Complaint

CYR-0168471

Created On
15/03/2023 1:35 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
15/03/2023 1:35 PM

Date Finalised
16/03/2023

Title: Reported to Twitter 15/03/2023 at 2:18 pm.
s 22 15/03/2023 3:28 PM

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
Yes

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Violent extremist material

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
Web Page

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions
s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path
https://twitter.com/s 47E(d)	Complained	Hosting Service				

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commer
	INV-2023-07860	Completed	https://twitter.com/s 47E(d)	No	Violent extremist material	RC9A	15/03/2023

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commer

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
15/03/2023 1:35 PM	Cyber Report	eSafety Cyber Report DL	Notification - Violent extremist material [SEC=OFFICIAL:Sensitive] CRM:0051360	Sent

1 - 1 of 1 (0 selected) Page 1

E-mail

s 47E(d), s 47F

Owner s 22

Date Sent/Received 16/03/2023 2:40 PM

E-mail

From
s 47E(d), s 47F@proton.me

To
Cyber Report; Cyber Report; Cyber Report; Cyber Report; Online Content; online online

Cc

Bcc

Subject
s 47E(d), s 47F

Classification
[SEC=OFFICIAL:Sensitive]

Regarding
CYR-0168905

s 47E(d), s 47F

Attachments

Drop files here.

File Name	File Size (Bytes)
No Attachment records are available in this view.	

0 - 0 of 0 (0 selected) Page 1

s 47E(d), s 47F

s 47E(d), s 47F

s 47E(d), s 47F

s 47E(d), s 47F

Approval

s 47E(d), s 47F

Approval Comments

Approving Officer **Escalating Officer**

Approval Date

Status Reason Received

Complaint

CYR-0168905

Created On
17/03/2023 8:43 AM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
17/03/2023 8:43 AM

Date Finalised
17/03/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
Yes

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Violent extremist material

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
WebPage

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	ServiceType	Service Category	P2P Filename	Newsgroup Name	Content Path
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.com	Complained	HostingService				

s 47E(d), s 47F

1 - 10 of 10 (0 selected) Page 1

Investigation Id	Status	Investigated Locator	Australian Conn...	ComplainantReason	Classification	Comme
INV-2023-08066	Discretion not to investig...	https://twitter.com	No	Violence	NA	17/03/202
INV-2023-08065	Discretion not to investig...	https://twitter.com	No	Violence	NA	17/03/202
INV-2023-08064	Discretion not to investig...	https://twitter.com/	No	Violent extremist material	NA	17/03/202
INV-2023-08049	Discretion not to investig...	https://twitter.com	No	Violent extremist material	NA	17/03/202

s 47E(d), s 47F

1 - 4 of 10 (0 selected) Page 1

17/03/2023 8:43 AM

Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commer
INV-2023-07256	Discretion not to investig...	https://twitter.com	No	Violence	NA	10/03/2023
INV-2023-07255	Discretion not to investig...	https://twitter.com	No	Violence	NA	10/03/2023
INV-2023-07254	Discretion not to investig...	https://twitter.com	No	Violence	NA	10/03/2023

s 47E(d), s 47F

1 - 3 of 3 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
16/03/2023 2:40 PM	s 47E(d), proton.me	Cyber Report; Cyber Report; Cyber ...	s 47E(d), s 47F	Received

1 - 1 of 1 (0 selected) Page 1

Complaint

CYR-0168907

Created On
17/03/2023 8:45 AM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
17/03/20238:45 AM

Date Finalised
17/03/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
Yes

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Violent extremistmaterial

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
Web Page

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path
https://twitter.com s 47E(d)	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
s 22	Complained	Hosting Service				

1 10 of 10 (0selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Comme
	INV 2023 08068	Discretion not to investig...	https://twitter.com s 47E(d)	No	Violence	NA	17/03/202
	INV 2023 08067	Discretion not to investig...	https://twitter.com	No	Violence	NA	17/03/202
	INV 2023 08064	Discretion not to investig...	https://twitter.com/	No	Violent extremist material	NA	17/03/202
	INV 2023 08057	Discretion not to investig...	s 22	No	Violent extremist material	NA	17/03/202

1 4 of 10 (0selected) Page 1

17/03/2023 8:46 AM

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commer
	INV 2023 07258	Discretion not to investig...	https://twitter.com s 47E(d)	No	Violence	NA	10/03/2023
	INV 2023 07257	Discretion not to investig...	https://twitter.com	No	Violence	NA	10/03/2023
	INV 2023 07256	Discretion not to investig...	https://twitter.com	No	Violence	NA	10/03/2023

1 3 of 3 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason

No E mailrecords are available inthis view.

0 - 0 of 0 (0 selected) Page1

Complaint

CYR-0168908

Created On
17/03/2023 8:49 AM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
17/03/2023 8:49 AM

Date Finalised
17/03/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complainant
Yes

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Violentextremistmaterial

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
Web Page

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path
https://twitter.com/ s 47E(d)	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.co	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				
s 22	Complained	Hosting Service				

1 - 6 of 6 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Comme
	INV-2023-08063	Discretion not to investig...	https://twitter.com s 47E(d)	No	Violent extremist material	NA	17/03/202
	INV-2023-08062	Discretion not to investig...	https://twitter.com	No	Violent extremist material	NA	17/03/202
	INV-2023-08061	Discretion not to investig...	https://twitter.com	No	Violent extremist material	NA	17/03/202
	INV-2023-08060	Discretion not to investig...	https://twitter.com	No	Violent extremist material	NA	17/03/202

1 - 4 of 4 (0 selected) Page 1

17/03/2023 8:49 AM

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commer

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason

No E-mail records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Complaint

CYR-0180240

Created On
28/04/20234:12 AM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
28/04/2023 4:12AM

Date Finalised
V05/2023

Complaint Status
Closed

Assig
s 22

Source Type
CYR

Dupl
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant
s 47E(d), s 47

Repeat Complainant
No

Complainant Reason
Advocates a terrorist act

Response Required
Yes

Response Method
Email

Submission

Method Received
Web Form

Presentation Type
Web Page

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions
s 47E(d), s 47F

Access Username

Access Password

Related Entities

URL	Type	Service Type	ServiceCategory	P2PFilename	NewsgroupName	Content Path
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				

s 47E(d), s 47F

1 - 5 of 5 (0selected) Page 1

S...	InvestigationId	Status	Investigated Locator	AustralianConn...	Complaint Reason	Classification	Commenced	Del
	INV-2023-12271	Discretion notto investigate	https://twitter.com	No	Advocates a terrorist act	NA	1/05/2023	1/05/2023
	INV-2023-12270	Discretion not to investigate	https://twitter.com	No	Advocates a terrorist act	NA	1/05/2023	1/05/2023
	INV-2023-12269	Discretion notto investigate	https://twitter.com	No	Advocates a terrorist act	NA	1/05/2023	1/05/2023
	INV-2023-12268	Discretion notto investigate	https://twitter.com	No	Advocates a terrorist act	NA	1/05/2023	1/05/2023

s 47E(d), s 47F

1 - 4 of 5 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	AustralianConn...	Complaint Reason	Classification	Commenced	Date Fi

No investigation recordsare available in thisview.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status
NA Content is not illegal or restricted notactionable	11/04/2019 12:44 PM	Active

1 - 1 of 1 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
1/05/2023 1:04 PM	Cyber Report	First Solak	Your report to the eSafety Commissioner: CY R0180240 [SEC=OFFICIAL:Sensitive] C RM:0104424	Sent
28/04/20234:12 AM	Cyber Report	eSafety Cyber Report DL	Notification Advocates a terrorist act [SEC=OFFICIAL:Sensitive] CRM:0051611	Sent

1 - 2 of 2 (0 selected) Page 1

Complaint

CYR-0196400

Created On
23/06/2023 2:05 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
23/06/2023 2:05 PM

Date Finalised
30/06/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Education Provider
No

Com
s 47E(d), s 47F

Repeat Complainant
Yes

Com
Violent extremist material

Response Required
Yes

Response Method
E-mail

Submission

Method Received
Web Form

Presentation Type
Web Page

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path
https://twitter.co	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				
https://twitter.com	Complained	Hosting Service				

s 47E(d), s 47F

1 - 4 of 4 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	AustralianConn...	Complainant Reason	Classification
	INV-2023-18030	Discretion not to investigate	https://twitter.com	No	Violentextremistmaterial	NA
	INV-2023-18028	Discretion not to investigate	https://twitter.com	No	Violentextremistmaterial	NA
	INV-2023-18026	Discretion not to investigate	https://twitter.com	No	Violentextremistmaterial	NA
	INV-2023-18024	Discretion not to investigate	https://twitter.com	No	Violentextremistmaterial	NA

s 47E(d), s 47F

1 - 4 of 4 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	AustralianConn...	Complainant Reason	Classification	Commer

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status
NA Content is not illegal or restricted - not actionable	11/04/2019 12:44 PM	Active

1 - 1 of 1 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
30/06/2023 8:38 AM	CyberReport	s 47E(d), s 47F	Your report to the eSafety Commissioner: CYR 0196400 [SEC=OFFICIAL:Sensitive] CRM:0054550	Sent
23/06/2023 2:06 PM	CyberReport	eSafety Cyber Report DL	Notification Violent extremist material [SEC=OFFICIAL:Sensitive] CRM:0051656	Sent

1 - 2 of 2 (0 selected) Page 1

Complaint

CYR-0199033

Created On
3/07/2023 12:38 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
3/07/2023 12:38 PM

Date Finalised
4/07/2023

Title:
s 22 and appears the Twitter accounts have been suspended.
/07/2023 2:37 PM

Complaint Status
Closed

Ass
s 22

Source Type
CYR

Der
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Education Provider
No

Com
s 47E(d)

Repeat Complainant
Yes

Complainant Reason
Violent extremist material

Response Required
Yes

Response Method
E-mail

Submission

Method Received
Web Form

Presentation Type
WebPage

Device Type
Mobile Phone

Device Name

Remote IP Address

Access Information

Additional access instructions
s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path
https://twitter.com/s 47E(d)	Complained	Hosting Service				
https://twitter.com/	Complained	Hosting Service				

1 - 2 of 2 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	ComplainantReason	Classification	C
	INV-2023 -19031	Terminated	https://twitter.com/s 47E(d)	No	Violentextremistmaterial	NC	3/07/2
	INV-2023 -19030	Terminated	https://twitter.com/	No	Violentextremistmaterial	NC	3/07/2

1 - 2 of 2 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	ComplainantReason	Classification	Commer

No Investigationrecords are available inthis view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status
NC Content is not available or has been removed	17/06/2013 4:15 PM	Active

1 - 1 of 1 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
4/07/2023 3:11 PM	Cyber Report	s 47E(d)	Your report to the eSafety Commissioner: CYR 0199033 [SEC=OFFICIAL:Sensitive] CRM:0066850	Sent
3/07/2023 12:38 PM	Cyber Report	eSafety Cyber Report DL	Notification Violent extremist material [SEC=OFFICIAL:Sensitive] CRM:0051660	Sent

1 - 2 of 2 (0 selected) Page 1

Complaint

CYR-0213095

Created On
21/08/2023 6:41 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
21/08/2023 6:41 PM

Complaint Status
Closed

Source Type
CYR

Date Finalised
25/08/2023

Assigned To
s 22

Derived
No

Title Access Info
s 37(2)(b), s 47E(d)
s 22 22/08/2023 12:25 PM

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Repeat Complainant
Yes

Education Provider
No

Client
s 47E(d)

Complaint Issue Reason
Violent extremist material

Response Required
Yes

Response Method
E-mail

Submission

Method Received
Web Form

Presentation Type
Social media

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information
s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path	Hosting Country
https://twitter.	Complained	SocialMedia Service (SMS)					
https://twitter.com	Complained	SocialMedia Service (SMS)					
https://twitter.co	Complained	Social Media Service (SMS)					
http://twitter.com	Complained	SocialMedia Service (SMS)					
https://twitter.co	Complained	SocialMedia Service (SMS)					
http://twitter.com	Complained	Social Media Service (SMS)					
http://twitter.co	Complained	SocialMedia Service (SMS)					
http://twitter.com	Complained	SocialMedia Service (SMS)					
https://twitter.com	Complained	SocialMedia Service (SMS)					
http://twitter.com	Complained	SocialMedia Service (SMS)					

s 47E(d)

1 - 10 of 10 (0 selected) Page 1

Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced	Date Finalised	Investi
INV-2023-23450	Completed		No	Violentextremistmaterial	RCNA	22/08/2023	24/08/2023	Scott Vallance
INV-2023-23448	Completed		No	Violentextremistmaterial	RCNA	22/08/2023	23/08/2023	Scott Vallance
INV-2023-23447	Discretion not to investig...	https://twitter.com	No	Violentextremist material	NA	22/08/2023	22/08/2023	Scott Vallance
INV-2023-23446	Discretion not to investig...	https://twitter.com	No	Violentextremistmaterial	NA	22/08/2023	22/08/2023	Scott Vallance

s 22

s 47E(d)

1 - 4 of 17 (0 selected) Page 1

Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced	Date Finalised	Investig

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

21/08/2023 6:41 PM

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
25/08/2023 30 5PM	Cyber Report	s 47E(d)	Your report to the eSafety Commissioner: CYR-0213095 [SEC=OFFICIAL:Sensitive] CRM:0104603	Sent
		s 22		Received
				Received
				Received

1 - 4 of 6 (0 selected) Page 1

Basic Online Safety Expectations

Regulatory Guidance

September 2023

Australian Government | eSafetyCommissioner

eSafety.gov.au

Contents

Overview of this guidance

This guidance is for online service providers (**providers**) and other stakeholders who require information about the Basic Online Safety Expectations, also known as 'the Expectations' and the functions of the eSafety Commissioner (**eSafety**) in assessing compliance with those Expectations.

The Expectations are determined under the Online Safety Act 2021 (**the Act**), and set out the Australian Government's expectations of the steps that should be taken by providers of social media services, messaging services, gaming services, file sharing services, apps and certain other sites accessible from Australia to keep Australians safe online. While compliance with the Expectations is not mandatory, eSafety has powers under the Act to obtain information from providers as to the steps they are taking to comply with the Expectations. eSafety can also publish statements about whether providers have or have not complied with the Expectations. The aim is to increase the transparency and accountability of providers, thereby helping to incentivise and improve safety standards.

More information on the Basic Online Safety Expectations is available on eSafety's website.[1]

This guidance provides information on:

- the legal framework for the Expectations (**Part 1**)
- eSafety's approach to the use of reporting powers (**Part 2**)
- eSafety's approach to assessing compliance with the Expectations (**Part 3**)
- examples of reasonable steps that can be taken by providers to ensure compliance with the Expectations (**Part 4**)

This guidance updates previous eSafety regulatory guidance published on 25 July 2022, including through the addition of Part 4.

[1] eSafety website, Basic Online Safety Expectations | eSafety Commissioner.

Part 1: The legal framework for the Expectations

Overview of the Expectations

The Act provides for the Minister for Communications to set online safety expectations through a legislative instrument called a determination. The Online Safety (Basic Online Safety Expectations) Determination 2022[2] (the Determination) was registered on 23 January 2022. An Explanatory Statement to the Determination was also published.

The Expectations include a range of foundational steps that providers are expected to take to ensure safety for their end-users, including:

- ensuring end-users can use online services in a safe manner
- ensuring safe use of certain features of a service, such as encrypted services and anonymous accounts
- minimising provision of unlawful and harmful material and activity
- enabling end-users to make reports and complaints about unlawful and harmful material and activity
- having terms of use, policies and procedures to ensure safe use, and enforcing these terms.

The Expectations are set out in the table at Annex A on page 68, and are dealt with in more detail under Part 4 of this guidance.

eSafety has a number of relevant powers under the Act.

- The power to require providers to report on how they are meeting any or all of the Expectations, either on a non-periodic or a periodic basis through a reporting notice or determination. The obligation to respond to a reporting notice or determination is enforceable and backed by civil penalties.
- The power to publish summaries of information, including from reporting notices or determinations.
- The power to publish statements regarding providers' compliance and non-compliance with the Expectations.

[2] Online Safety (Basic Online Safety Expectations) Determination 2022: Online Safety (Basic Online Safety Expectations) Determination 2022 (legislation.gov.au).

Who do the Expectations apply to?

The Expectations apply to social media services, relevant electronic services and designated internet services that can impact the online safety of Australians.

Table 1: The Expectations apply to three main sections of the online industry

Section of the online industry		Scope
Social media services	Providers of social media services	All providers of social media services that can impact the online safety of Australians, including: • social networks • media sharing networks • discussion forums • consumer review networks.
Relevant electronic services	Providers of relevant electronic services	All providers of relevant electronic services that can impact the online safety of Australians, including: • email services • instant messaging services • SMS and MMS services • chat services • online games where end-users can play with or against each other • online dating services.
Designated internet services	Providers of designated internet services	All providers of designated internet services, such as websites and file storage services that that can impact the online safety of Australians (unless a service is otherwise considered a social media service or a relevant electronic service).

Providers should also be aware of their obligations under other regulatory requirements under the Act including applicable industry codes and industry standards. Industry codes and industry standards can place enforceable requirements on sections of the online industry including social media services, relevant electronic services and designated internet services, in relation to class 1 and class 2 material. More detail is set out on page 9.

What harms are covered by the Expectations?

The Expectations apply to all harmful material and activity covered by the Act, as well as more broadly to address harms that impact on the online safety of Australians.

There are a wide range of potential harms that may arise on a service, impacting the online safety of Australians. It is expected that providers will have systems and processes in place to identify such harms, and take steps to ensure they are complying with the Expectations in relation to these harms.

What is 'unlawful' material and activity?

'Unlawful' material or activity is material or activity prohibited under law. For the purposes of the Determination, the term 'unlawful' refers to illegal material or activity dealt with under the Act and other unlawful material or activities that may have a negative impact on the online safety of Australians. Unlawful material and activity is therefore generally considered to also be harmful.

Examples of unlawful material and activity include:

- material that is illegal and has been refused classification under the *Classification (Publications, Films and Computer Games) Act 1995* including:
    - child sexual exploitation and abuse[3] (CSEA) material
    - material that advocates terrorism
    - material that depicts extreme crime and violence
    - material that incites or instructs or depicts, without justification, crime and violence or illicit drug use
      (known as class 1 material in the Act)
- grooming[4] of children
- the sharing of, or threatening to share, a non-consensual intimate image[5], including sexual extortion[6] (also known as sextortion).

[3] Child sexual exploitation and abuse (CSEA) can include both material and activity (for example, grooming). CSEA material is a broad category of material, normally referring to images and videos depicting the sexual abuse of a child, including sexual assault (child sexual abuse material or 'CSAM'), as well as content that sexualises and is exploitative of a child, but that does not necessarily show the child's sexual abuse (child sexual exploitation material or 'CSEM').

[4] Predatory conduct to prepare a child or young person for sexual activity at a later time.

[5] A non-consensual intimate image includes a still visual image or moving visual images. See section 15 of the Act.

[6] Sexual extortion, also known as sextortion, is a crime involving online blackmail, where victims are tricked into sending intimate images of themselves to someone who then threatens to share the images unless demands are met, usually for payment. Sextortion is currently an online child sexual exploitation trend, targeting teenage males in particular.

What is 'harmful' material and activity?

'Harmful' material or activity is material or activity that may not be unlawful but is covered within the scope of the Act. It is also material or activity that should fall under a provider's terms of use, policies and procedures and standards of conduct for end-users (as outlined in Section 14 of the Determination).

Some material or activity will be both unlawful and harmful, such as class 1 material, non-consensual intimate images and material depicting abhorrent violent conduct.

The Expectations specifically highlight the importance of minimising the extent to which the following material is available on a provider's service:

a. cyberbullying material targeted at an Australian child
b. adult cyber abuse material
c. a non-consensual intimate image of a person
d. class 1 material
e. material promoting, inciting, instructing in, or depicting abhorrent violent conduct.

Class 2 material is material that would be harmful for a child to see.[7] It is defined in the Act and is material[8] that is, or would likely be, classified as either:

- X18+ (or, in the case of publications, category 2 restricted),[9] or
- R18+ (or, in the case of publications, category 1 restricted)[10]

under the National Classification Scheme, because it is considered inappropriate for general public access and/or for children and young people under 18 years old.

The Expectations specifically require providers to take reasonable steps to prevent access by children to class 2 material.

Additional information on the classification of material under the National Classification Scheme is available in the Online Content Scheme Regulatory Guidance on eSafety's website.

[7] X18+, R18+ classifications require that the material be unsuitable for a child to see. In the case of Category 2 and Category 1 classification (which relate to publications), the material is either unsuitable for a child to see or read, or contains particular depictions likely to cause offence to a reasonable adult. More information on the approach to classifications can be found in the National Classification Code: National Classification Code (May 2005) (legislation.gov.au)
[8] Section 107 of the Act. This material includes films, publications, computer games and any other material that is not a film, publication or computer game.
[9] Section 107(1)(a) - (e) of the Act.
[10] Section 107(1)(f) - (l) of the Act.

The Explanatory Statement to the Determination provides further examples of harmful material.

- Hate against a person or group of people on the basis of race, ethnicity, disability, religious affiliation, caste, sexual orientation, sex, gender identity, serious disease, asylum seeker or refugee status, or age.
- Promotion of suicide and self-harm content, such as pro-anorexia content, that does not meet the threshold of class 1 or class 2 material.
- High volume, cross-platform attacks that have a cumulative effect that is damaging but does not meet the threshold of adult cyber-abuse when reported as singular comments or posts.
- Promotion of dangerous viral activities that have the potential to result in real injury or death.

eSafety's approach to exercising its powers in relation to the Expectations

eSafety will focus on a number of objectives when exercising its powers in relation to the Expectations.

- Enhancing providers' transparency and accountability, and improving insights into the effectiveness and impact of what providers are doing to keep end-users safe online.
- Tracking harms, safety interventions and technology over time through use of periodic reporting notices and improving understanding of where gaps and challenges exist.
- Incentivising proactive and systemic safety interventions, including by using statements of compliance or non-compliance with the Expectations to highlight good practice, as well as areas where insufficient action is being taken.

eSafety expects that providers regularly review their policies, procedures and practices to ensure alignment with the Expectations and that they put in place additional measures where a service is not compliant.

What are the reasonable steps a provider should take to comply with the Expectations?

The Determination does not prescribe how the Expectations must be met by providers but gives examples of reasonable steps that a provider may choose to take. This provides flexibility in the way providers can meet the Expectations. However, a provider's approach

should be informed by examples provided in the Determination and this guidance, and advice from eSafety.

Part 4 of this document sets out more detailed guidance for providers on steps that could be taken to comply with the Expectations but does not prescribe specific steps or the use of particular technology. This guidance also sets out where certain harms or safety issues are likely to require a more rigorous or particular response to meet the relevant Expectation.

Providers are expected to have regard to this guidance, as set out in section 7 of the Determination.

Further detail on the reasonable steps is also included in the Explanatory Statement to the Determination.

Providers must also comply with any other relevant legal obligations when implementing the Expectations, such as the *Privacy Act 1988 (Cth)*.

Interaction with industry codes and industry standards

What are industry codes and industry standards?

The industry codes and industry standards are mandatory requirements that apply to particular sections of the online industry. Industry codes are developed by industry associations that represent those sections of the online industry, and industry standards are determined by the eSafety Commissioner.

On 31 May 2023, the Commissioner decided to register industry codes for five sections of the online industry,[11] including social media services. On 7 September 2023, the Commissioner decided to register search engine services code. eSafety will develop standards for two sections of the online industry – relevant electronic services and designated internet services.

Unlike the Expectations, compliance with the industry codes and industry standards requires mandatory minimum compliance measures that eSafety is able to enforce through the courts, as well as through other means.

[11] See eSafety's register of industry codes: Register of industry codes and industry standards for online safety | eSafety Commissioner.

What do industry codes and industry standards address?

The Act provides for the introduction of industry codes and/or industry standards to address class 1 and class 2 material. The class 1 material covered under the first phase of industry codes and industry standards covers child sexual abuse material (CSAM),[12] child sexual exploitation material (CSEM)[13] and pro-terror material,[14] as well as material that deals with crime and violence and drug-related content.

The registered industry codes represent the mandatory and enforceable measures that industry must meet in order to comply with their legally binding obligations in relation to class 1 material.

Relationship between industry codes, industry standards and the Expectations

The obligations in industry codes and industry standards will be narrower in scope than the Expectations as they focus on class 1 material (and class 2 material in the future) rather than the broader unlawful and harmful material and activity covered by the Expectations.

In some cases, specific mandatory steps to address class 1 material required under an industry code (or an industry standard) will be directly relevant to an Expectation, including requirements under an industry code or industry standard to:

- undertake risk assessments and ensure safety by design (section 6 of the Determination)
- minimise the provision of class 1 material (sections 6 and 11(d) of the Determination)
- provide reporting and complaint mechanisms for end-users (sections 13 and 16 of the Determination)
- ensure terms of use, policies and procedures are in place to address class 1 material, and enforcing these rules (sections 14, 15, 17 and 18 of the Determination).

[12] For the purposes of industry codes, CSAM is a sub-category of class 1 material to the extent that it is comprised of visual depictions of child sexual abuse.

[13] For the purposes of industry codes, CSEM is a sub-category of class 1 material that is broader than CSAM, and includes material relating to the promotion or provision of instruction in paedophile activity, includes or contains descriptions or depictions of child sexual abuse or any other exploitative or offensive descriptions or depictions involving a person who is, or appears to be, a child under 18, or describes or depicts in a way that is likely to cause offence to a reasonable adult, a person who is or appears to be a child under 18 (whether or not the person is engaged in sexual activity).

[14] For the purposes of industry codes, pro-terror material is class 1 material that advocates the doing of a terrorist act.

Compliance with the requirements in an industry code or industry standard is relevant to a provider's implementation of certain expectations (in relation to class 1 material) but will not be determinative of meeting any particular Expectation.

This is because what is 'reasonable' for a provider to do to address unlawful and harmful material under the Expectations may extend beyond the *minimum* requirement in the mandatory (and enforceable) industry code or industry standard. Additional steps may be required to meet the applicable Expectations.

Additionally, the Expectations apply to a broader range of harmful material (beyond class 1 material), and to harmful activities.

Table 2: Differences between industry codes and industry standards, and the Expectations

	Applies to	Applies to unlawful and harmful 'material'	Applies to unlawful and harmful 'activity'	Consequences for failure to comply
Basic Online Safety Expectations	• Social media services • Relevant electronic services • Designated internet services	Yes	Yes	eSafety may prepare, and publish, a Statement of Non-Compliance with one or more Expectations. eSafety has a range of enforcement options in relation to ensuring compliance with a reporting notice or determination.
Industry Codes (Phase 1)	• Social media services • App distribution services • Hosting services • Internet carriage services • Manufacturers, maintenance and installation providers of equipment • Search engine services	Applies only to class 1 material	Applies to certain activities that affect the provision of class 1 material	eSafety may issue a formal warning, or written direction to comply with an industry code. Failure to comply with a direction may result in enforcement through an enforceable undertaking or injunction. It may also result in an infringement notice or civil penalty proceedings.
Industry Standards (Phase 1)	• Relevant electronic services • Designated internet services	Applies only to class 1 material	Applies to certain activities that affect the provision of class 1 material	Failure to comply with an industry standard may result in a formal warning, enforcement through an enforceable undertaking or injunction. It may also result in an infringement notice or civil penalty proceedings.

Interaction with other regulatory requirements in the Act

Failure to comply with an expectation under the Determination may result in other enforcement action by eSafety. For example, eSafety has the power to give providers a removal notice in relation to specific material under the four complaints based reporting schemes.[15] These powers may need to be exercised more frequently if a provider has failed to take reasonable steps to minimise the provision of certain material on their service (section 11 of the Determination). Failure to comply with a removal notice is a civil penalty provision and may result in a range of enforcement actions by eSafety. eSafety does not need to establish that a provider failed to comply with section 11 of the Determination (or an industry code or industry standard) prior to giving a removal notice.

Additional information about eSafety's regulatory schemes and powers is available on eSafety's website.

[15] eSafety can investigate reports of cyber-bullying of children, adult cyber abuse, image-based abuse (sharing, or threatening to share, intimate images without the consent of the person shown) and illegal and restricted content. More information on these schemes is available on the eSafety website: Report online harm | eSafety Commissioner.

Part 2: Reporting powers

Reporting and information gathering powers

A core element of the Act is to empower eSafety to seek information from providers on their compliance with the Expectations. This information is sought to improve transparency and accountability, and to assist eSafety determine whether a provider is compliant with the Expectations.

There are three ways eSafety can seek information from providers regarding compliance with the Expectations.

1. **Requests for information**
   As part of the Expectations (section 20 of the Determination), eSafety may request information about:
   - the number of complaints about breaches of a provider's terms of use
   - the time frame for responding to removal notices given to the provider by eSafety
   - measures taken to make sure people can use the service in a safe manner
   - the performance of online safety measures that providers have announced publicly or reported to eSafety.

   A failure to respond within 30 days is non-compliance with the Expectations. This gives the Commissioner discretion to prepare a statement that the provider is not complying with the Expectations. Providers should consider whether they have processes in place to respond to these requests. For more information on section 20, see Part 4 of this guidance.

2. **Reporting notices**
   eSafety may give a reporting notice to a provider requiring them to produce a report on their implementation in relation to one or more expectations. These notices are enforceable, backed by the power to seek civil penalties and other enforcement mechanisms.

   Reporting notices are specific to the provider, and can require:
   - non-periodic reporting
   - periodic reporting at regular intervals of between 6 and 24 months for as long as the Commissioner deems appropriate.

   eSafety intends to give periodic reporting notices to providers in order to track the development and improvement of tools, processes, and their effectiveness. Periodic reporting notices may focus on specific harms and issues that have already been

identified through eSafety's use of non-periodic reporting notices, or a range of other issues.

3. **Reporting determinations**
   eSafety can make a reporting determination – a legislative instrument – requiring periodic or non-periodic reporting for a specified class of services. Like the reporting notices, these are enforceable and backed by civil penalties and other enforcement mechanisms.

Type of Information Gathering	Can require reporting on	Periodic or non-periodic	Reporting period	Time to respond	Enforceable
Requests for information section 20 of the Expectations	**1.** Terms of service complaints; **2.** The timeframe for responding to removal notices **3.** Measures taken to make sure people can use the service in a safe manner **4.** The performance of online safety measures that providers have announced publicly or reported to the the Commissioner	Non-periodic	Not shorter than 6 months for reporting categories 1 and 2. N/A for reporting categories 3 and 4	Within 30 days	✗
Reporting notices to **individual providers**	Implementation of any part or the entirety of the Expectations	Either periodic or Non-periodic	6 to 24 months	28 days or longer as specified	✓
Reporting determinations to a **specified class of providers**	Implementation of any part or the entirety of the Expectations	Either periodic or Non-periodic	6 to 24 months	28 days or longer as specified	✓

eSafety's approach to the use of reporting and information gathering powers

eSafety is taking a phased approach in exercising its powers related to the Expectations, starting with the use of non-periodic reporting notices with a focus on specific expectations and acute issues of particularly high harm, such as CSEA. eSafety intends to expand the use of its statutory powers related to the Expectations over time, with the first periodic reporting notices intended to be given in 2023-24.

eSafety is committed to a number of principles.

- Applying eSafety's powers under Part Four of the Act in a fair and proportionate way, based on evidence and insights.
- Taking an open and transparent approach – both in exercising eSafety's powers, and in terms of the information obtained through notices. eSafety intends to make information obtained through use of reporting notices and determinations publicly available where appropriate in the interests of transparency and accountability.
- Recognising the importance of reducing regulatory requirements by considering information that:
    - providers already publish voluntarily
    - is provided as part of international transparency initiatives
    - is provided to eSafety under another regulatory scheme, including reporting obligations through an industry code or industry standard.
- Recognising that differences between providers in terms of resources, risk, technical architecture and user base, means that 'one size does not fit all'.
- Taking a consultative approach, seeking input and feedback from providers as well as from civil society organisations, academics and other experts to ensure implementation meets standards of good regulatory practice.
- Ensuring eSafety systems securely store information, including information which is commercial-in-confidence, personal information or information, which if disclosed would adversely affect public safety.

Complying with a notice, determination, or request for information from eSafety

Reporting notices may require information such as:

- qualitative information on safety tools, processes and policies, and why these are reasonable steps to implement the Expectations - these may be phrased as yes/no questions, multiple choice questions or worded to seek descriptive information
- quantitative information on the operation of safety tools, processes and policies - this may consist of metrics to determine the impact of interventions or information about the resources allocated.

Reporting notices will be related to specific expectations. Responses will be used to understand the extent to which a provider is compliant with one or more expectations as well as increasing transparency through building an understanding across different providers of common practices, trends and challenges. Given the breadth of some of the expectations, eSafety is likely to ask questions targeted at assessing how the provider's compliance with a particular expectation minimises specific types of harms. Targeted questions assist providers and eSafety by ensuring the provision of meaningful information. It also minimises the regulatory burden on providers and encourages transparency and accountability about issues that impact on the online safety of Australians.

Providers are required under the Act to respond to a reporting notice in the manner and form specified.[16] eSafety provides a response template as part of a notice and providers must respond to questions in the manner and form specified in that template. Providers should engage with eSafety if they cannot answer in the form specified. Providers are required to respond within the time frame specified. In line with the Act, the time to respond will be no shorter than 28 days from the giving of a notice, or from the end of the reporting period specified in the notice. eSafety will consider the appropriate length of time for a provider to respond to a notice on a case-by-case basis.

eSafety understands that not every expectation will apply equally to every service. If a provider is of the view that a particular expectation or question does not apply, they should contact eSafety **before** providing their response to the notice.

Where a provider does not collect and is not capable of obtaining the required information, they should endeavour to provide alternative relevant data. For example, where Australian data cannot be disaggregated from regional or global data, a broader dataset may be acceptable.

[16] Sections 49(2)(b) and 56(2)(b) of the Act.

Providers are required to provide information in response to a reporting notice even if that information is considered commercial-in-confidence or covered by a confidentiality obligation in a third-party contract. As set out on page 18, providers will be asked to clearly identify any information they believe should not be published.

eSafety will also endeavour to inform a provider of the intention to give a reporting notice, and the intended scope of the proposed notice, before it is given to them. The purpose of this is to enable the provider to identify any specific barriers to compliance within the proposed time frame of the notice and to confirm the appropriate entity for receipt of the notice. However, advance notice may not be possible in every circumstance. For example, this might not be appropriate where a provider has not previously engaged in a constructive or reasonable manner with eSafety or where there are factors leading to a degree of urgency.

If a provider does not respond to a notice or comply with its requirements, eSafety has civil enforcement powers,[17] and the power to issue a formal warning,[18] or prepare and publish a statement that the provider is non-compliant.[19]

In addition to the information provided in response to a specific question in a notice, providers can share additional information and context with eSafety as part of their response to the notice.

In the interests of consistency, enforceability and transparency, where eSafety has decided that a notice is the appropriate mechanism, eSafety will not normally agree to withhold a formal notice and agree to the same information being provided voluntarily.

[17] The maximum penalty for non-compliance with a reporting notice under sections 50 and 57 of the Act is 500 penalty units for an individual and can be multiplied by 5 for a body corporate (at the date of publication of this guidance, a single penalty unit is $313). In cases of non-compliance, eSafety may give an infringement notice, initiate civil penalty proceedings, apply for an injunction or enter into an enforceable undertaking under the *Regulatory Powers (Standard Provisions) Act 2014*.

[18] Sections 51 and 57 of the Act.

[19] Sections 55 and 62 of the Act.

How does eSafety decide which providers receive notices?

When deciding which providers to give a notice to, the Act requires eSafety to have regard to these specified criteria:[20]

- the number of complaints eSafety has received under the Act in relation to the service in the previous 12 months
- any deficiencies in the provider's safety practices and/or terms of use
- any previous contraventions of civil penalty provisions relating to the Expectations
- whether the provider has agreed to give the Secretary of the Department regular reports relating to safe use of their service[21]
- any other matters the Commissioner considers relevant.

