%
Voice and video calls (public and private)	N/A*	N/A*
Group video calls (public and private)	'Included in group chats'**	
Stories	65%	45%

[25] NBC News, 'Telegram ignored outreach outreach from child safety watchdogs before CEO's arrest, groups say', 28 August 2024, URL: https://www.nbcnews.com/tech/security/telegram-ceo-pavel-durov-child-safety-rcna168266.

s 47G(1)(b) [redacted]		

* In answer to a follow-up question from eSafety to clarify why its answer was 'N/A' for voice and video calls, Telegram referred to the answer it gave with respect to the percentage of TVE detected proactively and by reports on voice and video calls (see section 4E).

**Telegram stated that its group video call data was included in the relevant group chat statistics because 'information on resulting bans is not stored separately'.

E. Appeals against CSEA-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for CSEA, where Telegram was alerted by automated tools or user reports, and how many of those were successful, Telegram provided the following information:

Table DD

How Telegram was alerted to CSEA	Number of appeals made for accounts banned for CSEA breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for CSEA breach	Number of appeals that were successful for material removed
Automated tools	7,098	573	N/A*	
User reports	2,702	218		

*Telegram stated that because CSEA-related content violations resulted in the users and Communities involved being removed from Telegram, 'It is not generally possible to appeal for reinstatement of removed CSAM materials, so only account appeals are included'. Telegram stated that in some jurisdictions, such as the EU and the EU Terrorist Content Online Regulation, it may receive appeals against content removals from legally mandated contact lines. However, Telegram reported that there were 'no such appeals connected to removal of CSAM content' during the Report Period.

10. Questions about resources, expertise, and human moderation

A. Median time to reach an outcome to a user report of CSEA

Telegram was asked to provide the median time taken to reach an outcome[26] after receiving a user report about CSEA for the following parts of the service:

Table EE

Parts of the service	Reports from users globally	Reports from users in Australia *
Chats	11 hours	11 hours
Secret Chats	11 hours	11 hours
Group chats (public)	10 hours	10 hours
Group chats (private)	10 hours	10 hours
Channels (public)	10 hours	10 hours
Channels (private)	10 hours	10 hours

In response to a question asking how median time was calculated Telegram stated that to calculate these figures it registered 'the net time frames between the submission of each individual report and the moderator's decision in respect of that report'.

* Telegram stated that it 'currently doesn't have the technical means to provide separate statistics by country'.

11. Questions about steps to prevent recidivism

A. Measures and indicators

In response to a question asking if Telegram had measures in place to prevent recidivism for CSEA-related breaches on its service, Telegram responded 'yes' and stated:

> Given the severity of CSAM, any infringement related to it typically results in the permanent removal of related accounts and Communities. Owners of

[26] Defined in the Notice as a calculation from 'the time that a user report is made, to a content moderation outcome or decision, such as removing the content, banning the account, or deciding that no action should be taken.'

infringing groups and channels may also face removal, preventing them from creating new Communities or accounts on Telegram.

Telegram listed a minimal[27] number of indicators that it used to detect users that have previously been banned for CSEA breaches. eSafety has chosen not to publish these indicators to prevent the information being misused.

12. Additional information

In response to an opportunity to provide further information and context to any of its responses to the questions asked in the Notice, Telegram added that:

> Telegram was built to safeguard the privacy of individuals at risk – such as legitimate activists, journalists, and protesters – and preserve their right to private correspondence. While staying true to its core value of user privacy, Telegram actively engages in policy efforts and implements robust moderation tools to address abusive content.
>
> Telegram's exponential growth in the recent years has presented unique moderation challenges due to the sheer volume and diversity of content. Recognizing these challenges, Telegram is continuing to expand its moderation framework while enhancing existing solutions – leveraging advanced software, growing its dedicated teams and fostering key partnerships to mitigate harmful content effectively, as outlined in detail below.

Telegram outlined the following features, tools, and resources it used to address harmful material and activity on its service:

- **'Content review'** – Telegram stated content on the service was reviewed 24/7 through proactive detection, user reports, 'email referrals from users and trusted organizations', and monitoring media stories. Telegram stated that it 'relies on a combination of AI / ML customized recognition tools, manual search strategies, as well as prevention of reappearance of already removed items'. Telegram stated that '[p]ublic content flagged by

[27] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

algorithms is processed, validated, and assigned additional priority if needed, which allows moderators to receive relevant reports and properly sort miscategorized items'.

- **'Limited content discovery'** – Telegram stated that '[b]y design, Telegram does not employ recommendation algorithms or any other form of targeted amplification'. Telegram stated that this means that bad actors cannot exploit Telegram to 'spread harmful content rapidly or to reach a meaningful share of users'. Telegram said this was also true of Telegram Stories.

- **'State-of-the-art Software Solutions'** – Telegram stated that its proactive detection tools had been created by 'world-class engineers', and that it used a combination of hash and pattern matching tools and other 'state-of-the-art AI/ML models that span a wide array of technologies'. Telegram stated that models included:

    - fine-tuned self-supervised multilingual transformer-based language models;
    - fine-tuned vision transformer models;
    - multilingual transformer-based end-to-end ASR systems;
    - multimodal transformer-based models aligned on image-text datasets;
    - multilingual transformer-based large language models; and
    - custom data clustering algorithms.

  Telegram stated that 'several' of these models had been deployed by the end of the Report Period (29 February 2024), but it had 'since significantly expanded its use of AI and ML technologies'.

- **'Trained professionals'** – Telegram stated that its moderators are 'highly trained professionals that undergo regular quality-assurance checks including daily peer examinations'. Telegram stated that although its 'strict selection process ensures that only the most capable and suitable individuals are chosen for a moderator role', it conducted daily assessments of between 1 and 5% of all reports by randomly distributing them to moderators to calculate potential error rates. Telegram stated that 'Moderators with suboptimal error rates, or involved in systematic, gross, or material errors are replaced'. Telegram also stated that it has 'specialized

moderator task groups' responsible for responding to harms such as CSAM and TVE. Telegram stated that these task forces, and its escalation processes, have 'significantly reduced the response time for handling critical reports'.

- **'Key partnerships'** – Telegram cited its collaboration with the Global Center for Combating Extremist Ideology (or, 'Etidal'). Telegram stated that between February 2022 and June 2024, Telegram's partnership with Etidal had resulted in '93,99 million pieces of content related to spreading terrorist ideologies' being removed from the service.

  Telegram stated that it was 'consistently expanding its network amongst industry leaders and international community to stay up-to-day on the latest developments and best practices related to content moderation'. Telegram also provided the following as examples of organisations that its staff regularly engaged with:

  - UK Home Office
  - Etidal
  - EU Internet Forum
  - Europol
  - Ofcom
  - UNSC Counter-Terrorism Committee Executive Directorate

  Telegram stated that:

  > Reports of CSAM, terrorist content and violent propaganda received from trusted organizations are processed within 1 hour.

  Telegram also reiterated that, as at October 2024, it was in the process of joining the Internet Watch Foundation's safety programs to gain access to the IWF's hash lists of known CSEA material.

Telegram summary

Overview

Telegram FZ LLC was asked about its Telegram service.

s 47G(1)(b)

Part 1. Questions in relation to Terrorism and Violent Extremism (TVE)

1. Questions about Telegram's definitions of 'terrorist material and activity' and 'violent extremist material and activity'

A. Terrorist material and activity

In response to a question about how Telegram defines 'terrorist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, Telegram stated:

> For the purposes of its moderation procedures, Telegram is guided by the notion of "terrorist content" that comprises texts, imagery, recordings, and footage promoting and glorifying violence and terrorist ideology, soliciting funds for terrorist causes, instructing or advising on planning or carrying out of terror attacks.

B. Violent extremist material and activity

In response to a question about how Telegram defines 'violent extremist material and activity' or a different but equivalent term for the purposes of its terms of service and community guidelines, Telegram stated that, in its moderation procedures, it defines 'violent extremist content' as:

> texts, imagery, recordings, and footage advocating for violence against a person or a group (i.e., specific threats of physical harm, etc.), as well [sic] instructions for creating and obtaining weapons, explosives and other means of carrying out violent attacks.

Telegram also stated that in its moderation procedures, it defined 'violent content' as:

> [content which] also covers graphic, gruesome or shocking materials, like graphic details of torture—unless such content is clearly communicated for the purposes of news reporting or raising awareness of human rights violations.

2. Prohibiting 'illicit content' on private parts of Telegram

On 18 March 2024, when Telegram was issued the Notice, the Telegram Terms of Service stated:

s 47G(1)(b)

> by signing up for Telegram, you accept our Privacy Policy and agree not to:
>
> - Use our service to send spam or scam users
> - Promote violence on ***publicly viewable*** [emphasis added] Telegram channels, bots, etc.
> - Post illegal pornographic content on ***publicly viewable*** [emphasis added] Telegram channels, bots, etc.[1]

eSafety highlighted this in the Notice, and asked Telegram to specify whether these rules permitted end-users to promote violence or post illegal pornographic content on private parts of the service – namely, Private Channels, Groups, and Secret Chats.

Telegram stated that this was not the case, and that the 'Telegram Terms of Service … apply throughout the app, regardless of chat type.' Telegram stated that it 'doesn't tolerate illicit content' in these parts of the service, and that it will 'take appropriate action within its technical capabilities whenever it becomes aware of such content'.

Telegram stated that the references quoted by eSafety in Telegram's terms of service to publicly viewable parts of the service refer to the areas of its service that Telegram moderators proactively monitored, but that is not to say that Telegram 'tolerate[d] illicit content in private channels, groups, or secret chats'.