Examples of other matters that the Commissioner might consider relevant may include:

- aggregated evidence from eSafety's other regulatory schemes, such as types of complaints, a service's responsiveness to removal requests or notices, or other investigative insights regarding a service's safety issues
- a service's reach and the profile of its end-users, including whether the service is used by children
- higher risk design choices and features, such as livestreaming and end-to-end encryption (E2EE)
- the measures the service currently has in place to protect end-users from harm
- evidence of systemic harm, or evidence of key safety issues, including from victims, civil society organisations, media, academics, or other experts
- the information already published by a provider, as well as any lack of information regarding a service's safety policies, processes and tools, or limited information about the impact or effectiveness of these interventions.

The same requirements do not exist if eSafety makes a determination requiring reporting from a specified class of services. However, eSafety intends to take a similar approach to understanding risk and priority sectors prior to making any determination.

---

[20] Section 56(5) of the Act.
[21] This provision was included to ensure that eSafety takes into account other Australian Government reporting initiatives, and considers the burden on providers from any duplication.

Reporting on compliance with the Expectations and industry codes and standards

Certain providers will be required to provide reports to eSafety under an industry code or industry standard, either as a matter of course or at the request of eSafety, depending on the application of the particular code or standard.

eSafety will seek to reduce regulatory burden in reporting requirements where possible and where appropriate. For example, where a provider has reported information in response to a notice related to the Expectations, they may refer to this information - insofar as it is relevant – for the purposes of preparing an annual report under an industry code or industry standard.

However, in some instances, the public interest in transparency and accountability will outweigh any potential administrative burden in reporting certain information.

Importantly, information obtained through a reporting notice given in connection with the Expectations, could be considered by eSafety in considering a provider's compliance with an industry code or industry standard.

Is information received via reporting notices and determinations published?

The Explanatory Memorandum to the Act highlights the objective of the Expectations to 'improve the transparency and accountability of online service providers for the safety of their users and the mitigation of online harms'. It further notes that:

> **The transparency reporting obligation within the BOSE [Basic Online Safety Expectations] proposal would create greater transparency of the online safety practices for both government and the community, and encourage uplift through imposing reputational costs for non-compliance.**

eSafety considers that the transparency and accountability objectives of the Act are most effectively met by making information received from industry in response to a reporting notice public, where appropriate. This transparency improves and promotes the online safety of Australians by increasing awareness of online safety issues and the way that services respond to online harms.

Providers will be asked to:

- clearly identify in their response if any information is commercial-in-confidence or should otherwise not be published, for example because it would adversely affect public safety
- provide clear reasons in support of any claim that certain information should not be published.

eSafety considers these claims carefully. eSafety considers whether there are steps that can be taken to protect such information while ensuring the transparency and accountability objectives of the Act are still met. eSafety's approach to information that could impact public safety will be informed by its own expertise, engagement with external experts, and other sources.

In line with the transparency objectives of the Act, eSafety may disclose the names of providers given a notice at the time a reporting notice is given, along with a summary of the information sought in the reporting notice. The number and type of reporting notices given, and outcomes (such as whether a notice was complied with and whether any enforcement action was taken), will also be published in eSafety's annual report.

Review rights

A provider may seek either internal review or external review by the Administrative Appeals Tribunal of certain actions taken by eSafety relating to the Expectations. The purpose of these review rights is to ensure that eSafety has made the correct and preferable decision on a case-by-case basis.

Action which can be reviewed	Who can seek review
The giving of a non-periodic reporting notice (Section 49 of the Act)	The provider named in the non-periodic reporting notice
The giving of a periodic reporting notice (Section 56 of the Act)	The provider named in the periodic reporting notice

An internal review may not always be appropriate, particularly if the reporting notice has been given by the Commissioner. Additional information about seeking a review can be found on eSafety's website.

Part 3: Assessing compliance with the Expectations

Statements of compliance or non-compliance

If eSafety decides that a provider is not complying with one or more of the Expectations, the Act empowers eSafety to prepare and publish a statement to that effect. eSafety may also publish a statement that confirms that a provider is meeting the Expectations. This supports transparency and encourages best practice. These are referred to as 'service provider notifications' in the Act.[22] eSafety uses the terms 'statements of compliance' and 'statements of non-compliance' to differentiate them from other kinds of service provider notifications in the Act.

If a statement is prepared, eSafety will share this statement with the provider. If eSafety decides to publish the statement, the provider will be given the opportunity to make submissions including evidence to demonstrate that it is compliant with the relevant expectation(s) or reasons that it should not be published.

eSafety will have regard to this regulatory guidance in assessing whether a provider is compliant with one or more expectations.

eSafety's approach to assessing compliance

The Determination does not prescribe how the Expectations must be met, although it does contain examples of reasonable steps that could be taken within some sections of the Determination. The Determination affords flexibility to providers to determine the most appropriate method of complying with the Expectations, and eSafety supports this approach.

Additional examples of reasonable steps are provided in Part 4 of this guidance to assist providers in complying with each applicable expectation. Providers are expected to have regard to this guidance in ensuring they are compliant with each applicable Expectation.

[22] Section 48 of the Act.

How will eSafety decide whether to give and publish a statement of non-compliance?

eSafety will take a risk-based approach when assessing whether providers are taking reasonable steps to comply with the Expectations, taking into account the level of harm and extent of the safety issues relating to a service.

A statement of non-compliance can be published for a failure to comply with one or more expectations, although eSafety recognises that not all expectations will apply to all services. For example, if a service does not use encryption or permit anonymous accounts, then sections 8 or 9 may not apply. In some instances, where there is no appreciable risk of harm, it would also not be proportionate for eSafety to expect steps to be taken in relation to certain expectations.

The Commissioner will consider a number of factors[23] when assessing whether a provider of a service has complied with the Expectations or whether they have contravened an expectation, including the following:

- The risks related to the service, including:
    - the number of end-users, including Australian end-users
    - the user base and demographics of those end-users
    - risk and evidence of online harms
    - design features that may increase risk or limit the effective use or operation of any safety measures
    - other relevant factors.
- The effectiveness and proportionality of the steps taken by a provider in meeting an expectation.
- Whether there are any particular technical or practical limits which might prevent a provider from taking certain steps to meet the Expectations.
- The resources available to the provider and the costs or other burden to implement certain steps.
- Substantiated information establishing that a provider has plans to take further action or other steps in the short to medium term.

[23] This is not intended to be an exhaustive list of factors that the Commissioner may consider in assessing a provider's compliance with the Expectations.

- Whether the provider has engaged constructively with eSafety and responded to requests for information.
- How information provided in response to a notice compares with relevant evidence from other sources, such as eSafety's investigative insights, industry codes or industry standards reporting, as well as academic, civil society, or other expert evidence.

eSafety intends to publish statements of non-compliance on the eSafety website.

How will eSafety decide whether to give and publish a statement of compliance?

eSafety can only publish a statement of compliance if a provider has met all relevant expectations at all times during a specified period. This constitutes a higher bar than a statement of non-compliance which can be given for the failure to implement any individual expectation.

Similar to a statement of non-compliance, eSafety will take into account a number of factors when deciding whether a provider is complying with the Expectations, including the following:

- Evidence that a provider has implemented reasonable steps across all the relevant expectations, with evidence that these are operating effectively and consistently.
- Evidence that the reasonable steps have been taken and implemented for a reasonable time in order to evaluate their effectiveness.
- Whether the provider has engaged constructively with eSafety and responded positively to requests for information.
- How information provided by the service compares with evidence from other sources, such as investigative insights, academic, civil society or other expert evidence.

To support a decision that a provider has complied with all relevant expectations during a specified period, providers will need to demonstrate the effectiveness of their safety measures. Providers are encouraged to collect relevant information and metrics internally to evaluate the effectiveness of their safety interventions, and to provide these to eSafety – such as by responding to a non-periodic or periodic reporting notice.

eSafety intends to publish statements of compliance on the eSafety website.

Part 4: Examples of reasonable steps to comply with the Expectations

Overview

This part sets out examples of the reasonable steps that providers could take to comply with the Expectations.

The Determination does not prescribe how the Expectations must be met but includes non-exhaustive examples of reasonable steps. This guidance identifies further steps that eSafety considers would assist providers in complying with the Expectations. This is not an exhaustive list. eSafety recognises that each service is different and new technologies continue to emerge which may assist with complying with the Expectations. Providers may elect to take different steps to meet the Expectations that better suit their service and the risks posed. Providers should be prepared to report on these steps, why they are reasonable in light of the objectives of the Determination, and how these steps meet the relevant Expectations and keep Australians safe online.

As set out in the Explanatory Statement to the Determination, the Commissioner will take a risk-based approach towards assessing compliance, noting that what is 'reasonable' to comply with the Expectations may differ depending on the nature and severity of the harms and risks on a service.

Providers are expected to prioritise responding to the most harmful risks on their service, particularly where these involve unlawful material or activity, or where they impact on groups at higher risk. However, providers are also expected to take reasonable steps to address other harmful material and activity occurring, or likely to occur, on their service.

Unlawful and harmful material and activity may arise online as a result of human-generated content and conduct, but may also be generated artificially, and shared or otherwise misused in similar ways. The Act recognises this in relation to class 1 material (which includes material that describes or depicts a child under 18 or a person who 'appears to be' a child under 18 in relation to child sexual exploitation and abuse),[24] and in relation to image-based abuse (the non-consensual sharing of intimate images)[25] by including images that have been digitally or artificially generated.

The Expectations apply to material and activity that is unlawful and harmful, regardless of how it is generated. Providers should therefore take steps to address and mitigate the harms of the emerging technologies, including the ability to generate synthetic material, and where providers introduce or integrate features into their existing services which involve artificial intelligence (such as chatbots, among others). The Expectations also apply where services enable end-users to post synthetic material generated that was generated elsewhere.

For more information on eSafety's position on emerging technologies and trends, including Generative AI and how to take a safety-by-design approach to these issues, see eSafety's Position Statements.[26]

Reasonable steps

The Expectations require providers to take 'reasonable steps' to address various safety issues.

The term 'reasonable' is not defined in the Act or the Determination and bears the ordinary meaning as being based upon or according to reason, and capable of sound explanation.

What steps are reasonable is a question of fact in each individual case and is an objective test that has regard to how a reasonable person, who is properly informed, would be expected to act in the circumstances. What is reasonable can be influenced by current standards and practices, the nature and extent of the harms involved that require mitigation, as well as by other legislative requirements or obligations that apply to each provider.

It is the responsibility of each provider to be able to justify why the steps they are taking are reasonable, and how these steps amount to compliance with the Expectations.

[24] As defined in section 106 of the Act.
[25] As defined in section 15 and 16 of the Act.
[26] eSafety website, Tech Trends and Challenges, Tech trends and challenges | eSafety Commissioner.

Consultation

Section 7 of the Determination sets out the expectation that providers will consult with the Commissioner in determining the reasonable steps to ensure safe use.

eSafety has engaged with industry on online safety issues and on the development of updated guidance. Providers are also encouraged to engage with eSafety regarding their specific services, as the reasonable steps are likely to differ depending on factors outlined above under 'reasonable steps', as well as a service's risks, business model, user base, technical architecture and design.

eSafety intends to update this guidance as needed in response to new harms, technologies and safety issues, or in response to other events.

The Determination sets out an expectation that providers will have regard to any relevant guidance material made available by the Commissioner (section 7(2)).

Chapter 1: Expectations regarding safe use

Division 2 of the Determination sets out expectations in relation to ensuring safe use of a service in the following sections.

- Section 6: take reasonable steps to ensure that end-users are able to use the service in a safe manner and take reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful.
- Section 7: consult with the Commissioner in determining what reasonable steps are for the purpose of section 6(1) and refer to the Commissioner's guidance in determining such reasonable steps to ensure safe use.
- Section 8: on an encrypted service, take reasonable steps to develop and implement processes to detect and address material and activity that is unlawful or harmful.
- Section 9: take reasonable steps to prevent anonymous accounts from being used to deal with material, or for activity, that is unlawful or harmful.
- Section 10: take reasonable steps to consult and cooperate with other service providers to promote the ability of end-users to use all services in a safe manner.

Further guidance on steps that providers may take to ensure compliance with these expectations is set out in the following pages.

Section 6 of the Determination – Ensuring safe use and proactive minimisation of unlawful and harmful material and activity

Determination, section 6:

Core expectation

1. The provider of the service will take reasonable steps to ensure that end-users are able to use the service in a safe manner.

Additional expectation

2. The provider of the service will take reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful.

Examples of reasonable steps that could be taken

3. Without limiting subsection (1) or (2), reasonable steps for the purposes of this section could include the following:

a. developing and implementing processes to detect, moderate, report and remove (as applicable) material or activity on the service that is unlawful or harmful;

b. if a service or a component of a service (such as an online app or game) is targeted at, or being used by, children (the children's service)—ensuring that the default privacy and safety settings of the children's service are robust and set to the most restrictive level;

c. ensuring that persons who are engaged in providing the service, such as the provider's employees or contractors, are trained in, and are expected to implement and promote, online safety;

d. continually improving technology and practices relating to the safety of end-users;

e. ensuring that assessments of safety risks and impacts are undertaken, and safety review processes are implemented, throughout the design, development, deployment and post-deployment stages for the service.

The intention of section 6(1) is to uplift how services develop and implement products, policies and terms in a way that has regard for the safety of Australian end-users.[27]

[27] Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 12, URL: https://www.legislation.gov.au/Details/F2022L00062/Explanatory%20Statement/Text

Importantly, providers should continually assess and evaluate the effectiveness of online safety measures deployed on a service or designed into a service, and update, refine and adjust these measures accordingly to ensure safe use.[28]

Section 6(2) requires providers to take proactive steps to identify and address existing and emerging harms online, and section 6(3) outlines a range of steps that providers could take to meet these Expectations.

Risk and impact assessments

Undertaking safety risk and impact assessments and reviews are listed as examples of a reasonable step in section 6(3)(e). Assessments should:

- be a **priority** throughout the service or feature lifecycle. It is especially important when a new feature is designed, developed, and deployed to ensure harms are mitigated from the earliest stages
- be undertaken routinely, clearly documented, and updated regularly
- be informed by a human rights approach – meaning that the likelihood and severity or impact of harms occurring should be considered from the point of end-users and the community more broadly, and take into account other applicable human rights
- be informed by community and victims' groups and other expert insights to ensure all relevant risks are understood, and the impacts of any proposed safety mitigations are also assessed and mitigated
- not be limited to consideration of how a risk or a harm impacts the provider as a business, or from a narrow compliance perspective (although providers should ensure they assess whether they are complying with the Expectations as part of this process).

eSafety recognises that complete mitigation of all harms may not be possible, and the Expectations do not require this outcome. However, providers should be prepared to report on the nature of the safety risk assessments undertaken, what safety risks were identified, how the risk assessment recommends the risks be mitigated, and what steps the provider has taken to implement these recommendations.

Providers may already undertake other risk assessments, for example privacy or human rights impact assessments. While safety risks and impacts could be considered as part of

[28] Providers should note that the Commissioner may request a report on the performance of online safety measures that a provider has publicly announced or otherwise reported to the Commissioner, under section 20 of the Determination. The Commissioner may also require information on the performance of online safety measures through mandatory reporting notices under the Act.

these broader processes, eSafety expects that providers will thoroughly identify and address the specific safety issues.

Relevant industry code and industry standard measures

Providers may be required to undertake certain safety by design risk assessments under the applicable industry code or industry standard in relation to class 1 content to determine each service's risk profile, which informs how a relevant code or standard applies. Where these risk assessments (including any risk assessments carried out as a result of a change to the risk profile) result in the identification and addressing of harms related to certain material and activity (such as class 1 content), this may constitute a reasonable step for the purposes of section 6(2) in relation to unlawful material and activity.

However, eSafety expects that risk assessments will be undertaken to identify, address and mitigate a broader range of harms and material in order to comply with the Expectations.

The section 6 expectations require providers to take steps in relation to both material and activity. It is important for providers to consider how certain material, or certain activity, may be harmful in some circumstances and less so in others. The severity or impact of a harm may vary for different individuals, or groups within the community.

Additionally, providers of services that permit children or young people to use their service or that are likely to be accessed by children should ensure that risk assessments involve consideration of the risks faced by this younger cohort. For example, risk assessments should consider risks related to content (a child or young person engaging with, or being exposed to, certain content), contact (experiencing, or being targeted by, potentially harmful contact, including by adults) and conduct (witnessing, participating in, or being a victim of harmful conduct).

For a structured framework to consider and mitigate safety risks in the design, development and deployment of services, see eSafety's Safety by Design tools.

- There are two tools – one designed for early-stage companies and another for mid-tier and enterprise organisations.
- For each tool, end-users are provided with an educative module on online harms, and are taken through a series of question and response options which culminate in a tailored end report, guiding and supporting providers to enhance online safety practices.

Resourcing of safety interventions and teams

Another key example of a reasonable step to comply with section 6 is ensuring that a service's safety interventions are resourced proportionate to the risks identified and to enable compliance with the Expectations. This should involve:

- appropriately resourcing trust and safety teams, to ensure that appropriate safety interventions are in place, that interventions are working effectively, and that safety issues are responded to as a priority
- ensuring all relevant staff are suitably trained and supported, including through training on Safety by Design principles – there should be specialist training for trust and safety teams, and trust and safety functions should be subject to oversight and accountability by senior management
- trust and safety teams engaging with experts in online safety and technology, as well as victims, to inform policies and processes
- having clear and effective escalation processes to refer complex or specialist cases to expert teams.

Providers should invest in the development of tools and processes to support their compliance with the Expectations. This includes research and development into technology to detect, disrupt and deter unlawful and harmful material and activity. Investment should be proportionate to the resources of the provider, and the risks posed by the service.

Moderation

Content moderation, where provided by a service, should be provided in a range of relevant languages to support the demographics of a service's end-users. This is particularly important for harms that require context to identify, such as grooming or hate speech. This helps ensure that unlawful and harmful content is properly identified, and the accuracy of content moderation decisions.

Community moderation may be a useful mechanism to support alignment of material and activity on a service with the terms of use, standards of conduct and other service policies. However, it is important that the burden of enforcing terms of use, standards of conduct and otherwise addressing unlawful and harmful material and activity is not delegated solely to community moderation.

Where community moderation is used, it is important that community moderators are properly supported and equipped with information and tools from the service, and this should include requirements to escalate certain issues to the provider and professional trust and safety staff. This escalation is important so that trust and safety staff can take appropriate action including banning accounts across all parts of a service (not just the

section that the violating conduct was identified within) and making onward reports to appropriate authorities.

Community moderated services must always retain an appropriate level of visibility over the activity on their service. This responsibility should never sit solely with community moderators or other end-users.

Proactively minimising the extent to which material or activity on the service is unlawful or harmful (section 6(2))

There is considerable cross-over between this expectation (section 6(2)) and the section 11 and 12 expectations to minimise certain material and class 2 material. Many of the reasonable steps to comply with those expectations will support compliance with section 6(2). However, section 6(2) is broader than section 11 and 12, including by capturing unlawful or harmful **activity** as well as **material**.

Importantly, this section expects '**proactive**' minimisation of unlawful and harmful material and activity. This means providers are expected to take reasonable steps upfront to reduce the likelihood of such material being made available, or activity taking place, on the service. The key example of how this can be achieved is via the use of technologies or other tools. Proactive steps can be contrasted with reactive or responsive measures such as user reporting mechanisms or community moderation which should work alongside proactive steps, but which may be insufficient to demonstrate compliance with section 6(2).

Recommender systems

Providers that use a recommender system on any part of a service should consider the safety risks that currently exist or may arise as a result of these systems.

Recommender systems, also known as content curation algorithms, are the systems that prioritise content or make personalised content suggestions to users of online services.

The different inputs and end goals for recommender systems can lead to both positive and negative outcomes. For example, recommender algorithms that prioritise user engagement and then serve up similar content in the future may result in people seeing things they find interesting, entertaining or valuable. But equally, if an end-user spends time engaging with potentially harmful content, those same metrics may lead to them seeing more of the same material or increasingly extreme material in their feeds.

In addition to risks and harms at an individual level, recommender systems have the potential to cause new, or exacerbate existing harms on a societal level – for example, content promoting hate or inciting violence can cause harm to the people targeted and can also spill over into violence and discrimination affecting the broader community.

The following safety interventions may be used in relation to recommender systems,[29] but also apply more broadly in relation to ensuring safe use of a service.

- Providing opt-in or opt-out measures for end-users to maintain choice, ownership and control of the types of content they receive.
- Adjusting recommender algorithms to focus on other metrics such as authoritativeness or diversity of content as an alternative, or in addition to, user-engagement. These metrics should be subject to consultation, public scrutiny and testing.
- Offering end-users alternative curation models for their news feeds.
- Using human review as a safety check for content that is being rapidly disseminated or promoted.
- Introducing additional safeguards through design features, such as prompts to read an article linked before sharing it, which may reduce the likelihood of it being shared.
- Labelling content as potentially harmful or likely to include certain themes or topics, particularly where content may be sensitive to some higher risk groups and communities and not others. Where content warnings are provided to some end-users and not others, consideration should be given to the data which informs these choices and the risk of bias.
- Including behavioural cues and prompts that can help end-users establish positive patterns of behaviour – for example, that help end-users reconsider posting harmful content or manage their time spent online.
- Enhancing transparency reporting and auditing practices.
- Curating recommendations so they are age appropriate, including friend or follower suggestions between adults and children.
- Offering parental controls to allow parents and carers to limit and/or monitor what material and activity their child is exposed to and engages with, with the ability to adjust these settings as children develop and their capacity evolves.
- Employing measures to test and update recommender systems with the objective of improving overall safety – for example, internal audits, external audits, risk and impact assessments, a/b testing.
- Preventing autocomplete searches of phrases that are likely to be associated with unlawful or harmful content.

[29] For more information, see eSafety's Position Paper on recommender systems and algorithms: Recommender systems and algorithms – position statement | eSafety Commissioner.

For more detailed guidance on reasonable steps to minimise unlawful or harmful material or activity, see Chapter 2 on section 11 (Minimise provision of certain material) and section 12 (Preventing children's access to class 2 material).

Section 7 of the Determination – Consulting with the Commissioner and referring to the Commissioner's guidance

Determination, section 7:

Core expectation

1. In determining what are reasonable steps for the purposes of subsection 6(1), the provider of the service will consult the Commissioner.

Additional expectation

2. In addition, in determining what are reasonable steps for the purposes of subsection 6(1), the provider of the service will have regard to any relevant guidance material made available by the Commissioner.

Section 7(1) intends to establish a dialogue between the Commissioner and service providers. It gives providers the opportunity to outline and justify the steps they take to ensure safe use, including in circumstances where the examples included in section 6(3) are not appropriate for a service and alternative steps are taken. Section 7(1) also establishes a means for information sharing between the Commissioner and industry to improve online safety outcomes.

Providers can contact eSafety at industrybose@esafety.gov.au if they have specific questions regarding reasonable steps and their ability to comply with the Expectations (although eSafety cannot provide legal advice). Providers are also expected to engage with eSafety if specific safety issues related to a service are identified, and a provider's willingness to engage and implement or consider eSafety's recommendations may be reflected upon when deciding whether a provider is complying with the Expectations.

Section 7(2) requires that in determining what are reasonable steps for the purposes of complying with section 6(1), a provider will have regard to any relevant guidance material made available by the Commissioner.

This guidance is made available to providers to assist them in meeting the Expectations. Providers are expected to have regard to this guidance material in implementing the Expectations, alongside the Safety by Design tools on the eSafety website, and other relevant materials published by eSafety.

This guidance may be updated in the future where additional guidance is required in relation to new harms, technologies and safety issues or in response to other events, or to include the responses to common questions from providers raised during section 7 engagement.

Further opportunities for consultation will be afforded to providers if they receive a non-periodic or periodic reporting notice which requires a provider to produce a report on their compliance with any or all of the Expectations. Further information is set out in Part 3 of this guidance.

Section 8 of the Determination – Detecting and addressing unlawful or harmful material or activity on encrypted services

Determination, section 8:

Additional expectation

1. If the service uses encryption, the provider of the service will take reasonable steps to develop and implement processes to detect and address material or activity on the service that is unlawful or harmful.
2. Subsection 8(1) does not require the provider of the service to undertake steps that could do the following:
    a. implement or build a systematic weakness, or a systematic vulnerability, into a form of encrypted service;
    b. build a new decryption capability in relation to encrypted services; or
    c. render methods of encryption less effective.

Encryption is a way to prevent unauthorised access to information. Encryption is not new and, in its modern form, has been used for more than 40 years as an essential tool for privacy and security. It is primarily employed for the secure transmission and storage of information, and can help to prevent data breaches and hacking.

Section 8 applies to services that are encrypted in any form, including those using 'in transit' encryption such as Transport Layer Security, encryption at rest, and those using end-to-end encryption (E2EE). The reasonable steps that a provider should take to develop and implement processes to detect and address material or activity that is unlawful or harmful may depend on the nature of the encryption implemented on the service, and whether encryption is used on some, or all, parts of a service.

Services that use encryption in transit and/or at rest should take reasonable steps to detect unlawful and harmful material and activity on their service. This may involve the use of both

automated tools such as hash matching or Artificial Intelligence (AI) classifiers, and human review. Further details are set out in the guidance on the section 6 and 11 expectations.

For providers that use E2EE on all or part of a service, there is a higher risk of unlawful and harmful material and activity going undetected, given the limitations E2EE creates for widely used detection technologies and interventions. Services that allow large groups, live streaming or video calling, and E2EE services that enable end-users to connect to other unknown users on the basis of shared interests, are also likely to pose greater risks.

While section 8 makes it clear that the Expectations do not require providers make E2EE less effective,[30] providers are required to take reasonable steps to develop and implement processes to both detect and address material or activity that is unlawful and harmful.

Reasonable steps to **detect** unlawful and harmful material and activity on E2EE services may include a number of options.

- Using hashing, machine learning, artificial intelligence and other detection technologies on any parts of the service that are not E2EE (such as profile pictures, content in user reports, group names).
- Using technology that enables unlawful and harmful material and activity to be detected at the device level or prior to upload on the service, where this can be done without building a systematic weakness or vulnerability (such as client-side scanning using hashing, AI classifiers, natural language processing of text to detect patterns indicative of grooming of children and sexual extortion).
- Using classifiers to detect signals and metadata relevant to unlawful and harmful content (such as behavioural signals related to private group membership, frequency of joining or leaving groups, engagement with children or young people using the service).

Reasonable steps to **address** unlawful and harmful material and activity on E2EE services may also include a number of options.

- Introducing obstacles to accessing E2EE services for the purpose of engaging in unlawful and harmful activity, such as:

[30] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 19, Explanatory Statement to the Determination, which states 'The Determination does not require or expect service providers to undertake actions inconsistent with obligations under the *Privacy Act 1988*, the *Telecommunications Act 1997* or *Telecommunications and Other Legislation Amendment (Assistance and Access) Act 2018*. Any adherence to expectations around anonymous (or pseudonymous) accounts and encrypted services are not to conflict with obligations under a Commonwealth Act.'

  - working with law enforcement and relevant experts (for example, experts in relation to CSEA and terrorism) to identify and block access to E2EE channels associated with illegal activity
  - limiting the use of joining links[31] shared on unencrypted services (for example, the Terrorist Content Analytics Platform[32] can support this by alerting the encryption provider to join links shared on unencrypted spaces).

- Introducing registration requirements such as requiring end-users to register for the service using a phone number, email address or other identifier. If these identifiers are authenticated (for example, through an authentication link or code), this can help prevent recidivism where accounts have been identified as breaching the law or terms of use. This links to sections 9 and 14 of the Determination.
- Introducing obstacles to storing or sharing unlawful and harmful material, such as:
  - taking steps to ensure that unlawful and harmful material that is detected is not uploaded, shared, or hosted on the service (for example, referring to law enforcement, blocking or reporting the end-user, or advising the end-user that the material might be unlawful, harmful or inappropriate and in breach of the service's terms of use)
  - incorporating safety features (for example, interstitial warnings, blurring or blocking content, providing safety information to end-users)
  - restricting or limiting a end-user's ability to share material with large numbers of people instantaneously (for example, restricting the ability to forward a message to many other users or groups at once).
- Providing end-users with reporting tools. Given some technologies may be challenging to implement on E2EE services, a particularly important step should be to provide end-users with clear and readily identifiable tools to report unlawful and harmful content on E2EE services to the service. Examples of clear and readily identifiable reporting mechanisms are outlined on page 50 of this document.

It may be difficult for a provider to demonstrate compliance with section 8 if they are taking limited or no steps to detect and address material or activity on the service that is unlawful or harmful, noting that the service is already likely vulnerable to exploitation by those seeking to engage in unlawful and harmful conduct without detection.

Providers should ensure that risks are fully considered and steps are built into a service's design before E2EE or other forms of encryption are implemented, rather than considered

[31] Links, often shared on unencrypted services, driving users to encrypted spaces.
[32] See Terrorist Content Analytics Platform (terrorismanalytics.org).

afterwards when harms arise. By adopting a holistic combination of the most suitable measures in a proportionate manner, providers can help to mitigate risks occurring on end-to-end-encrypted services.

Section 9 of the Determination – Preventing anonymous accounts being used for unlawful or harmful material or activity

Determination, section 9:

Additional expectation

1. If the service permits the use of anonymous accounts, the provider of the service will take reasonable steps to prevent those accounts being used to deal with material, or for activity, that is unlawful or harmful.

Examples of reasonable steps that could be taken

2. Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following:
   a. having processes that prevent the same person from repeatedly using anonymous accounts to post material, or to engage in activity, that is unlawful or harmful;
   b. having processes that require verification of identity or ownership of accounts.

'Anonymous accounts' are accounts that hide or disguise the identity of an end-user.[33]

There are many ways of appearing anonymous online. They include the following examples.

- Full anonymity – where an end-user does not provide any personal information or identifiers, and neither the online service nor other users can identify the end-user at the time of a particular interaction, or subsequently. This may be a result of a service design, or as a result of an end-user taking active identity shielding steps to prevent the collection of their data (for example, the use of a virtual private network (VPN) or other technologies that prevent disclosure of their geo-location or Internet Protocol (IP) address).[34]
- Public anonymity – where an end-user may appear anonymous to others, however the provider collects and holds some information about the end-user (for example, personal information such as their name, email or phone number, or their

[33] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 15, Online Safety (Basic Online Safety Expectations) Determination 2022 (legislation.gov.au).
[34] It is important to note that there are legitimate reasons for users to employ tools such as VPNs – for example, to keep their information secure when using public Wi-Fi.

geo-location, or their IP address, or the way they have engaged with the service and other users).[35]

- Pseudonymity – where an end-user has registered for a service using a username, handle or avatar that is not their real name, however the service collects and holds some information about the end-user (for example, many services require end-users to provide an email address or phone number at sign up).[36]

Section 9 applies to both anonymous accounts and pseudonymous accounts.[37]

There are many benefits in and valid reasons for maintaining a level of anonymity or practicing identity shielding online, including the right to privacy and protection from violence or unwanted contact.

However, anonymity and identity shielding can also enable harmful behaviours, particularly against people and communities who are at higher risk. eSafety's investigations teams regularly see anonymity being used as a tactic by those who seek to harm or abuse others online, for example:

- in the cyberbullying of children and in adult cyber abuse
- in the non-consensual sharing of intimate images (image-based abuse)
- in creating, storing and sharing unlawful content such as child sexual exploitation and abuse material.

The section 9 expectation does not require services that permit users of anonymous accounts to stop doing so, for example by employing a 'real name' policy or otherwise 'unmasking' their identities. Rather, it states that providers are expected to take reasonable steps to prevent anonymous accounts from being used to deal with material or for activity that is unlawful or harmful.

eSafety supports a balanced approach to this issue, which minimises the potential to disrupt the positive outcomes that online anonymity can afford. Providers are expected to have measures in place that allow them to effectively prevent and respond to harms perpetrated by anonymous account holders, for example:

---

[35] It is important to note that there are legitimate reasons for users to be publicly anonymous – for example, to protect their privacy and confidentiality when seeking out information and assistance online about sensitive topics.

[36] It is important to note that there are legitimate reasons for users to choose pseudonyms rather than using their real names online – for example, eSafety advises children not to use their real names online due to safety and privacy risks associated with sharing their personal details with people they do not know.

[37] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 14, Online Safety (Basic Online Safety Expectations) Determination 2022 (legislation.gov.au).

- ensuring the provider is able to identify and engage with accounts that are engaging in unlawful or harmful activity or material including by taking enforcement action when terms of use or policies are breached
- ensuring that end-users are not able to evade enforcement action by registering for a new account and continuing to cause harm.

Additionally, providers of services that do **not** permit anonymous or pseudonymous accounts should ensure they are taking reasonable steps to effectively enforce this rule. If a service's prohibition on use of anonymous accounts is being circumvented by end-users and that enables harms to occur on the service, the provider should consider whether section 9 applies to the service and comply if it does.

Verification of identity or ownership of accounts (section 9(2)(b))

Section 9(2)(b) of the Determination provides a key example of a reasonable step that can be taken to meet the section 9 expectation: implementing processes that require verification of identity or ownership of accounts.

However, providers are not required to 'unmask' an end-user's identity in order to demonstrate compliance with this expectation, although this may be a step that some providers take for their own purposes or for the safety or comfort of their end-users. For example, some business networking sites or dating sites may require real identities.

Providers are instead expected to take reasonable steps to prevent accounts from being used to deal with activity or material that is unlawful or harmful, which could include the following options.

- Requiring end-users to authenticate their accounts on sign-up by sending an authentication code or message or link to an email address or phone number used to create an account (including multi-factor authentication). This means that an account must be linked to a valid email or phone number. This may reduce instances of individuals seeking to create multiple accounts for harmful purposes, and may act as a deterrent against misuse and abuse as end-users know the service will be able to take appropriate enforcement action against them.
- Collecting appropriate identifiers from end-users on registration or sign up which enable the provider to deal effectively with that end-user (for example, to contact the end-user, to enforce terms of use and take other enforcement action, to respond to complaints about that end-user, to respond to legal requests for end-user details from eSafety and other regulators or law enforcement bodies). This could include collecting personal information such as name and date of birth, or using device identifiers or other identifiers.

- Using tools outlined elsewhere in this guidance, to prevent and detect abuse.

Processes that prevent the same person from repeatedly using anonymous accounts to post material, or engage in activity that is unlawful or harmful (recidivism – section 9(2)(a))

One of the significant safety risks and harms in relation to the use of anonymous accounts is the ability for individuals to engage in **repeated** activity or conduct that is unlawful or harmful (recidivism).

The Explanatory Statement to the Determination identifies a number of suggested steps to comply with section 9. Specifically, it suggests providers could have processes that uses web identifiers (such as cookies, IP addresses, browser fingerprinting), device or hardware identifiers, or other identifiers (such as account or behavioural analysis, metadata and traffic signals) to identify and stop re-registrations or alternative accounts in appropriate circumstances.

Other steps to **address recidivism** through the use of anonymous accounts may include:

- using other identifiers, in addition to those listed in the previous paragraph, to identify and stop re-registrations or alternative accounts, including personal information provided by the account holder (such as their name, address, date of birth, phone number, email, account photos, credit card details or other payment information), or behavioural indicators (such as their registration date, email alias, posting behaviour, usernames, or key phrases they use)
- using technology to detect previously banned end-users (for example, hash-matching that detects the profile pictures of banned end-users when an attempt is made to use them again)
- scanning for indicators of known or suspected offenders across all of the services operated by a provider, and implementing effective cross-service bans for offenders where appropriate
- providing end-users with clear communication advising if they are engaging in unlawful or harmful conduct, including conduct that violates terms of use, standards of conduct or other policies (for example, providing a warning via a pop-up)
- enabling end-users to block content from unverified or unauthenticated accounts
- imposing a strike system to determine appropriate action in response to repeated conducted (for example, warnings, penalties, bans, requiring identity verification to continue using the service)

- taking effective and appropriate enforcement action where necessary, such as implementing a device block to prevent an account from re-registering on the same device, or blocking an IP address.

Section 10 of the Determination – Consulting and cooperating with other service providers to ensure safe use

Determination, section 10:

Additional expectation

1. The provider of the service will take reasonable steps to consult and cooperate with providers of other services to promote the ability of end-users to use all of those services in a safe manner.

Examples of reasonable steps that could be taken

2. Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following:
    a. working with other service providers to detect high volume, cross-platform attacks (also known as volumetric or 'pile-on' attacks);
    b. sharing information with other service providers on material or activity on the service that is unlawful or harmful, for the purpose of preventing and dealing with such material or activity.

Providers are expected to take reasonable steps to cooperate with other members of industry to identify and respond to new harms, trends and issues that impact the safety of end-users. The intent of information sharing and cooperation is to allow providers to prevent and deal with unlawful and harmful material and activity in an effective manner, that suits their circumstances.

Importantly, it is expected that providers will take all reasonable steps to ensure there is information sharing and cooperation across **their own services**, as well as with third party service providers. The barriers to sharing information across a provider's own services will be lower than those sharing with other providers' services.

It is not expected that providers would cooperate in a way that puts a service's intellectual property at risk or involves the sharing of commercial-in-confidence information. The focus is on consultation and cooperation which aims to minimise unlawful and harmful material or activity that adversely impacts online safety for Australians.

High volume, cross-platform attacks

Section 10(2)(a) suggests that a reasonable step that a provider could take to cooperate with other providers or services is to detect and share information regarding high volume and/or cross-platform attacks (also known as volumetric or 'pile-on' attacks).

High volume attacks occur when a person is named in, tagged, or linked to an abusive post, which others 'like', share, re-post with additional commentary, and/or link to via other services. The volume of material can proliferate rapidly across services.

Cooperating to promote safe use in this way could include making other services aware of a volumetric attack by sharing information like URLs, hashtags or account names, as well as information on the people or groups being targeted, and insights on sources and trends. This information would assist a service to respond, subject to its own terms of use and policies.

Information sharing

Section 10(2)(b) suggests that a reasonable step could be to share information with other providers about material or activity that is unlawful or harmful with a view to preventing and dealing with it. For example, providers or services could share information about a section of the community that is being targeted with abuse due to an identifying characteristic (such as sexuality, ethnicity or disability), or linked to a specific event (such as a sporting or political event). Providers or services that receive this information could then take appropriate actions to prevent and deal with unlawful or harmful material or activity targeted at that group or event.

There are a number of additional reasonable steps that could be taken.

- Wherever possible, providers should take part in regular forums organised or facilitated by an industry association to discuss and evaluate effectiveness of safety tools and features that promote and ensure compliance with the Expectations and any other applicable safety laws.
- Providers could consider the off-platform behaviour of end-users of their services when making internal decisions affecting end-users. For example, when considering whether an end-user or account has violated terms of use, community guidelines or other policies, or whether an end-user poses an unacceptable safety risk to a service, services could take into account credible information (such as that published, provided or validated by another service or provider) about significant threats related to that end-user, such as those related to child sexual exploitation and abuse or terrorism.

- Providers could consider collaborating or partnering with organisations that seek to work with industry to address particular online harms.

Relevant industry code and industry standard measures

Certain providers will be required under the social media services industry code to take part in an annual forum to discuss online safety and evaluate the effectiveness of measures implemented under the code and share best practice with other industry participants.[38] Additionally, certain providers will be required under the social media services industry code to collaborate and contribute to expert groups that tackle child sexual exploitation and abuse and pro-terror material.[39] Similar obligations may be in place once industry standards are determined for the relevant electronic services and designated internet services sections of the online industry.

[38] Minimum compliance measure 15. The Social Media Services Industry Code can be accessed on eSafety's Register of Industry Codes.
[39] Minimum compliance measure 16. The Social Media Services Industry Code can be accessed on eSafety's Register of Industry Codes.

Chapter 2: Expectations regarding certain material

Division 3 of the Determination sets out expectations regarding certain material and activity, including that reasonable steps will be taken to minimise the extent to which the following material is provided on a service.

- Section 11: child cyberbullying material, adult cyber abuse material, non-consensual intimate images, class 1 material, and material that promotes, incites, instructs and depicts abhorrent violent conduct.
- Section 12: class 2 material.

Section 11 of the Determination - Minimising provision of certain material

Determination, section 11:

The provider of the service will take reasonable steps to minimise the extent to which the following material is provided on the service:

a. cyber-bullying material targeted at an Australian child;

b. cyber-abuse material targeted at an Australian adult;

c. a non-consensual intimate image of a person;

d. class 1 material;

e. material that promotes abhorrent violent conduct;

f. material that incites abhorrent violent conduct;

g. material that instructs in abhorrent violent conduct;

h. material that depicts abhorrent violent conduct.

Section 11 relates specifically to material set out in sections 11(a)-(h) (**certain material**). eSafety has published regulatory guidance on eSafety's powers in relation to these categories of material, separate to the Expectations. For more detail on the nature of each category of material, see eSafety's other regulatory guidance documents.[40]

The reasonable steps taken to minimise the extent to which certain material is provided on a service may differ, depending on each category of material and the way in which this material is provided, or able to be provided, on a service.

[40] See eSafety webpage for regulatory guidance: Regulatory schemes | eSafety Commissioner.

Providers should assess the risks of this certain material being provided on their service, and tailor their steps to address the risks.

For example, the risks for a service may include:

- end-users storing certain material on a service
- end-users generating certain material of themselves or others
- facilitating the creation of certain material (for example, through generative AI)
- end-users sharing certain material with other users, or sharing links to certain material
- end-users advertising the sale of, or access to, certain material.
- end-users encouraging other users to produce, share, store or otherwise access certain material
- end-users finding and connecting with victims or potential victims (including children) to obtain certain material
- repeated harassment, threatening, bullying, intimidating or abuse of a person, including through anonymous accounts or by creating multiple accounts to continue the behaviour.

Reasonable steps to minimise the provision of certain material should include both organisational and technical measures, to ensure that this material is:

- communicated to end-users as material that is not permitted on a service, or is subject to moderation (for more detail, see guidance on section 14 regarding terms of use and certain policies regarding reports, complaints and conduct)
- proactively detected by the provider, where appropriate (see examples in the following paragraph)
- able to be reported to the provider by end-users and trusted flaggers (see guidance on user reporting in section 13 for more detail)
- prioritised for review and action expeditiously by the provider.

A key step to minimising provision of certain material is the ability to detect it – either before it is uploaded or shared on a service, or immediately after it is provided on the service. A number of steps may be used to proactively detect certain material, including the following options.

- Hash matching technology to detect known images and videos of unlawful material such as CSEA and terrorism material.

- Hash matching technology to detect non-consensual intimate images shared on a service (see, for example, the National Center for Missing and Exploited Children's (NCMEC) Take It Down hash list for images of under 18 year-olds and StopNCII hash data base for images of people 18 years and older). Additionally, providers could use hash matching technology internally to hash content or material that is reported to them from end-users or otherwise detected by the provider, and scan for these internal hashes across their service.
- AI classifiers to identify new material that is likely to be unlawful (such as CSEA and terrorism material) or harmful, and prioritise for human review, including where this material is livestreamed on a service (for example, broadcast to a wide audience or occurring in a private video chat or call).
- Technologies such as language or text analysis which can identify a wide range of unlawful or harmful activity occurring on online services. These technologies and processes should be regularly evaluated and updated to respond to evolving use of language by end-users, including deliberate attempts to avoid detection through the use of new words, phrases, symbols and text.

Where content is unlawful it should be removed and reported to appropriate authorities. It may also be appropriate to ban the account holder and prevent them from re-registering on the service.

Providers could also use proactive nudges or prompts to end-users that the material they are attempting to upload, save, send or otherwise share may be unlawful or harmful, including whether such material is prohibited in terms of use or other policies. For more serious content, end-users should also be notified that the material may be unlawful.

Additionally, providers are expected to exercise vigilance in detecting ongoing patterns of abuse against end-users once abuse has been reported to the service. Material set out in section 11 may be provided on a service by end-users in a manner that demonstrates repeated abuse of other users, and providers should ensure they are taking reasonable steps to minimise the repeated provision of material.

It is important that tools are used on all appropriate parts of a service in order to detect certain material. Subject to technical or other constraints, eSafety considers that a provider is unlikely to be meeting the section 11 expectation (and section 6) if a service is only using relevant tools on one part of its service, but leaves other at-risk parts of a service without any intervention.

Additionally, eSafety will have regard to the extent to which these tools are implemented and relevant processes are updated. For example, it is unlikely to be sufficient to deploy a hash matching tool to detect CSEA, but only update the list of available hashes once a year.

Relevant industry code and industry standard measures

eSafety notes that certain providers will be required under the social media services industry code to deploy systems, technologies or processes to proactively detect known CSEA and terrorism material.[41] The industry standards to be determined for relevant electronic services and designated internet services may also contain requirements on certain service providers to proactively detect known CSAM or pro-terror material.

The use of technological tools to proactively detect certain material should be supported by human moderators who review content flagged and take steps to remove and report or otherwise deal with the material. Appropriately resourcing systems and processes to ensure that user reports of unlawful and harmful content are responded to, and actioned, in a timely manner support compliance with this expectation.

It is particularly important that end-users are provided with clear and readily identifiable mechanisms to report certain material and make complaints. For more detailed guidance on reporting and complaint mechanisms, see Chapter 3.

Section 12 of the Determination – Preventing children's access to class 2 material

Determination, section 12:

Core expectation

1. The provider of the service will take reasonable steps to ensure that technological or other measures are in effect to prevent access by children to class 2 material provided on the service.

Examples of reasonable steps that could be taken

2. Without limiting subsection (1) of this section, reasonable steps for the purposes of that subsection could include the following:
   a. Implementing age assurance mechanisms;
   b. conducting child safety risk assessments.

[41] Minimum compliance measures 8 and 9. Social Media Service providers should also have regard to minimum compliance measure 10 which requires certain providers to take specific actions that aim to disrupt/deter users from creating, posting or disseminating CSAM and pro-terror material on the service. The Social Media Service Industry Code can be accessed on eSafety's Register of Industry Codes.

What is class 2 material?

Class 2 material is defined earlier in this guidance on page 6.

Why should children be prevented from accessing this material?

There are risks for children and young people under the age of 18[42] as a result of intended, unintended, non-consensual or coerced access to class 2 material. Therefore, a range of interventions should be adopted by providers to suit the evolving developmental needs of children and young people.

More information on the risks and harms related to children and young people's access to pornography can be found in eSafety's Age Verification Roadmap and background report.[43]

This guidance will be updated in the future to address any overlap between the section 12 expectation and industry codes or industry standards relating to class 2 material.

Technological and other measures that may be used to prevent access by children to class 2 material

In determining what reasonable steps should be taken to prevent access by children and young people to class 2 material, it is important to consider the extent to which class 2 material is provided on a service. For example, providers may operate services that:

1. deliberately host or provide access to class 2 material for end-users (for example, porn sites),
2. permit class 2 material, or do not actively enforce prohibition of this material, but it is not a core aspect of the service (for example, end-users can share material or distribute links to class 2 material, advertisements may be placed that contain or link to class 2 material), or
3. prohibit class 2 material.

Section 12(2) of the Determination provides two examples of reasonable steps that can be taken to ensure compliance with section 12 – implementing age assurance mechanisms and conducting child safety risk assessments.

Age assurance is not defined in the Determination, and is an umbrella term which includes both age verification and age estimation solutions.

[42] References to 'children and young people' generally means children and young people under the age of 18.
[43] See eSafety's website: Age verification | eSafety Commissioner.

- Age verification measures determine a person's age to a high level of accuracy, and can involve the use of physical or digital government identity documents to establish a person's age.
- Age estimation technologies provide an approximate age to allow or deny access to age-restricted online content or services. Age estimation can involve the use of biometric data, such as a facial scan or voice recording, to infer a person's age or age range.

By identifying 'age assurance mechanisms' as an example of a reasonable step, providers have a degree of flexibility as to how they protect children and young people from access to class 2 material. For example, age assurance mechanisms may:

- ensure that underage or prohibited end-users are not able to access services (for example, many services do not permit children who are under 13 – which relates to the section 6 expectation on ensuring safe use of a service)
- assist providers in enforcing their minimum age requirements and terms of use (also relevant to section 14)
- provide an indication to a service that an end-user is of a certain (or approximate) age, which enables high privacy and safety settings to be implemented by default for that end-user, including preventing access or exposure to certain content on a service (also relevant to section 6).

Providers can consider the elements of a Restricted Access System,[44] as set out in the Online Safety (Restricted Access Systems) Declaration 2022[45] in terms of measures that may be adopted to prevent children and young people from accessing class 2 material on their service, although additional steps may be required, depending on the nature of the service. These elements include:

- requiring an end-user to apply for access to relevant class 2 material, with a declaration that they are at least 18 years old
- giving warnings and safety information for class 2 material
- incorporating reasonable steps to confirm the age of applicants.

Measures to prevent children and young people from accessing this material should not unduly restrict the rights of adults to create, access and share lawful content, and it is

[44] A restricted access system is a means of limiting access to material that is inappropriate to children and young people under 18. The Commissioner may give remedial notices to certain providers requiring the recipient to take all reasonable steps to remove class 2B material from a service, or place the material behind a restricted access system. See eSafety's Online Content Scheme Regulatory Guidance for more information: Online Content Scheme Regulatory Guidance.pdf (esafety.gov.au).

[45] Online Safety (Restricted Access Systems) Declaration 2022, Online Safety (Restricted Access Systems) Declaration 2022 (legislation.gov.au).

important that steps to achieve this be balanced against the need to preserve age-appropriate access to sexual health and wellbeing information and support.

For services that **deliberately permit class 2 material as a core part of the service,** it is important that robust measures are in place to prevent children and young people under 18 from accessing the service.

This may include:

- clearly communicating to end-users that the service contains class 2 material and is intended for adult access (over 18 years old)
- applying meta-tags to the site, such as the Restricted to Adults label, to ensure the service or platform is blocked by any filters that may be in place for children on accounts or devices
- implementing age assurance or age verification mechanisms to prevent access to the service, and to prevent account registration if accounts are required.
- ensuring that landing pages or first point of contact with a service do not contain class 2 material and that this material is placed behind an age-gate.

For services that do not have class 2 material as a core part of their service but **permit class 2 material**, steps should be taken to prevent access to that material by children and young people under 18. For example, the service may:

- take the same steps listed for services that intentionally permit class 2 material (communicating to end-users that the service may contain class 2 material, using meta-tags to ensure the service is blocked by filters in place for children, and using age assurance mechanisms where appropriate)
- limit the searchability or discoverability of class 2 content by children and young people under 18, for example by preventing autocomplete or predictive entries for searching for terms that are known to be associated with class 2 material, and filtering out search responses for children and young people under 18
- blur class 2 material by default for all end-users to prevent unintentional access or exposure
- deploy technology or other tools to minimise the risk that class 2 material is provided, promoted or otherwise accessible to children and young people via the service, either as content or in advertisements
- deploy technology or tools to ensure that any permitted class 2 material, and any accounts dedicated to or commonly providing class 2 material, are appropriately tagged and that end-users are provided with appropriate warnings and options not to view the tagged content

- provide support to children and young people where they are specifically seeking out class 2 material – for example, pop up messages, tools or resources that explain why this material is not available to them (or is otherwise inappropriate for their age) or direct them to appropriate resources or support
- provide clear and accessible guidelines for end-users about access to class 2 material on the service and what safety measures are in place for children and young people under 18
- provide clear and readily identifiable reporting tools for children and young people (or their parents or carers) to flag class 2 material that they encounter, and ensure that flagged or reported material is not provided to the child or young person again
- provide strong parental controls, filtering and other supervision tools to support parents in ensuring that class 2 material is not accessible to a child or young person.

For services that **do not permit class 2 material**, steps should be taken to ensure that this policy is known to end-users and enforced. For example, the service may:

- set out this prohibition clearly in terms of use, community guidelines and/or other relevant policies
- take steps to enforce these terms of use – for example by warning, suspending or banning end-users who breach the terms of use, or preventing them from re-registering where appropriate
- enable end-users to report class 2 material to the service, and respond to these reports
- provide proactive detection of class 2 material
- provide strong parental controls, filtering and other supervision tools to support parents in ensuring that class 2 material is not accessible by children and young people
- use AI classifiers to detect nudity, combined with human moderation.

Chapter 3: Expectations regarding reports and complaints

Division 4 of the Determination sets out expectations in relation to:

- Section 13: mechanisms to report and make complaints
- Section 14: terms of use, certain policies etc.
- Section 15: mechanisms to report and make complaints about breaches of terms of use
- Section 16: accessible information on how to complain to the Commissioner

Section 13 of the Determination – Providing mechanisms to report and make complaints about certain material

Determination, section 13:

Core expectation

1. The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable end users to report, and make complaints about, any of the following material provided on the service:
   a. cyber-bullying material targeted at an Australian child;
   b. cyber-abuse material targeted at an Australian adult;
   c. a non-consensual intimate image of a person;
   d. class 1 material;
   e. class 2 material;
   f. material that promotes abhorrent violent conduct;
   g. material that incites abhorrent violent conduct;
   h. material that instructs in abhorrent violent conduct;
   i. material that depicts abhorrent violent conduct.

Additional expectation

2. The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable any person ordinarily resident in australia to report, and make complaints about, any of the following material provided on the service:
   a. cyber-bullying material targeted at an australian child;
   b. cyber-abuse material targeted at an australian adult;
   c. a non consensual intimate image of a person;
   d. class 1 material;
   e. class 2 material;
   f. material that promotes abhorrent violent conduct;
   g. material that incites abhorrent violent conduct;
   h. material that instructs in abhorrent violent conduct;
   i. material that depicts abhorrent violent conduct.

The intention of this section is to ensure that services have appropriate complaints processes for all Australians to report certain material regulated under the Act to a service, without the requirement to have an account with that service.[46]

Reporting and complaint mechanisms should be clear and readily identifiable to end-users and others at all relevant points in time when they are engaging with material, activity or other users.

Providers should conduct a safety risk and impact assessment of what harms and risks end-users and individuals ordinarily resident in Australia are likely to experience on their services, and design intuitive reporting options for end-users accordingly. This assessment should include accessibility requirements to ensure all end-users are able to effectively use the reporting options.

Additionally, providers should ensure that report and complaint mechanisms on their services are designed in a way that enables for the prioritisation of reports for escalation and rapid response – for example, reports that are likely to relate to unlawful material or activity or present a serious threat to life, health or safety.

Clear and readily identifiable reporting and complaint mechanisms are particularly critical as a safety intervention where providers are limited in their ability to deploy technologies on their services that proactively detect unlawful and harmful material and activity.

What is a 'clear' mechanism for reporting and making a complaint?

A reporting or complaint mechanism is more likely to be 'clear' if individuals are presented with a menu which contains an appropriate category or description of the issue that they want to report.

Issue-specific reporting options are important to empower individuals to clearly identify the reason they are concerned with the content, and to enable the provider to respond appropriately including by prioritising certain reports. For example, a specific CSEA reporting option is critical to ensuring that this extremely harmful, unlawful material is reported and able to be prioritised for review and action (such as banning the account and referral to law enforcement). This might be provided alongside a 'general' reporting category to ensure those who want to make a report that is not harm-specific also have the opportunity to do so.

Providers should offer a clear mechanism for individuals who do not have an account with the service to report material or other end-users, without the need to create an account

[46] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 17, Online Safety (Basic Online Safety Expectations) Determination 2022 (legislation.gov.au).

themselves. This is important where material or activity may be impacting an individual who is not an end-user of the service – for example, cyberbullying or other abusive material where the victim is not an end-user of the service where the material is being shared.

If a service is known, or likely, to be used in a way that facilitates extremely harmful, unlawful activity such as CSEA and the promotion of terrorism, it is unlikely that the provider can demonstrate compliance with section 13 if they do not have a specific reporting option for these categories (for example, if a service requires individuals to rely on broad reporting options such as 'inappropriate content' or 'sexual activity' to report this unlawful content).

Individuals should also be provided with relevant information, at the time of reporting, about how their personal information will be used (if at all) as a result of making a report or a complaint, to ensure individuals feel comfortable, informed and empowered to make a genuine report or complaint without fear of consequences. Providers should consider eliminating barriers to reporting and complaints, such as requirements to provide personal information or to follow multiple steps to locate reporting options.

What is a 'readily identifiable' mechanism for reporting and making a complaint?

A reporting option is 'readily identifiable' if it can be quickly and easily accessed and used by an individual without barriers, at every part of the user experience. For example, reporting and complaint mechanisms should:

- be provided on all aspects of a service so that an individual can report all relevant material and activity - including material they have seen in a post, a livestream, a video chat or direct communication, or activity by another end-user or by a group or forum
- enable individuals to report and complain about material that an individual has knowledge of but does not have direct access to (for example, an intimate image that they know has been shared on a service, but the individual does not know where on the service, or who has access to it)
- be accessible in-service at the point in which the individual wishes to flag material, meaning they can report content without needing to navigate to a separate part of the service or exit the service to report via email or complaint form
- be available to all end-users of a service, regardless of whether they have an account, or are logged in or not
- be consistently accessible for individuals where a service may be accessed via an app or browser or via desktop

- ensure a seamless process for material of concern to be identified to the provider (for example, report and complaint mechanisms should be designed so they automatically flag and preserve the material in question for review by the service)
- not require individuals to take screenshots, save links or otherwise create their own copy of the material in order to make a report or complaint to the service, although this additional functionality may be useful to individuals.

Section 14 of the Determination – Providing terms of use and certain policies and procedures regarding reports, complaints and conduct

Determination, section 14:

1. The provider of the service will ensure that the service has:
   a. terms of use; and
   b. policies and procedures in relation to the safety of end-users; and
   c. policies and procedures for dealing with reports and complaints mentioned in section 13 or 15; and
   d. standards of conduct for end-users (including in relation to material that may be posted using the service by end-users, if applicable), and policies and procedures in relation to the moderation of conduct and enforcement of those standards.

Note 1: see section 17 in relation to making this information accessible to end-users.

Note 2: for paragraph (b), the policies and procedures might deal with the protection, use and selling (if applicable) of end users personal information.

2. The provider of the service will take reasonable steps to ensure that penalties for breaches of its terms of use are enforced against all accounts held or created by the end-user who breached the terms of use of the service.

Terms of use, standards of conduct, policies and procedures are key mechanisms for providers to communicate what is and is not allowed on their service (in terms of both material and activity). They are also important mechanisms for providing a clear and transparent rationale for action a provider may take to address unlawful and harmful material and activity on the service.

Some providers refer to the relevant parts of terms of use, standards of conduct, policies or procedures as community guidelines, community standards or rules.

eSafety considers these important mechanisms to be interrelated and core to ensuring safe use of a service. A provider should set out clearly how standards of conduct and/or relevant policies are linked to terms of use of a service.

It is expected that terms of use and policies will be clear, explicit and easy to understand.

One of the factors eSafety is required by the Act to consider in determining whether to give a reporting notice is 'whether there are deficiencies in a service's terms of use, so far as they relate to the capacity of end-users to use the service in a safe manner'.[47]

Relevant industry code and industry standard measures

Certain social media service providers will also be required to ensure their service's policies and terms of use, in their treatment of class 1 content, meet the requirements set out in the social media services industry code.[48] The industry standards to be determined for relevant electronic services and designated internet services may also contain requirements for the policies and terms of use for certain service providers.

What online safety harms should terms of use and policies and procedures cover?

Terms of use should prohibit activity and material that is unlawful and harmful. At a minimum, providers should ensure that their terms of use and other policies align generally with the unlawful and harmful matters dealt with under the Act (the matters specified in section 13 of the Determination). Additional harms suggested in the Explanatory Statement to be covered by terms of use and other policies include, but are not limited to:

- hate against a person or group of people on the basis of race, ethnicity, disability, religious affiliation, caste, sexual orientation, sex, gender identity, serious disease, disability, asylum seeker/refugee status, or age
- promotion of suicide and self-harm content, such as pro-anorexia content, that does not meet the threshold of class 1 or class 2 material
- high volume, cross-platform attacks that have a cumulative effect that is damaging but does not meet the threshold of adult cyber abuse when reported as singular comments or posts
- promotion of dangerous 'viral' activities that have the potential to result in real injury or death.

[47] See sections 56(5)(d) (non-periodic reporting notice) and 49(5)(d) (periodic reporting notice) of the Act.
[48] Minimum compliance measures 2, 3, 11 and 12. The Social Media Services Industry Code can be accessed on eSafety's Register of Industry Codes.

Providers should consider whether their terms of use, policies, procedures and/or standards of conduct effectively address the range of harms and risks that currently do, or may, arise on their service. Providers are best placed to identify emerging forms of harmful end-user conduct or material, and are afforded flexibility by the Determination to choose the best and most responsive way to address them on their service. Providers should update their terms of use, standards of conduct and other policies and procedures as new risks and harms emerge over time.

Where a provider provides multiple services, there should be service-specific terms of use, policies and procedures that are tailored to the service and any particular safety risks or harms posed by, or to, end-users of that service. It may not be sufficient for a service to rely on high-level, broad terms of use that do not clearly and explicitly set out what material and activity is prohibited or restricted, and how the service enforces these rules.

Policies and procedures for dealing with reports and complaints

Providers should have clear policies and procedures for dealing with reports and complaints and should take steps to communicate these to individuals. For example:

- users should be provided with confirmation that their report or complaint has been received, and an indication of when they will receive a response from the provider - this could include providing the user with a receipt, reference or report number in relation to the report or complaint
- users should be advised of the outcome of reports and complaints
- policies and procedures should include clear guidance on when reporting to external bodies is required – for example, to law enforcement bodies or in response to a request from eSafety.

Providers should also have internal policies and procedures for prioritising and responding to reports or complaints that are likely to relate to unlawful material or activity, or present a serious threat to life, health or safety. It is important for providers to have the resources commensurate with the size and risk of their service, and allow for prompt and accurate response to user reports and complaints.

Reasonable steps to enforce breaches of terms of use

In addition to setting out clear and comprehensive terms of use and policies relating to the safety of end-users, it is expected that providers will also have in place effective systems to enforce terms of use. It is also expected that providers will enforce any standards of conduct and policies included or incorporated in the terms of use.

This would include providers making appropriate enquiries into any suspected breaches of terms of use, standards of conduct or other relevant policies.

Providers should consider a range of enforcement options and apply these in a manner that is proportionate to the nature of the breach. Enforcement against breaches should also have regard to issues such as minimising the risk of material or activity occurring again, including by banning accounts where there are severe breaches of the terms of use. More serious breaches are likely to require a more significant response.

Providers should also be able to explain these steps to eSafety in relation to an investigation or other escalation.

Options to enforce breaches of terms of use may include:

- warnings and strikes, nudges and prompts to end-users
- requiring an end-user to review certain safety information
- removing certain privileges or functionality for an end-user (such as the ability to monetise or livestream content, or removal of a 'credibility' or similar badge)
- account blocking or account limiting (or blocking or limiting content)
- removal of an account, or content
- requiring an end-user to apologise, in appropriate circumstances
- account suspension – accounts may be de-activated or suspended for a temporary period of time, and alerts may be sent to give the end-user time to address the issue
- disabling an account – accounts may be permanently disabled so they are no longer visible or active
- down-rank content – demote content visibility for some or all content posted by an end-user
- geo-blocking or geo-IP-blocking.

Reasonable steps which support the effective and consistent enforcement of penalties for breaches of terms of use may include:

- ensuring content moderation staff – including community or volunteer moderators - are trained to apply these terms of use, content guidelines and other internal guidelines consistently and objectively
- ensuring transparency regarding these enforcement processes and outcomes, and publish relevant information in a regular transparency report or other safety report
- ensuring terms of use and policies and procedures are regularly reviewed and updated as needed - this could be done as part of regular safety risk assessments

- ensuring effective measures are in place to detect end-users who attempt to re-register or regain access to a service when they have been banned, or had other enforcement action taken against them, and to prevent this recidivism (see chapter 1, section 9 on anonymous accounts for examples of steps that may be taken to address recidivism)
- appropriately resourcing trust and safety teams and content moderation teams.

It is unlikely to be sufficient for a service to only refer individuals who make reports or complaints about breaches of terms of use to external sources of support and to take no further steps to address the content and/or account that is the subject of the report or complaint, including to prevent future harm on the service. For example, for a severe or repeated breach of terms of use, the service should also take action such as banning the account.

Relevant industry code and industry standard measures

Certain social media service providers will also be required under the social media services industry code to enforce their terms of use including taking reasonable steps to prevent recidivism.[49] The industry standards to be determined for relevant electronic services and designated internet services may contain similar requirements for certain services.

Section 15 of the Determination – Providing mechanisms to report and make complaints about breaches of terms of use

Determination, section 15:

Core expectation

1. The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable end users to report, and make complaints about, breaches of the service's terms of use.

Additional expectation

2. The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable any person ordinarily resident in Australia to report, and make complaints about, breaches of the service's terms of use.

[49] Minimum compliance measures 2, 3, 11 and 12. The Social Media Services Industry Code can be accessed on eSafety's Register of Industry Codes.

The purpose of this section is to provide an avenue for all Australians to have material or activity that breaches a service's terms of use removed or otherwise dealt with in an appropriate manner by the service without requiring them to have an account with a particular service.[50]

For example, an individual may be aware that harmful material relating to them, which breaches the terms of use of a service, is accessible on a service that they do not have an account with or otherwise engage with. Providers should ensure that individuals (and, in certain circumstances, their parent or guardian) are not prevented from reporting or complaining about a breach of a service's terms of use because they do not have an account.

Providers should consider the list of steps set out in this guidance in relation to section 13 (reporting and complaints about certain material) as these are also relevant to providing reporting and complaint mechanisms in relation to breaches of terms of use.

Section 16 of the Determination – Providing access to information on how to complain to the eSafety Commissioner

Determination, section 16:

The provider of the service will ensure that information and guidance on how to make a complaint to the Commissioner, in accordance with the Act, about any of the material mentioned in section 13 provided on the service, is readily accessible to end-users.

The purpose of this expectation is to make end-users in Australia aware that they can make complaints to the Commissioner regarding material included in section 13 of the Determination.

It is at the discretion of providers to decide how they provide this information, and providers have flexibility to design their services in a way that best supports end-users with important safety information, including that a complaint can be made to eSafety in relation to certain material and activity. Providers may choose to make this information accessible at all points of the end-user experience, or at the point of account creation or first use, or at regular intervals, or in a sequence appropriate for that services' complaints process.

However, end-users should be provided with this information in a clear and readily accessible manner at the point when they report material to the service and when they complain to the service. This is important because complaining to a service is a necessary

[50] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 19, Online Safety (Basic Online Safety Expectations) Determination 2022 (legislation.gov.au)

first step for end-users who are seeking removal of cyberbullying material directed at a child[51] or of adult cyber abuse, under eSafety's complaint schemes.

Additionally, this information (including direct links to information on the eSafety website about how to make a complaint) should be clearly set out in appropriate documents and links on a service, such as in the terms of use, community guidelines, safety centre or other safety resources.

Chapter 4: Expectations regarding accessible information

Section 17 of the Determination – Providing access to information on terms of use, policies, complaints and similar topics

Determination, section 17:

1. The provider of the service will ensure that the information specified in subsection (2) is:
   a. readily accessible to end-users; and
   b. in relation to the information mentioned in paragraph (2)(b)—accessible at all points in the end-user experience, including, but not limited to, point of purchase, registration, account creation, first use and at regular intervals (as applicable); and
   c. regularly reviewed and updated; and
   d. written in plain language.

2. For the purposes of subsection (1), the information is the following:
   a. the terms of use, policies and procedures and standards of conduct mentioned in section 14;
   b. information regarding online safety and parental control settings, including in relation to the availability of tools and resources published by the Commissioner.

[51] For more information, see eSafety's regulatory guidance on the Cyberbullying Scheme: Regulatory schemes | eSafety Commissioner.

This expectation relates to the provision, accessibility, review and presentation of information regarding a service's terms of use and information about online safety and parental control settings – including in relation to the availability of tools and resources published by eSafety.

eSafety has published a suite of tools and resources on the eSafety website[52] that providers could provide to end-users to supplement their own safety information such as terms of use, policies, procedures and standards of conduct.

The information specified in section 14(2) should be simple and as easy as possible for users to locate and to use, to make their (or their children's) user experience as safe and age-appropriate as possible. This is particularly important when a user is registering to use a service or using the service for the first time, but it is also important that the information is easy to find throughout a user's experience of the service.

Where a user has indicated to a service that they are seeking specific information, such as information for parents, services should provide relevant eSafety resources at that point in time to assist end users.

For the purpose of section 17(1)(b), provision of this information at 'regular intervals' may be satisfied through adhering to the section 18 expectation.

Information should be written in plain language and should be provided in multiple languages to ensure end-users are able to understand key safety information. Information should also be age-appropriate to suit the developmental needs of children if a service permits or has child-users.

Section 18 of the Determination – Providing updates about changes in policies, terms and conditions

Determination, section 18:

The provider of the service will ensure that end-users receive updates written in plain language in relation to changes in the information specified in subsection 17(2), including through targeted in-service communications.

Providers should ensure that end-users receive updates in plain language regarding changes to the terms of use, policies, procedures and standards of conduct and information available about online safety and parental control settings, including through targeted in-service communications.

[52] eSafety website, Online safety | eSafety Commissioner.

Depending on the nature of the update, end-users could be required to confirm that they understand the changes and how they will be impacted – for example, if terms of use are updated to prohibit certain activity or material, end-users should be required to confirm that they have read and understood this and agree to abide by this rule.

These updates should be provided in multiple languages to support end-users and should also be age-appropriate to suit the developmental needs of children and young people, if a service permits younger users.

Providers may choose how to best present these updates to end-users, including through age-appropriate means to young people and children. Infographics, videos, tiered notices and other measures to ensure end-users are able to understand the updates and how this impacts their safety experience may all be appropriate.

Chapter 5: Expectations regarding record keeping

Section 19 of the Determination – Keeping records regarding certain matters

Determination, section 19:

The provider of the service will keep records of reports and complaints about the material mentioned in section 13 provided on the service for 5 years after the making of the report or complaint to which the record relates.

The purpose of this expectation is to ensure providers can provide the Commissioner with information on complaints about the material in section 13 and how the provider actioned the complaints.

This information will help the Commissioner assess the effectiveness of complaint and moderation practices over time and point out areas where services are doing this well, as well as areas where improvements could be made.

Providers should retain an appropriate amount of detail in these records to assist the Commissioner in assessing the adequacy of a service's response to reports and complaints.

For example, where a report or complaint is made to the service about certain material, the service should include in its record:

- the mechanism by which the end-user made the report – such as through in-service reporting, or via a webform or email

- the specific category of material reported – both as reported by the end-user and as designated or established by the service
- the service's response to the report or complaint (such as any content moderation decision like material removed, report made to a law enforcement body, or enforcement action taken against the offending end-user)
- the time taken to respond to the report or complaint
    - this should include an overall indication of the time taken, from the point at which an end-user made a report to the point where action was completed by the service
    - this could also include more specific information such as the time taken for a report to be flagged to a specialist team for review and action, and any re-review required or other escalation.

Where certain enforcement action is taken against end-users as a result of a report or complaint, such as a permanent ban from the service, records could include details about offending end-users to ensure they are prevented from re-registering or accessing the service.

Where records of reports and complaints contain personal information, including sensitive information or information that is likely to be perceived as sensitive to end-users, providers are expected to ensure this information is subject to robust privacy protections.