Telegram provided the following information about how it detected violative material in the private parts of Telegram where moderators cannot proactively check messages:

[1] Telegram, 'Terms of Service', accessed 9 February 2024, URL: https://telegram.org/tos

- User reporting – Telegram stated that content in private Communities[2] can be reported by users. Telegram also stated that 'regular users, non-registered viewers, and organisations' can report material using in-service reporting options, or through dedicated email addresses.[3] Telegram stated that messages reported by end-users in private Communities (i.e., Channels and Groups) 'are forwarded to moderators', but messages reported in Secret Chats are not (see section 2B).
- Proactive detection measures – Telegram stated that it used 'algorithmic detection measures that prevent abuse' on its service. Telegram referred to its use of hash-matching to detect known TVE and CSEA material on public and private Communities, matched against previously identified TVE and CSEA on Telegram's public content. Telegram also referred to its use of 'automated content detection and manual search strategies tailored to locate Communities engaged in violations of Telegram Terms of Service'.

eSafety notes that responses captured in sections 5 and 9 highlight that there was inconsistent use of proactive detection tools across the 'private' parts of Telegram's service, s 47G(1)(b) Telegram also did not take any external hashes from external organisations which share hashes of terrorism and violent extremism.

s 47G(1)(b)

s 47G(1)(b)

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing TVE material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material has already been identified by other online service providers and hashed in extensive shared databases like those run by the GIFCT or Tech Against Terrorism.

s 47G(1)(b)

A. Moderating 'private' Communities using 'invite links'

Telegram stated in circumstances where a 'private' Community is made accessible to the broader public via an 'invite link', it also changes Telegram's ability to monitor that Community. Telegram gave the example that if an 'invite link' to a

[2] Telegram used the term 'Communities' to refer generally to groups and channels on the service.
[3] Telegram provided the following e-mail addresses for reporting violative material on the service: abuse@telegram.org and StopCA@telegram.org.

private Community is shared on a public part of Telegram or another social media service, then

> the Community is considered to be public for content moderation purposes (thanks to the fact that moderators can follow the link and view the messages within before any user reports are made).

Telegram stated that it routinely detects such 'invite links' when the public channels or groups hosting them are taken down, and removes the associated Communities 'if appropriate based on the content they publish'.

B. Moderating E2EE 'Secret Chats'

Telegram added that it has 'no technical means of verifying the accuracy of user reports regarding content stored' inside Secret Chats because these messages are protected by E2EE. Telegram stated that messages in Secret Chats were not 'forwarded' to moderators when they were reported by an end-user.

Without access to the messages being reported, Telegram reported that it relies on alternative signals or indicators to determine if 'the reported user is not otherwise engaging in harmful or malicious behaviour'. eSafety has chosen not to publish these alternative signals to prevent them being misused.

eSafety notes that there are alternative measures that enable content moderators to review E2EE messages that have been reported by end-users as harmful or otherwise violative. For example, WhatsApp (which is E2EE) has processes in place that enable its moderators to receive the last 5 messages sent to an end-user from the account they are reporting.[4] eSafety considers that having measures in place that enable moderators to review the material being reported by end-users is key to ensuring that these reports can be responded to effectively.

C. Changes to Telegram's FAQs in September 2024

On 18 March 2024, when Telegram was issued the Notice, Telegram's 'frequently asked questions' web page stated:

> Q: There's illegal content on Telegram. How do I take it down?

s 47G(1)(b) [redacted]

[4] WhatsApp, 'About reporting and blocking someone on WhatsApp', accessed 15 October 2024, URL: https://faq.whatsapp.com/414631957536067/

> All Telegram chats and group chats are private amongst their participants. We do not process any requests related to them.
>
> But sticker sets, channels, and bots on Telegram are publicly available. If you find sticker sets or bots on Telegram that you think are illegal, please ping us at abuse@telegram.org.

eSafety pointed to this in the Notice and asked Telegram to specify the steps it was taking to comply with the Expectations in relation to the safe use of private chats and private group chats, including after a user report was made about illegal or harmful content, given this statement .Telegram responded by referring to its use of proactive moderation tools, user reporting tools, and specialised moderation teams to address abuse on its service. Telegram stated

> If a report is confirmed, Telegram acts with due regard to all known circumstances. Disseminating terrorist and violent content or CSAM on any parts of Telegram service leads to permanent removal of associated accounts and Communities.

Regarding the statement on its 'frequently asked questions' page that it would not process removal requests for 'illegal content' on certain parts of its service, Telegram stated this information was outdated and was the result of statements about Telegram's stance on copyright infringement having been mistakenly copied to a section dealing with Telegram's stance on illegal content. Specifically, Telegram stated

> As at February 29, 2024, certain portions of the Telegram FAQ may have featured outdated information, some of which was updated in September 2024 (with further updates planned in the coming months). This included the item quoted under "Context" in this question, which mistakenly included text from an earlier revision of the FAQ. Namely, the mention of "not processing requests regarding private chats" had been erroneously copied from the FAQ section related to copyright infringement, where it is present to this day.[5]...
>
> The quoted paragraph was aimed at law-abiding Telegram users who rely on Telegram for the privacy for their personal communication. It was meant to emphasize that, as Telegram moderators cannot proactively inspect the

[5] Telegram, 'Telegram FAQ - Q: A bot or channel is infringing on my copyright. What do I do?', URL supplied by Telegram on 13 September 2024, URL: https://telegram.org/faq#q-abot-or-channel-is-infringing-on-my-copyright-what-do-i-do

private messages of its users, they cannot act on unsupported requests which do not rely on reporting mechanisms.

The text was never meant to imply that Telegram's Terms of Service could be violated in private chats. This is evidenced by the fact that users could always report both incoming private and secret chats, and messages in private groups and channels to moderators.

eSafety notes that online media outlet, The Verge, first reported this change to Telegram's FAQs page as having occurred on 6 September 2024.[6]

eSafety also notes that the Internet Archive's Wayback Machine shows that the item Telegram states was 'mistakenly included text from an earlier version of the FAQ' was present on Telegram's FAQ page as far back as 15 March 2016, and that this item appears to pre-date any reference to copyright infringement in Telegram's FAQs.[7]

3. Thresholds/criteria to determine action on TVE breaches

Telegram was asked if it had criteria or thresholds in place to determine what action would be taken when TVE was identified on Telegram. Telegram provided the following information:

Table A

Actions taken on accounts or content when TVE was identified	Criteria/thresholds reported for Telegram
Permanent account ban	Telegram stated the following: • Disseminating material that calls for violence in the form of text, image, recordings, footage or otherwise. Telegram specified this means material 'like concrete and specified threats of physical harm'. • Disseminating material that is gruesome or shockingly graphic. Telegram gave such examples as 'graphic details of torture,

[6] The Verge, 'Telegram changes its tone on moderating private chats after CEO's arrest', 6 September 2024, URL: https://www.theverge.com/2024/9/5/24237254/telegram-pavel-durov-arrest-private-chats-moderation-policy-change

[7] Internet Archive Wayback Machine, 'Telegram FAQ – 15 March 2016', accessed 16 October 2024, URL: http://web.archive.org/web/20160315182715/https://telegram.org/faq

	accident photos' or material that 'glorif[ies] or promote[s] violent or terrorist ideologies'. • Soliciting funds for terrorist organisations or causes. • Owning or being an administrator of a Community involved in the above activities.
Account strikes	Telegram stated that if a Community, or an account belonging to a 'journalist' or 'researcher', reposts TVE with the intention of sharing 'legitimate scientific research, historical records, or news', then Telegram may either: • grant an exception; or • apply up to two warnings before terminating the Community or account. Telegram stated the decision on enforcement depends on the 'severity, purpose and relevance of the posted content under applicable law'.

Telegram also stated that 'where appropriate', it will remove publications in Communities and remove associated groups and channels.

s 47G(1)(b)

s 47G(1)(b)

4. Questions about reporting of TVE

A. In-service reporting of TVE on different parts of Telegram

In response to questions about whether users could report instances of TVE to Telegram within the service (as opposed to navigating to a separate webform or email address), Telegram responded:

Table B

Parts of the service	In-service reporting option?	Reporting category
Chats	Yes	'Block user > Report Spam'*
Secret Chats	Yes	
Group chats (public)	Yes	'Violence'
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Voice calls	No**	N/A
Video calls	No**	

*In response to a follow up question from eSafety, which highlighted that in eSafety's testing on the Telegram iOS app, for Chats and Secret Chats the option to 'Report spam' was not present in all cases. Telegram subsequently clarified that the 'Block + Report Spam' reporting flow is only available when the Chat or Secret Chat is 'initiated by non-contacts and strangers'. eSafety understands that when an end-user wishes to report a message from an account they have already added as contact, s 47G(1)(b)

s 47G(1)(b)

Telegram provided the following reason for this discrepancy in reporting functionality

> In the extremely unlikely event that a user's friend or acquaintance began sending them TVE content, Telegram contends that it would be more reasonable and effective for said user to contact authorities directly, providing all relevant proof and contact information.

eSafety considers that limiting reporting tools to scenarios where the account sending harmful or violative material is not a contact of the end-user risks preventing Telegram from identifying and preventing bad actors from continuing to perpetrate harm on the platform even after they have been blocked by an end-user on the service.

Telegram stated that the single reporting option 'Block + Report Spam' for private and Secret Chats intended to simplify the user experience and minimise the length of time and number of interactions necessary for a user to end the chat. Telegram stated that 'once the report is processed by moderators, it is escalated as necessary – including via AI / ML if appropriate'.

eSafety notes that in response to other questions in the Notice, Telegram stated that it had no means of accessing messages reported by end-users from Secret Chats (see Section 2). Instead, Telegram stated it relies on alternative signals to assess and prioritise reports made about material in E2EE parts of the service.

eSafety notes that this may limit Telegram's ability to review, assess, prioritise, and respond to reports about harmful and illegal material or activity occurring in Telegram's Secret Chats.

**Telegram's original response to the Notice stated that end-users could make in-service reports about voice calls and video calls using a 'Violence (via the community info section)' reporting category. In response to a follow-up question from eSafety, Telegram subsequently stated that in-service reporting for voice and video calls was not available during the Report Period. Instead, Telegram stated that 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

s 47G(1)(b)

s 47G(1)(b)

B. Reporting of TVE by third party services that use Telegram's API

eSafety asked whether Telegram had minimum safety requirements for third party services that use Telegram's APIs to access its service. Telegram responded that it did have minimum safety requirements and that this includes the requirement for user reporting functions on third party apps to notify Telegram of breaches of its terms of service.

Telegram provided links to its Telegram API Terms of Service[8] and Security Guidelines[9], and stated that the API Terms of Service expect that third-party services make available all basic functionalities of the Telegram service, including the reporting tools, and that third-party services are accountable for ensuring that these features function correctly.

Telegram stated that it would 'from time to time re-confirm that it correctly receives user reports from the most popular third-party clients, including reports of potentially terrorist and violent content'. Telegram also stated that third-party services that fail to comply with Telegram's Terms of Service are 'routinely flagged for removal to third-party app stores and blocked from accessing Telegram's core APIs'.

[8] Telegram, 'Telegram API Terms of Service', URL supplied by Telegram on 13 September 2024, URL: https://core.telegram.org/api/terms
[9] Telegram, 'Security Guidelines', URL supplied by Telegram on 13 September 2024, URL: https://core.telegram.org/mtproto/security_guidelines

C. Reporting mechanisms for other entities to report TVE

In answer to questions about having separate reporting mechanisms for certain entities to report TVE, Telegram stated that it did have dedicated reporting mechanisms for:

- Law enforcement;
- Trusted Flaggers;
- Regulatory and public authorities; and
- 'International organizations'.

Telegram stated that these reporting mechanisms enable '[f]aster processing times whenever possible; processing by a dedicated team member / task group; deeper review if needed'. Telegram pointed to its collaboration with the Global Center for Combating Extremist Ideology, or 'Etidal', as an example of a specialised reporting mechanism. Telegram stated that between February 2022 and June 2024, Telegram removed '93,99 million pieces of TVE content' through its collaboration with Etidal.

Telegram stated that during the Report Period, it did not have a separate reporting mechanism for civil society groups to report TVE to the service. Telegram reported

> While engagement with these groups regarding content reporting has been minimal as at 29 February 2024, Telegram recognizes the potential benefits of collaborating with more external experts. To this end, Telegram is actively considering the introduction of dedicated contact points and other initiatives to enhance its responsiveness to valid concerns raised by civil society groups and remains open to dialogue in furtherance of that goal.

D. Percentage of TVE sent for human review

Telegram was asked to provide the percentage of TVE reports it sent for human review and the criteria and thresholds used to determine when reports were sent for human review.

Table C

Percentage of user reports of TVE sent for human review	75%

Criteria and thresholds used to determine when a user report is sent for human review	Telegram stated some reports were not reviewed by humans because: • Reported content had already been removed by proactive measures. • Reported content had already been removed because the channel/group it was posted in was removed.
Percentage of TVE detected through automated tools sent for human review	65%
Criteria and thresholds used to determine when a report of TVE is detected through automated tools is sent for human review	Telegram provided a select number of scenarios where content would be sent for human review. eSafety has chosen not publish these specific scenarios to prevent this information being misused. Telegram stated that some of these criteria were intended to prevent Telegram's systems from automatically banning 'researchers, human rights activists, legitimate news sources etc' as well to prevent bad actors from attempting to 'silence' Telegram communities by deliberately posting violative material in them as 'abusive spam'. Telegram also stated that when its automated tools detect material that is a '100% hash match', in a private Community, human moderators are notified 'but do not receive a copy of the media item itself'. In response to other questions in the Notice, Telegram stated that detections of hash matched TVE material result in resulted in the automated removal 'of all users, Communities and publications involved' except for 'Communities or users that are likely to yield false positives' (see section 5Av).

E. Percentage of TVE detected proactively

Telegram was asked what percentage of TVE was detected proactively, compared to TVE reported by users, trusted flaggers or through other channels for the following parts of its service:

Table D

Parts of Telegram	Percentage of TVE detected proactively	Percentage of TVE reported by users, trusted flaggers or other
Chats	N/A	100%
Secret Chats	N/A	100%
Group chats (public)	67%	33%
Group chats (private)	82%	18%

Channels (public)	69%	31%
Channels (private)	79%	21%
Voice and video calls (public and private)	N/A*	N/A*
Group video calls (public and private)	'Included in group chats'**	
Stories	60%	40%

* In answer to a follow-up question from eSafety to clarify why its answer was 'N/A' for voice and video calls Telegram stated that voice and video calls could not be directly reported by end-users using in-service reporting tools. Instead, 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

**Telegram stated that its video group call data was included in the relevant group chat statistics because 'information on resulting bans is not stored separately'. In response to other questions in the Notice, Telegram stated that it did not use any proactive detection tools to detect livestreamed TVE in group video calls. eSafety therefore understands that 100% of any TVE detected in group video calls during the Report Period was reported by users.

F. Appeals against TVE-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for TVE, where Telegram was alerted by automated tools or user reports, and how many of those were successful, Telegram provided the following information:

Table E

How Telegram was alerted to TVE	Number of appeals made for accounts banned for TVE breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for TVE breach	Number of appeals that were successful for material removed
Automated tools	3,420	110	N/A*	
User reports	1,107	26		

*Telegram stated that because TVE-related content violations result in the users and Communities involved being removed from Telegram, 'It is not generally possible to appeal for reinstatement of removed TVE materials, so only account appeals are included'. Telegram stated that in some jurisdictions, such as the EU and the EU Terrorist Content Online Regulation, it may receive appeals against content removals from legally mandated contact lines. However, Telegram

reported that there were 'no actionable appeals connected to removal of terrorist, violent or extremist content' during the Report Period.

5. Questions about proactive detection

s 47G(1)(b)

s 47G(1)(b)

In response to various questions in the Notice, Telegram stated that when TVE material was confirmed, the material was removed along with 'users, Communities and publications involved'.

s 47G(1)(b)

A. Detecting known material using hash-matching

i. Known TVE images

In response to questions about hash matching for known TVE images, Telegram provided the following information:

Table F

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram Hash Matching System
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	

User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why hash matching tools were not used on Chats or Secret Chats user reports, Telegram stated that Telegram was 'founded on the principle of defending user privacy and their right to private communication' and that 'this commitment prioritizes user privacy above all'. Telegram stated that because of this commitment to user privacy

> encrypted contents of private chats are always protected, ensuring that the confidentiality of private correspondence is never compromised.

eSafety notes that Telegram stated that it does use hash matching tools on other 'private' parts of the service – namely, private groups and private channels. eSafety further notes that Chats, Private Groups, and Private Channels all use the same form of encryption – which is not E2EE.

It is unclear to eSafety why tools capable of detecting known TVE, verified as harmful and/or violative by Telegram's own trust and safety staff, are not being used on Chats given Telegram stated that they are used on other private parts of Telegram's service, namely private groups and private channels. In relation to Secret Chats user reports, as noted at section 2B, alternative methods also exist which could enable hash-matching tools to review content reported in E2EE messages.

In response to a question about the alternative steps Telegram took to detect known TVE images on Chats and Secret Chats user reports, Telegram stated that it relied on users reporting messages via its 'Block + Report Spam' reporting tool (which as eSafety notes above, appears to only exist for messages from users that have not been added as contacts by the reporting users). Telegram stated these reports are 'processed by Telegram's tools and moderators…including via AI / ML if appropriate'. Telegram also stated that it

> employs extensive automated rate-limiting and spam-preventive measures, ensuring that no user or software is able to share content in bulk or to a significant number of users via any chat, irrespective of the nature of said content. In so doing, by design and without compromising user privacy, Telegram prevents malicious actors from effectively using its private messaging component to spread their messages.

ii. Known TVE video

In response to questions about hash matching for known TVE video, Telegram provided the following information:

Table G

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram Hash Matching System*
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Content in user reports	Yes	

*In response to a follow-up question, Telegram subsequently stated that it considered that 'there is no material difference in the way in which video and image hashing is performed on a technical level'.

In response to why hash matching tools are not used to detect known TVE videos in Chats and Secret Chats user reports, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known TVE videos on Chats and user reports about Secret Chats, Telegram referred to the alternative measures it took for known TVE images (see section 5Ai).

iii. Known TVE written material

In response to questions about hash matching for known TVE written material on Telegram, such as manifestos or text promoting, inciting, or instructing in TVE, Telegram provided the following information:

Table H

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	No	N/A

Group chats (private)	No	N/A
Channels (public)	No	N/A
Channels (private)	No	N/A
Stories	No	N/A
Content in user reports	Yes	Internal Telegram Hash Matching System*

*In response to a follow-up question, Telegram stated that it considered that 'there is no material difference in the way in which text and image hashing is performed on a technical level'.

In response to why hash matching tools are only used to detect TVE written material in content referred to Telegram in user reports, Telegram stated

> Relying on hash matching on all text content at scale is generally not advisable as any one message can be expressed in numerous ways across different languages and formats. Instead, Telegram relies on ML models finetuned on known TVE written material.... Conversely, hash matching on reported content provides a useful and efficient preliminary layer to the existing moderation pipeline.