Records should be kept for five years. Providers are not expected to have five years of records until at least five years following the making of the Determination.[53]

eSafety recognises that some jurisdictions may prevent providers from storing relevant data for this period of time, and will have regard to this when assessing compliance with this expectation.

Taking steps to ensure an appropriate level of detail is retained in records under this section is likely to support providers in responding to Commissioner information requests, as set out in section 20.

[53] The Online Safety (Basic Online Safety Expectations) Determination 2022 was registered on 23 January 2023: Online Safety (Basic Online Safety Expectations) Determination 2022 (legislation.gov.au).

Relevant industry code and industry standard measures

Social media service providers are also required to keep records in relation to the measures they have adopted to comply with the social media service industry code. Similar requirements may be placed on relevant electronic services and designated internet service providers under the industry standards.

Chapter 6: Dealings with the Commissioner

Section 20 of the Determination – Providing requested information to the Commissioner

Determination, section 20:

Core expectation

1. If the Commissioner, by written notice given to the provider of the service, requests the provider to give the Commissioner a statement that sets out the number of complaints made to the provider during a specified period (not shorter than 6 months) about breaches of the service's terms of use, the provider will comply with the request within 30 days after the notice of request is given.
2. If the Commissioner, by written notice given to the provider of the service, requests the provider to give the Commissioner a statement that sets out, for each removal notice given to the provider during a specified period (not shorter than 6 months), how long it took the provider to comply with the removal notice, the provider will comply with the request within 30 days after the notice of request is given.
3. If the Commissioner, by written notice given to a provider of the service, requests the provider to give the Commissioner specified information relating to the measures taken by the provider to ensure that end users are able to use the service in a safe manner, the provider will comply with the request within 30 days after the notice of request is given.

Additional expectation

4. If the Commissioner, by written notice given to a provider of the service, requests the provider to give the Commissioner a report on the performance of online safety measures that relevant providers have announced publicly or reported to the Commissioner, the provider will comply with the request within 30 days after the notice of request is given.

The information which the Commissioner may request from a provider under section 20 can be directly relevant to how services are meeting the Expectations. For example, information requested under section 20(1) (number of complaints) can provide the Commissioner with an indication of how effectively terms of use are communicated to users and enforced by a provider. It is also relevant to how a provider is ensuring safe use of a service, including by taking reasonable steps to proactively minimise the provision of certain material (section 6).

For more information on the various reporting powers and options, including a section 20 request for information, see page 12 of this guidance.

It is at the discretion of the provider to provide additional information regarding complaints (for example, how many were deemed vexatious, how many did not meet a threshold for action, how complaints were resolved), however providers should consider what additional information or context they could include in response to a section 20 request, as this would assist in better understanding and assessing how a provider is ensuring safe use of their service and meeting the Expectations.

For example, where complaints about breaches of terms of use indicate an increase in a specific type of harmful activity or trend, it is useful to provide additional, relevant information (such as improved reporting options, updated terms of use or introduction of new safety features) which may be relevant to an increase in the number of reports.

Under section 20(2), the Commissioner may request a statement that, for each removal notice given to the provider during a specified period, sets out how long it took a provider to comply with the removal notice. This information will help the Commissioner assess how rapidly providers are complying with removal notices given under the Act's schemes.

Under section 20(3), the Commissioner may request information relating to the measures taken by the provider to ensure that end-users are able to use the service in a safe manner. The purpose of this expectation is to enable the Commissioner to request specified information concerning online safety measures being taken by a provider.

The Commissioner may also request a report on the performance of safety measures that it has publicly announced or reported to the Commissioner (section 20(4)). In practice, when a provider announces a significant new safety feature, that provider should expect to be asked by the Commissioner to report on the impact of that safety feature on the experience of end-users. The intention of this expectation is to address the scenario of a provider announcing a safety feature, but failing to disclose whether the feature was effective. Providers should ensure they continually evaluate and assess safety features and collect relevant information about the performance of such measures, in order to comply with a section 20(4) request.

eSafety may use section 20 requests for information as part of an escalation of regulatory engagement with providers. In the first instance eSafety may seek some of the information included in section 20 on an informal basis, including through regular engagement and specific queries. This reflects the regular and ongoing engagement that eSafety has with providers, and that some information can be shared through these mechanisms. This informal engagement helps inform eSafety regarding providers' practices, trends and specific risks.

However, where information is required for specific regulatory purposes, or if eSafety intends to publish relevant information for the purpose of improving transparency, eSafety may make a formal request through section 20.

Providers are not required to respond or comply with requests through section 20, but failure to do so would provide the Commissioner with grounds to publish a statement to that effect. The Commissioner may also seek the information through a non-periodic or periodic reporting notice instead, which would carry civil penalties for non-compliance.

Section 21 of the Determination – Providing a designated contact point

Determination, section 21:

1. The provider of the service will ensure that there is an individual who is:
   a. an employee or agent of the provider; and
   b. designated as the service's contact point for the purposes of the Act.
2. The provider will ensure that the following: contact details of the contact point are notified to the Commissioner:
   a. an email address; and
   b. a phone number or voice chat address.
3. If there is a change to the identity or contact details of the individual designated as the service's contact point for the purposes of the Act, the provider will give the Commissioner written notice of the change within 14 days after the change.

Section 21 requires providers to notify eSafety of a designated contact point. Any changes must be notified to eSafety in writing within 14 days after the change.

In order to facilitate the sharing of contact details, and also to enable the sharing of other information, eSafety has established a webform for relevant providers. Providers are

encouraged to use this webform. By completing and maintaining information via this form, eSafety will regard a provider as meeting the expectation under section 21. Contact details will not be made public without the consent of providers.

Contact details may be used for engagement on implementation of the Expectations, and on other online safety issues, as well as a point of contact for eSafety for communications related to the enforcement of the Act. Where eSafety has existing contacts, particularly those used for content removal notices and other engagement under the Act, these are likely to continue to be used. Providers may want to nominate these existing contacts for the purposes of section 21 to ensure consistency, or may choose alternative points.

The webform includes some voluntary questions that providers may answer (for example, details of terms of use and reporting processes). Where appropriate, this information may be published in the interests of transparency.

To access the webform link and share the relevant information, please contact industrybose@esafety.gov.au.

Annex A

Table 3: Summary of the Expectations

Division	Headline Expectation	Expectations	Examples of reasonable steps that could be taken (where provided in the Determination) or qualifications
2. Safe use[54]	(S 6) Reasonable steps to ensure safe use	The provider of the service will take reasonable steps to ensure that end-users are able to use the service in a safe manner. The provider of the service will take reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful.	**Examples of reasonable steps:** a. Developing and implementing processes to detect, moderate, report and remove (as applicable) material or activity on the service that is unlawful or harmful. b. Ensuring that the default privacy and safety settings of the children's service are robust and set to the most restrictive level - if a service or a component of a service (such as an online app or game) is targeted at, or being used by, children (the children's service). c. Ensuring that persons who are engaged in providing the service, such as the provider's employees or contractors, are trained in, and are expected to implement and promote, online safety. d. Continually improving technology and practices relating to the safety of end-users. e. Ensuring that assessments of safety risks and impacts are undertaken, and safety review processes are implemented, throughout the design, development, deployment and post- deployment stages for the service.
	(S 7) Consult with the eSafety Commissioner and refer to the Commissioner's guidance in determining such reasonable steps to ensure safe use	The provider will consult the Commissioner in determining the reasonable steps to ensure safe use. The provider will also have regard to any relevant guidance material made available by the Commissioner.	

[54] Division 1 provides an overview of the purpose of the Determination.

Division	Headline Expectation	Expectations	Examples of reasonable steps that could be taken (where provided in the Determination) or qualifications
2. Safe use	(S 8) Reasonable steps regarding encrypted services	If the service uses encryption, the provider will take reasonable steps to develop and implement processes to detect and address material and activity on the service that is unlawful or harmful.	**Qualifications** This expectation does not create a requirement to: 1. implement or build a systemic weakness, or systemic vulnerability, into an encrypted service 2. build a new decryption capability into an encrypted service 3. render methods of encryption less effective.
	(S 9) Reasonable steps regarding anonymous accounts	If the service permits the use of anonymous accounts, the provider will take reasonable steps to prevent those accounts being used to deal with material, or for activity, that is unlawful or harmful.	**Examples of reasonable steps** 1. Having processes that prevent the same person from repeatedly using anonymous accounts to post material, or engage in activity, that is unlawful or harmful. 2. Having processes in place that require verification of identity or ownership of accounts.
	(S 10) Consult and cooperate with other services to promote safe use	The provider will take all reasonable steps to consult and cooperate with other service providers to promote the ability of end-users to use all those services in a safe manner	**Examples of reasonable steps** 1. Working with other service providers to detect high volume, cross-platform attacks (also known as 'pile-on' or 'volumetric' attacks). 2. Sharing information with other service providers about unlawful or harmful material and activity for the purpose preventing and dealing with such nmaterial or activity.
3. Certain material	(S 11) Reasonable steps to minimise provision of certain material	The provider will take reasonable steps to minimise the extent to which the following material is provided on the service: 1. Cyberbullying material targeted at an Australian child. 2. Adult cyber abuse material. 3. Non-consensual intimate images of a person. 4. Class 1 material. 5. Material promoting, inciting, instructing in or depicting abhorrent violent conduct.	
	(S 12) Reasonable steps to prevent access by children to class 2 material	The provider will take reasonable steps to ensure that technological and other measures are in effect to prevent access by children to class 2 material provided on the service.	**Examples of reasonable steps** 1. Implementing age assurance mechanisms. 2. Conducting child safety risk assessments.

Division	Headline Expectation	Expectations	Examples of reasonable steps that could be taken (where provided in the Determination) or qualifications
4. Reports and complaints	(S 13) Mechanisms to report and make complaints about certain material	The provider will ensure that the service has clear and readily identifiable mechanisms that enable end-users and any person ordinarily resident in Australia to report and make complaints about certain material provided on the service.	
	(S 14) Service has terms of use, policies, procedures to deal with complaints	The provider will ensure that the service has: 1. terms of use 2. policies and procedures relating to end-user safety 3. policies and procedures for dealing with complaints and reports 4. standards of conduct for end-users 5. policies and procedures relating to content moderation and the enforcement of conduct standards. Providers will take reasonable steps so that penalties for breaches of terms of use are enforced against all accounts held or created by the end-user who breached the terms of service.	
	(S 15) Service will have mechanisms to report and make complaints about breaches of terms of use	The provider will ensure that the service has clear and readily identifiable mechanisms that enable: 1. end-users, and 2. any person ordinarily residing in Australia, to report, and make complaints about, breaches of the service's terms of use.	
	(S 16) Accessible information on how to complain to eSafety	The provider will ensure that there is readily accessible information and guidance provided to end-users on how to make a complaint to eSafety, in accordance with the Online Safety Act 2021, about any of the 'certain material' listed above – including class 2 material.	
5. Accessible information	(S 17) Information on terms of use, policies and complaints made accessible	The provider will provide information on: 1. terms of use, policies and procedures, and standards of conduct 2. online safety and parental control settings – including the availability of tools and resources published by eSafety. The provider will ensure that that this information is: 1. readily accessible to end-users 2. accessible at all points in the end-user experience (for online safety settings, parental controls, and eSafety resources) 3. regularly reviewed and updated 4. written in plain language.	
	(S 18) End-users receive updated information about changes to policies, terms and conditions, or similar documents	The provider will ensure that end-users receive plain language updates about any changes to the information listed above. Such updates include targeted in-service communications.	

Division	Headline Expectation	Expectations	Examples of reasonable steps that could be taken (where provided in the Determination) or qualifications
6. Record Keeping	(s19) Records of end- user-reports and complaints to be kept for five years	The provider will keep records of reports and complaints about certain material provided on the service for five years after the report or complaint is made.	
7. Dealings with the Commissioner	(S 20) Provider will provide requested information to the Commissioner	The provider must comply within 30 days if the Commissioner gives them a written notice requesting: 1. A statement that sets out the number of complaints made to the provider during a specified period (not shorter than six months) about breaches of the service's terms of use. 2. A statement that sets out, for each removal notice given to the provider during a specific period (not shorter than six months), how long it took the provider to comply with the removal notice. 3. Specified information relating to the measures taken by the provider to ensure that end-users are able to use the service in a safe manner. 4. A report on the performance of online safety measures that the provider has announced publicly or reported to the Commissioner.	
	(S 21) Provider will have a designated contact point	1. The provider will ensure that there is an employee, or agent of the provider, that is designated as the service's contact point for the purposes of the Online Safety Act 2021. 2. The provider will ensure that this contact person's e-mail address and phone number are given to the Commissioner. 3. If there is a change to the identity or contact details of the contact point, the provider will give the Commissioner written notice of the change within 14 days.	

eSafety.gov.au

Complaint

CYR-0228040

Created On
14/10/2023 8:11 AM

Complaint Status
Closed

Assigned To
s 22

Response Required

Summary

Date Received
14/10/2023 8:11 AM

Date Finalised
14/10/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant
s 22, s 47E(d), s 47F

Repeat Complainant
Yes

Complainant Reason
Violent extremist material

Response Required

Response Method
E-mail

Submission

Method Received
Web Form

Presentation Type
Social media

Device Type
Mobile Phone

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Narr
https://twitter.com/s 47E(d), s 47F	Complained	Social Media Se...			

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Comme
	INV-2023-27890	Discretion not to investig...	https://twitter.co s 47E(d), s 47F	No	Violentextremistmaterial	NA	14/10/202

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commer

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason

No E-mail records are available in this view.

0 - 0 of 0 (0 selected) Page 1

E-mail

FOR URGENT ATTENTION: DETECTION OF ONLINE INCIDENT - Shooting,...

Owner s 22

Date Sent/Received 17/10/2023 9:29 AM

E-mail

From
National Situation Room

To
s 22 notify_oce; s 22

Cc
Cyber Report; Cyber Report; Cyber Report; Cyber Report; National Situation Room; Online Content; online online

Bcc

Subject
FOR URGENT ATTENTION: DETECTION OF ONLINE INCIDENT - Shooting, Brussels, Belgium [SEC=OFFICIAL]

Classification
[SEC=OFFICIAL:Sensitive]

Regarding
CYR-8000441

Attachments

Drop files here.

File Name	File Size (Bytes)
Appendix C - NSR Notification to eSafety.docx	641,778
image001.jpg	17,127

1 - 2 of 2 (0 selected) Page 1

OFFICIAL

Good morning safety,

PFA a report from the Online Terrorism and Violent Extremism Policy Section of the Department of Home Affairs. For convenience please see the reported links below.

https://twitter.com s 47E(d)
https://twitter.com

Will we await further advise from you, regarding further action.

Kind regards,

s 22
Senior Information Officer | National Situation Room | National Crisis Operations Branch
Emergency Management and Response Group
National Emergency Management Agency

s 22 @nema.gov.au

Australian Government
National Emergency Management Agency

The National Emergency Management Agency acknowledges the Traditional Owners of Country throughout Australia. We recognise the deep continuing connection Aboriginal and Torres Strait Islander peoples have to land, sea, culture and community. We pay our respects to the elders past, present and emerging.

OFFICIAL

Type
General

Approval

Approval Comments

Approving Officer **Escalating Officer**

Approval Date

Status Reason Received

Complaint

CYR-8000441

Created On
17/10/2023 9:47 AM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
17/10/2023 9:46 AM

Date Finalised
18/10/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
Yes

Education Provider
No

Complainant
National Situation Room

Repeat Complainant
Yes

Complainant Reason
Violentextremist material

Response Required
No

Response Method
Email

Submission

Method Received
E-mail

Presentation Type
Web Page

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	Service Type	ServiceCategory	P2P Filename	Newsgroup Name	Content Path	HostingCountry
https://twitter.com/s 47E(d)	Complained	Social Media Se...					
https://twitter.co	Complained	Social Media Se...					

1 - 2 of 2 (0 selected) Page 1

S...	InvestigationId	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced
	INV-2023-28120	Discretion not to investigate	https://twitter.com/s 47E(d)	No	Violentextremistmaterial	NA	17/10/2023
	INV-2023-28119	Discretion not to investigate	https://twitter.com	No	Violent extremist material	NA	17/10/2023

1 - 2 of 2 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	ComplainantReason	Classification	Commenced

No investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
18/10/2023 4:45 AM	Twitter	Cyber Report; Cyber Report; Cyber ...	Case# (0344767885: [SEC=OFFICIAL:Sensitive] From the eS...	Received
17/10/2023 11:14 AM	notify_oce	National Situation Room	FW: FOR URGENT ATTENTION: DETECTION OF ONLINE IN...	Received
17/10/2023 11:04 AM	Twitter	Cyber Report; Cyber Report; Cyber ...	We have an update about 0344767885 [ref:!00DA0000!0D...	Received
17/10/2023 10:52 AM	Cyber Report	Twitter Escalations	[SEC=OFFICIAL:Sensitive] CRM0035971	Sent

1 - 4 of 5 (0 selected) Page 1

Complaint

CYR-0230272

Created On
23/10/2023 10:08 AM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
23/10/2023 10:08 AM

Date Finalised
24/10/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
Yes

Australian Resident
Yes

PoliceInvolved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Advocates a terroristact

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
Social media

Device Type
Mobile Phone

Device Name

Remote IP Address

Access Information

Additional access instructions
s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path	Hosting Country
https://twitter.com/s 47E(d)	Complained	Social Media Se...					

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced
	INV-2023-28524	Discretion not to investig...	https://twitter.com/s 47E(d)	No	Advocates a terroristact	X 18+	24/10/2023

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced

No investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
23/10/2023 10:08 AM	Cyber Report	eSafety Cyber Report DL	Notification - Advocates a terroristact [SEC=OFFICIAL:Sa...	Sent

1 - 1 of 1 (0 selected) Page 1

E-mail

FOR URGENT ATTENTION: DETECTION OF ONLINE INCIDENT - NOTIFICA...

Owner s 22

Date Sent/Received 14/10/2023 9:16 AM

E-mail

From
National Situation Room

To
Cyber Report; Cyber Report; Cyber Report; Cyber Report; s 22
s 22 notify_oce; Online Content; online online

Cc
National Situation Room

Bcc

Subject
FOR URGENT ATTENTION: DETECTION OF ONLINE INCIDENT - NOTIFICATION 14OCT23 [SEC=OFFICIAL:Sensitive]

Classification
[SEC=OFFICIAL:Sensitive]

Regarding
CYR-8000445

Attachments

Drop files here.

File Name	File Size (Bytes)
image007.jpg	876
image008.jpg	873
image009.jpg	912
image010.jpg	906
image011.jpg	873
image012.jpg	7,005

1 - 6 of 6 (0 selected) Page 1

Some people who received this message don't often get email from nsr@nema.gov.au. Learn why this is important

OFFICIAL: Sensitive

Good morning e-safety,

We have detected Terrorist or Extreme Violent Material (TEVM) on the internet that is accessible in Australia and spreading or likely to spread virally. Known information is detailed below.

Please perform your initial assessment as soon as possible and advise if this is an Online Crisis Event:

- **If yes: eSafety must notify Online Content Incident Arrangement (OCIA) participants of steps taken and which legislative powers will be invoked to contain the threat.**

- **If no: eSafety must advise nsr@nema.gov.au and we will notify OCIA stakeholders.**

Summary:

Likely bystander footage of the terrorist attack in Arras, France posted on twitter at 9.10pm on 13 October 2023. It appears content has been flagged on twitter from public reports. 54.5k views at 8.15am. 36 second clip.

Where is the TEVM being distributed (*all that apply*):

☐ **WEBSITE**

☐ **SOCIAL MEDIA PLATFORM**
X (FORMELY Twitter): *https://twitter.com* s 7(2)

Type of Content (*all that apply*):

☐ **Audio-video** ☐ **Audio only**

☐ **Video only** ☐ **Image(s)**
☐ **Manifesto/Threat** ☐ **Other (*Please describe below*)**

Other:

-

Depiction of TEVM (*all that apply*):

☐ **Terrorist or violent extremist act leading to death or serious injury**
☐ **Murder or attempted murder** ☐ **Rape**
☐ **Kidnapping**

Regards,

s 22
Shift Leader | National Situation Room | National Crisis Operations Branch
National Emergency Management Agency

PH: 1300 243 222
s 22 @nema.gov.au

Level 1/Basement, Edmund Barton Building, Kings Ave, (PO Box 133 Canberra) Barton ACT 2600

nema.gov.au | @nemagovau

Australian Government
National Emergency Management Agency

The National Emergency Management Agency acknowledges Traditional Owners of Country throughout Australia and recognises the continuing connection to lands, waters and communities. We pay our respect to Aboriginal and Torres Strait Islander cultures; and to Elders past and present.

OFFICIAL: Sensitive

Type
General

Approval

Approval Comments

Approving Officer **Escalating Officer**

Approval Date

Status Reason	**Received**

Complaint

CYR-8000445

Created On
27/10/2023 3:28 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
14/10/2023 9:00 AM

DateFinalised
27/10/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant
National Situation Room

Repeat Complainant
Yes

Complainant Reason
Advocates a terrorist act

Response Required
Yes

Response Method
E-mail

Submission

Method Received
E-mail

Presentation Type
Web Page

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Nam
https://twitter.com/s 47E(d)	Complained	Hosting Service			

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Comme
	INV-2023-28863	Discretion not to investig...	https://twitter.com/s 47E(d)	No	Advocates a terrorist act	NA	27/10/2023

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	AustralianConn...	Complainant Reason	Classification	Commer

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
14/10/2023 10:00 AM	s 22	National Situation Room	Re: Online Incident Update Footage of Terrorist Attack in Arras, France - 14OCT23 [SEC=OFFICIAL:Sensitive]	Received
14/10/2023 9:44 AM	National Situation Room	s 37(2)(b)	Online Incident Update Footage of Terrorist Attack in Arras, France - 14OCT23 [SEC=OFFICIAL:Sensitive]	Received
14/10/2023 9:31 AM	notifyoca	Cyber Report; Cyber Report; Cyber ...	RE: FOR URGENT ATTENTION: DETECTION OF ONLINE INCIDENT - NOTIFICATION 14OCT23We have reviewed the footage and it is not abhorrent violent conduct material and [SEC=OFFICIAL]	Received
14/10/2023 9:30 AM	National Situation Room	s 37(2)(b)	Potential Online Incident - Footage of Terrorist Attack in Arras, France - 14OCT23 [SEC=OFFICIAL:Sensitive]	Received

1 - 4 of 5 (0 selected) Page 1

Complaint

CYR-0233320

Created On
3/11/2023 4:55 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
3/11/2023 4:55 PM

Date Finalised
6/11/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
Yes

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Advocates a terrorist act

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
Social media

Device Type
Mobile Phone

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path	Ho
https://twitter.co s 47E(d)	Complained	Social Media Se...					

1 - 1 of 1 (0selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenc
	INV-2023-29534	Terminated	https://twitter.com s 47E(d)	No	Advocates a terrorist act	NC	6/11/2023

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
3/11/2023 4:56 PM	Cyber Report	eSafety Cyber Report DL	Notification - Advocates a terrorist act [SEC=OFFICIAL:Sa..	Sent

1 - 1 of 1 (0 selected) Page 1

Complaint

CYR-0233353

Created On
3/11/2023 8:43 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
3/11/2023 8:43 PM

Date Finalised
6/11/2023

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
Yes

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Advocates a terrorist act

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
Social media

Device Type
MobilePhone

Device Name

Remote IP Address

Access Information

s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path	Hosting C
https://twitter.com s 47E(d)	Complained	Social Media Se...					

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced
	INV-2023-29547	Discretion not to investig...	https://twitter.com s 47E(d)	No	Advocates a terrorist act	NA	6/11/2023

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced

No investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
3/11/2023 8:43 PM	Cyber Report	eSafety Cyber Report DL	Notification A dvocates a terrorist act [SEC=OFFICIALSe...	Sent

1 - 1 of 1 (0 selected) Page 1

Complaint

CYR-0269445

Created On
19/12/2023 6:21 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
19/12/2023 6:21 PM

Date Finalised
22/12/2023

Complaint Status
Closed

Asset
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant
s 47E(d)

Repeat Complainant
Yes

Complainant Reason
Violent extremist material

Response Required
Yes

Response Method
E-mail

Submission

Method Received
Web Form

Presentation Type
Social media

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional Access Instructions
s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	NewsgroupName	Content Path	Hosting Country	Parent Locator	Cre
s 22	Complained	Social Media Service (SNS)							19/12/2023 6:22 P
	Complained	Social Media Service (SNS)							19/12/2023 6:22 P
https://x.com/s 47E(d)	Complained	Social Media Service (SNS)							19/12/2023 6:21 P

1 - 3 of 3 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Com...	Complainant Reason	Classification	Commenced	Date Finalised	Investigator	Typ
	INV-2023-3293	Discretion not to investigate	s 22	No	Violent extremist material	NA	19/12/2023	22/12/2023	s 22	OSASection 36
	INV-2023-32933	Discretion not to investigate		No	Violent extremist material	NA	19/12/2023	22/12/2023		OSASection 36
	INV-2023-32932	Discretion not to investigate	https://x.com/s 47E(d)	No	Violent extremist material	RC1(a)	19/12/2023	22/12/2023		OSASection 36

1 - 3 of 3 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Com...	Complainant Reason	Classification	Commenced	Date Finalised	Investigator	Typ

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Responses/RFI Templates

Name	Created On	Status
Content is RC1(a) (complaint reason not exact)	10/11/2022 12:10 PM	Active

1 - 1 of 1 (0 selected) Page 1

Email Activities

Date Sent/ Received	From	To	Subject	Status Reason
28/02/2024 1:51 PM	s 37(2)(b)	Cyber Report; Cyber Report; Cyber ...	RW: Controlled Violent Extortion CRM:0066913 [SEC=OFFICIAL:Sensitive]	Received
28/02/2024 10:36 AM	Cyber Report	s 37(2)(b) s 47E(d)	Controlled Violent Extortion [SEC=OFFICIAL:Sensitive] CRM:0066913	Sent
23/12/2023 11:10 AM	Cyber Report		Your report to the eSafety Commissioner CYR-0269445 [SEC=OFFICIAL:Sensitive] CRM:0066842	Sent
19/12/2023 6:22 PM	Cyber Report	eSafety Cyber Report DL	Not Illegal – Violent extremist material [SEC=OFFICIAL:Sensitive] CRM:0051798	Sent

1 - 4 of 4 (0 selected) Page 1

Complaint

CYR-0315275

Created On
23/12/2023 2:33 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
23/12/2023 2:33 PM

Date Finalised
2/01/2024

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

PoliceInvolved
No

Education Provider
No

Com
s 47E(d), s 47F

Repeat Complainant
No

Complainant Reason
Violent extremist material

Response Required
Yes

Response Method
E-mail

Submission

Method Received
WebForm

Presentation Type
SocialMedia

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path	Hosting Cou
https://twitter.com/	Complained	Social Media Se...					
https://twitter.com/	Complained	Social Media Se...					

s 47E(d), s 47F

1 - 2 of 2 (0 selected) Page 1

S...	InvestigationId	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced
	INV-2023-33347	Terminated	https://twitter.com	No	Violent extremist material	NC	2/01/2024
	INV-2023-33346	Discretion not to investig...	https://twitter.com	No	Violent extremist material	NA	2/01/2024

s 47E(d), s 47F

1 - 2 of 2 (0 selected) Page 1

S...	InvestigationId	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
2/01/2024 10:54 AM	Cyber Report	s 47E(d), s 47F	Your report to the eSafety Commissioner: CYR-0315275 [...	Sent
23/12/2023 2:33 PM	Cyber Report	eSafety Cyber Report DL	Notification Violent extremist material [SEC=OFFICIAL:S...	Sent

1 - 2 of 2 (0 selected) Page 1

E-mail

s 47E(d)

Owner s 22

Date Sent/Received 24/12/2023 7:55 PM

E-mail

From s 47E(d)

To Cyber Report; Cyber Report; Cyber Report; Cyber Report; Online Content; online online

Cc

Bcc

Subject s 47E(d)

Classification
[SEC=OFFICIAL:Sensitive]

Regarding
CYR-8000468

You don't often get email from s 47E(d) Learn why this is important

s 47E(d), s 7(2)

Attachments

Drop files here.

File Name	File Size (Bytes)
No Attachment records are available in this view.	
0 - 0 of 0 (0 selected)	Page 1

s 47E(d), s 7(2)

Type
General

Approval

Approval Comments

Approving Officer

Escalating Officer

Approval Date

Status Reason Received

Complaint

CYR-8000468

Created On
2/01/2024 10:54 AM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
24/12/2023 7:30 AM

Date Finalised
5/01/2024

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant
s 47E(d)

Repeat Complainant
Yes

Complaint Reason
Advocates a terrorist act

Response Required
Yes

Response Method
E-mail

Submission

Method Received
E-mail

Presentation Type
Web Page

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path	Hosting Country	Parent Le
https://twitter.co	Complained	Hosting Service						
https://twitter.co	Complained	Hosting Service						
https://twitter.co	Complained	Hosting Service						
https://twitter.co	Complained	Hosting Service						
https://twitter.co	Complained	Hosting Service						
https://twitter.com	Complained	Hosting Service						
https://twitter.co	Complained	Hosting Service						
https://twitter.co	Complained	Hosting Service						

s 47E(d)

1 - 8 of 8 (0 selected) Page 1

Investigation Id	Status	InvestigatedLocator	Australian Com...	Complainant Reason	Classification	Commenced	Date Finalised
INV-2024-00115	Completed	https://twitter.com/	No	Advocates a terrorist act	RCOA	3/01/2024	5/01/2024
INV-2024 -00114	Discretion not to investigate	https://twitter.com/	No	Advocates a terrorist act	NA	3/01/2024	3/01/2024
INV-2024-00113	Terminated	https://twitter.com	No	Advocates a terrorist act	NC	3/01/2024	3/01/2024
INV-2024-00112	Discretion not to investigate	https://twitter.com/	No	Advocates a terrorist act	NA	3/01/2024	3/01/2024

s 47E(d)

1 - 4 of 39 (0 sel bonl

s 22

Investigation Id	Status	InvestigatedLocator	Australian Com...	Complainant Reason	Classification	Commenced	DateFinalised	Investig

No Investigation records are available in this view.

0 - 0 of 0 (0selected) Page1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
5/01/2024 9:55 AM	CyberReport	s 47E(d)	Your report to the eSafety Commissioner: CYR-8000468 [SEC=OFFICIAL:Sensitive] CRM:0104746	Sent
5/01/2024 9:07 AM	Twitter	Cyber Report; Cyber Report; Cyber ...	We have an update for you about @ 035578764 [ref:!00DA0010A8.500Vp01WiNF:ref]	Received
4/01/2024 11:24 AM	Cyber Report	Twitter AUEscalations	Graphic terrorist videos removal request CYR-8000468 [SEC=OFFICIAL:Sensitive] CRM:0104742	Sent
24/12/2023 7:55 PM	s 47E(d)	Cyber Report; Cyber Report; Cyber ...	s 47E(d)	Received

1 - 4 of 4 (0 selected) Page 1

Complaint

CYR-0404296

Created On
4/01/2024 7:29 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
Yes

Summary

Date Received
4/01/2024 7:29 PM

Date Finalised
5/01/2024

Complaint Status
Closed

Ass
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
No

Australian Resident
Yes

Police Involved
No

Education Provider
No

Com
s 47E(d), s 47F

Repeat Complainant
No

Complainant Reason
Violent extremist material

Response Required
Yes

Response Method
Email

Submission

Method Received
Web Form

Presentation Type
Socialmedia

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

s 47E(d), s 47F
Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path	Hosting
https://twitter.com s 47E(d), s 47F	Complained	SocialMediaSe...					

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced	
	INV-2024-00484	Terminated	https://twitter.com s 47E(d), s 47F	No	Violent extremist material	NC	5/01/2024	5/01/2024

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced

No investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status
NC - Content is not available or has been removed	17/06/2013 4:15 PM	Active

1 - 1 of 1 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
5/01/2024 8:36 AM	Cyber Report	s 47E(d), s 47F	Your report to the eSafety Commissioner: CYR-0404296 [SEC=OFFICIAL:Sensitive] CRM:0066854	Sent
4/01/2024 7:29 PM	Cyber Report	eSafety Cyber Report DL	Notification - Violent extremist material [SEC=OFFICIAL:Sensitive] CRM:0051817	Sent

1 - 2 of 2 (0 selected) Page 1

Complaint

CYR-0481787

Created On
15/01/2024 9:30 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
15/01/2024 9:30 PM

Date Finalised
16/01/2024

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
Yes

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Advocates a terrorist act

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
Social media

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path	Hostin
https://twitter.com s 47E(d)	Complained	Social Media Se...					

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australia...	Complainant Reason	Classification	Commenced
	INV-2024-0173	Discretion not to investigate	https://twitter.com/s 47E(d)		Advocates a terrorist act	NA	16/01/2024

1 - 1 of 1 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	AustralianConn...	Complainant Reason	Classification	Commenced

No investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
15/01/2024 9:30 PM	Cyber Report	eSafety Cyber Report DL	Notification Advocates a terrorist act [SEC=OFFICIAL:Sensitive] CRM:0151822	Sent

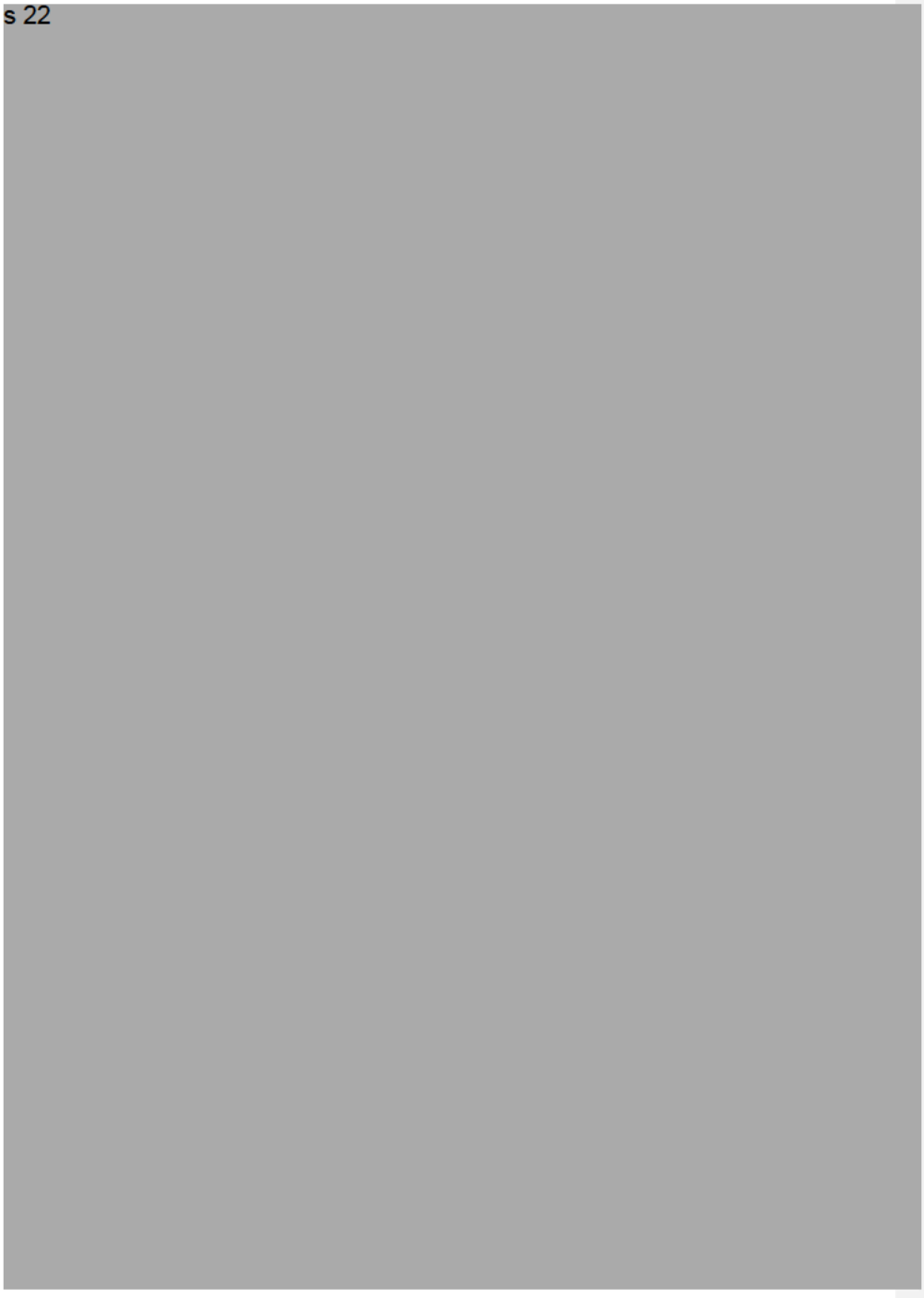
1 - 1 of 1 (0 selected) Page 1

Standard Operating Procedures: Planned BOSE Non-Periodic Reporting Notice

Style Definition: TOC 2: Tab stops: 15.9 cm, Right,Leader: ...