In response to a question about the alternative steps Telegram took to detect known TVE written material on chats, secret chats, private group chats, and private group channels, Telegram referred to the measures it took for known TVE images on Chats and Secret Chats (section 5Ai). For 'Other chats', which eSafety assumes refers to Public Group Chats and Public Channels, Telegram stated that it used machine learning models 'finetuned on known TVE written material to check a subset of text messages sampled from all relevant chats according to reasonable criteria'. eSafety has chosen not to publish these criteria to prevent the information being misused.

iv. Sources of TVE hashes

Telegram reported that it sourced its hashes of known TVE images, videos, and text from **internal databases of hashes of TVE material that had previously been identified on Telegram and removed by its human moderators.** In answer to a question about how often it updated this database, Telegram stated that the database was updated every time a human moderator removed an item of new, or previously 'unknown', TVE material from the service.

Telegram noted that its '[e]xclusive reliance' on human moderators to compile its TVE hash database is designed to mitigate risks of 'circular data pollution' posed by automated systems being trained on decisions by previous automated systems.

Telegram stated that it 'purposefully avoids reintroduction of machine-labeled or synthetic data into datasets, because such measures can degrade the quality and reliability of the datasets'.

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing TVE material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material has already been identified by other online service providers and hashed in extensive shared databases like those run by the GIFCT or Tech Against Terrorism.

v. Action taken on known TVE

In response to questions about what action was taken when known TVE images and video were detected by its tools, Telegram stated that it resulted in the automated removal 'of all users, Communities and publications involved' except for 'Communities or users that are likely to yield false positives'. Telegram provided further information on the measures it took to address known TVE which eSafety has chosen not to publish to prevent the information being misused.

Telegram stated that when hashes of known TVE written material were detected in user reports, Telegram removed the material and the user who posted it, 'unless there is reason to believe the match may be a false positive … in which case a human moderator reviews the match and takes action accordingly.'

B. Detecting new TVE material

i. New or 'unknown' TVE images

In response to questions about the detection of new (or 'previously unknown') TVE images, Telegram provided the following information:

Table I

Parts of service	Used tools for images?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A

Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[10]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

When asked why it did not use technology to detect new TVE images on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using hash matching tools to detect known TVE images (see section 5Ai).

In a follow-up question to Telegram, eSafety noted that Telegram had stated that it used tools to detect *known* TVE on Private Group Chats and Private Channels. In light of this, eSafety asked Telegram to provide the reason it did not use tools to detect *new* TVE images on these parts of the service. Telegram stated that the 'technical architecture and access rules of private groups and channels' prevent anyone who is not a member of those groups from accessing them and seeing the content being shared inside. Telegram reported that moderators could only access such content when it was reported by an end-user, or the community became accessible via a public invite link. Telegram stated that it used hash matching tools to detect known TVE on these parts of the service because when its tools detected the material, moderators would receive a notification that a 100% match with violative content had occurred – rather than the actual image or video being disclosed to them for specific review. Telegram stated that it considered that this process should not be applied to detections of new TVE material because

> Telegram contends that notifying moderators of such matches would be insufficient, as matches for new content, while accurate, cannot be verified with absolute certainty without checking the content itself. Even assuming that such matches were taken at face value, moderators would then be unable to process potential appeals by the deleted accounts.

[10] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

eSafety considers that not using proactive detection tools to identify and review potential TVE material increases the likelihood that such material will remain undetected and continue to circulate on these parts of the service.

eSafety notes that Chats, private group chats, and private channels are not E2EE – This means technical options are available for content detection and review by human moderators. Telegram has stated it uses such tools on other parts of its service.

In response to a question about the alternative steps Telegram took to detect new TVE images on these parts of the service, Telegram referred to the measures it took to detect likely TVE in text (see section 5Biii).

ii. New or 'previously unknown' TVE videos

In response to questions about the detection of new (or 'previously unknown') TVE videos, Telegram provided the following information:

Table J

Parts of service	Used tools for videos?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[11]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
Content in user reports	Yes	

In response to why it did not use technology to detect new TVE videos on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect likely TVE in images (see section 5Bi).

[11] In response to a follow-up question seeking the names of the tools Telegram uses to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

In response to a question about the alternative steps Telegram took to detect new TVE videos on these parts of the service, Telegram referred to measures it took to detect likely TVE in text (see section 5Biii).

iii. Text analysis to detect TVE

In response to questions about technology to detect phrases, codes or hashtags, indicating likely TVE in text (for example manifestos or text promoting, inciting, instructing TVE), Telegram provided the following information:

Table K

Parts of service	Used text analysis tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[12]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
Profile username	Yes	
Profile description	Yes	
Group username	Yes	
Group description	Yes	
Channel username	Yes	
Channel description	Yes	
Content in user reports	Yes	

In response to why it did not use technology to scan Chats, Secret Chats, Private Group Chats, and Private Channels for indications of likely TVE in text, Telegram referred to the reasons it gave for not using hash matching tools to detect known TVE images (see section 5Ai).

In response to what alternative steps Telegram took to detect known phrases, codes, or hashtags indicating likely TVE on these parts of the service, Telegram

[12] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

stated that it provided reporting options for users to report such material on these parts of the service. Telegram noted that when users report such TVE material in 'private groups and channels', this results in the material being forwarded to Telegram moderators for their review. However, for what Telegram describes as 'private 1-on-1 chats' (i.e. chats and Secret Chats), Telegram reiterated that user reports in these parts of the service are 'processed by Telegram's tools and moderators...including via AI / ML if appropriate'.

Telegram also referred to the 'automated rate limiting and spam-preventative measures' it used to prevent bad actors from spreading known TVE material in private messages (see section 5Ai).

iv. Sources of phrases, codes, and hashtags

Telegram stated that it sourced phrases, codes, and hashtags indicating likely TVE from 'samples ... based on content items' that had been removed from Telegram by its human moderators.

v. Action taken when new TVE was detected

In response to questions about what action was taken when new TVE images, video, and likely TVE in text was detected by its tools, Telegram stated that the material was sent for human review and if the content was confirmed as TVE it resulted in the removal of 'users, Communities and publications involved'. Telegram reported that following removal, new TVE images and video were then added to Telegram's internal hash database so they could be actioned in the future as known TVE.

Telegram provided further information on the measures it took to address new TVE which eSafety has chosen not to publish to prevent the information being misused.

Telegram also noted that when a Community is removed, human moderators review the most common search terms that were used to find that Community in order for them to be 'possibly removed from Telegram's public search to limit future spread and reach of similar content'.

In answer to a question asking if the detection of phrases, codes, or hashtags indicating likely TVE in text resulted in Telegram blocking these words or phrases to users searching for them, Telegram responded that 'yes', it blocked 'certain keywords or text patterns' from its search results.

C. Languages covered by language analysis tools

i. Detecting TVE in text

In response to questions about the languages covered by Telegram's language analysis tools, Telegram did not provide a list of languages.

Telegram stated that its models and tools for detecting TVE text 'perform reasonably well in most languages', and that these technologies 'aim to abstract away the concept of language when deriving embeddings from text'.

In response to follow-up questions from eSafety, Telegram stated

> Telegram did not maintain a specific list of all languages included in the training sets of the underlying models, which is why it was unable to provide a list relevant for the Report Period. Telegram must note that a significant share of messages in its datasets contain text in multiple languages (i.e., multiple languages within the same message). As well, messages are often too short to automatically identify a specific language with absolute certainty.

ii. Detecting TVE in video

In response to a question about the languages covered by the tools Telegram used to detect new TVE in video, Telegram did not provide a list of languages.

Telegram stated the tools did not 'have particular regard for the specific language of the content in question' but focussed on detecting patterns the model had previously learned to associate with TVE from an existing dataset.

In response to follow-up questions from eSafety, Telegram referred to the answer it gave to eSafety's follow-up question regarding the languages covered by the tools Telegram used to detect likely TVE in text (see section 5Ci).

D. Livestreamed TVE

i. Detecting livestreamed TVE

The Notice specified that livestreaming includes one-on-one video calls and video calls where one or more multiple people stream material to a group of any size.

In response to questions about the measures Telegram had in place to detect the livestreaming of TVE on its service, Telegram provided the following information:

Table L

Parts of service	Measures in place to detect TVE in livestreams?	Interventions used	Name of tools used
Group video calls	No	N/A	N/A
Channel livestreams	No		

When asked why it did not have any measures in place to detect livestreamed TVE, Telegram stated

> While livestreaming functionalities are supported on Telegram, they represent a generally insignificant share of the service's overall usage, particularly so as it concerns the spread of harmful content. As such, Telegram finds that immediate user reports already provide reliable coverage to detect and address such incidents effectively.

ii. Reducing the likelihood of livestreamed TVE

In response to questions about the steps taken by Telegram to reduce the likelihood that TVE could occur in livestreams, Telegram stated that it used the following measures:

- Restrictions for those who have previously violated terms of service or community guidelines/standards.
- User reports – Telegram stated that it 'relies on immediate user reports' to detect livestreamed TVE **but did not indicate whether or how it prioritises reviews of reports of livestreamed content.** In response to other questions in the Notice, Telegram stated that it did not provide in-service reporting tools for video calls (see section 4A) during the Report Period.

s 47G(1)(b)

E. Blocking links to TVE material

i. Detection and sources of URLs

Telegram was asked about its use of lists or databases to proactively detect and block URLs linking to TVE on other platforms. Specifically, Telegram was asked about:

- Known URLs linking to websites/services operated by individuals/organisations dedicated to the creation, promotion, or dissemination of TVE
- URLs linking to known TVE material on other services/websites (which may not be dedicated to TVE)
- Join-links to groups, Channels, communities, or forums on other services that were known to be associated with TVE.

Table M

Parts of service	Blocked URLs to websites/services dedicated to TVE?	Blocked URLs linking to known TVE material on other services/websites?	Blocked join-links to groups/channels on other services known to be associated with TVE?	URL sources
Chats	No	No	No	N/A
Secret chats (E2EE)	No	No	No	
Group chats (public)	No	No	No	
Group chats (private)	No	No	No	
Channels (public)	No	No	No	
Channels (private)	No	No	No	
Profile description	No	No	No	
Group description	No	No	No	
Channel description	No	No	No	

When asked why URLs to TVE material were not blocked and whether alternative steps were taken to block URLs, Telegram stated that 'focusing its efforts on ML-based classification tends to yield better results when compared to static link blacklists'. Telegram stated that links to harmful material tended to be 'routinely taken down by all hosting providers and tend to either rotate constantly or be hidden behind URL shorteners, proxies etc'. Telegram also stated that it used Internal Telegram AI and Machine Learning Models (see Section 12) that are trained on previous detections of TVE material, including material that may contain external links.

F. Off-platform monitoring

Telegram was asked if it used off-platform monitoring[13] either provided internally or by third-party services, to identify accounts, groups or channels on its service that were dedicated to TVE. Telegram stated it performs '[e]xtensive monitoring' of media sources as well as reviewing referrals sent by 'non-registered users and trusted organizations' to Telegram via email.

6. Questions about resources, expertise, and human moderation

A. Trust and Safety

i. Trust and Safety and other staff

Telegram was asked to provide the number of staff that were employed or contracted by Telegram to carry out certain functions at the end of the Report Period.

Table N

Category of staff	Number of staff*
Engineers employed by Telegram focussed on trust and safety	5
Content moderators employed by Telegram	0
Content moderators contracted by Telegram	150
Trust and safety staff employed by Telegram (other than engineers and content moderators)	4

* Telegram stated that these figures represented the number of staff who 'may from time to time be involved with decisions regarding content or reports from Australia and do not reflect or approximate the total number of global content

[13] Monitoring of activity on other services.

moderation and trust and safety personnel contracted by Telegram.' Telegram also stated that Australian end-users make up less than 0.2% of its monthly active users.

eSafety notes that Telegram did not provide its global resourcing. In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

eSafety notes that TVE is a global harm and the resources that a service has in place to respond to 'content or reports' internationally is highly relevant to the online safety of Australians, and implementation of the Expectations.

s 47G(1)(b) [redacted]

ii. Trust and Safety dedicated to minimising TVE

In response to a question asking if Telegram had a dedicated trust and safety team responsible for minimising TVE on the service, Telegram answered 'yes', reporting that '[t]he relevant team members are high-performing professionals trained in team management, threat mitigation and legal affairs'. Telegram stated that this team was responsible for 'overseeing content moderation and reviewing reports from external stakeholders, such as trusted flaggers and international organizations'.

Telegram provided the following information about the composition of its team:

Table O

Name of role/area of expertise	Number of staff	Number of contractors
Trust and Safety managers*	4**	0

* Telegram stated that it did not have specific titles for these positions because Telegram's 'hierarchy is informal and horizontal'.

** Telegram stated that this figure was specific to 'staff that may from time to time be involved in decisions regarding content or reports from Australia and do not reflect or approximate the total number of global trust and safety personnel contracted by Telegram'.

In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be

involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

s 47G(1)(b)

[redacted]

iii. Surge teams to respond to a TVE crisis

Telegram was asked if it had a surge team(s) to respond to TVE crises, such as a livestreamed attack with content disseminated on the service. Telegram answered 'yes' and stated that it 'maintains crisis mitigation protocols in accordance with industry standards'. Telegram stated that on occasion it could establish task forces of '2-3 people tailored to resolving specific situations, e.g., appearance of terrorist or violent content in significant quantities'.

Telegram provided the following information about the composition of this team:

Table Q

Name of role/area of expertise	Number of staff	Number of contractors
Trust and Safety managers and contractors*	3**	13**

* Telegram stated that it did not have specific titles for these positions because Telegram's 'hierarchy is informal and horizontal'.

** Telegram stated that these figures were specific to 'staff that may from time to time be involved in decisions regarding content or reports from Australia and do not reflect or approximate the total number of global trust and safety personnel contracted by Telegram'.

In response to follow-up questions from eSafety seeking the total numbers of staff in the categories, rather than the numbers that 'may from time to time be involved with decisions regarding content or reports from Australia', Telegram did not provide the information.

s 47G(1)(b)

[redacted]

B. Languages human moderators operate across

In response to a question about the languages that its human moderators operated across (both employees and contractors), Telegram provided the following:

Table R

Languages covered by employees (all languages)	Languages covered by contractors (all languages)[14]	
N/A*	• English • Amharic • Arabic • Azerbaijani • Bulgarian • Chinese (traditional and simplified) • Croatian • Czech • Danish • Estonian • Farsi • Filipino • Finnish • French • Georgian • German • Greek • Hindi • Icelandic • Indonesian • Italian • Japanese	• Kazakh • Korean • Kyrgyz • Luganda • Lunyakore • Lusoga • Malay • Moldavian • Norwegian • Polish • Portuguese (Brazil) • Portuguese (Europe) • Romanian • Russian • Serbian • Shona • Spanish • Swahili • Swedish • Tajik • Turkish • Ukrainian • Urdu

[14] Telegram also advised that since the Report Period, it had expanded the languages covered by its contracted content moderators by adding Afrikaans, Bengali (Bangladesh), Chichewa (Zambia), Dhivehi (Maldives), Dutch, Gujarati, Kabyle (Algeria), Kinyarwanda, Lithuanian, Macedonian, Punjabi, Sinhalese (Sri Lanka), and Thai.

		• Uzbek • Yoruba

*Telegram stated that '[a]ll ordinary moderators' on Telegram are contractors.

eSafety notes that the top 5 languages, other than English, spoken in Australian homes are Arabic, Cantonese, Mandarin, Vietnamese and Punjabi.[15] Telegram's human moderators do not cover Vietnamese or Punjabi.

s 47G(1)(b)

C. Median time to reach an outcome to a user report of TVE

Telegram was asked to provide the median time taken to reach an outcome after receiving a user report about TVE for the following parts of the service:

Table S

Parts of the service	Reports from users globally	Reports from users in Australia *
Chats	18 hours	18 hours
Secret Chats	18 hours	18 hours
Group chats (public)	15 hours	15 hours
Group chats (private)	15 hours	15 hours
Channels (public)	15 hours	15 hours
Channels (private)	15 hours	15 hours

In response to a question asking how median time was calculated Telegram stated that to calculate these figures it registered 'the net time frames between the submission of each individual report and the moderator's decision in respect of that report'.

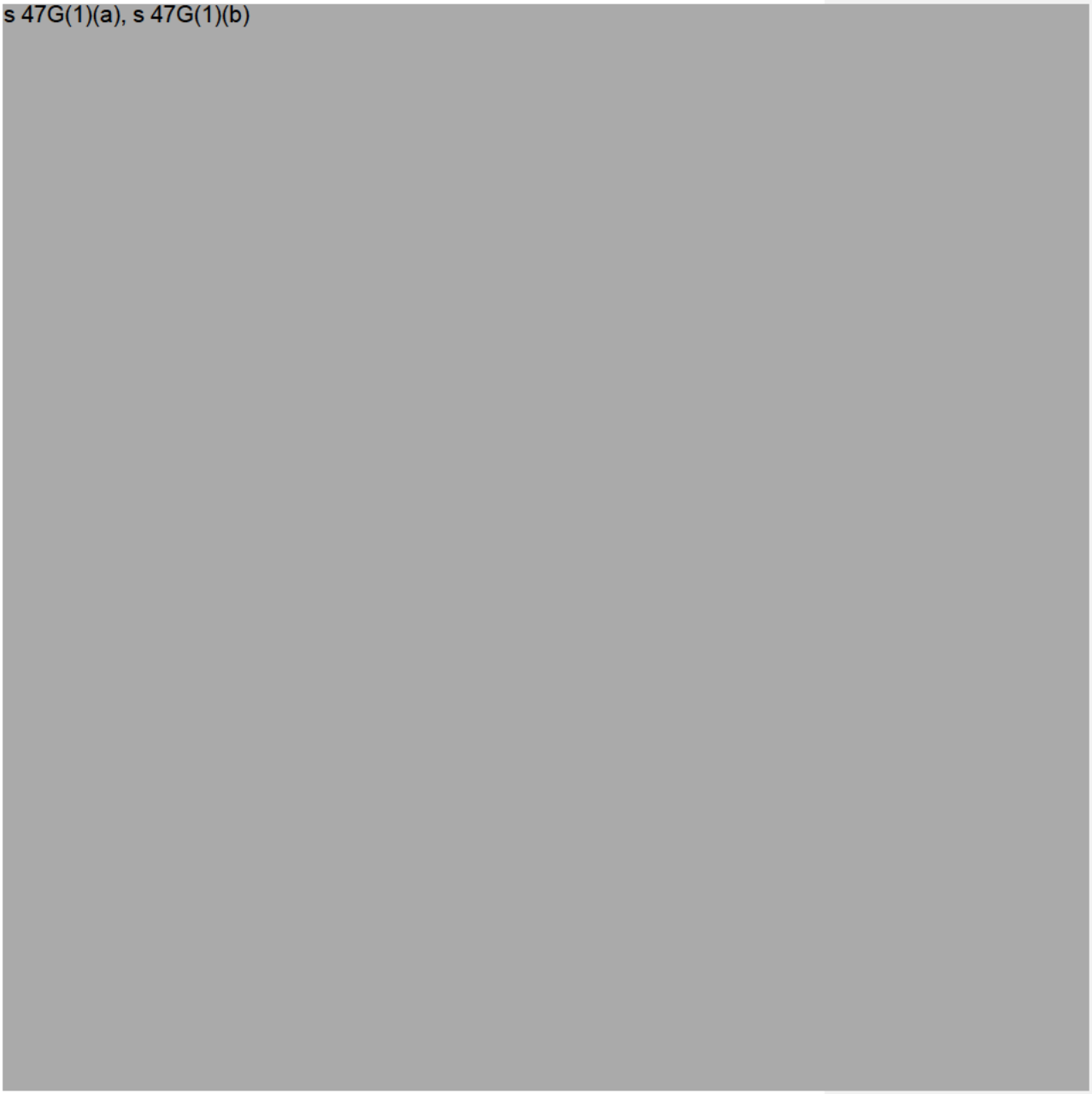
* Telegram stated that it 'currently doesn't have the technical means to provide separate statistics by country'.