These operating procedures outline the steps that eSafety staff should take when developing and executing a non-periodic reporting notices under section 56(2) of the Act.

s 22


s 22

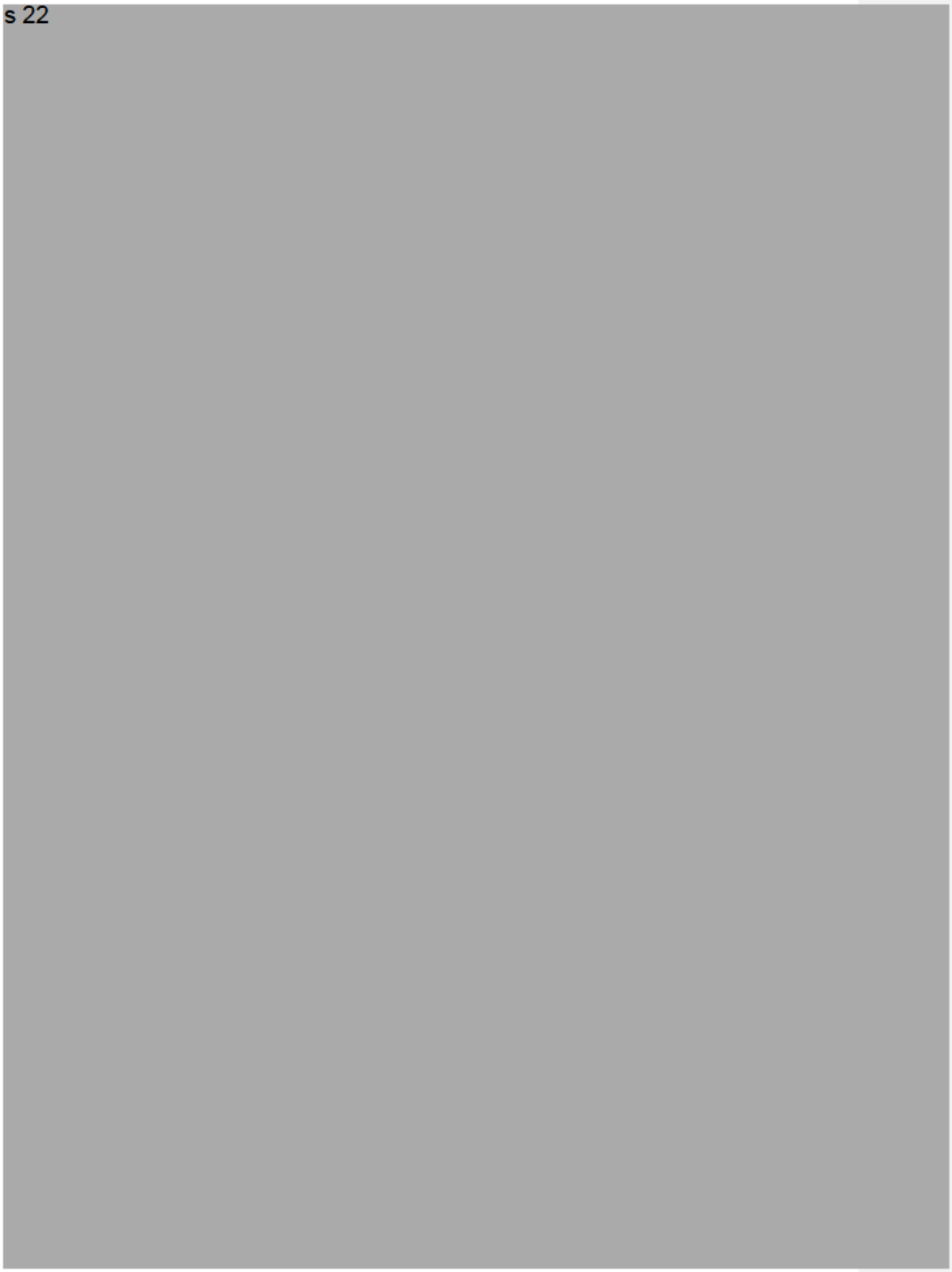
Step 1: Establish the harms or issues to focus on

The scope of BOSE goes beyond the harms captured by eSafety's investigative schemes; (cyberbullying, image-based abuse, adult cyber abuse, abhorrent violent material or child sexual exploitation), and the industry codes (Class 1 and Class 2). Through its section 6 "safe use" expectation, and its section 14 "terms of use enforcement" expectation, the BOSE can target additional harms such as tech-facilitated abuse, hate speech as well as key features that have online safety implications – such as recommender algorithms or generative AI. This broad scope is a key strength of the BOSE but makes prioritisation essential.

s 22

s 22

s 22


s 22

s 22

s 22

NOTE A: When prioritising – criteria in Act.

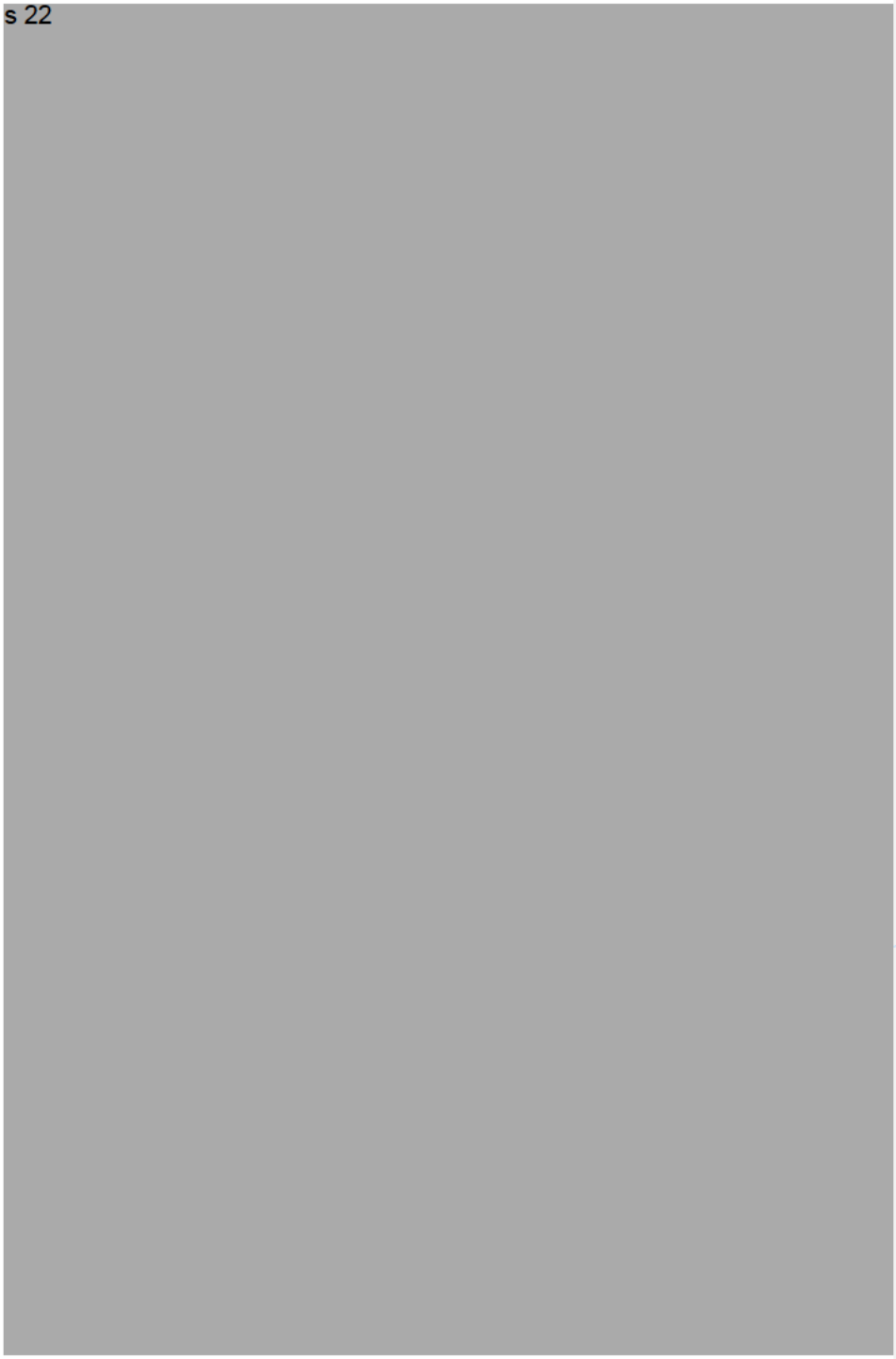
The Act provides considerations that eSafety must have regards to when determining whether to issue a non-periodic reporting notice to a provider. Set out in in s56(5) of the Act, these 6 considerations provide the essential, baseline criteria that staff must consider when prioritising a service for a reporting notice:

- 56(5)(a) the number of occasions during the previous 12 months on which complaints about material provided on the service were made to the Commissioner under the Act.
- 56(5)(b) whether the provider has previously contravened a civil penalty provision of Division 3 of the Act (reporting under the BOS)
- 56(5)(c) whether there are deficiencies in the provider's practices, so far as those practices relate to the capacity of end-users to use the service in a safe manner
- 56(5)(d) whether there are deficiencies in the provider's terms of use, so far as they relate to the capacity of end-users to use the service in a safe manner
- 56(5)(e) whether the provider has agreed to give the Secretary regular reports relating to the capacity of end-users to use the service in a safe manner
- 56(5)(f) such other matters (if any) as the Commissioner considers relevant

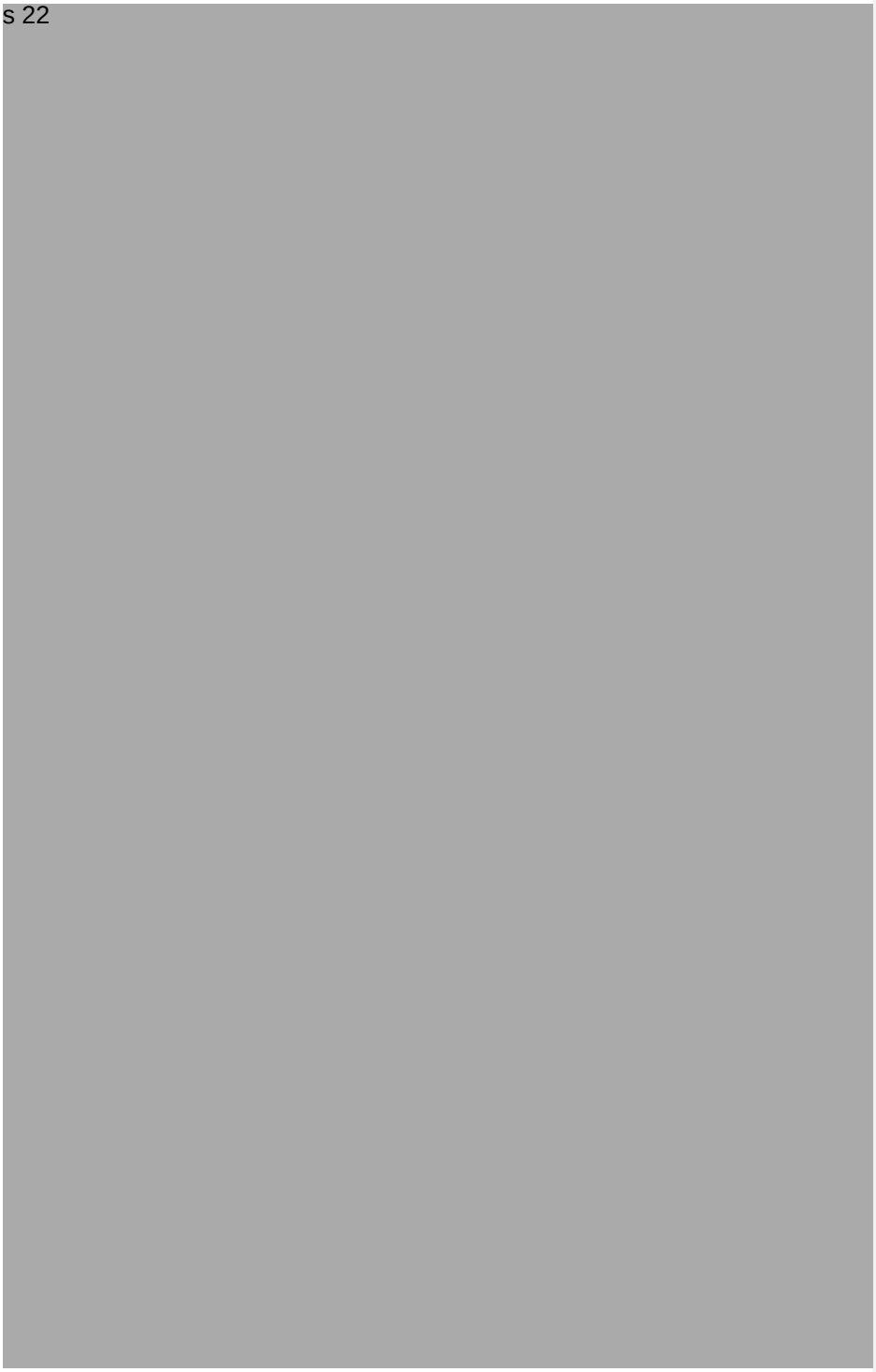
Note: eSafety set out other matters that we may consider in our BOSE regulatory guidance.

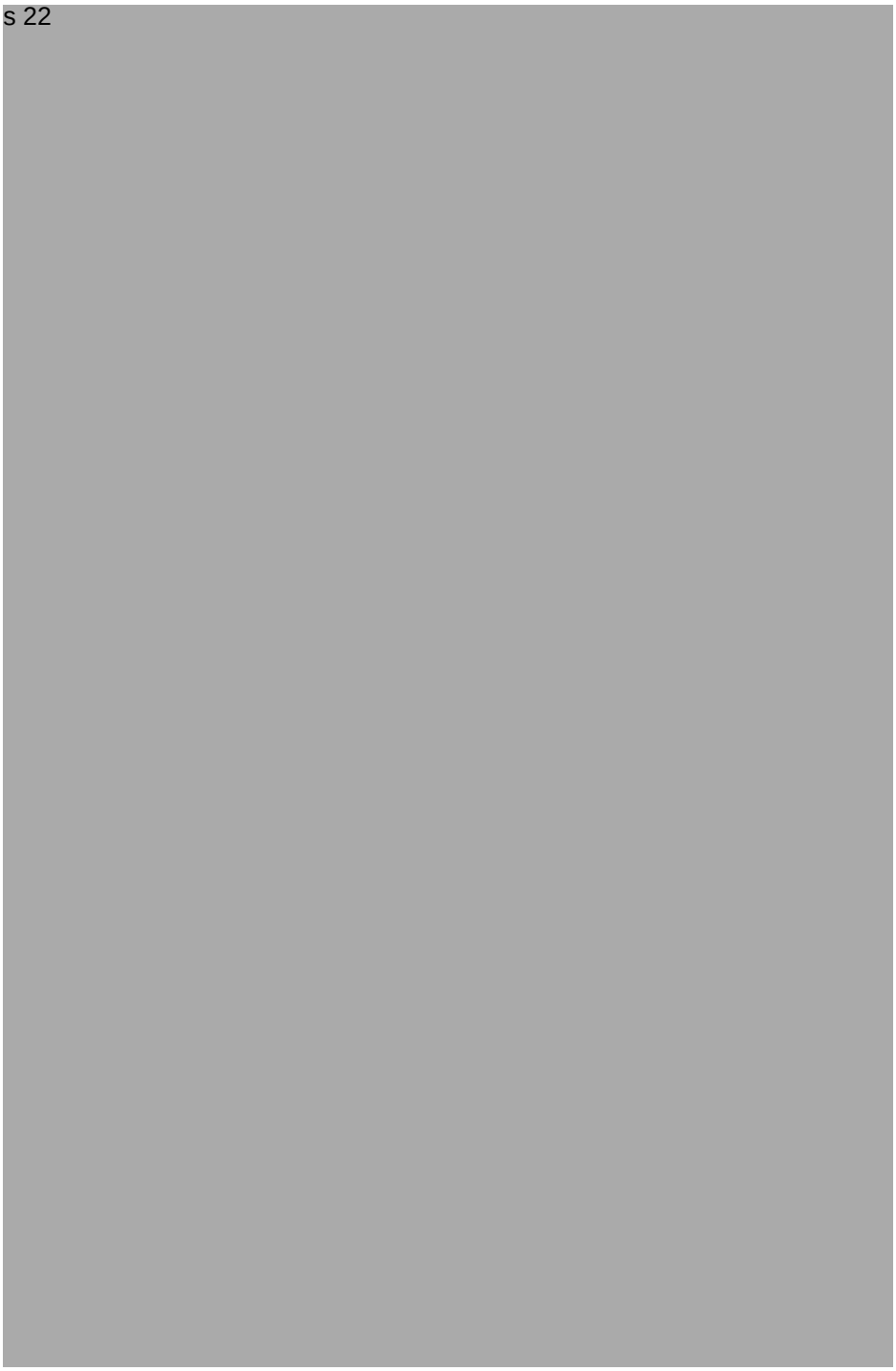
s 22

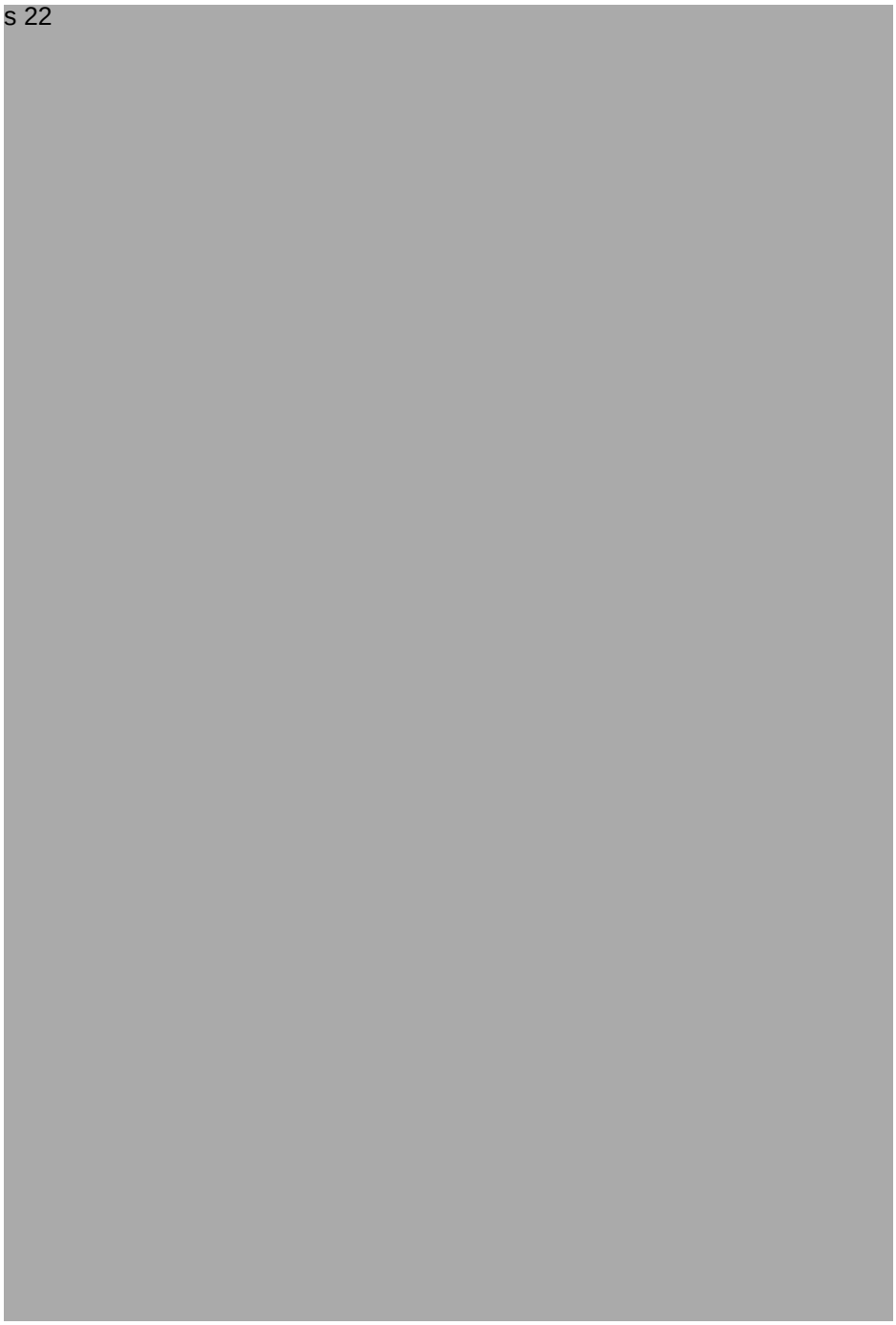
[1] Under subsection 56(5)(e) of the Act, when considering issuing a non-periodic notice to a service provider, the eSafety Commissioner must have regard to the number of complaints that eSafety has received about material on that service. s 22



s 22


s 22


s 22


s 22

s 22 [redacted]

Drafting effective questions

Providers are required to respond to notices in the "manner and form" specified by eSafety. This enables eSafety to determine how companies should answer questions. s 22 [redacted]

s 22 [redacted]

Instructions attached to Schedule B Questions

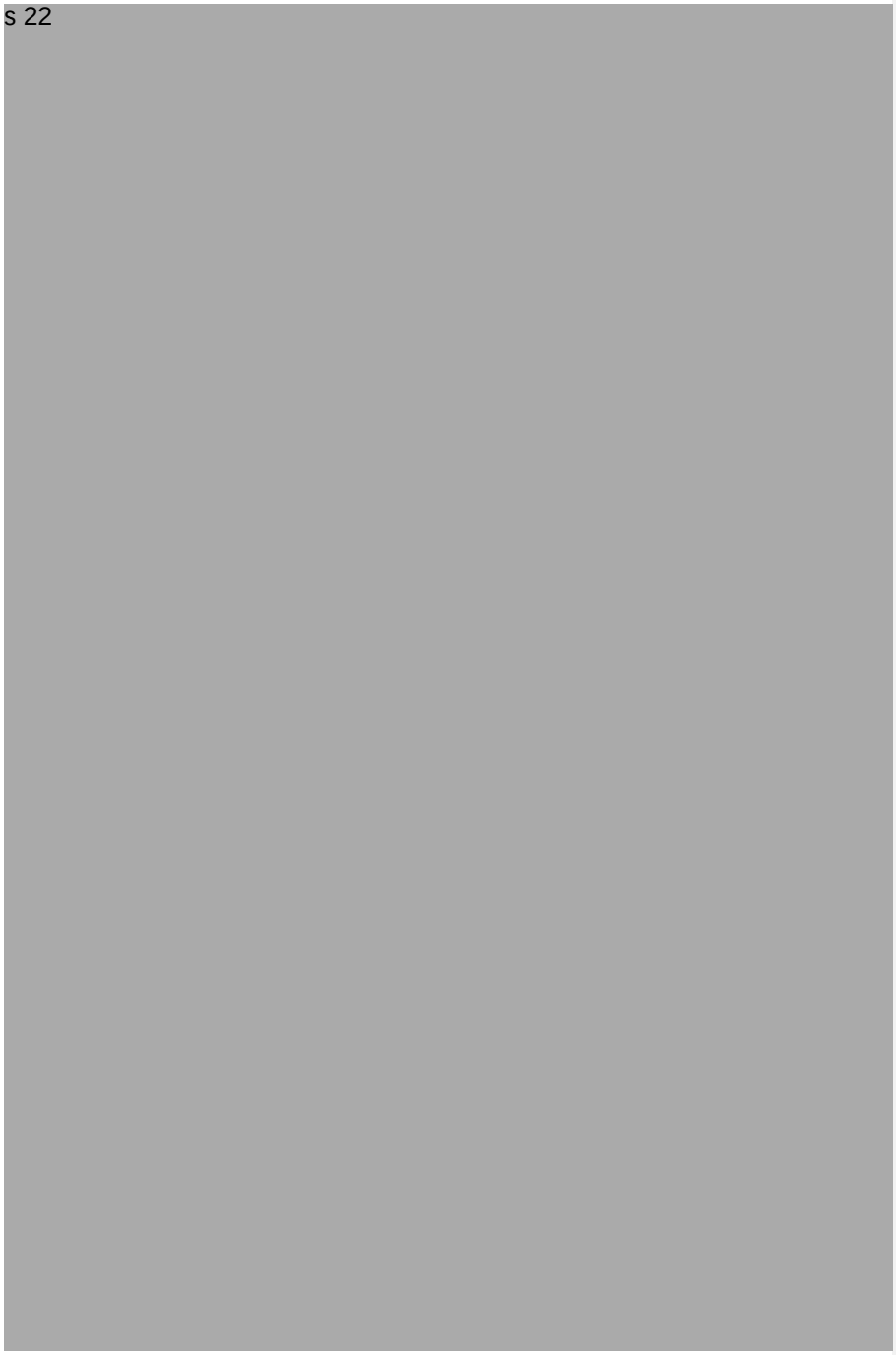
Ensure that the following is stated clearly in the instructions:

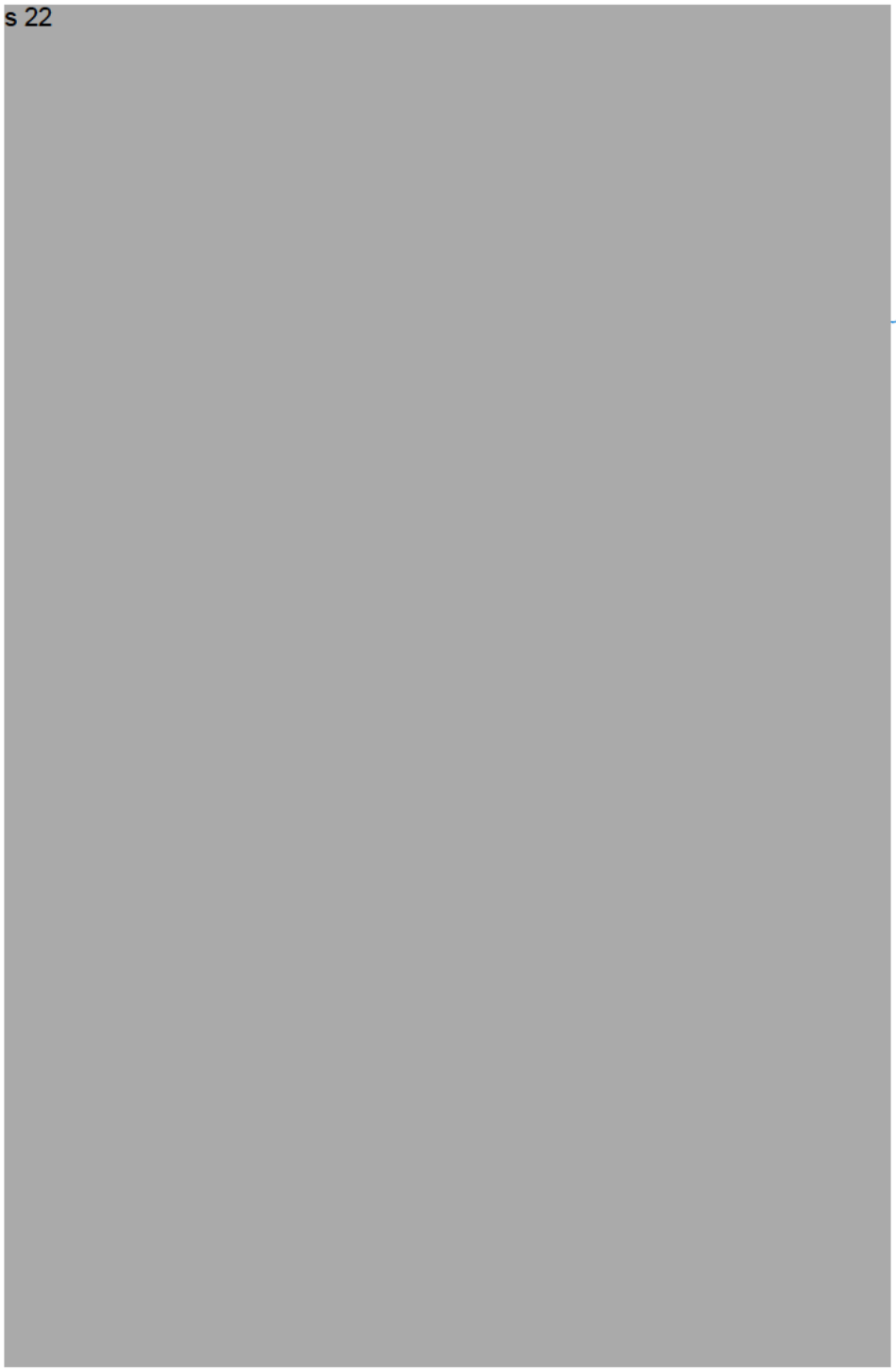
- The questions relate to the specific Expectations in the *Online Safety (Basic Online Safety Expectations) Determination 2022 (the Determination)* (Cth) (**Determination**) referenced below.
- You are required to respond to all the questions in the manner and form specified, including using this template for your response. Responding in an alternative manner or form is likely to be non-compliance with the Notice.
- You are required to answer truthfully and accurately.
- Where information is required, including data, provide all the relevant information from xx to xx (the **Report Period**), unless a question requires information on a specific time period within the Report Period.
- You are required to comply with the Notice, including providing the information required by this document, to the extent you are capable. If you cannot obtain the relevant information, provide alternative relevant information and a detailed explanation of why you cannot provide the information requested. For example, where Australian data cannot be disaggregated from global data, provide the broader dataset. However, you must ensure you consult with eSafety before providing alternative information to that sought by the Notice to confirm whether any alternative information is appropriate.
- If you think a question is not applicable, consult with eSafety before finalising your response and explain why you do not consider a question to be applicable. The questions have been drafted to be specific to your service so should be relevant. Leaving a response to a question blank without first consulting with eSafety is likely to be non-compliance with the Notice.

s 22

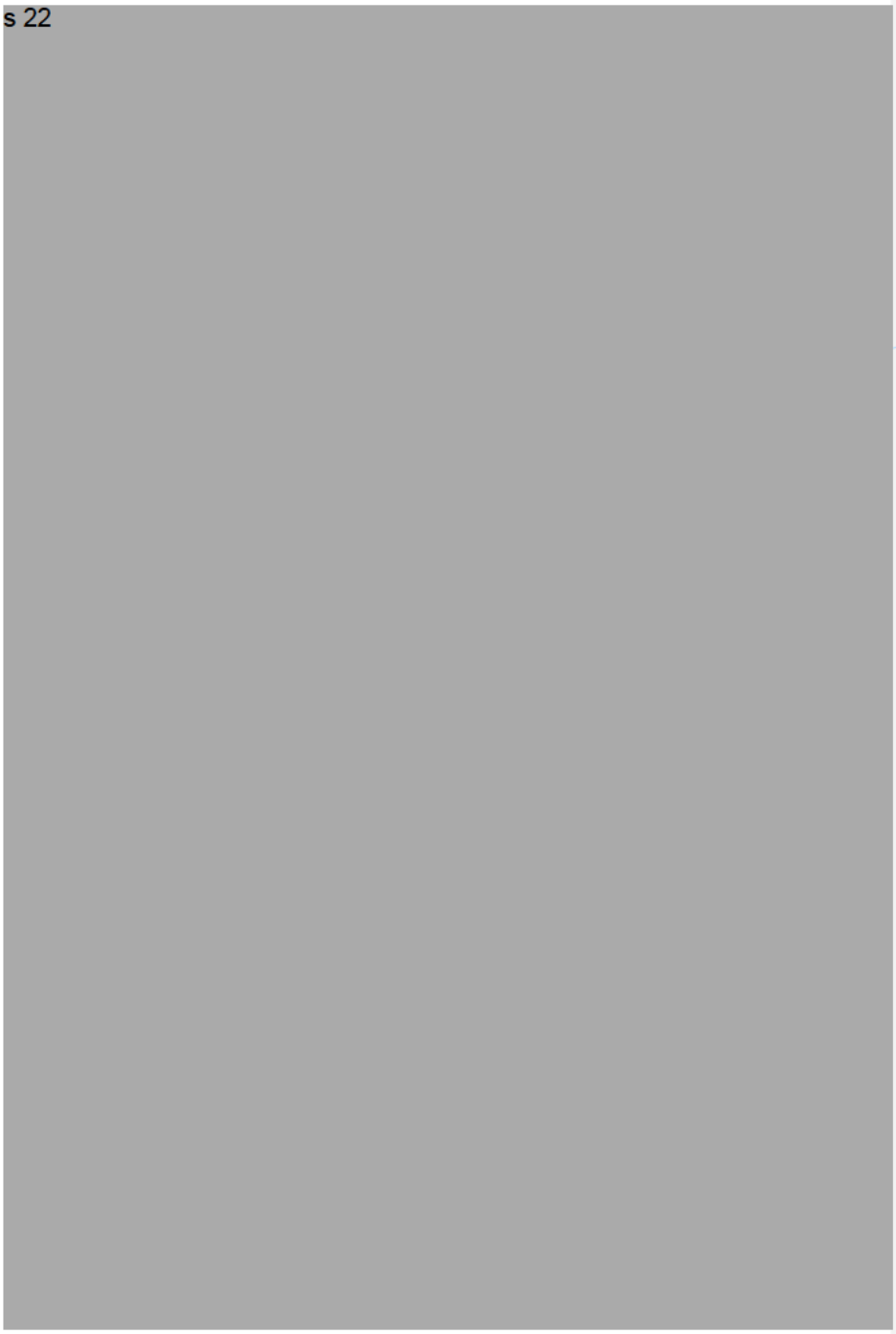
- The timeframe for return of the Notice has been determined following careful consideration of the information you are required to provide. However, if you require additional time, you should contact eSafety to request an extension to the deadline. If requesting additional time, provide details as to why it is required to ensure your compliance with the Notice. This will allow eSafety to properly assess your request.

s 22


s 22


s 22


s 22


s 22


s 22

[redacted]