It is unclear to eSafety how Telegram determines that it has reached an outcome for Secret Chats when it has stated that it does not review the

[15] Australian Bureau of Statistics, 'Cultural diversity: Census', 28 June 2021, URL: https://www.abs.gov.au/statistics/people/people-and-communities/cultural-diversity-census/latest-release#:~:text=Top%205%20languages%20used%20at,Punjabi%20(0.9%20per%20cent).

contents of the messages being reported on this part of the service (see section 2B).

s 47G(1)(a), s 47G(1)(b)



s 47G(1)(a), s 47G(1)(b)

s 47G(1)(b)

7. Questions about steps to prevent recidivism

A. Measures and indicators

In response to a question asking if Telegram had measures in place to prevent recidivism for TVE-related breaches on its service, Telegram responded 'yes' and

[16] Telegram, 'Channels FAQ', accessed 12 November 2024, URL: https://telegram.org/faq_channels?setln=en
[17] Telegram, 'Telegram FAQ', accessed 12 November 2024, URL: https://telegram.org/faq
[18] Telegram, 'Supergroups 10,000 admin tools and more', accessed 12 November 2024, URL: https://telegram.org/blog/admin-revolution?setln=tr

listed a minimal[19] number of indicators that it used to detect users that have previously been banned for TVE breaches. eSafety has chosen not to publish these indicators to prevent the information being misused.

B. Preventing banned TVE groups and channels from being recreated

In response to a question about the measures Telegram took to prevent banned TVE groups and channels from being recreated, Telegram stated that '[o]wners and administrators of infringing Communities may also face removal alongside the Communities themselves, preventing them from creating new Communities or accounts on Telegram'. Telegram also referred to a minimal number of signals it may use to 'routinely detect' Communities that bear similarities to previously banned Communities which eSafety has chosen not to publish to prevent the information being misused.

C. Applying TVE-related bans to associated accounts

Telegram was asked, when it took action against an account for a TVE-related breach, whether it applied bans to associated accounts. eSafety defined 'associated accounts' as 'other users who are associated with the banned user'. Telegram answered 'yes' and stated that when it identified a user disseminating TVE material, it reviewed 'further reports linked to this user, as well as to any Communities which the user owns or administrates'. Telegram stated that any Communities found to be involved in disseminating TVE would also be removed.

s 47G(1)(b) [redacted]

Telegram stated that Channel subscribers or group members 'who are neither managing nor directly spreading or promoting prohibited content, even if they are part of Communities that may contain such content, are not subject to automatic bans'. Telegram stated that this approach was adopted to avoid inadvertently disrupting law enforcement, journalists, activists and others who may be part of these groups for legitimate reasons.

Telegram further reported that while it 'does not compile data or operate tools conducive to internal investigative work on private user interactions', Telegram has

[19] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

developed tools to assist its moderators to identify 'interconnected management' between 'flagged Communities and their administrators'. eSafety has chosen not to publish further detail on the measures Telegram reported to prevent the information being misused.

D. Sharing of banned account details with other entities

Telegram was asked if it shared details of accounts banned for TVE with the following entities:

Table U

Entity	Shared details of accounts banned for TVE?
Other service providers	No
Law enforcement	Yes*
Regulatory or other public authorities	No
Global Internet Forum to Counter Terrorism	No
Civil society groups	No

* Telegram stated that it provided information to law enforcement in response to 'valid legal requests submitted by law enforcement agencies through designated channels'.

Telegram stated that as at 29 February 2024, Telegram had not received any legal requests from Australian law enforcement agencies 'via dedicated formal channels (e.g., mutual legal assistance requests to the governments in jurisdictions in which the relevant Telegram companies are located)'.

Part 2. Questions in relation to chid sexual exploitation and abuse (CSEA)

8. Questions about reporting of CSEA

A. In-service reporting of CSEA on different parts of Telegram

In response to questions about whether users could report instances of CSEA to Telegram within the service (as opposed to navigating to a separate webform or email address), Telegram responded:

Table V

Parts of the service	In-service reporting option?	Reporting category
Chats	Yes	'Block user > Report Spam'
Secret Chats	Yes	
Group chats (public)	Yes	'Child Abuse'
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Voice calls	No*	N/A
Video calls	No*	
Stories	Yes	'Child Abuse'

*Telegram's original response to the Notice stated that end-users could make in-service reports about voice calls and video calls using a 'Child Abuse (via the Community's info section)' reporting category. In response to a follow-up question from eSafety, Telegram subsequently stated that in-service reporting for voice and video calls was not available during the Report Period. Instead, Telegram stated that 'calls are reported together with their respective community (via the community info section and by additionally including a subset of objectionable sample messages)'.

s 47G(1)(b)

s 47G(1)(b)

As noted at section 4Ai, Telegram subsequently clarified that the 'Block + Report Spam' reporting flow is only available when the Chat or Secret Chat is 'initiated

by non-contacts and strangers'. eSafety understands that when an end-user wishes to report a message from an account they have already added as contact, s 47G(1)(b)

s 47G(1)(b)

With respect to TVE (which eSafety understands is applicable to CSEA), Telegram stated that this was because

> In the extremely unlikely event that a user's friend or acquaintance began sending them TVE content, Telegram contends that it would be more reasonable and effective for said user to contact authorities directly, providing all relevant proof and contact information.

As noted at section 4Ai, eSafety considers that limiting reporting tools to scenarios where the account sending harmful or violative material is not a contact of the end-user risks preventing Telegram from identifying and preventing bad actors from continuing to perpetrate harm on the platform even after they have been blocked by an end-user on the service.

Telegram stated that the single reporting option 'Block + Report Spam' for private and Secret Chats was intended to simplify the user experience and minimise the length of time and number of interactions necessary for a user to end the chat. Telegram stated that 'once the report is processed by moderators, it is escalated as necessary – including via AI / ML if appropriate'.

As noted at section 4Ai, eSafety notes that in response to other questions in the Notice, Telegram stated that it had no means of accessing messages reported by end-users from Secret Chats (see Section 2). Instead, Telegram stated it relies on alternative signals to assess and prioritise reports made about material in E2EE parts of the service.

eSafety notes that this may limit Telegram's ability to review, assess, prioritise, and respond to reports about harmful and illegal material or activity occurring in Telegram's Secret Chats.

9. Questions about proactive detection of CSEA

s 47G(1)(b)

s 47G(1)(b)

In response to various questions in the Notice, Telegram stated that when CSEA material was confirmed, the material was removed along with 'users, Communities and publications involved'.

s 47G(1)(b)

A. Detecting known material using hash matching

i. Known CSEA images

In response to questions about hash matching for known CSEA images, Telegram provided the following information:

Table W

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram Hash Matching System
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why hash matching tools were not used on Chats or Secret Chats user reports, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known CSEA images on Chats and Secret Chats user reports, Telegram referred to the alternative measures it took for known TVE images (see section 5Ai). Telegram also noted that it maintains a dedicated hotline StopCA@telegram.org for reporting any content related to CSAM.

ii. Known CSEA video

In response to questions about hash matching for known CSEA video, Telegram provided the following information

Table X

Parts of service	Used image hash matching tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram Hash Matching System
Group chats (private)	Yes	
Channels (public)	Yes	
Channels (private)	Yes	
Stories	Yes	
Content in user reports	Yes	

When asked why hash matching tools were not used to detect known CSEA videos in Chats and user reports about Secret Chats, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to a question about the alternative steps Telegram took to detect known TVE videos on Chats and user reports about Secret Chats, Telegram referred to the alternative measures it took for known TVE images and known CSEA images (see section 5Ai).

iii. Sources of CSEA hashes

Telegram reported that it sourced its hashes of known CSEA images and videos from internal databases of hashes of CSEA material that had previously been identified on Telegram and removed by its human moderators. In answer to a question about how often it updated this database, Telegram stated that the

database was updated every time a human moderator removed an item of new, or previously 'unknown', CSEA material from the service.

Telegram also referred again to the reasons it gave for its 'exclusive reliance' on human moderators to compile its TVE hash database (see section 5Aiv).

eSafety notes that limiting hash matching exclusively to material that Telegram itself has previously seen and removed risks missing CSEA material that Telegram has not detected yet, and this material continuing to circulate on the platform even when such material has already been identified by other online service providers and hashed in extensive shared databases like those run by the IWF or NCMEC.

This means that to the extent that it used hash matching tools, Telegram did not have access to NCMEC's hash database, which contains more than 5 million hashes of verified CSEA material.[20] eSafety notes media reporting that NCMEC and the IWF both claimed to have made past efforts to contact Telegram that went ignored prior to the CEO of Telegram's arrest on 27 August 2024.[21]

s 47G(1)(b) [redacted]

iv. Action taken when CSEA hashes are detected

Telegram stated that detections of known CSEA images and videos through hash-matching resulted in the automated removal 'of all users, Communities and publications involved'. Telegram also referred to further information it had provided on the measures it took to address known TVE which eSafety has chosen not to publish to prevent the information being misused.

B. Detecting new CSEA material

i. Text analysis to detect CSEA

In response to questions about technology to detect terms, abbreviations, codes and hashtags indicating likely CSEA (for example grooming, sexual extortion, or the trading and sale of CSEA material), Telegram provided the following information:

Table Y

[20] Google Safety Center, 'NCMEC, Google and Image Hashing Technology', accessed 15 November 2024, URL: https://safety.google/stories/hash-matching-to-help-ncmec/

[21] NBC News, 'Telegram ignored outreach outreach from child safety watchdogs before CEO's arrest, groups say', 28 August 2024, URL: https://www.nbcnews.com/tech/security/telegram-ceo-pavel-durov-child-safety-rcna168266 .

Parts of service	Used text analysis tools?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[22]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
Profile username	Yes	
Profile description	Yes	
Group username	Yes	
Group description	Yes	
Channel username	Yes	
Channel description	Yes	
Content in user reports	Yes	

In response to why it did not use technology to scan Chats, Secret Chats, private group chats, and private channels for indications of likely CSEA, Telegram referred to its reasons for not using such tools to detect known TVE images (see section 5Ai).

In response to what alternative steps Telegram took to detect known terms, abbreviations, codes or hashtags indicating likely CSEA on these parts of the service, Telegram referred to the alternative steps it took to detect text indicating likely TVE (see section 5Bi). Telegram also referred to the dedicated hotline it maintains for receiving reports about CSAM (see section 9Ai).

ii. Sources of terms, abbreviations, codes, and hashtags

Telegram stated that it sourced phrases, codes, and hashtags indicating likely CSEA from samples of CSEA material that had been removed from Telegram by its human moderators.

[22] In response to a follow-up question seeking the names of the tools Telegram uses to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

iii. Languages covered by language analysis tools

When asked what languages were covered by technology used to detect terms, abbreviations, codes and hashtags indicating likely CSEA, Telegram did not provide a list of languages.

Telegram referred to the answer it gave in response to questions about the languages covered by the tools Telegram used to detect likely TVE in text (see section 5Ci).

iv. New or 'unknown' CSEA images

In response to questions about the detection of new (or 'previously unknown') CSEA images, Telegram provided the following information:

Table Z

Parts of service	Used tools for images?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[23]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
User profile picture	Yes	
Group profile picture	Yes	
Channel profile picture	Yes	
Content in user reports	Yes	

In response to why it did not use technology to detect new CSEA images on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect new TVE images (see section 5Bi).

In response to a question about the alternative steps Telegram took to detect new CSEA images on these parts of the service, Telegram referred to measures it took

[23] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

to detect known terms, abbreviations, codes and hashtags that indicate likely CSEA on the service and likely TVE in text (see sections 9Bi and 5Bi).

As noted at section 5Bi, eSafety considers that not using proactive detection tools to identify and review potential CSEA material increases the likelihood that such material will remain undetected and continue to circulate on these parts of the service.

eSafety notes that Chats, Private Group chats, and Private Channels are not E2EE – leaving alternative technical options available for content detection and review by human moderators.

v. New or 'previously unknown' CSEA videos

In response to questions about the detection of new (or 'previously unknown') CSEA videos, Telegram provided the following information:

Table AA

Parts of service	Used tools for videos?	Names of tools used
Chats	No	N/A
Secret chats (user reports)	No	N/A
Group chats (public)	Yes	Internal Telegram AI and Machine Learning Models[24]
Group chats (private)	No	N/A
Channels (public)	Yes	Internal Telegram AI and Machine Learning Models
Channels (private)	No	N/A
Stories	Yes	Internal Telegram AI and Machine Learning Models
Content in user reports	Yes	

In response to why it did not use technology to detect new CSEA videos on Chats, Secret Chats, Private Group Chats, and Private Channels, Telegram referred to the reasons it gave for not using proactive detection tools to detect new TVE images (see section 5Bi).

In response to a question about the alternative steps Telegram took to detect new CSEA videos on these parts of the service, Telegram referred to measures it took

[24] In response to a follow-up question seeking the names of the tools Telegram used to detect new forms of TVE and CSEA material, Telegram referred to the 'state-of-the-art AI/ML models that span a wide array of technologies' which it described in its original response to the Notice. These models are described at Section 12 of the Summary.

to detect known terms, abbreviations, codes and hashtags that indicate likely CSEA on the service and likely TVE in text (see sections 9Bi and 5Bi).

vi. Action taken when new CSEA is detected

Telegram stated that when likely new CSEA material was detected, it was either immediately processed by Telegram's automated tools or sent for human review 'depending on the degree of confidence to which the relevant model is able to issue a judgement, combined with other factors'. eSafety has chosen not to disclose these additional factors due to public safety reasons. If the detection was confirmed, it resulted in the removal of all 'users, Communities and publications involved'. Telegram reported that when new CSEA images and videos were removed, they were then added to Telegram's internal hash database.

C. Blocking links to CSEA material

i. URLs linking to known CSEA

In response to a question about whether Telegram blocked URLs linking to known CSEA, Telegram provided the following information:

Table BB

Parts of service	Blocked URLs linking to known TVE material on other services/websites?	URL sources
Chats	No	N/A
Secret chats (E2EE)	No	
Group chats (public)	No	
Group chats (private)	No	
Channels (public)	No	
Channels (private)	No	
Profile description	No	
Group description	No	
Channel description	No	

In response to why URLs to known CSEA material were not blocked and whether alternative steps were taken to block such URLs, Telegram reiterated that 'focusing its efforts on ML-based classification tends to yield better results when compared to static link blacklists'. Telegram stated that links to harmful material tend to be taken down routinely or hidden behind URL shorteners. Telegram stated that it used proactive detection tools that are trained on previous detections of CSEA material, including material that may contain external links.

Telegram also stated that, as at October 2024, it was 'in the process of joining the Internet Watch Foundation's safety programs involving *inter alia* access to URL lists containing links to known CSAM websites'.

As noted above, eSafety is aware of public statements made by the Internet Watch Foundation (IWF) asserting that prior to the arrest of Telegram's CEO in August 2024, the IWF had made repeated efforts to reach out to Telegram and that Telegram had refused to 'take any of its services to block, prevent, and disrupt the sharing of child sexual abuse imagery'.[25] It is unclear why Telegram did not take the opportunity to work with the IWF sooner.

D. Percentage of CSEA detected proactively

Telegram was asked what percentage of CSEA was detected proactively, compared to CSEA reported by users, trusted flaggers or through other channels for the following parts of its service:

Table CC

Parts of Telegram	Percentage of CSEA detected proactively	Percentage of CSEA reported by users, trusted flaggers or other
Chats	N/A	100%
Secret Chats	N/A	100%
Group chats (public)	71%	29%
Group chats (private)	85%	15%
Channels (public)	74%	26%
s 47G(1)(b)		
Channels (private)	80%	20%
Voice and video calls (public and private)	N/A*	N/A*
Group video calls (public and private)	'Included in group chats'**	
Stories	65%	45%

Commented [A23]: s 47G(1)(b)

[25] NBC News, 'Telegram ignored outreach outreach from child safety watchdogs before CEO's arrest, groups say', 28 August 2024, URL: https://www.nbcnews.com/tech/security/telegram-ceo-pavel-durov-child-safety-rcna168266.

s 47G(1)(b)

Commented [A24]: s 47G(1)(b)

* In answer to a follow-up question from eSafety to clarify why its answer was 'N/A' for voice and video calls, Telegram referred to the answer it gave with respect to the percentage of TVE detected proactively and by reports on voice and video calls (see section 4E).

**Telegram stated that its group video call data was included in the relevant group chat statistics because 'information on resulting bans is not stored separately'.

E. Appeals against CSEA-related moderation

In response to a question about how many appeals were made by users for accounts banned or content removed for CSEA, where Telegram was alerted by automated tools or user reports, and how many of those were successful, Telegram provided the following information:

Table DD

How Telegram was alerted to CSEA	Number of appeals made for accounts banned for CSEA breach	Number of appeals that were successful for accounts banned	Number of appeals made for material removed for CSEA breach	Number of appeals that were successful for material removed
Automated tools	7,098	573	N/A*	
User reports	2,702	218		

*Telegram stated that because CSEA-related content violations resulted in the users and Communities involved being removed from Telegram, 'It is not generally possible to appeal for reinstatement of removed CSAM materials, so only account appeals are included'. Telegram stated that in some jurisdictions, such as the EU and the EU Terrorist Content Online Regulation, it may receive appeals against content removals from legally mandated contact lines. However, Telegram reported that there were 'no such appeals connected to removal of CSAM content' during the Report Period.

10. Questions about resources, expertise, and human moderation

A. Median time to reach an outcome to a user report of CSEA

Telegram was asked to provide the median time taken to reach an outcome[26] after receiving a user report about CSEA for the following parts of the service:

Table EE

Parts of the service	Reports from users globally	Reports from users in Australia *
Chats	11 hours	11 hours
Secret Chats	11 hours	11 hours
Group chats (public)	10 hours	10 hours
Group chats (private)	10 hours	10 hours
Channels (public)	10 hours	10 hours
Channels (private)	10 hours	10 hours

In response to a question asking how median time was calculated Telegram stated that to calculate these figures it registered 'the net time frames between the submission of each individual report and the moderator's decision in respect of that report'.

* Telegram stated that it 'currently doesn't have the technical means to provide separate statistics by country'.

11. Questions about steps to prevent recidivism

A. Measures and indicators

In response to a question asking if Telegram had measures in place to prevent recidivism for CSEA-related breaches on its service, Telegram responded 'yes' and stated:

> Given the severity of CSAM, any infringement related to it typically results in the permanent removal of related accounts and Communities. Owners of

[26] Defined in the Notice as a calculation from 'the time that a user report is made, to a content moderation outcome or decision, such as removing the content, banning the account, or deciding that no action should be taken.'