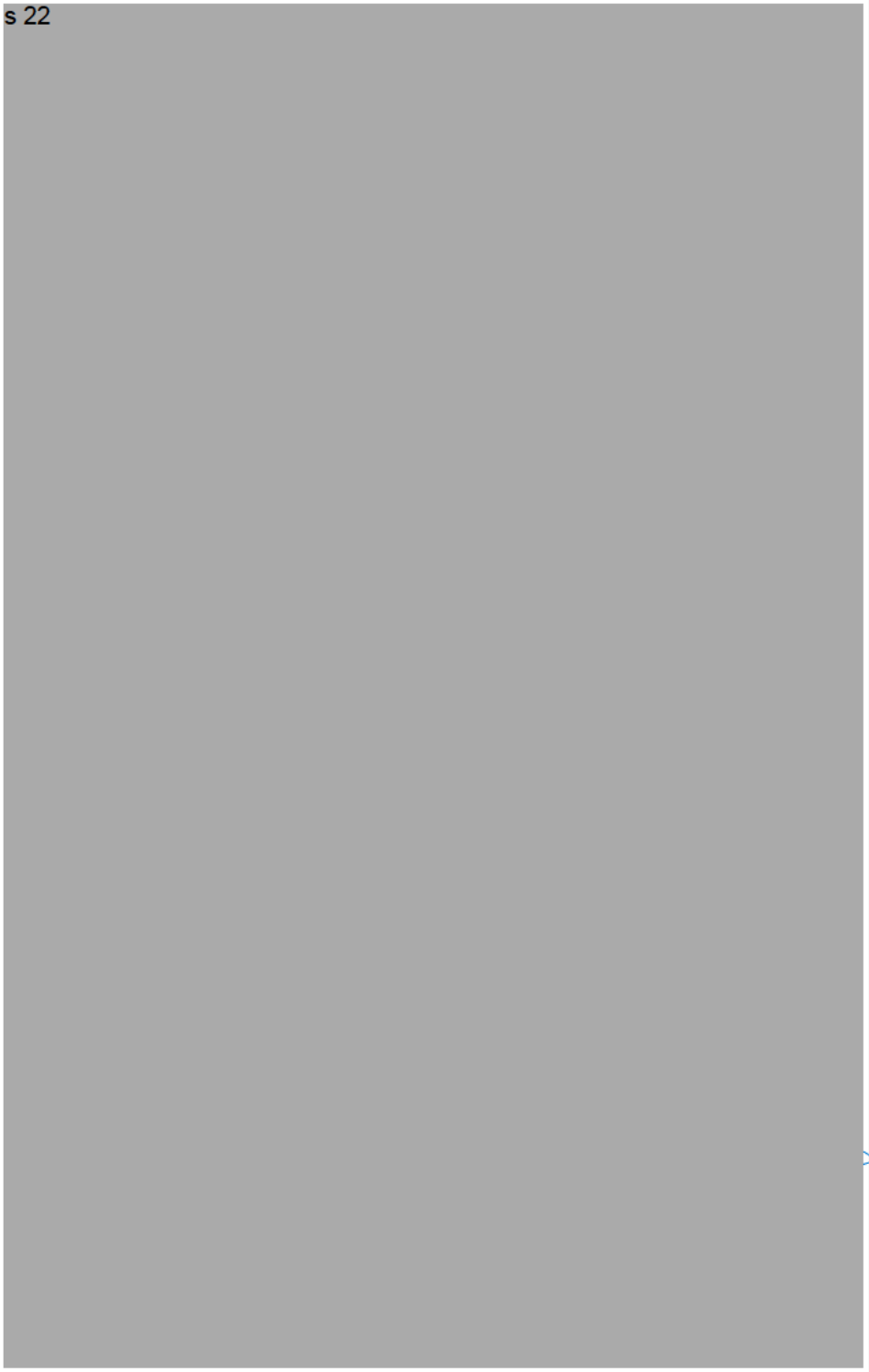
Timeframe

Providers are entitled to a minimum 28-day period to provide eSafety with a response – see subsection 56(4) of the Act. eSafety requires a response from all providers within this timeframe.

s 22

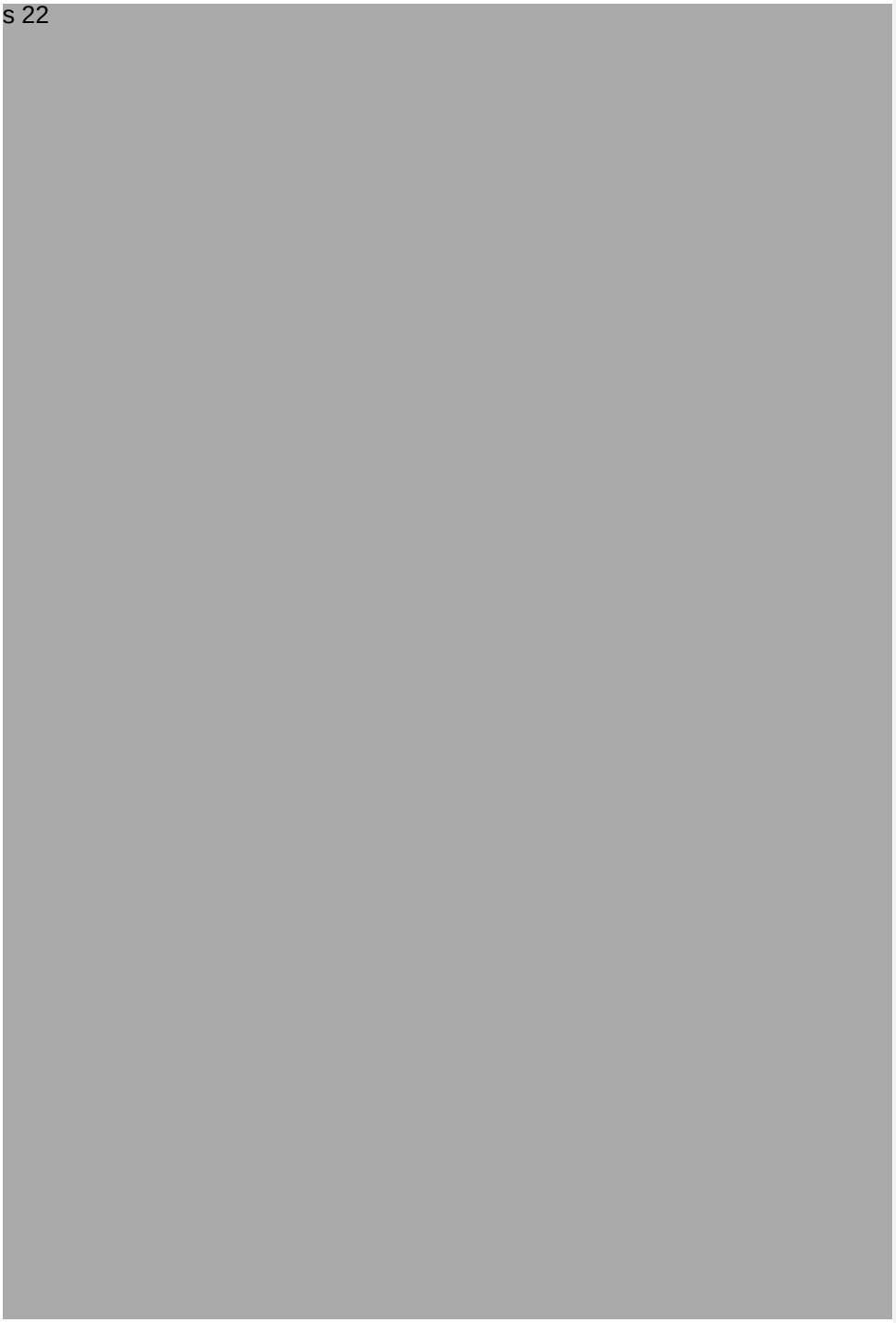
[redacted]

Commented [MS10]: To discuss further


s 22


s 22

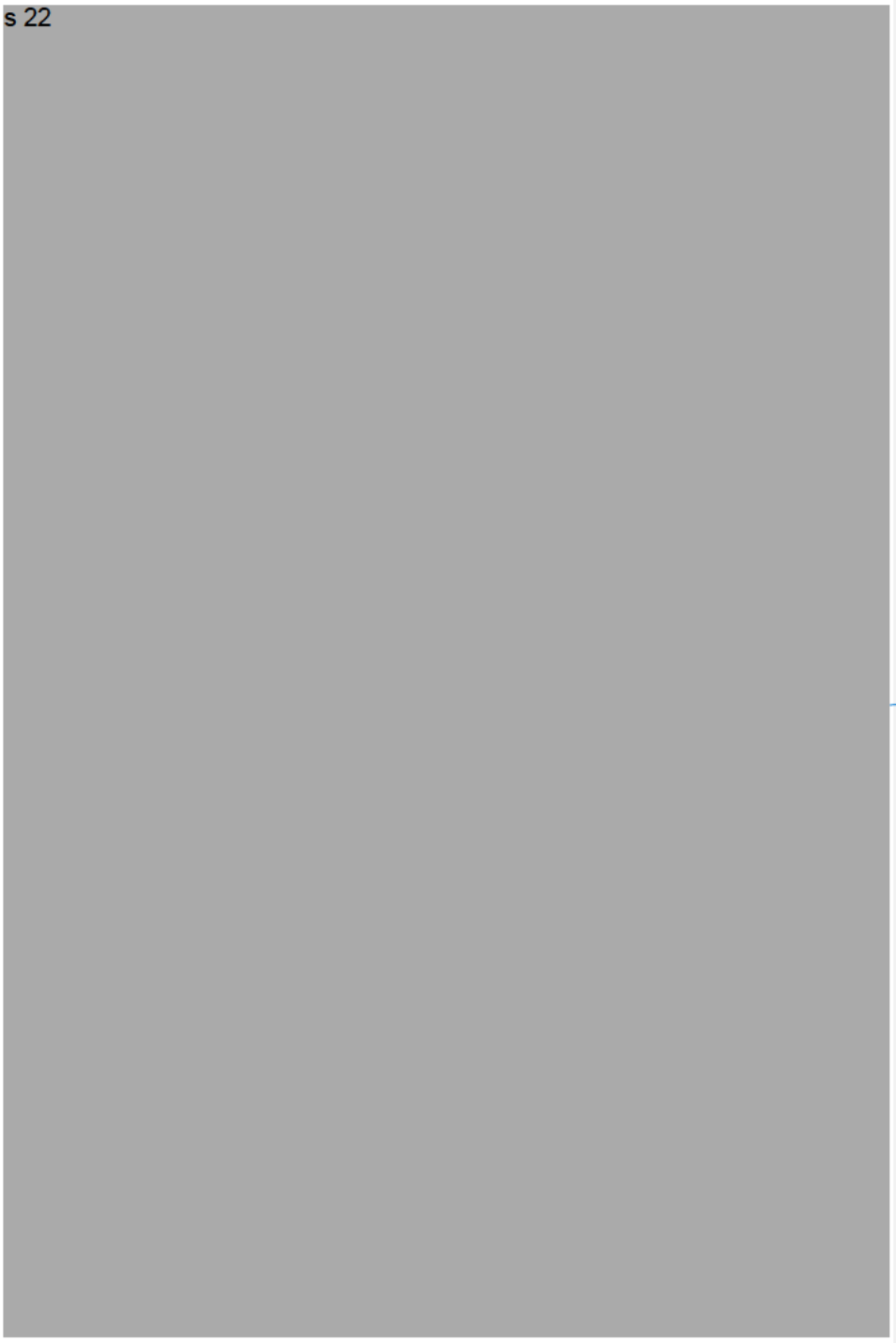
s 22



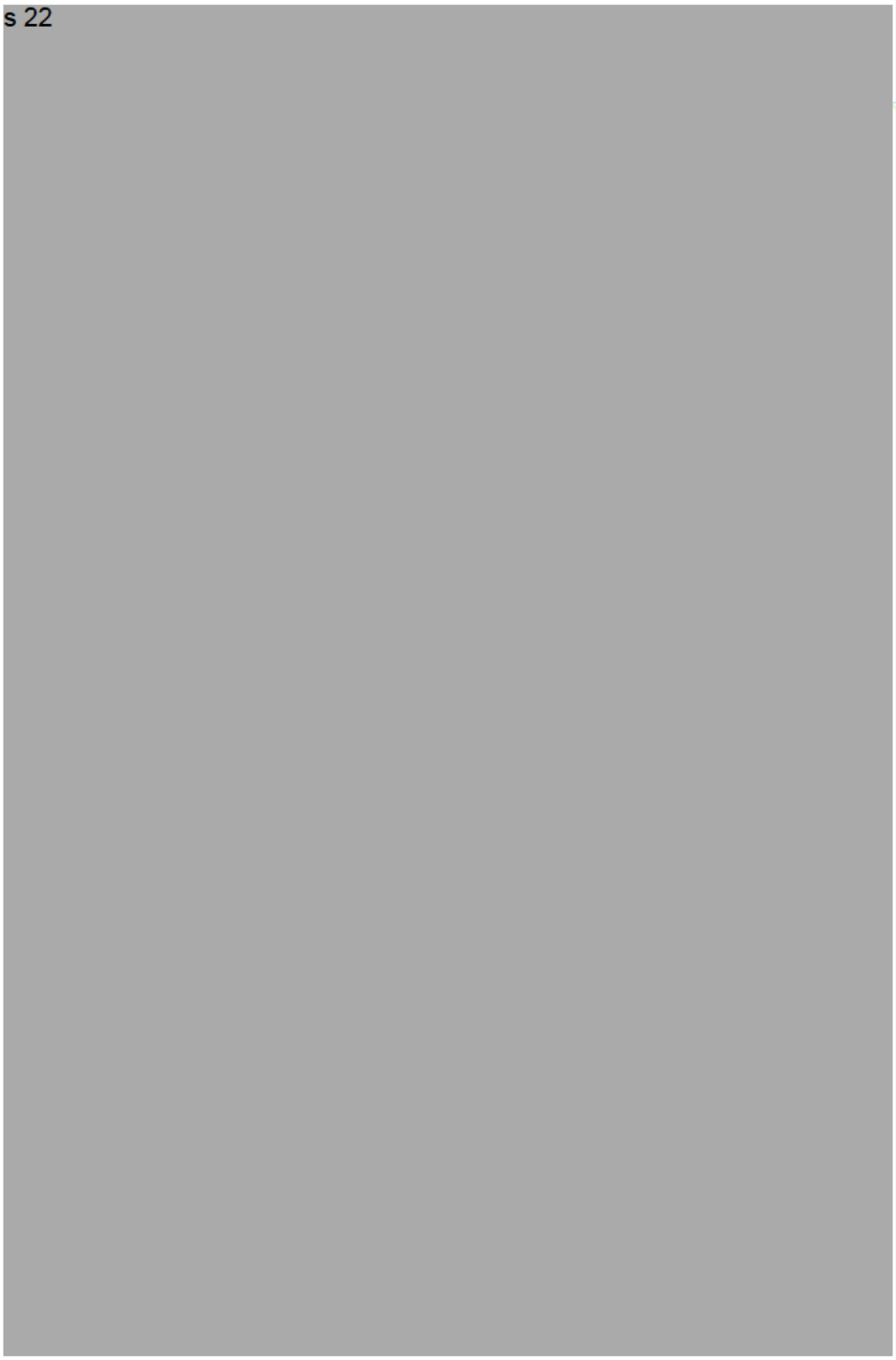
s 22

Formatted Table

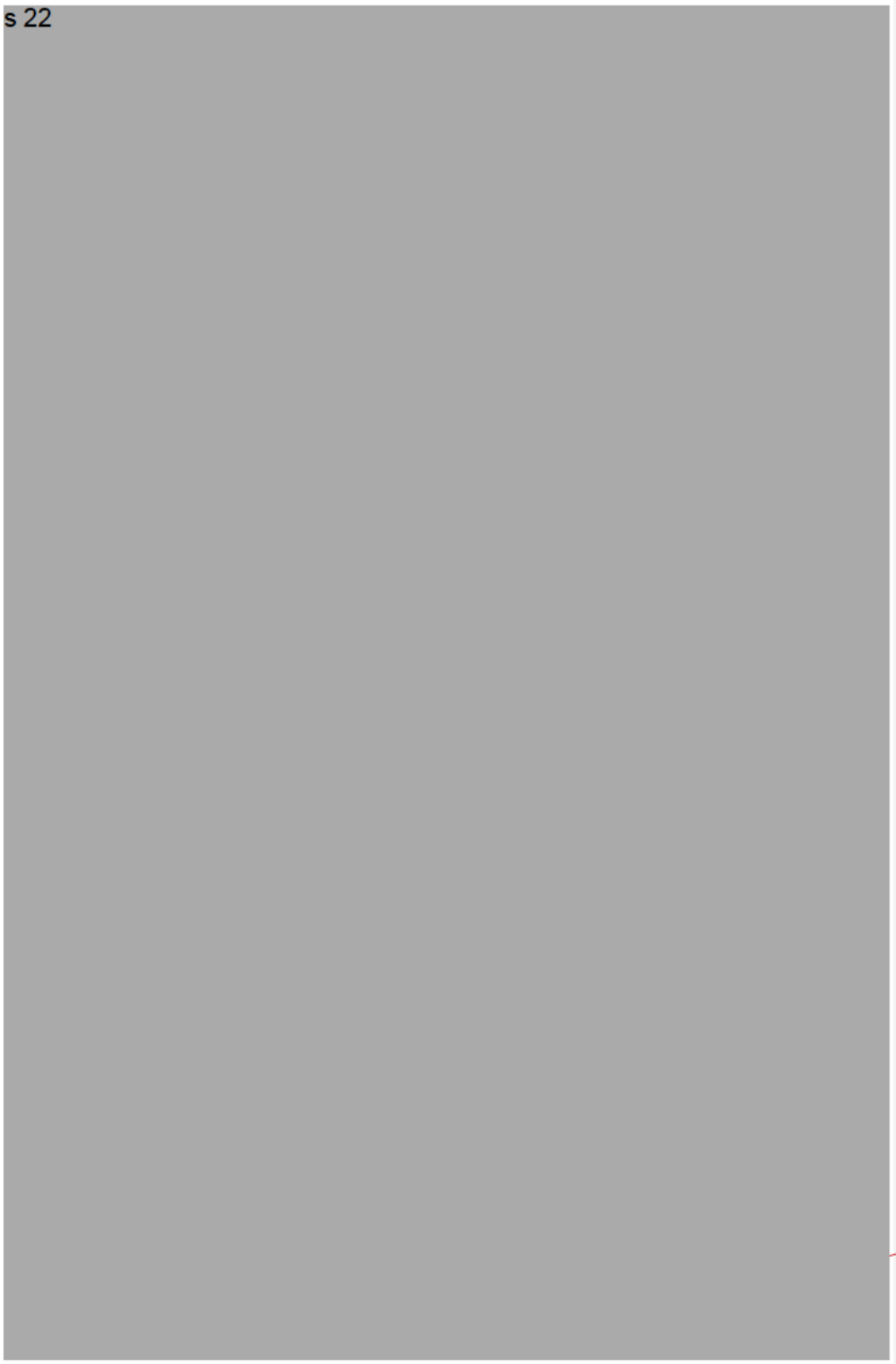
s 22



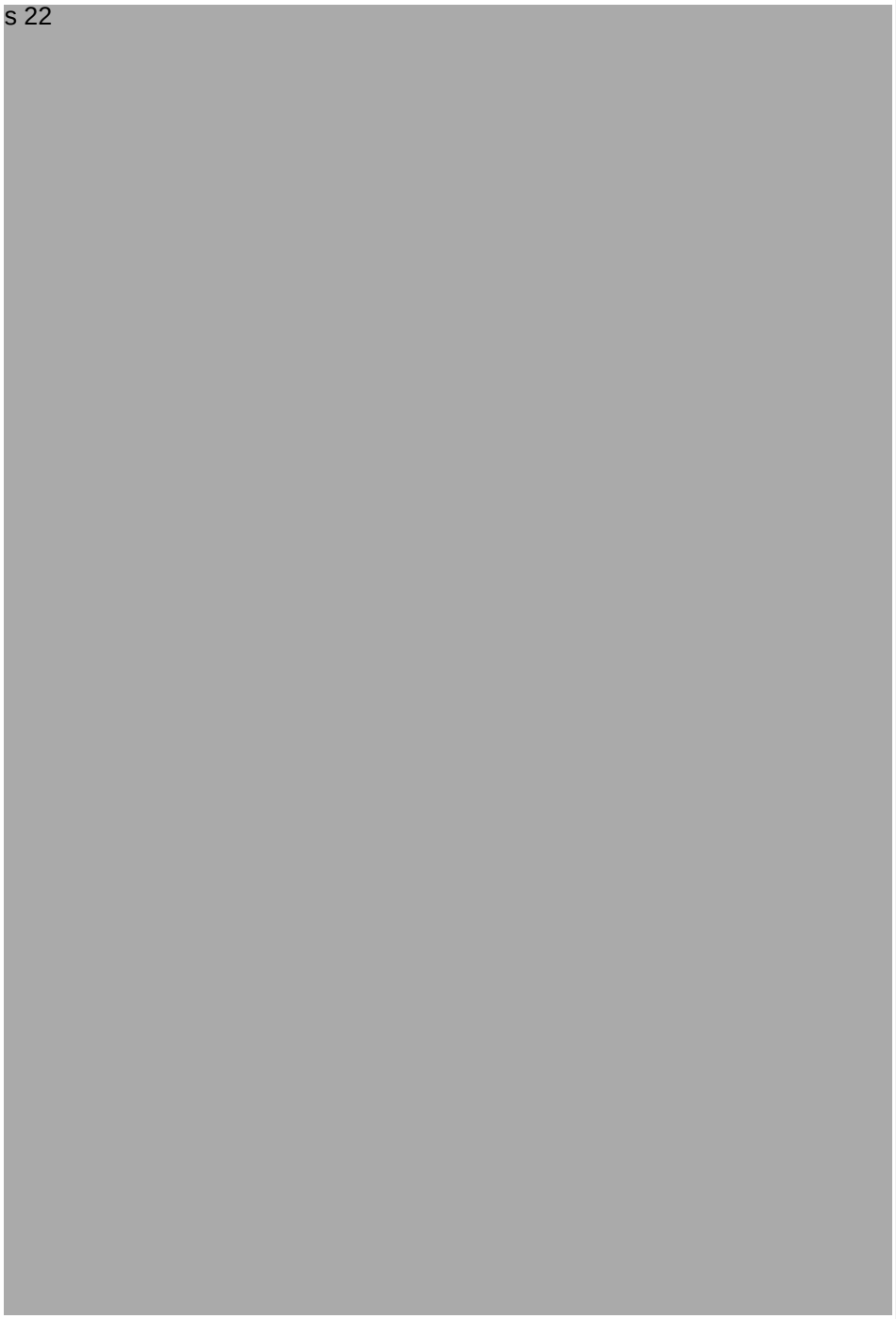
Formatted Table

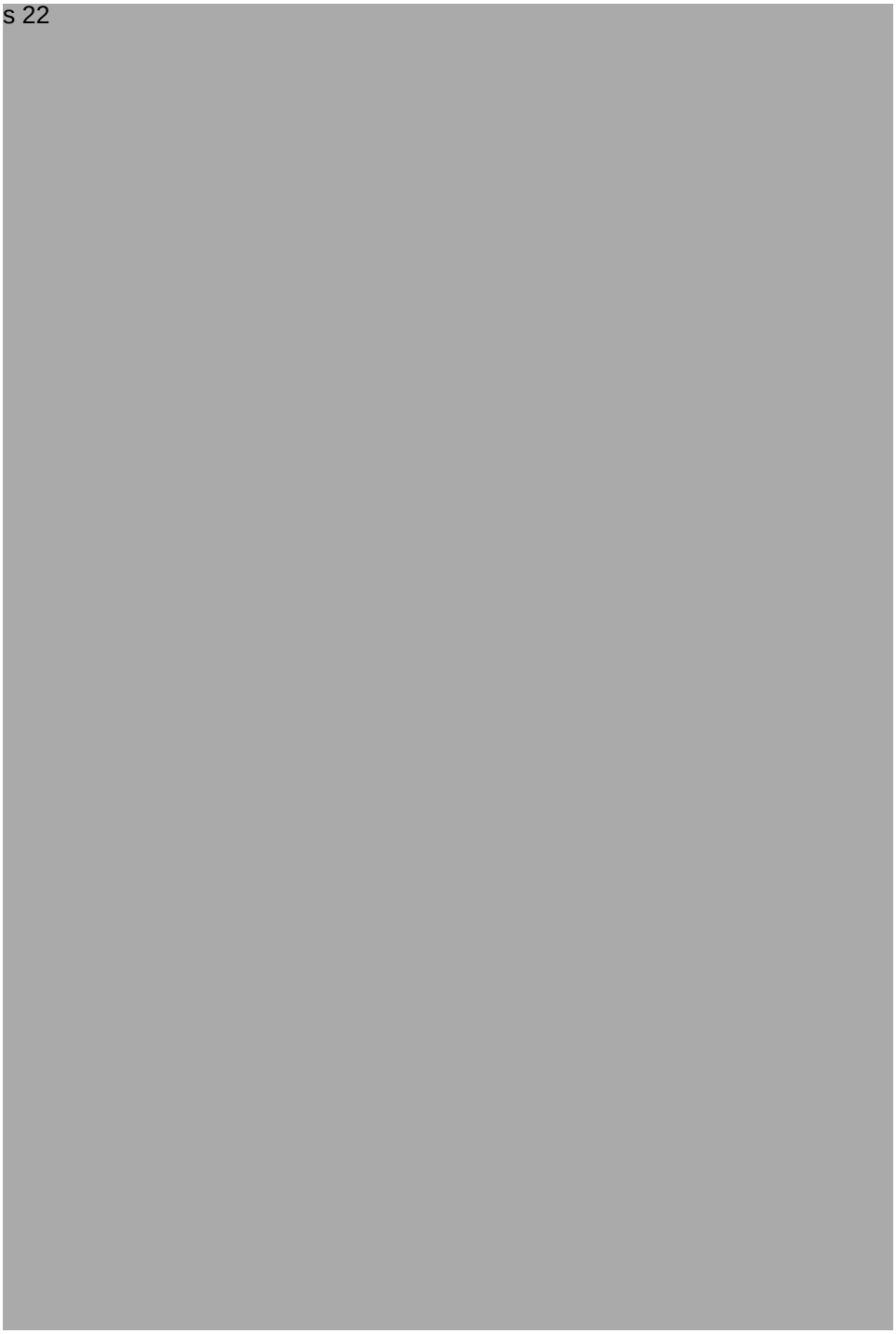

s 22

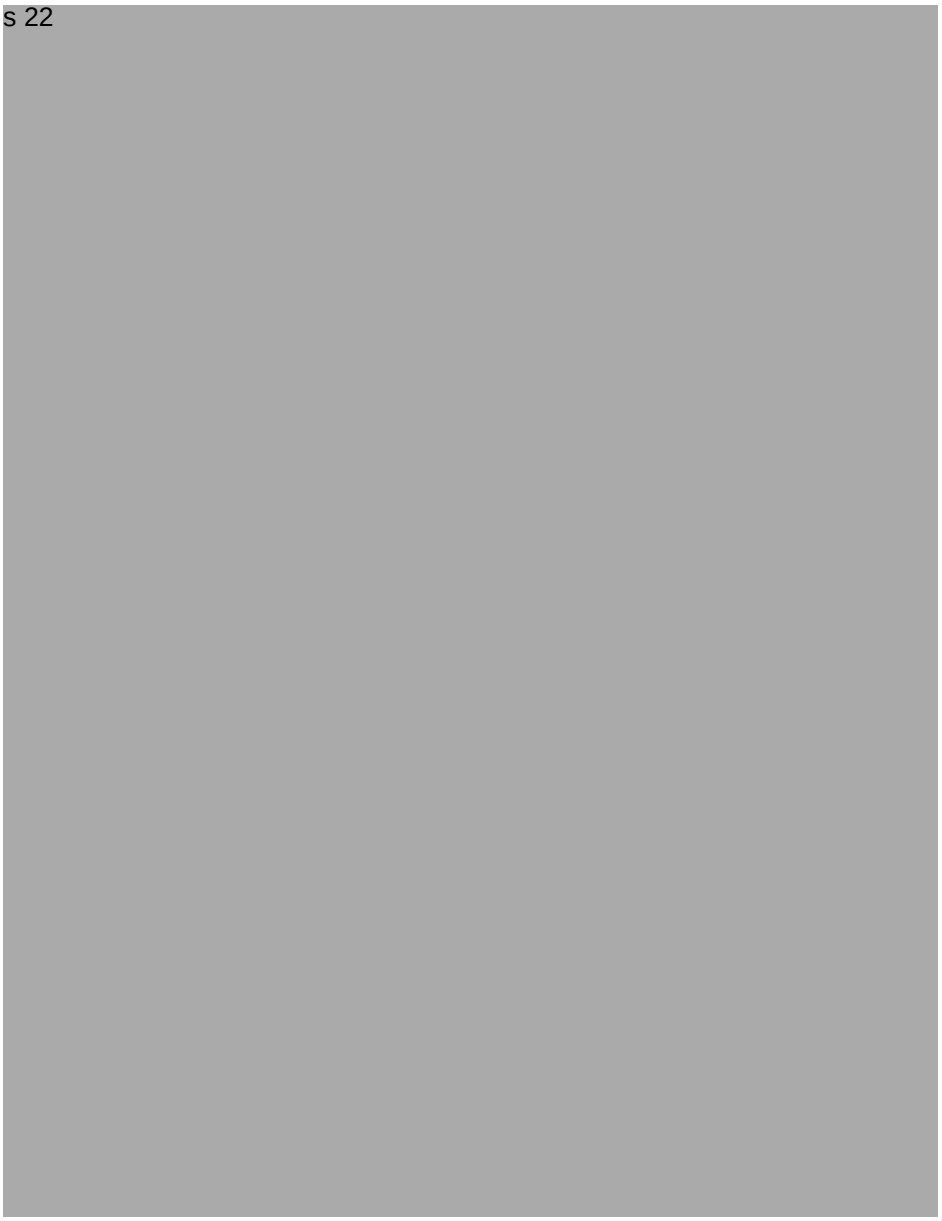

s 22


s 22


s 22


s 22


s 22


s 22

Complaint

CYR-0485528

Created On
28/01/2024 6:44 PM

Complaint Status
Closed

Assigned To
s 22

Response Required
No

Summary

Date Received
28/01/2024 6:44 PM

Date Finalised
29/01/2024

Complaint Status
Closed

Assigned To
s 22

Source Type
CYR

Derived
No

Complainant

Anonymous Complaint
Yes

Australian Resident
Yes

Police Involved
No

Education Provider
No

Complainant

Repeat Complainant
No

Complainant Reason
Violent extremist material

Response Required
No

Response Method

Submission

Method Received
Web Form

Presentation Type
Social media

Device Type
PC/Laptop

Device Name

Remote IP Address

Access Information

Additional access instructions

s 47E(d)

Access Username

Access Password

Related Entities

URL	Type	Service Type	Service Category	P2P Filename	Newsgroup Name	Content Path
https://twitter.com/s 47E(d)	Complained	SocialMedia Service (SMS)				
https://twitter.com/	Complained	SocialMedia Service (SMS)				

1 - 2 of 2 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced
	INV-2024-02849	Discretion not to investigate	https://twitter.com/s 47E(d)	No	Violentextremistmaterial	NA	29/01/2024
	INV-2024-02850	Discretion not to investigate	https://twitter.com/	No	Violentextremist material	NA	29/01/2024

1 - 2 of 2 (0 selected) Page 1

S...	Investigation Id	Status	Investigated Locator	Australian Conn...	Complainant Reason	Classification	Commenced

No Investigation records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Response

Response/RFI Templates

Name	Created On	Status

No Complaint Response records are available in this view.

0 - 0 of 0 (0 selected) Page 1

Email Activities

Date Sent/Received	From	To	Subject	Status Reason
28/01/2024 6:44 PM	Cyber Report	eSafetyCyber Report DL	Notification - Violent extremist material [SEC=OFFICIALS...	Sent

1 - 1 of 1 (1 selected) Page 1

Basic Online Safety Expectations

Regulatory Guidance

July 2024

Australian Government | eSafetyCommissioner

eSafety.gov.au

Contents

Overview of this guidance

This guidance is for online service providers (**providers**) and other stakeholders who require information about the Basic Online Safety Expectations, also known as 'the Expectations' and the functions of the eSafety Commissioner (**eSafety**) in assessing compliance with those Expectations.

The Expectations are determined under the Online Safety Act 2021 (**the Act**) and set out the Australian Government's expectations of the steps that should be taken by providers of social media services, messaging services, gaming services, file sharing services, apps and certain other sites accessible from Australia to keep Australians safe online. While compliance with the Expectations is not mandatory, eSafety has powers under the Act to obtain information from providers as to the steps they are taking to comply with the Expectations. eSafety can also publish statements about whether providers have or have not complied with the Expectations. The aim is to increase the transparency and accountability of providers, thereby helping to incentivise and improve safety standards.

More information on the Basic Online Safety Expectations is available on eSafety's website.[1]

This guidance provides information on:

- the legal framework for the Expectations (**Part 1**)
- eSafety's approach to the use of reporting powers (**Part 2**)
- eSafety's approach to assessing compliance with the Expectations (**Part 3**)
- examples of reasonable steps that can be taken by providers to ensure compliance with the Expectations (**Part 4**)

This guidance updates previous eSafety regulatory guidance published in September 2023 and July 2022.

Document history

Version	Date	Descriptions
01	**July 2022**	
02	**September 2023**	Guidance updated, including addition of Part 4 – examples of reasonable steps that can be taken to comply with the Expectations
03 (current)	**July 2024**	Guidance updated to reflect amendments as a result of the Online Safety (Basic Online Safety Expectations) Amendment Determination 2024 ('2024 Amendment Determination').

[1] eSafety website, Basic Online Safety Expectations | eSafety Commissioner.

Part 1: The legal framework for the Expectations

Overview of the Expectations

The Act provides for the Minister for Communications to set online safety expectations through a legislative instrument called a determination. The Online Safety (Basic Online Safety Expectations) Determination 2022[2] (the Determination) was registered on 23 January 2022. An Explanatory Statement to the Determination was also published.

On 30 May 2024 the Minister for Communications amended the Determination to address changing online safety challenges and strengthen the Expectations.[3]

The Expectations include a range of foundational steps that providers are expected to take to ensure safety for their end-users and Australians more broadly, including:

- ensuring all end-users can use online services in a safe manner
- that the best interests of the child is a primary consideration in the design and operation of services likely to be used by children
- ensuring safe use of certain features of a service, such as encrypted services, anonymous accounts, generative artificial intelligence (AI) and recommender systems
- minimising provision of unlawful and harmful material and activity
- enabling end-users to make reports and complaints about unlawful and harmful material and activity and reviewing and responding to these reports
- having terms of use, policies and procedures to ensure safe use, and enforcing these terms.

eSafety has a number of relevant powers under the Act.

- The power to require providers to report on how they are meeting any or all of the Expectations, either on a non-periodic or a periodic basis through a reporting notice or determination. The obligation to respond to a reporting notice or determination is enforceable and backed by civil penalties.
- The power to publish summaries of information, including from reporting notices or determinations.
- The power to publish statements regarding providers' compliance and non-compliance with the Expectations.

[2] Online Safety (Basic Online Safety Expectations) Determination 2022: Online Safety (Basic Online Safety Expectations) Determination 2022 (legislation.gov.au).
[3] Online Safety (Basic Online Safety Expectations) Amendment Determination 2024: Federal Register of Legislation - Online Safety (Basic Online Safety Expectations) Amendment Determination 2024.

Who do the Expectations apply to?

The Expectations apply to social media services, relevant electronic services and designated internet services that can impact the online safety of Australians.

Table 1: The Expectations apply to three main sections of the online industry

Section of the online industry	Scope
Social media services	Social media services include, but are not limited to: • social networks • media sharing networks • discussion forums • consumer review networks.
Relevant electronic services	Relevant electronic services include, but are not limited to: • email services • instant messaging services • SMS and MMS services • chat services • online games where end-users can play with or against each other • online dating services.
Designated internet services	Designated internet services include but are not limited to: • websites and file/photo storage services and some services which deploy or distribute generative AI models (unless a service is otherwise considered a social media service or a relevant electronic service).