> infringing groups and channels may also face removal, preventing them from creating new Communities or accounts on Telegram.

Telegram listed a minimal[27] number of indicators that it used to detect users that have previously been banned for CSEA breaches. eSafety has chosen not to publish these indicators to prevent the information being misused.

12. Additional information

In response to an opportunity to provide further information and context to any of its responses to the questions asked in the Notice, Telegram added that:

> Telegram was built to safeguard the privacy of individuals at risk – such as legitimate activists, journalists, and protesters – and preserve their right to private correspondence. While staying true to its core value of user privacy, Telegram actively engages in policy efforts and implements robust moderation tools to address abusive content.
>
> Telegram's exponential growth in the recent years has presented unique moderation challenges due to the sheer volume and diversity of content. Recognizing these challenges, Telegram is continuing to expand its moderation framework while enhancing existing solutions – leveraging advanced software, growing its dedicated teams and fostering key partnerships to mitigate harmful content effectively, as outlined in detail below.

Telegram outlined the following features, tools, and resources it used to address harmful material and activity on its service:

- **'Content review'** – Telegram stated content on the service was reviewed 24/7 through proactive detection, user reports, 'email referrals from users and trusted organizations', and monitoring media stories. Telegram stated that it 'relies on a combination of AI / ML customized recognition tools, manual search strategies, as well as prevention of reappearance of already removed items'. Telegram stated that '[p]ublic content flagged by

[27] eSafety uses the following terms to give an impression of the extent of the indicators used by services in the table below, rather than publishing the specific indicators which could be misused:
- Minimal: A small number
- Several: A moderate number
- Multiple: A significant number.

algorithms is processed, validated, and assigned additional priority if needed, which allows moderators to receive relevant reports and properly sort miscategorized items'.

- **'Limited content discovery'** – Telegram stated that '[b]y design, Telegram does not employ recommendation algorithms or any other form of targeted amplification'. Telegram stated that this means that bad actors cannot exploit Telegram to 'spread harmful content rapidly or to reach a meaningful share of users'. Telegram said this was also true of Telegram Stories.
- **'State-of-the-art Software Solutions'** – Telegram stated that its proactive detection tools had been created by 'world-class engineers', and that it used a combination of hash and pattern matching tools and other 'state-of-the-art AI/ML models that span a wide array of technologies'. Telegram stated that models included:
    - fine-tuned self-supervised multilingual transformer-based language models;
    - fine-tuned vision transformer models;
    - multilingual transformer-based end-to-end ASR systems;
    - multimodal transformer-based models aligned on image-text datasets;
    - multilingual transformer-based large language models; and
    - custom data clustering algorithms.

  Telegram stated that 'several' of these models had been deployed by the end of the Report Period (29 February 2024), but it had 'since significantly expanded its use of AI and ML technologies'.
- **'Trained professionals'** – Telegram stated that its moderators are 'highly trained professionals that undergo regular quality-assurance checks including daily peer examinations'. Telegram stated that although its 'strict selection process ensures that only the most capable and suitable individuals are chosen for a moderator role', it conducted daily assessments of between 1 and 5% of all reports by randomly distributing them to moderators to calculate potential error rates. Telegram stated that 'Moderators with suboptimal error rates, or involved in systematic, gross, or material errors are replaced'. Telegram also stated that it has 'specialized

moderator task groups' responsible for responding to harms such as CSAM and TVE. Telegram stated that these task forces, and its escalation processes, have 'significantly reduced the response time for handling critical reports'.

- **'Key partnerships'** – Telegram cited its collaboration with the Global Center for Combating Extremist Ideology (or, 'Etidal'). Telegram stated that between February 2022 and June 2024, Telegram's partnership with Etidal had resulted in '93,99 million pieces of content related to spreading terrorist ideologies' being removed from the service.

  Telegram stated that it was 'consistently expanding its network amongst industry leaders and international community to stay up-to-day on the latest developments and best practices related to content moderation'. Telegram also provided the following as examples of organisations that its staff regularly engaged with:

  - UK Home Office
  - Etidal
  - EU Internet Forum
  - Europol
  - Ofcom
  - UNSC Counter-Terrorism Committee Executive Directorate

  Telegram stated that:

  > Reports of CSAM, terrorist content and violent propaganda received from trusted organizations are processed within 1 hour.

  Telegram also reiterated that, as at October 2024, it was in the process of joining the Internet Watch Foundation's safety programs to gain access to the IWF's hash lists of known CSEA material.

Telegram's note:

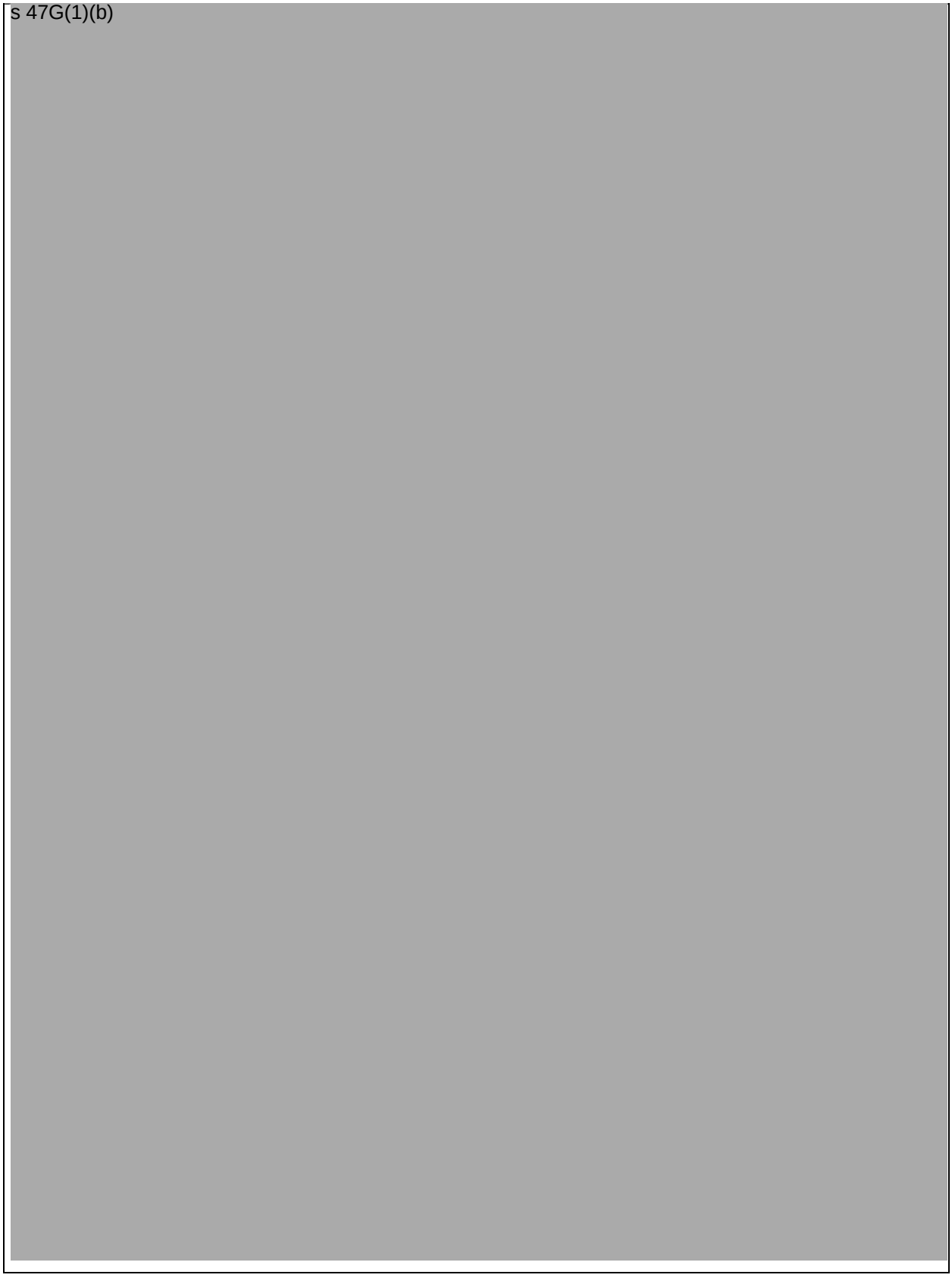
s 47G(1)(b)

27. s 47E(d)

a. s 47E(d) s 47G(1)(a), s 47G(1)(b)

b. s 47E(d)

s 47G(1)(a)


s 47G(1)(b)

s 47G(1)(b)

[redacted]

c. s 47E(d)

[redacted]

s 47G(1)(a), s 47G(1)(b)

[redacted]

d. s 47E(d)

[redacted]

s 47G(1)(b)

[redacted]

28. s 47E(d)

s 47G(1)(a), s 47G(1)(b)

29. s 47E(d)

a. s 47E(d)

s 47G(1)(a), s 47G(1)(b)

b. s 47E(d)

s 47G(1)(a), s 47G(1)(b)

s 47G(1)(a)

c. s 47E(d)

s 47G(1)(b)

30. s 47E(d)

a. s 47E(d) s 47G(1)(a), s 47G(1)(b)

b. s 47E(d)

s 47G(1)(a), s 47G(1)(b)

c. s 47E(d)

s 47G(1)(b)

s 47E(d)

31. s 47E(d)

a. s 47E(d) s 47G(1)(a), s 47G(1)(b)

b. s 47E(d)

s 47G(1)(b)

s 47G(1)(a), s 47G(1)(b)