Providers should also be aware of their obligations under other regulatory requirements under the Act including applicable industry codes and industry standards. Industry codes and industry standards place enforceable requirements on sections of the online industry including social media services, relevant electronic services and designated internet services, in relation to class 1 and class 2 material. More detail is set out on page 8.

What harms are covered by the Expectations?

The Expectations apply to all unlawful and harmful material and activity covered by the Act, as well as more broadly to address harms that impact on the online safety of Australians.

There are a wide range of potential harms that may arise on a service, impacting the online safety of Australians. It is expected that providers will have systems and processes in place to identify such harms, and take steps to ensure they are complying with the Expectations in relation to these harms.

What is 'unlawful' material and activity?

'Unlawful' material or activity is material or activity prohibited under law. For the purposes of the Determination, the term 'unlawful' refers to illegal material or activity dealt with under the Act and other unlawful material or activities that may have a negative impact on the online safety of Australians. Unlawful material and activity is therefore generally considered to also be harmful.

Examples of unlawful material and activity include:

- material that is illegal and has been refused classification under the *Classification (Publications, Films and Computer Games) Act 1995* including:
    - child sexual exploitation and abuse[4] (CSEA) material
    - material that advocates terrorism
    - material that depicts extreme crime and violence
    - material that incites or instructs or depicts, without justification, crime and violence or illicit drug use
      (known as class 1 material in the Act)
- grooming[5] of children
- the sharing of, or threatening to share, a non-consensual intimate image[6], including sexual extortion[7] (also known as sextortion).

[4] Child sexual exploitation and abuse (CSEA) can include both material and activity (for example, grooming). CSEA material is a broad category of material, normally referring to images and videos depicting the sexual abuse of a child, including sexual assault (child sexual abuse material or 'CSAM'), as well as content that sexualises and is exploitative of a child, but that does not necessarily show the child's sexual abuse (child sexual exploitation material or 'CSEM').

[5] Predatory conduct to prepare a child or young person for sexual activity at a later time.

[6] A non-consensual intimate image includes a still visual image or moving visual images. See section 15 of the Act.

[7] Sexual extortion, also known as sextortion, is a crime involving online blackmail, where victims are tricked into sending intimate images of themselves to someone who then threatens to share the images unless demands are met, usually for payment. Sextortion is currently an online child sexual exploitation trend, targeting teenage males in particular.

What is 'harmful' material and activity?

'Harmful' material or activity is material or activity that may not be unlawful but is covered within the scope of the Act. It is also material or activity that should fall under a provider's terms of use, policies and procedures and standards of conduct for end-users (as outlined in Section 14 of the Determination).

Some material or activity will be both unlawful and harmful, such as class 1 material, non-consensual intimate images and material depicting abhorrent violent conduct.

The Expectations specifically highlight the importance of minimising the extent to which the following material is available on a provider's service:

a. cyberbullying material targeted at an Australian child

b. adult cyber abuse material

c. a non-consensual intimate image of a person

d. class 1 material

e. material promoting, inciting, instructing in, or depicting abhorrent violent conduct.

Class 2 material is material that would be harmful for a child to see.[8] It is defined in the Act and is material[9] that is, or would likely be, classified as either:

- X18+ (or, in the case of publications, category 2 restricted),[10] or
- R18+ (or, in the case of publications, category 1 restricted)[11]

under the National Classification Scheme, because it is considered inappropriate for general public access and/or for children and young people under 18 years old.

The Expectations specifically require providers to take reasonable steps to prevent access by children to class 2 material.

Additional information on the classification of material under the National Classification Scheme is available in the Online Content Scheme Regulatory Guidance on eSafety's website.

[8] X18+, R18+ classifications require that the material be unsuitable for a child to see. In the case of Category 2 and Category 1 classification (which relate to publications), the material is either unsuitable for a child to see or read, or contains particular depictions likely to cause offence to a reasonable adult. More information on the approach to classifications can be found in the National Classification Code: National Classification Code (May 2005) (legislation.gov.au).

[9] Section 107 of the Act. This material includes films, publications, computer games and any other material that is not a film, publication or computer game.

[10] Section 107(1)(a) - (e) of the Act.

[11] Section 107(1)(f) - (l) of the Act.

The Explanatory Statement to the Determination provides further examples of harmful material.

- Hate against a person or group of people on the basis of race, ethnicity, disability, religious affiliation, caste, sexual orientation, sex, gender identity, disease, immigrant status, asylum seeker or refugee status, or age.[12]
- Promotion of suicide and self-harm, such as pro-anorexia content, that does not meet the threshold of class 1 or class 2 material.
- High volume, cross-platform attacks that have a cumulative effect that is damaging but does not meet the threshold of adult cyber-abuse when reported as singular comments or posts.
- Promotion of dangerous viral activities that have the potential to result in real injury or death.

eSafety's approach to exercising its powers in relation to the Expectations

eSafety will continue to focus on a number of objectives when exercising its powers in relation to the Expectations.

- Enhancing providers' transparency and accountability, and improving insights into the effectiveness and impact of what providers are doing to keep end-users safe online.
- Tracking harms, safety interventions and deficiencies, and technology over time through use of periodic reporting notices, and improving understanding of where gaps and challenges exist.
- Incentivising proactive and systemic safety interventions, including through publishing appropriate information and using statements of compliance or non-compliance with the Expectations to highlight good practice, as well as areas where insufficient action is being taken.

eSafety expects that providers regularly review their policies, procedures and practices to ensure compliance with the Expectations and that they put in place additional measures where a service is not compliant.

[12] See definition of 'hate speech' in the Explanatory Statement to the Online Safety (Basic Online Safety Expectations) Amendment Determination 2024: Federal Register of Legislation - Online Safety (Basic Online Safety Expectations) Amendment Determination 2024.

What are the reasonable steps a provider should take to comply with the Expectations?

The Determination does not prescribe how the Expectations must be met by providers but gives examples of reasonable steps that a provider may choose to take. This provides flexibility in the way providers can meet the Expectations. However, a provider's approach should be informed by examples provided in the Determination and this guidance, and advice from eSafety.

Part 4 of this document sets out more detailed guidance for providers on steps that could be taken to comply with the Expectations but does not prescribe specific steps or the use of particular technology. This guidance also sets out where certain harms or safety issues are likely to require a more rigorous or particular response to meet the relevant Expectation.

Providers are expected to have regard to this guidance, as set out in section 7 of the Determination.

Further detail on the reasonable steps is also included in the Explanatory Statement to the Determination and the Explanatory Statement to the 2024 Amendment Determination.

Providers must also comply with any other relevant legal obligations when implementing the Expectations, such as the *Privacy Act 1988 (*Cth*)*.

Interaction with industry codes and industry standards

What are industry codes and industry standards?

The industry codes and industry standards are mandatory requirements that apply to particular sections of the online industry. Industry codes are developed by industry associations that represent those sections of the online industry, and industry standards are determined by the eSafety Commissioner.[13]

There are six industry codes in effect which focus on class 1A and 1B material:

- Social Media Services Online Safety Code (Class 1A and Class 1B Material) (referred to throughout this Guidance as the 'SMS Code')
- App Distribution Services Online Safety Code (Class 1A and Class 1B Material)
- Hosting Services Online Safety Code (Class 1A and Class 1B Material)

[13] See eSafety's register of industry codes and industry standards: Register of industry codes and industry standards for online safety | eSafety Commissioner.

- Internet Carriage Services Online Safety Code (Class 1A and Class 1B Material)
- Equipment Online Safety Code (Class 1A and Class 1B Material)
- Internet Search Engine Services Online Safety Code (Class 1A and Class 1B Material).

Two industry standards which focus on class 1A and 1B material have been registered and will come into effect in December 2024:

- Online Safety (Relevant Electronic Services—Class 1A and Class 1B Material) Industry Standard 2024 (referred to throughout this Guidance as the 'RES Standard')
- Online Safety (Designated Internet Services—Class 1A and Class 1B Material) Industry Standard 2024 (referred to throughout this Guidance as the 'DIS Standard').

The process for developing codes to address class 1C and class 2 material is underway. Unlike the Expectations, the industry codes and industry standards include mandatory minimum compliance measures that are enforceable through various means.

What do industry codes and industry standards address?

The Act provides for the introduction of industry codes and/or industry standards to address class 1 and class 2 material. The class 1 material covered under the first phase of industry codes and industry standards focuses on the most harmful forms of online material referred to as 'class 1A' and 'class 1B' material. These are:

- class 1A material, such as child sexual exploitation material[14] (including child sexual abuse material[15]) and pro-terror material[16]
- class 1B material, such as crime and violence material and drug related material.

The registered industry codes and industry standards represent the mandatory and enforceable measures that industry must meet in order to comply with their legally binding obligations in relation to class 1A and 1B material.

[14] For the purposes of industry codes and standards, CSEM is a sub-category of class 1 material that is broader than CSAM, and includes material relating to the promotion or provision of instruction in paedophile activity, includes or contains descriptions or depictions of child sexual abuse or any other exploitative or offensive descriptions or depictions involving a person who is, or appears to be, a child under 18, or describes or depicts in a way that is likely to cause offence to a reasonable adult, a person who is or appears to be a child under 18 (whether or not the person is engaged in sexual activity).
[15] For the purposes of industry codes and standards, CSAM is a sub-category of class 1 material to the extent that it is comprised of visual depictions of child sexual abuse.
[16] For the purposes of industry codes and standards, pro-terror material is class 1 material that advocates the doing of a terrorist act.

Relationship between industry codes, industry standards and the Expectations

The obligations in industry codes and industry standards will be narrower in scope than the Expectations as they focus on specific categories of class 1 material (and class 2 material in the future) rather than the broader unlawful and harmful material and activity covered by the Expectations.

In some cases, specific mandatory steps to address class 1 material required under an industry code or an industry standard will be directly relevant to an Expectation, including requirements under an industry code or industry standard to:

- undertake risk assessments and ensure safety by design (section 6 of the Determination)
- minimise the provision of certain material, including class 1 material (sections 6 and 11(d) of the Determination)
- incorporate safety measures in relation to generative AI capabilities (section 8A of the Determination)
- provide reporting and complaint mechanisms for end-users and review and respond to reports and complaints (sections 13, 14, 15 and 16 of the Determination)
- ensure the implementation and enforcement of terms of use, policies and procedures that address class 1 material (sections 14, 15, 17 and 18 of the Determination).

Compliance with the requirements in an industry code or industry standard is relevant to a provider's implementation of certain expectations (in relation to class 1 material, and class 2 material in the future) but will not be determinative of meeting any particular Expectation.

This is because what is 'reasonable' for a provider to do to address unlawful and harmful material under the Expectations may extend beyond the **minimum** requirement in the mandatory (and enforceable) industry code or industry standard. Additional steps may be required to meet the applicable Expectations. Additionally, the Expectations apply to a broader range of harmful material (beyond class 1 material), and to harmful activities.

eSafety will have regard to where compliance with an industry code or standard supports compliance with the Expectations and will assess each service on a case-by-case basis.

Table 2: Differences between industry codes and industry standards, and the Expectations

	Applies to	Applies to unlawful and harmful 'material'	Applies to unlawful and harmful 'activity'	Consequences for failure to comply
Basic Online Safety Expectations	• Social media services • Relevant electronic services • Designated internet services	Yes	Yes	eSafety may prepare, and publish, a Statement of Non-Compliance with one or more Expectations. eSafety has a range of enforcement options in relation to ensuring compliance with a reporting notice or determination.
Industry Codes (Phase 1)	• Social media services • App distribution services • Hosting services • Internet carriage services • Manufacturers, maintenance and installation providers of equipment • Search engine services	Applies only to certain categories of class 1 material	Applies to certain activities that affect the provision of certain categories of class 1 material	eSafety may issue a formal warning, or written direction to comply with an industry code. Failure to comply with a direction may result in enforcement through an enforceable undertaking or injunction. It may also result in an infringement notice or civil penalty proceedings.
Industry Standards (Phase 1)	• Relevant electronic services • Designated internet services	Applies only to certain categories of class 1 material	Applies to certain activities that affect the provision of certain categories of class 1 material	Failure to comply with an industry standard may result in a formal warning, enforcement through an enforceable undertaking or injunction. It may also result in an infringement notice or civil penalty proceedings.

Interaction with other regulatory requirements in the Act

Failure to comply with an expectation under the Determination may result in other enforcement action by eSafety. For example, eSafety has the power to give providers a removal notice in relation to specific material under the four complaints-based reporting schemes.[17] These powers may need to be exercised more frequently if a provider has failed to take reasonable steps to minimise the provision of certain material on their service (section 11 of the Determination). Failure to comply with a removal notice is a civil penalty provision and may result in a range of enforcement actions by eSafety. eSafety does not need to establish that a provider failed to comply with section 11 of the Determination (or an industry code or industry standard) prior to giving a removal notice.

Additional information about eSafety's regulatory schemes and powers is available on eSafety's website.

[17] eSafety can investigate reports of cyber-bullying of children, adult cyber abuse, image-based abuse (sharing, or threatening to share, intimate images without the consent of the person shown) and illegal and restricted content. More information on these schemes is available on the eSafety website: Report online harm | eSafety Commissioner.

Part 2: Reporting powers

Reporting and information gathering powers

A core element of the Act is to empower eSafety to seek information from providers on their compliance with the Expectations. This information is sought to improve transparency and accountability, and to assist eSafety to determine whether a provider is compliant with the Expectations.

There are three prescribed ways eSafety can seek information from providers regarding compliance with the Expectations.

1. **Requests for information**

   As part of the Expectations (section 20 of the Determination), eSafety may request information about:

   - the number of complaints about breaches of a provider's terms of use
   - the time frame for responding to removal notices given to the provider by eSafety
   - measures taken to make sure people can use the service in a safe manner
   - the performance of online safety measures that providers have announced publicly or reported to eSafety
   - the number of active end-users of the service in Australia.

   A failure to respond within 30 days is non-compliance with the Expectations. This gives the Commissioner discretion to prepare a statement that the provider is not complying with the Expectations. Providers should ensure they have processes in place to respond to these information requests. For more information on section 20, see Part 4 of this guidance.

2. **Reporting notices**

   eSafety may give a reporting notice to a provider requiring them to produce a report on their implementation in relation to one or more expectations. These notices are enforceable, backed by the power to seek civil penalties and other enforcement mechanisms.

   Reporting notices are specific to the provider, and can require:

   - non-periodic reporting
   - periodic reporting at regular intervals of between 6 and 24 months for as long as the Commissioner deems appropriate.

3. **Reporting determinations**
   eSafety can make a reporting determination – a legislative instrument – requiring periodic or non-periodic reporting for a specified class of services. Like the reporting notices, these are enforceable and backed by civil penalties and other enforcement mechanisms.

Type of Information Gathering	Can require reporting on	Periodic or non-periodic	Reporting period	Time to respond	Enforceable
Requests for information section 20 of the Expectations	**1.** Terms of service complaints; **2.** The timeframe for responding to removal notices **3.** Measures taken to make sure people can use the service in a safe manner **4.** The performance of online safety measures that providers have announced publicly or reported to the the Commissioner **5.** The number of active end-users of the service in Australia (disaggregated into active end-users who are children and those who are adult end-users)	Non-periodic	Not shorter than 6 months for reporting categories 1 and 2. N/A for reporting categories 3, 4 and 5	Within 30 days	✗
Reporting notices to **individual providers**	Implementation of any part or the entirety of the Expectations	Either periodic or Non-periodic	6 to 24 months	28 days or longer as specified	✓
Reporting determinations to a **specified class of providers**	Implementation of any part or the entirety of the Expectations	Either periodic or Non-periodic	6 to 24 months	28 days or longer as specified	✓

eSafety's approach to the use of reporting and information gathering powers

eSafety is taking a phased approach in exercising its powers related to the Expectations, starting with the use of non-periodic reporting notices with a focus on specific expectations and acute issues of particularly high harm, such as CSEA. eSafety intends to expand the use of its statutory powers related to the Expectations over time, with the first periodic reporting notices intended to be given in 2024.

eSafety is committed to a number of principles.

- Applying eSafety's powers under Part Four of the Act in a fair and proportionate way, based on evidence and insights.
- Taking an open and transparent approach – both in exercising eSafety's powers, and in terms of the information obtained through notices. eSafety intends to make information obtained through use of information requests made under section 20 of the Determination, reporting notices and determinations publicly available where appropriate in the interests of transparency and accountability.
- Recognising the importance of reducing regulatory requirements by considering information that:
    - providers already publish voluntarily
    - is provided as part of international transparency initiatives
    - is provided to eSafety under another regulatory scheme, including reporting obligations through an industry code or industry standard.
- Recognising that differences between providers in terms of resources, risk, technical architecture and user base, means that 'one size does not fit all'.
- Taking a consultative approach, seeking input and feedback from providers as well as from civil society organisations, academics and other experts to ensure implementation meets standards of good regulatory practice.
- Ensuring eSafety systems securely store information, including information which is commercial-in-confidence, personal information, or information, which if disclosed, would adversely affect public safety.

Complying with a notice, determination, or request for information from eSafety

Reporting notices may require information such as:

- qualitative information on safety tools, processes and policies, and why these are reasonable steps to implement the Expectations – these may be phrased as yes/no questions, multiple choice questions or worded to seek descriptive information.
- quantitative information on the operation of safety tools, processes and policies – this may consist of metrics to determine the impact of interventions or information about the resources allocated.

Reporting notices will be related to specific expectations. Responses will be used to understand the extent to which a provider is compliant with one or more expectations as well as increasing transparency through building an understanding across different providers of common practices, trends and challenges. Given the breadth of some of the expectations, eSafety is likely to ask questions targeted at assessing how the provider's compliance with a particular expectation minimises specific types of harms. Targeted questions assist providers and eSafety by ensuring the provision of meaningful information. It also minimises the regulatory burden on providers and encourages transparency and accountability about issues that impact on the online safety of Australians.

Providers are required under the Act to respond to a reporting notice in the manner and form specified and to the extent that they are capable of doing so.[18] eSafety provides a response template as part of a notice and providers must respond to all questions in the manner and form specified in that template.

- For example, if a question requires a yes/no answer, a provider must respond accordingly. If a question requires a response of 'all' examples or 'all relevant indicators/steps/tools' (or similar), providers must provide all relevant information.
- Providers must respond truthfully and accurately to each question.

Providers should engage with eSafety if they cannot answer in the form specified. Providers are required to respond within the timeframe specified. In line with the Act, the time to respond will be no shorter than 28 days from the giving of a notice, or from the end of the reporting period specified in the notice. eSafety will consider the appropriate length of time for a provider to respond to a notice on a case-by-case basis.

eSafety understands that not every expectation will apply equally to every service. If a provider is of the view that a particular expectation or question does not apply, they must contact eSafety **before** providing their response to the notice. eSafety will be available throughout the

[18] Sections 49(2)(b), 50, 56(2)(b) and 57 of the Act.

process to answer questions and provide clarification.

Where a provider does not collect and is not capable of obtaining the required information, they should provide alternative relevant data. However, a provider must engage with eSafety to confirm whether an alternative response is acceptable.

Providers are required to provide information in response to a reporting notice even if that information is considered commercial-in-confidence or covered by a confidentiality obligation in a third-party contract. As set out on page 20 providers will be asked to clearly identify any information they believe should not be published.

eSafety will also endeavour to inform a provider of the intention to give a reporting notice, and the intended scope of the proposed notice, before it is given to them. The purpose of this is to enable the provider to identify any specific barriers to compliance within the proposed time frame of the notice and to confirm the appropriate entity for receipt of the notice. However, advance notice may not be possible in every circumstance. For example, this might not be appropriate where a provider has not previously engaged in a constructive or reasonable manner with eSafety or where there are factors leading to a degree of urgency.

If a provider does not respond to a notice or comply with its requirements, eSafety has civil enforcement powers[19] and the power to issue a formal warning,[20] or prepare and publish a statement that the provider is non-compliant (referred to as a Service Provider Notification in the Act).[21]

In addition to the information provided in response to a specific question in a notice, providers can share additional information and context with eSafety as part of their response to the notice.

In the interests of consistency, enforceability and transparency, where eSafety has decided that a notice is the appropriate mechanism, eSafety will not normally agree to withhold a formal notice and agree to the same information being provided voluntarily.

[19] The maximum penalty for non-compliance with a reporting notice under sections 50 and 57 of the Act is 500 penalty units for an individual and can be multiplied by 5 for a body corporate (at the date of publication of this guidance, a single penalty unit is $313). In cases of non-compliance, eSafety may give an infringement notice, initiate civil penalty proceedings, apply for an injunction or enter into an enforceable undertaking under the *Regulatory Powers (Standard Provisions) Act 2014*.

[20] Sections 51 and 58 of the Act.

[21] Sections 55 and 62 of the Act.

Periodic notices

Under subsection 49(2) of the Act, eSafety can give periodic notices requiring providers to report on their compliance with the Expectations at regular intervals.

eSafety intends to give periodic reporting notices to providers in order to track the development and improvement of tools, processes, and their effectiveness. Periodic reporting notices may also focus on specific harms and issues that have already been identified through eSafety's use of non-periodic reporting notices, or a range of other issues. eSafety intends to give periodic notices to certain online service providers in 2024.

The Act requires that the reporting interval of periodic notices is no shorter than six months, and no longer than 24 months. eSafety intends to first use periodic notices to require certain online service providers to prepare four reports over a 24-month report period, each covering a reporting interval of six months.

The following table shows an example of the reporting intervals for periodic notices.

Report number	Example reporting interval	Example report due date
Report one	1 January 2024 to 1 July 2024	1 August 2024
Report two	2 July 2024 to 2 January 2025	2 February 2025
Report three	3 January 2025 to 3 July 2025	3 August 2025
Report four	4 June 2025 to 4 January 2026	4 February 2026

How does eSafety decide which providers receive notices?

When deciding which providers to give a notice to, the Act requires eSafety to have regard to these specified criteria:[22]

- the number of complaints eSafety has received under the Act in relation to the service in the previous 12 months
- any deficiencies in the provider's safety practices and/or terms of use

[22] Section 56(5) of the Act.

- any previous contraventions of civil penalty provisions relating to the Expectations
- whether the provider has agreed to give the Secretary of the Department regular reports relating to safe use of their service[23]
- any other matters the Commissioner considers relevant.

Examples of other matters that the Commissioner might consider relevant may include:

- evidence from eSafety's other regulatory schemes, such as types of complaints, a service's responsiveness to removal requests or notices, or other investigative insights regarding a service's safety issues
- a service's reach and the profile of its end-users, including whether the service is used by children
- higher risk design choices and features, such as livestreaming and end-to-end encryption (E2EE)
- the measures the service currently has in place to protect end-users from harm
- evidence of systemic harm, or evidence of key safety issues, including from victims, civil society organisations, media, academics, or other experts
- the information already published by a provider, as well as any lack of information regarding a service's safety policies, processes and tools, or limited information about the impact or effectiveness of these interventions.

The same requirements do not exist if eSafety makes a determination requiring reporting from a specified class of services. However, eSafety intends to take a similar approach to understanding risk and priority sectors prior to making any determination.

Reporting on compliance with the Expectations and industry codes and standards

Certain providers will be required to provide reports to eSafety under an industry code or industry standard, either as a matter of course or at the request of eSafety, depending on the application of the particular code or standard.

eSafety will seek to reduce regulatory burden in reporting requirements where possible and where appropriate. For example, where a provider has reported information in response to a notice related to the Expectations, they may refer to this information – insofar as it is relevant – for the purposes of preparing a report under an industry code or industry standard.[24]

[23] This provision was included to ensure that eSafety takes into account other Australian Government reporting initiatives, and considers the burden on providers from any duplication.

[24] See clause 7.3(3) of the Consolidated Industry Codes of Practice for the Online Industry (Class 1A and Class 1B Material) Head Terms.

Information obtained through a reporting notice given in connection with the Expectations, may be considered by eSafety in assessing a provider's compliance with an industry code or industry standard.

Is information received via reporting notices and determinations published?

The Explanatory Memorandum to the Act highlights the objective of the Expectations to 'improve the transparency and accountability of online service providers for the safety of their users and the mitigation of online harms'. It further notes that:

> **The transparency reporting obligation within the BOSE [Basic Online Safety Expectations] proposal would create greater transparency of the online safety practices for both government and the community, and encourage uplift through imposing reputational costs for non-compliance.**

eSafety considers that the transparency and accountability objectives of the Act are most effectively met by making information received from industry in response to a reporting notice or information request under section 20 of the Determination public, where appropriate. This transparency promotes the online safety of Australians by increasing awareness of online safety issues and the way that services respond to online harms, and incentivises improvements in the safety measures taken by industry.

eSafety will also provide notice recipients with information to assist them in making clear submissions in relation to information that should not be published, and the criteria that eSafety will have regard to in deciding what information should not be published. An example of this is at Annex A. Notice recipients will be asked to:

- clearly identify in their response if any information is commercial-in-confidence or should otherwise not be published, for example, because it would adversely affect public safety
- provide clear reasons in support of any claim that certain information should not be published.

eSafety considers these claims carefully, and prepares a summary of the information that it considers is appropriate to publish. eSafety also considers whether there are steps that can be taken to protect such information while ensuring the transparency and accountability objectives of the Act are still met. eSafety's approach to information that could impact public safety will be informed by its own expertise, engagement with external experts, and other sources.

In line with the transparency objectives of the Act, eSafety may disclose the names of

providers given a notice at the time a reporting notice is given, along with a summary of the information sought in the reporting notice. The number and type of reporting notices given, and outcomes (such as whether a notice was complied with and whether any enforcement action was taken), will also be published in eSafety's annual report.

eSafety may also publish summaries of information obtained through section 20 information requests where appropriate, and may follow a similar process as set out above for notices.

Review rights

A provider may seek either internal review or external review by the Administrative Appeals Tribunal of certain actions taken by eSafety relating to the Expectations. The purpose of these review rights is to ensure that eSafety has made the correct and preferable decision on a case-by-case basis.

Action which can be reviewed	Who can seek review
The giving of a non-periodic reporting notice (Section 49 of the Act)	The provider named in the non-periodic reporting notice
The giving of a periodic reporting notice (Section 56 of the Act)	The provider named in the periodic reporting notice

An internal review may not always be appropriate, particularly if the reporting notice has been given by the Commissioner. Additional information about seeking a review can be found on eSafety's website.

Part 3: Assessing compliance with the Expectations

Statements of compliance or non-compliance

If eSafety decides that a provider is not complying with one or more of the Expectations, the Act empowers eSafety to prepare and publish a statement to that effect. eSafety may also publish a statement that confirms that a provider is meeting the Expectations. This supports transparency and encourages best practice. These are referred to as 'service provider notifications' in the Act.[25] eSafety uses the terms 'statements of compliance' and 'statements of non-compliance' to differentiate them from other kinds of service provider notifications in the Act.

[25] Section 48 of the Act.

If the decision is made to give a statement of compliance or non-compliance, eSafety will share this statement with the provider. If eSafety decides to publish the statement, the provider will be given the opportunity to make submissions including evidence to demonstrate that it is compliant with the relevant Expectation(s) or reasons that it should not be published.

eSafety's approach to assessing compliance

The Determination does not prescribe how the Expectations must be met, although it does contain examples of reasonable steps that could be taken within some sections of the Determination. The Determination affords flexibility to providers to determine the most appropriate method of complying with the Expectations, and eSafety supports this approach.

Additional examples of reasonable steps are provided in Part 4 of this guidance to assist providers in complying with each applicable expectation. Providers are expected to have regard to this guidance in ensuring they are compliant with each applicable Expectation.

How will eSafety decide whether to give and publish a statement of non-compliance?

eSafety will take a risk-based approach when assessing whether providers are taking reasonable steps to comply with the Expectations, taking into account the level of harm and extent of the safety issues relating to a service.

A statement of non-compliance can be given and published for a failure to comply with one or more expectations, although eSafety recognises that not all expectations will apply to all services. For example, if a service does not use encryption or permit anonymous accounts, then sections 8 or 9 may not apply. In some instances, where there is no appreciable risk of harm, it would also not be proportionate for eSafety to expect steps to be taken in relation to certain expectations.

The Commissioner will consider a number of factors[26] when assessing whether a provider of a service has complied with the Expectations or whether they have contravened an expectation, including the following:

- The risks related to the service, including:
    - the number of end-users, including Australian end-users
    - the user base and demographics of those end-users
    - risk and evidence of online harms

[26] This is not intended to be an exhaustive list of factors that the Commissioner may consider in assessing a provider's compliance with the Expectations.

    - design features that may increase risk or limit the effective use or operation of any safety measures
    - other relevant factors.

- The effectiveness and proportionality of the steps taken by a provider in meeting an expectation.
- Whether there are any particular technical or practical limits which might prevent a provider from taking certain steps to meet the Expectations.
- The resources available to the provider and the costs or other burden to implement certain steps.
- Substantiated information establishing that a provider has plans to take further action or other steps in the short to medium term.
- Whether the provider has engaged constructively with eSafety and responded to requests for information.
- How information provided in response to a notice compares with relevant evidence from other sources, such as eSafety's investigative insights, industry codes or industry standards reporting, as well as academic, civil society, or other expert evidence.

eSafety intends to publish statements of non-compliance on the eSafety website.

How will eSafety decide whether to give and publish a statement of compliance?

eSafety can only decide to give a statement of compliance if a provider has met all relevant expectations at all times during a specified period. This constitutes a higher bar than a statement of non-compliance which can be given for the failure to implement any individual expectation.

Similar to a statement of non-compliance, eSafety will take into account a number of factors when deciding whether a provider is complying with the Expectations, including the following:

- Evidence that a provider has implemented reasonable steps across all the relevant expectations, with evidence that these are operating effectively and consistently.
- Evidence that the reasonable steps have been taken and implemented for a reasonable time in order to evaluate their effectiveness.
- Whether the provider has engaged constructively with eSafety and responded positively to requests for information.
- How information provided by the service compares with evidence from other sources, such as investigative insights, academic, civil society or other expert evidence.

To support a decision that a provider has complied with all relevant expectations during a specified period, providers will need to demonstrate the effectiveness of their safety measures. Providers are encouraged to collect relevant information and metrics internally to evaluate the effectiveness of their safety interventions, and to provide these to eSafety – such as by responding to a non-periodic or periodic reporting notice.

eSafety intends to publish statements of compliance on the eSafety website.

Part 4: Examples of reasonable steps to comply with the Expectations

Overview

This part sets out examples of the reasonable steps that providers could take to comply with the Expectations.

The Determination does not prescribe how the Expectations must be met but includes non-exhaustive examples of reasonable steps. This guidance identifies further steps that eSafety considers would assist providers in complying with the Expectations. This is not an exhaustive list. eSafety recognises that each service is different and new technologies continue to emerge which may assist with complying with the Expectations. Providers may elect to take different steps to meet the Expectations that better suit their service and the risks posed. Providers should be prepared to report on these steps, why they are reasonable in light of the objectives of the Determination, and how these steps meet the relevant Expectations and keep Australians safe online.

As set out in the Explanatory Statement to the Determination, the Commissioner will take a risk-based approach towards assessing compliance, noting that what is 'reasonable' to comply with the Expectations may differ depending on the nature and severity of the harms and risks on a service.

Providers are expected to prioritise responding to the most harmful risks on their service, particularly where these involve unlawful material or activity, or where they impact on groups at higher risk. However, providers are also expected to take reasonable steps to address other harmful material and activity occurring, or likely to occur, on their service.

Unlawful and harmful material and activity may arise online as a result of human-generated content and conduct, but may also be generated artificially, and shared or otherwise misused in similar ways. The Act recognises this in relation to class 1 material (which includes material that describes or depicts a child under 18 or a person who 'appears to be' a child under 18 in relation to child sexual exploitation and abuse)[27] and in relation to image-based abuse (the non-consensual sharing of intimate images)[28] by including images that have been digitally or artificially generated.

The Expectations apply to material and activity that is unlawful and harmful, regardless of how it is generated. Providers should therefore take steps to address and mitigate the harms of the emerging technologies, including the ability to generate synthetic material, and where providers introduce or integrate features into their existing services which involve artificial intelligence (such as chatbots, among others). The Expectations also apply where services enable end-users to post synthetic material that was generated elsewhere.

For more information on eSafety's position on emerging technologies and trends, including Generative AI and how to take a safety-by-design approach to these issues, see eSafety's Position Statements.[29]

Reasonable steps

The Expectations, in some cases, require providers to take 'reasonable steps' to address various safety issues.

The term 'reasonable' is not defined in the Act or the Determination. It bears the ordinary meaning as being based upon or according to reason, and capable of sound explanation.

What steps are reasonable is a question of fact in each individual case and is an objective test that has regard to how a reasonable person, who is properly informed, would be expected to act in the circumstances. What is reasonable can be influenced by current standards and practices, the nature and extent of the harms involved that require mitigation, as well as by other legislative requirements or obligations that apply to each provider.

It is the responsibility of each provider to be able to justify why the steps they are taking are reasonable, and how these steps amount to compliance with the Expectations.

[27] As defined in section 106 of the Act.
[28] As defined in section 15 and 16 of the Act.
[29] eSafety website, Tech Trends and Challenges, Tech trends and challenges | eSafety Commissioner.

Consultation

Section 7 of the Determination sets out the expectation that providers will consult with the Commissioner in determining the reasonable steps to ensure safe use.

eSafety has engaged with industry on online safety issues and on the development of updated guidance. Providers are also encouraged to engage with eSafety regarding their specific services, as the reasonable steps are likely to differ depending on factors outlined above under 'reasonable steps', as well as a service's risks, business model, user base, technical architecture and design.

eSafety intends to update this guidance as needed in response to new harms, technologies and safety issues, or in response to other events.

The Determination sets out an expectation that providers will have regard to any relevant guidance material made available by the Commissioner (subsection 7(2)).

Chapter 1: Expectations regarding safe use

Division 2 of the Determination sets out expectations in relation to ensuring safe use of a service in the following sections.

- Section 6: take reasonable steps to ensure that end-users are able to use the service in a safe manner, including by taking reasonable steps to:
    - proactively minimise the extent to which material or activity on the service is unlawful or harmful
    - ensure that the best interests of the child are a primary consideration in the design and operation of any service that is likely to be accessed by children
    - make controls available that give end-users choice and autonomy to support safe online interactions.
- Section 7: consult with the Commissioner in determining what reasonable steps are for the purpose of section 6(1) and refer to the Commissioner's guidance in determining such reasonable steps to ensure safe use.
- Section 8: on an encrypted service, take reasonable steps to develop and implement processes to detect and address material and activity that is unlawful or harmful.
- Section 8A: take reasonable steps regarding safety of generative artificial intelligence capabilities.
- Section 8B: take reasonable steps regarding safety of recommender systems.
- Section 9: take reasonable steps to prevent anonymous accounts from being used to deal with material, or for activity, that is unlawful or harmful.

- Section 10: take reasonable steps to consult and cooperate with other service providers and ensure consultation and cooperation between the provider's services to promote the ability of end-users to use all those services in a safe manner.

Further guidance on steps that providers may take to ensure compliance with these expectations is set out in the following pages.

Section 6 of the Determination – Ensuring safe use and proactive minimisation of unlawful and harmful material and activity

Determination, section 6:

Core expectation

1. The provider of the service will take reasonable steps to ensure that end-users are able to use the service in a safe manner.

Additional expectation

2. The provider of the service will take reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful.

Additional expectation

2A. The provider of the service will take reasonable steps to ensure that the best interests of the child are a primary consideration in the design and operation of any service that is likely to be accessed by children.

Examples of reasonable steps that could be taken

3. Without limiting subsection (1),(2) or (2A), reasonable steps for the purposes of those subsections could include the following:

    a. developing and implementing processes to detect, moderate, report and remove (as applicable) material or activity on the service that is unlawful or harmful;

    b. if a service or a component of a service (such as an online app or game) is likely to be accessed by children (the children's service)—ensuring that the default privacy and safety settings of the children's service are robust and set to the most restrictive level;

    c. ensuring that persons who are engaged in providing the service, such as the provider's employees or contractors, are trained in, and are expected to implement and promote, online safety;

d. continually improving technology and practices relating to the safety of end-users;

e. ensuring that assessments of safety risks and impacts are undertaken (including child safety risk assessments), identified risks are appropriately mitigated, and safety review processes are implemented, throughout the design, development, deployment and post-deployment stages for the service;

f. assessing whether business decisions will have a significant adverse impact on the ability of end-users to use the service in a safe manner and in such circumstances, appropriately mitigating the impact;

g. having staff, systems, tools and processes to action reports and complaints within a reasonable period of time in accordance with subsection 14(3);

h. investing in systems, tools and processes to improve the prevention and detection of material and activity on the service that is unlawful or harmful;

i. having processes for detecting and addressing hate speech which breaches a service's terms of use and, where applicable, breaches a service's policies and procedures and standards of conduct mentioned in section 14;

j. preparing and publishing regular transparency reports that outline the steps the service is taking to ensure that end-users are able to use the service in a safe manner, including:

    i. the use of online safety tools and processes,

    ii. providing metrics on the prevalence of material or activity on the service that is harmful,

    iii. the service's responsiveness to reports and complaints, and

    iv. how the service is enforcing its terms of use, policies and procedures and standards of conduct mentioned in section 14.

Additional expectation

5. The provider of the service will take reasonable steps to make available controls that give end-users the choice and autonomy to support safe online interactions.

Examples of reasonable steps that could be taken

6. Without limiting subsection (5), reasonable steps for the purpose of that subsection could include the following:

a. making available blocking and muting controls for end-users;

b. making available opt-in and opt-out measures regarding the types of content that end-users can receive;

c. enabling end-users to make changes to their privacy and safety settings.

Additional guidance on subsections 6(1), (2), (2A) and (3)

- The intention of subsection 6(1) is to uplift how services develop and implement products, policies and terms in a way that has regard for the safety of Australian end-users.[30] Importantly, providers should continually assess and evaluate the effectiveness of online safety measures deployed on a service or designed into a service, and update, refine and adjust these measures accordingly to ensure safe use.[31]

Subsection 6(2) requires providers to take proactive steps to identify and address existing and emerging harms online.

- There is considerable cross-over between this expectation and the section 11 and 12 expectations to minimise certain material and class 2 material. Many of the reasonable steps to comply with those expectations will support compliance with subsection 6(2). However, subsection 6(2) is broader than section 11 and 12, including by capturing unlawful or harmful **activity** as well as **material**.

- Importantly, this subsection expects '**proactive**' minimisation of unlawful and harmful material and activity. This means providers are expected to take reasonable steps upfront to reduce the likelihood of such material being made available, or activity taking place, on the service. The key example of how this can be achieved is via the use of technologies or other tools. Proactive steps can be contrasted with reactive or responsive measures such as user reporting mechanisms or community moderation which should work alongside proactive steps, but which may be insufficient to demonstrate compliance with subsection 6(2).

[30] Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 12: Federal Register of Legislation - Online Safety (Basic Online Safety Expectations) Determination 2022.

[31] Providers should note that the Commissioner may request a report on the performance of online safety measures that a provider has publicly announced or otherwise reported to the Commissioner, under section 20 of the Determination. The Commissioner may also require information on the performance of online safety measures through mandatory reporting notices under the Act.

Considering the best interests of the child

Subsection 6(2A) requires providers to take reasonable steps to ensure that the best interests of the child are a primary consideration in the design and operation of any service that is likely to be accessed by children.

- Providers are expected to design and implement services that are likely to be accessed by children in a manner consistent with the objectives underlying Article 3 of the *Convention of the Rights of the Child* (UNCRC) that '[i]n all actions concerning children… the best interests of the child shall be the primary consideration'.
- The Explanatory Statement to the 2024 Amendment Determination[32] states that providers are expected to give high priority to protecting and promoting the full enjoyment by children of all of their rights, recognising their particular vulnerabilities and state of development.

Best interests of the child

- The 'best interests of the child' implies 'the full and effective enjoyment of rights… and the holistic development of the child' in both the immediate and longer term.[33]
- Children's rights are broad and indivisible and promote and protect safety, health, wellbeing, relationships, physical, psychological and emotional development, identity, freedom of expression, privacy and agency to form their own views and to have their views heard.
- Providers should carefully consider existing guidance on the best interests of the child and ensuring it is a primary consideration, including:
    - General comment No. 14 (2013) on the right of the child to have his or her best interests taken as a primary consideration[34]
    - General comment No. 25 (2021) on children's rights in relation to the digital environment.[35]
- Services are expected to consider the bests interests of the child in a manner which enables practical, effective outcomes for children. In the context of service providers with a large cohort of online users, providers may consider the best interests of the child generally, including by having regard to the physical, psychological and emotional wellbeing of children in certain age groups.
- It is likely that different services, functions and features will pose different safety risks to children of varied ages and capacities, so services should consider and identify potential risks early and take appropriate steps to ensure that children can use the functions and features of a service safely.

[32] See: Explanatory Statement, Online Safety (Basic Online Safety Expectations) Amendment Determination 2024.
[33] See: General comment No. 14, para. 51, United Nations Committee on the Rights of the Child (2013).
[34] See: General comment No. 14, United Nations Committee on the Rights of the Child (2013).
[35] See: General comment No. 25 (2021) on children's rights in relation to the digital environment.

Primary consideration

- The best interests of the child must be a primary consideration in the design and operation of a service. In the digital environment and in the design and operation of services, it is likely that there will be other considerations including commercial and business considerations.
- The Committee on the Rights of the Child has stated that a child's interests should have high priority and a larger weight attached to what serves the child best if other considerations are in conflict.[36] The Explanatory Statement to the 2024 Amendment Determination emphasises that providers are expected to give high priority to protecting and promoting the full enjoyment by children of all of their rights, recognising their particular vulnerabilities and state of development.
- Providers are expected to ensure the best interests of the child is a primary consideration throughout the life cycle of a service, including in the design of all aspects of the service and any new features or functionalities and in the operation of the service, including continual improvements and reviews.

Likely to be accessed by children

- A service is likely to be accessed by children if it:
    - is designed for, and aimed at children under the age of 18, or
    - is likely to be accessed by children, regardless of who it is designed for.
- In considering whether or not a service is likely to be accessed by children, regardless of the design or intention of the service, the following factors, as set out in the Explanatory Statement to the 2024 Amendment Determination, are highly relevant:
    - The nature and content of the service, including whether it is particularly appealing to children.
    - Market research, current evidence on user behaviour, the user base of similar or existing services and service types.
    - The way in which users access the service and whether any measures put in place are effective in preventing children from accessing the service.
- It is important to consider that children can use and navigate the internet to access services in similar ways to adults and that, unless children are prevented from access, they could be considered to be 'likely to access' a service.
- If a service contains material and activity which is regarded as inappropriate for children or which prohibits children from accessing the service, but the service does not provide any safeguards to prevent children's access, then subsection 6(2A) applies.

[36] See: General comment No. 14, paragraph 39, United Nations Committee on the Rights of the Child (2013).

- A risk assessment may assist providers to determine whether children currently access the service, or whether it is likely to be accessed by children.

Reasonable steps

- There are a range of reasonable steps that can be taken to ensure the best interests of the child are a primary consideration, including the steps set out in the rest of this guidance about Section 6.[37]
- Providers are accountable for the way in which the design and operation of services impacts children. eSafety expects that the best interests of the child will be directly linked to, and evident in, the outcomes and experiences for children on the service.

Subsection 6(3) outlines a range of steps that providers could take to meet the Expectations.

Risk and impact assessments

Undertaking safety risk and impact assessments and reviews are listed as examples of a reasonable step throughout the Determination. Assessments should:

- be a **priority** throughout the service or feature lifecycle. It is especially important when a new feature is designed, developed, and deployed to ensure harms are mitigated from the earliest stages
- be undertaken routinely, clearly documented, and updated regularly
- be informed by a human rights approach – meaning that the likelihood and severity or impact of harms occurring should be considered from the point of end-users and the community more broadly, and take into account other applicable human rights
- be informed by community and victims' groups and other expert insights to ensure all relevant risks are understood, and the impacts of any proposed safety mitigations are also assessed and mitigated
- not be limited to consideration of how a risk or a harm impacts the provider as a business, or from a narrow compliance perspective (although providers should ensure they assess whether they are complying with the Expectations as part of this process).
- consider the risks faced by younger users, for providers of services that permit children or young people to use their service or that are likely to be accessed by children. For example, risk assessments should consider risks related to content (a child or young person engaging with, or being exposed to, certain content), contact (experiencing, or being targeted by, potentially harmful contact, including by adults)

[37] Providers can also consider the Institute of Electric and Electronics Engineers (IEEE) Standard for an Age Appropriate Digital Services Empowerment Framework Based on the 5Rights Principles for Children, available on the IEEE website.

and conduct (witnessing, participating in, or being a victim of harmful conduct).The Explanatory Statement to the 2024 Amendment Determination notes that providers are expected to proactively assess the likelihood that their service is accessed by children and undertake child safety risk assessments for the purposes of subsection 6(2A).

eSafety recognises that complete mitigation of all harms may not be possible, and the Expectations do not require this outcome. However, providers should be prepared to report on the nature of the safety risk assessments undertaken, what safety risks were identified, how the risk assessment recommends the risks be mitigated, and what steps the provider has taken to implement these recommendations.

Providers may already undertake other risk assessments, for example privacy or human rights impact assessments. While safety risks and impacts could be considered as part of these broader processes, eSafety expects that providers will thoroughly identify and address the specific safety issues.

Relevant industry code and industry standard measures

The steps that providers may take to comply with the section 6 Expectations may also be relevant to compliance with certain industry codes and industry standards.

For example, undertaking risk assessments may support compliance with the Expectations and may also be required under industry codes or industry standards to inform how a code or standard applies. However, eSafety expects that risk assessments will be undertaken to identify, address and mitigate a broader range of harms and material in order to comply with the Expectations, including the particular risks posed to children on services.

Additionally, processes to detect and address unlawful and harmful material and activity, resourcing teams and ensuring staff are trained in online safety, assessing the impact of business decisions and investing in systems, tools and processes may also be steps that are directly relevant to compliance with industry codes and standards.

Providers are encouraged to identify to eSafety, if requested, how steps taken by the provider in relation to a service supports compliance across both schemes.

The section 6 expectations require providers to take steps in relation to both material **and** activity. It is important for providers to consider how certain material, or certain activity, may be harmful in some circumstances and less so in others. The severity or impact of a harm may vary for different individuals, or groups within the community.

For a structured framework to consider and mitigate safety risks in the design, development and deployment of services, see eSafety's Safety by Design tools.

- There are two tools – one designed for early-stage companies and another for mid-tier and enterprise organisations.
- For each tool, users are provided with an educative module on online harms, and are taken through a series of question and response options which culminate in a tailored end report, guiding and supporting providers to enhance online safety practices.

Resourcing of safety interventions and teams

Several of the subsection 6(3) examples relate to a service's staff and teams, including in relation to online safety training and implementation (6(3)(c)), safety risk assessments (6(3)(e)), assessing whether business decisions will have a significant adverse impact on safe use (6)(3)(f)), having staff, systems, tools and processes to action complaints and reports (6(3)(g)) and investing in systems, tools and processes (6(3)(h)).

It is important that a service's safety interventions are resourced proportionate to the risks identified and to enable compliance with the Expectations. This should involve:

- appropriately resourcing trust and safety teams, to ensure that appropriate safety interventions are in place, that interventions are working effectively, and that safety issues are responded to as a priority
- ensuring all relevant staff are suitably trained and supported, including through training on Safety by Design principles – there should be specialist training for trust and safety teams, and trust and safety functions should be subject to oversight and accountability by senior management
- trust and safety teams engaging with experts in online safety and technology, as well as victims, to inform policies and processes
- having clear and effective escalation processes to refer complex or specialist cases to expert teams.

Providers should also invest in the development of tools and processes to support their compliance with the Expectations and efforts of trust and safety staff. This includes research and development into technology to detect, disrupt and deter unlawful and harmful material and activity. Investment should be proportionate to the resources of the provider, and the risks posed by the service.

Moderation

In addition to the examples provided in subsection 6(3), further guidance is provided below in relation to the importance of content moderation.

Content moderation, where provided by a service, should be provided in a range of relevant languages to support the demographics of a service's end-users. This is particularly

important for harms that require context to identify, such as grooming or hate speech. This helps ensure that unlawful and harmful content is properly identified, and the accuracy of content moderation decisions.

Community moderation may be a useful mechanism to support alignment of material and activity on a service with the terms of use, standards of conduct and other service policies. However, it is important that the burden of enforcing terms of use, standards of conduct and otherwise addressing unlawful and harmful material and activity is not delegated solely to community moderation.

Where community moderation is used, it is important that community moderators are properly supported and equipped with information and tools from the service, and this should include requirements to escalate certain issues to the provider and professional trust and safety staff. This escalation is important so that trust and safety staff can take appropriate action including banning accounts across all parts of a service (not just the section that the violating conduct was identified within) and making onward reports to appropriate authorities.

Providers of community-moderated services must always retain an appropriate level of visibility over the activity on their service. This responsibility should never sit solely with community moderators or other end-users.

Significant business decisions

The Determination provides the example of assessing whether business decisions will have a significant adverse impact on the ability of end-users to use the service in a safe manner and in such circumstances, appropriately mitigating the impact.

The Explanatory Statement to the 2024 Amendment Determination provides relevant guidance on this step, including that relevant decisions which should be assessed for safety implications may include (but are not limited to):

- significant changes to a service's terms of use, policies and procedures and standards of conduct
- the creation of different subscription tiers or account types with different safety features
- major staffing changes, such as reductions in trust and safety staff
- changes to a service's technical architecture, features or functions that affect a service's ability to detect and address unlawful or harmful content.

Decisions made by service providers which affect the operation of the service should not increase the prevalence of unlawful or harmful material or activity, adversely affect vulnerable users such as children, or otherwise make the service less safe.

Investing in systems, tools and processes to improve prevention and detection

The Determination provides the example of investing in systems, tools and processes to improve the prevention and detection of material or activity on the service that is unlawful or harmful. As the Explanatory Statement to the 2024 Amendment Determination notes, 'investment' is not necessarily limited to financial investment but could include a broad range of initiatives such as participation in and support for research, pilot projects, and collaboration with law enforcement, non-government and government organisations or cross-industry collaboration.

Detecting and addressing hate speech

The Determination provides the example of having processes for detecting and addressing hate speech.

The Explanatory Statement to the 2024 Amendment Determination states that hate speech is communication or conduct by an end-user that breaches a service's terms of use and, where applicable, breaches a service's policies and procedures or standards of conduct mentioned in section 14, and can include communication or conduct which expresses hate against a person or group of people. Expressions of hate against a person or group of people can be on the basis of race, ethnicity, disability, religious affiliation, caste, sexual orientation, sex, gender identity, disease, immigrant status, asylum seeker or refugee status, or age.

This definition is non-exhaustive and provides broad guidance on how to meet this step. Services may vary in how they define hate speech or hateful conduct in their terms, policies or standards of conduct. It is expected that providers will take reasonable steps to detect and address hate as it is defined by the service.

Publishing transparency reports

The Determination provides the example of preparing and publishing regular transparency reports that outline the steps the service is taking to protect Australians online, including:

i. the use of online safety tools and processes
ii. providing metrics on the prevalence of material or activity on the service that is harmful
iii. the service's responsiveness to reports and complaints
iv. how the service is enforcing its terms of use, policies and procedures and standards of conduct mentioned in section 14.

Transparency should involve the provision of meaningful information about the use of online safety tools, processes and resources as well as the effectiveness of these measures on each specific service and on all relevant parts of a service. Proactive publication of meaningful, specific information may reduce the need for eSafety to give notices in relation to the Expectations to those providers.

Additional guidance on subsection 6(5)

Providers are expected to take reasonable steps to make available controls that give end-users the choice and autonomy to support safe online interactions. The Determination sets out the following examples of reasonable steps:

a. making available blocking and muting controls for end-users
b. making available opt-in and opt-out measures regarding the types of content that end-users can receive
c. enabling end-users to make changes to their privacy and safety settings.

In addition to the above examples, providers could also consider:

- promoting the availability of user controls to ensure end-users are aware of, and understand how to use and adjust controls to support safe online experiences
- providing users with a mechanism to provide feedback to the service in relation to the efficacy of user controls (such as an escalation pathway if user controls fail to address a safety issue)
- undertaking audits of user controls.

Section 7 of the Determination – Consulting with the Commissioner and referring to the Commissioner's guidance

Determination, section 7:

Core expectation

1. In determining what are reasonable steps for the purposes of subsection 6(1), the provider of the service will consult the Commissioner.

Additional expectation

2. In addition, in determining what are reasonable steps for the purposes of subsection 6(1), the provider of the service will have regard to any relevant guidance material made available by the Commissioner.

Subsection 7(1) intends to establish a dialogue between the Commissioner and service providers. It gives providers the opportunity to outline and justify the steps they take to ensure safe use, including in circumstances where the examples included in subsection 6(3) are not appropriate for a service, and alternative steps are taken. Subsection 7(1) also establishes a means for information sharing between the Commissioner and industry to improve online safety outcomes.

Providers can contact eSafety at industrybose@esafety.gov.au if they have specific questions

regarding reasonable steps and their ability to comply with the Expectations (although eSafety cannot provide legal advice). Providers are also expected to engage with eSafety if specific safety issues related to a service are identified, and a provider's willingness to engage and implement or consider eSafety's recommendations may be reflected upon when deciding whether a provider is complying with the Expectations.

Subsection 7(2) requires that in determining what are reasonable steps for the purposes of complying with subsection 6(1), a provider will have regard to any relevant guidance material made available by the Commissioner.

This guidance is made available to providers to assist them in meeting the Expectations. Providers are expected to have regard to this guidance material in implementing the Expectations, alongside the Safety by Design tools on the eSafety website, and other relevant materials published by eSafety.

This guidance may be updated in the future where additional guidance is required in relation to new harms, technologies and safety issues or in response to other events, or to include the responses to common questions from providers raised during section 7 engagement.

Further opportunities for consultation will be afforded to providers if they receive a non-periodic or periodic reporting notice which requires a provider to produce a report on their compliance with any or all of the Expectations. Further information is set out in Part 3 of this guidance.

Section 8 of the Determination – Detecting and addressing unlawful or harmful material or activity on encrypted services

Determination, section 8:

Additional expectation

1. If the service uses encryption, the provider of the service will take reasonable steps to develop and implement processes to detect and address material or activity on the service that is unlawful or harmful.
2. Subsection 8(1) does not require the provider of the service to undertake steps that could do the following:
    a. implement or build a systemic weakness, or a systemic vulnerability, into a form of encrypted service;
    b. build a new decryption capability in relation to encrypted services; or
    c. render methods of encryption less effective.

Encryption is a way to prevent unauthorised access to information. Encryption is not new and, in its modern form, has been used for more than 40 years as an essential tool for privacy and security. It is primarily employed for the secure transmission and storage of information, and

can help to prevent data breaches and hacking.

Section 8 applies to services that are encrypted in any form, including those using 'in transit' encryption such as Transport Layer Security, encryption at rest, and those using end-to-end encryption (E2EE). The reasonable steps that a provider should take to develop and implement processes to detect and address material or activity that is unlawful or harmful may depend on the nature of the encryption implemented on the service, and whether encryption is used on some, or all, parts of a service.

Services that use encryption in transit and/or at rest should take reasonable steps to detect unlawful and harmful material and activity on their service. This may involve the use of both automated tools such as hash matching or artificial intelligence (AI) classifiers, and human review. Further details are set out in the guidance on the section 6 and 11 expectations.

For providers that use E2EE on all or part of a service, there is a higher risk of unlawful and harmful material and activity going undetected, given the limitations E2EE creates for widely used detection technologies and interventions. Services that allow large groups, live streaming or video calling, and E2EE services that enable end-users to connect to other unknown users on the basis of shared interests, are also likely to pose greater risks.

While section 8 makes it clear that the Expectations do not require providers make E2EE less effective,[38] providers are required to take reasonable steps to develop and implement processes to both detect and address material or activity that is unlawful and harmful.

Reasonable steps to **detect** unlawful and harmful material and activity on E2EE services may include a number of options.

- Using hashing, machine learning, artificial intelligence and other detection technologies on any parts of the service that are not E2EE (such as profile pictures, content in user reports, group names).
- Using technology that enables unlawful and harmful material and activity to be detected at the device level or prior to upload on the service, where this can be done without building a systemic weakness or vulnerability (such as client-side scanning using hashing, AI classifiers, natural language processing of text to detect patterns indicative of grooming of children and sexual extortion).
- Using classifiers to detect signals and metadata relevant to unlawful and harmful content (such as behavioural signals related to private group membership, frequency of joining or leaving groups, engagement with children or young people using the service).

Reasonable steps to **address** unlawful and harmful material and activity on E2EE services may also include a number of options.

[38] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 19: Federal Register of Legislation - Online Safety (Basic Online Safety Expectations) Determination 2022. It states 'The Determination does not require or expect service providers to undertake actions inconsistent with obligations under the *Privacy Act 1988*, the *Telecommunications Act 1997* or *Telecommunications and Other Legislation Amendment (Assistance and Access) Act 2018*. Any adherence to expectations around anonymous (or pseudonymous) accounts and encrypted services are not to conflict with obligations under a Commonwealth Act.'

- Introducing obstacles to accessing E2EE services for the purpose of engaging in unlawful and harmful activity, such as:
    - working with law enforcement and relevant experts (for example, experts in relation to CSEA and terrorism) to identify and block access to E2EE channels associated with illegal activity
    - limiting the use of joining links[39] shared on unencrypted services (for example, the Terrorist Content Analytics Platform[40] can support this by alerting the encryption provider to join links shared on unencrypted spaces).
- Introducing registration requirements such as requiring end-users to register for the service using a phone number, email address or other identifier. If these identifiers are authenticated (for example, through an authentication link or code), this can help prevent recidivism where accounts have been identified as breaching the law or terms of use. This links to sections 9 and 14 of the Determination.
- Introducing obstacles to storing or sharing unlawful and harmful material, such as:
    - taking steps to ensure that unlawful and harmful material that is detected is not uploaded, shared, or hosted on the service (for example, referring to law enforcement, blocking or reporting the end-user, or advising the end-user that the material might be unlawful, harmful or inappropriate and in breach of the service's terms of use)
    - incorporating safety features (for example, interstitial warnings, blurring or blocking content, providing safety information to end-users)
    - restricting or limiting an end-user's ability to share material with large numbers of people instantaneously (for example, restricting the ability to forward a message to many other users or groups at once).
- Providing end-users with reporting tools. Given some technologies may be challenging to implement on E2EE services, a particularly important step should be to provide end-users with clear and readily identifiable tools to report unlawful and harmful content on E2EE services to the service. Examples of clear and readily identifiable reporting mechanisms are outlined on page 50 of this document.

It may be difficult for a provider to demonstrate compliance with section 8 if they are taking limited or no steps to detect and address material or activity on the service that is unlawful or harmful, noting that the service is already likely vulnerable to exploitation by those seeking to engage in unlawful and harmful conduct without detection.

Providers should ensure that risks are fully considered and steps are built into a service's design before E2EE or other forms of encryption are implemented, rather than considered afterwards when harms arise. By adopting a holistic combination of the most suitable measures in a proportionate manner, providers can help to mitigate risks occurring on E2EE services.

[39] Links, often shared on unencrypted services, driving users to encrypted spaces.
[40] See Terrorist Content Analytics Platform.

Section 8A of the Determination—providers will take reasonable steps regarding the safety of generative artificial intelligence capabilities

(1) If the service uses or enables the use of generative artificial intelligence capabilities, the provider of the service will take reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of generative artificial intelligence capabilities on the service.

(2) If the service uses or enables the use of generative artificial intelligence capabilities, the provider of the service will take reasonable steps to proactively minimise the extent to which generative artificial intelligence capabilities may be used to produce material or facilitate activity that is unlawful or harmful.

Examples of reasonable steps that could be taken

(3) Without limiting subsection (1) or (2), reasonable steps for the purposes of this section could include the following:

(a) ensuring that assessments of safety risks and impacts are undertaken, identified risks are appropriately mitigated, and safety review processes are implemented throughout the design, development, deployment and post-deployment stages of generative artificial intelligence capabilities;

(b) providing educational or explanatory tools (including when new features are integrated) to end-users that promote understanding of generative artificial intelligence capabilities on the service and any risks associated with the capabilities;

(c) ensuring, to the extent reasonably practicable, that training material for generative artificial intelligence capabilities and models do not contain unlawful or harmful material;

(d) ensuring, to the extent reasonably practicable, that generative artificial intelligence capabilities can detect and prevent the execution of prompts that generate unlawful or harmful material.

The Explanatory Statement to the 2024 Amendment Determination specifies that generative AI is distinguished from other applications of AI by its capacity to generate novel material such as text, images, videos, audio or a combination of these.

Providers are expected to take reasonable steps to consider end-user safety at all stages of the life cycle of a generative AI capability, and to proactively minimise the extent to which generative artificial intelligence capabilities may be used to produce material or facilitate activity that is unlawful or harmful. This expectation recognises the increased risks of generative AI adversely affecting online safety for individuals as well as to society more broadly, and the importance of ensuring safety interventions are considered from the earliest stages of development.

Examples of reasonable steps are set out in subsection 8A(3).

The Explanatory Statement notes, and eSafety recognises, that the nature of the steps taken by each provider may differ depending on the nature of the generative artificial intelligence capabilities provided on a service, where a service sits in the cycle of development and deployment of generative AI and the matters within control of the service. Providers should be prepared to report on how the steps taken support compliance with subsections 8A(1) and (2).

Additional examples of reasonable steps could include the following:

- Addressing generative AI capabilities in relevant policies such as terms of use, policies and procedures and standards of conduct.
- Consulting with relevant groups to ensure that specialist knowledge on the various harms and risks posed to the community is obtained and incorporated into safety risk assessments and safety interventions.
- Red-teaming, violet-teaming and/or stress-testing generative AI capabilities.
- Using educative prompts and nudges when users attempt to misuse generative AI capabilities.
- Using warnings or disclaimers to advise users that certain outputs may be inaccurate, misleading or harmful.
- Incorporating reporting mechanisms for generative AI capabilities, including feedback loops for users to track the status of their reports.
- Using digital watermarking or other methods of content provenance to identify where material is AI-generated.
- Considering any specific risks posed to children through the availability or use of generative AI capabilities on a service, and considering the best interests of the child as per subsection 6(2A).
- Providing transparency in relation to generative AI capabilities through processes such as model cards, system cards and value alignment cards which document the capabilities, limitations, intended uses and prohibitive uses of a capability.

Relevant industry standard measures

Certain providers are also required under the DIS Standard to comply with enforceable obligations in relation to generative AI. Those obligations apply to specific types of services and focus on class 1 material. Section 8A applies more widely to DIS with a generative AI capability, and RES and SMS with a generative AI capability, and to unlawful and harmful material and activity more broadly.

For additional information on generative AI, see eSafety's Position Paper.[41]

Section 8B of the Determination—providers will take reasonable steps regarding the safety of recommender systems

(1) If the service uses recommender systems, the provider of the service will take reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of recommender systems on the service.

(2) If the service uses recommender systems, the provider of the service will take reasonable steps to proactively minimise the extent to which recommender systems amplify material or activity on the service that is unlawful or harmful.

Examples of reasonable steps that could be taken

(3) Without limiting subsection (1) or (2), reasonable steps for the purposes of this section could include the following:

(a) ensuring that assessments of safety risks and impacts are undertaken, identified risks are appropriately mitigated, and safety review processes are implemented throughout the design, development, deployment and post-deployment stages of recommender systems;

(b) providing educational or explanatory tools (including when new features are integrated) to end-users that promote understanding of recommender systems on the service, their objectives, and any risks associated with such systems;

(c) enabling end-users to make complaints or enquiries about the role recommender systems may play in presenting material or activity on the service that is unlawful or harmful;

(d) where technically feasible, enabling end-users to opt-out of receiving recommended content, or providing alternative curation options.

Recommender systems, also known as content curation algorithms, are the systems that prioritise content or make personalised suggestions to users of online services, including regarding material, other users, accounts or profiles.[42]

The different inputs and end goals for recommender systems can lead to both positive and negative outcomes. For example, recommender algorithms that prioritise user engagement and then serve up similar content in the future may result in people seeing things they find interesting, entertaining or valuable. But equally, if an end-user spends time engaging with potentially harmful content, those same metrics may lead to them seeing more of the same material or increasingly extreme material in their feeds.

In addition to risks and harms at an individual level, recommender systems have the

[41] eSafety Position Paper on generative AI: Generative AI – position statement | eSafety Commissioner.

[42] For more information, see eSafety's Position Paper on recommender systems and algorithms: Recommender systems and algorithms – position statement | eSafety Commissioner.

potential to cause new, or exacerbate existing harms on a societal level – for example, content promoting hate or inciting violence can cause harm to the people targeted and can also spill over into violence and discrimination affecting the broader community.

Examples of reasonable steps are set out in subsection 8B(3). Additional steps may include the following:

- Providing opt-in or opt-out measures for end-users to maintain choice, ownership and control of the types of content they receive.
- Adjusting recommender algorithms to focus on other metrics such as authoritativeness or diversity of content as an alternative, or in addition to, user-engagement. These metrics should be subject to consultation, public scrutiny and testing.
- Using human review as a safety check for content that is being rapidly disseminated or promoted.
- Introducing additional safeguards through design features, such as prompts to read an article linked before sharing it, which may reduce the likelihood of it being shared.
- Labelling content as potentially harmful or likely to include certain themes or topics, particularly where content may be sensitive to some higher risk groups and communities and not others. Where content warnings are provided to some end-users and not others, consideration should be given to the data which informs these choices and the risk of bias.
- Including behavioural cues and prompts that can help end-users establish positive patterns of behaviour – for example, that help end-users reconsider posting harmful content or manage their time spent online.
- Enhancing transparency reporting and auditing practices.
- Curating recommendations so they are age appropriate, including friend or follower suggestions between adults and children.
- Offering parental controls to allow parents and carers to limit and/or monitor what material and activity their child is exposed to and engages with, with the ability to adjust these settings as children develop and their capacity evolves.
- Employing measures to test and update recommender systems with the objective of improving overall safety – for example, internal audits, external audits, risk and impact assessments, a/b testing.
- Preventing autocomplete searches of phrases that are likely to be associated with unlawful or harmful content.

Section 9 of the Determination – Preventing anonymous accounts being used for unlawful or harmful material or activity

Determination, section 9:

Additional expectation

1. If the service permits the use of anonymous accounts, the provider of the service will take reasonable steps to prevent those accounts being used to deal with material, or for activity, that is unlawful or harmful.

Examples of reasonable steps that could be taken

2. Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following:
   a. having processes, including proactive processes, that prevent the same person from repeatedly using anonymous accounts to post material, or to engage in activity, that is unlawful or harmful;
   b. having processes that require verification of identity or ownership of accounts.

'Anonymous accounts' are accounts that hide or disguise the identity of an end-user.[43]

There are many ways of appearing anonymous online. They include the following examples:

- Full anonymity – where an end-user does not provide any personal information or identifiers, and neither the online service nor other users can identify the end-user at the time of a particular interaction, or subsequently. This may be a result of a service design, or as a result of an end-user taking active identity shielding steps to prevent the collection of their data (for example, the use of a virtual private network (VPN) or other technologies that prevent disclosure of their geo-location or Internet Protocol (IP) address).[44]
- Public anonymity – where an end-user may appear anonymous to others, however the provider collects and holds some information about the end-user (for example, personal information such as their name, email or phone number, or their geo-location, or their IP address, or the way they have engaged with the service and other users).[45]
- Pseudonymity – where an end-user has registered for a service using a username, handle or avatar that is not their real name, however the service collects and holds some information about the end-user (for example, many services require end-users

[43] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 15: Federal Register of Legislation - Online Safety (Basic Online Safety Expectations) Determination 2022.

[44] It is important to note that there are legitimate reasons for users to employ tools such as VPNs – for example, to keep their information secure when using public Wi-Fi.

[45] It is important to note that there are legitimate reasons for users to be publicly anonymous – for example, to protect their privacy and confidentiality when seeking out information and assistance online about sensitive topics.

to provide an email address or phone number at sign up).[46]

Section 9 applies to both anonymous accounts and pseudonymous accounts.[47]

There are many benefits in and valid reasons for maintaining a level of anonymity or practicing identity shielding online, including the right to privacy and protection from violence or unwanted contact.

However, anonymity and identity shielding can also enable harmful behaviours, particularly against people and communities who are at higher risk. eSafety's investigations teams regularly see anonymity being used as a tactic by those who seek to harm or abuse others online, for example:

- in the cyberbullying of children and in adult cyber abuse
- in the non-consensual sharing of intimate images (image-based abuse)
- in creating, storing and sharing unlawful content such as child sexual exploitation and abuse material.

The section 9 expectation does not require services that permit users of anonymous accounts to stop doing so, for example by employing a 'real name' policy or otherwise 'unmasking' their identities. Rather, it states that providers are expected to take reasonable steps to prevent anonymous accounts from being used to deal with material or for activity that is unlawful or harmful.

eSafety supports a balanced approach to this issue, which minimises the potential to disrupt the positive outcomes that online anonymity can afford. Providers are expected to have measures in place that allow them to effectively prevent and respond to harms perpetrated by anonymous account holders, for example:

- Ensuring the provider is able to identify and engage with accounts that are engaging in unlawful or harmful activity or material including by taking enforcement action when terms of use or policies are breached.
- Ensuring that end-users are not able to evade enforcement action by registering for a new account and continuing to cause harm.

Additionally, providers of services that do **not** permit anonymous or pseudonymous accounts should ensure they are taking reasonable steps to effectively enforce this rule. If a service's prohibition on use of anonymous accounts is being circumvented by end-users and that enables harms to occur on the service, the provider should consider whether section 9 applies to the service and comply if it does.

[46] It is important to note that there are legitimate reasons for users to choose pseudonyms rather than using their real names online – for example, eSafety advises children not to use their real names online due to safety and privacy risks associated with sharing their personal details with people they do not know.

[47] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 14: Federal Register of Legislation – Online Safety (Basic Online Safety Expectations) Determination 2022.

Verification of identity or ownership of accounts (subsection 9(2)(b))

Subsection 9(2)(b) of the Determination provides a key example of a reasonable step that can be taken to meet the section 9 expectation: implementing processes that require verification of identity or ownership of accounts.

However, providers are not required to 'unmask' an end-user's identity in order to demonstrate compliance with this expectation, although this may be a step that some providers take for their own purposes or for the safety or comfort of their end-users. For example, some business networking sites or dating sites may require real identities.

Providers are instead expected to take reasonable steps to prevent accounts from being used to deal with activity or material that is unlawful or harmful, which could include the following options.

- Requiring end-users to authenticate their accounts on sign-up by sending an authentication code or message or link to an email address or phone number used to create an account (including multi-factor authentication). This means that an account must be linked to a valid email or phone number. This may reduce instances of individuals seeking to create multiple accounts for harmful purposes, and may act as a deterrent against misuse and abuse as end-users know the service will be able to take appropriate enforcement action against them.
- Collecting appropriate identifiers from end-users on registration or sign up which enable the provider to deal effectively with that end-user (for example, to contact the end-user, to enforce terms of use and take other enforcement action, to respond to complaints about that end-user, to respond to legal requests for end-user details from eSafety and other regulators or law enforcement bodies). This could include collecting personal information such as name and date of birth, or using device identifiers or other identifiers.
- Using tools outlined elsewhere in this guidance, to prevent and detect abuse.

Processes that prevent the same person from repeatedly using anonymous accounts to post material, or engage in activity that is unlawful or harmful (recidivism – subsection 9(2)(a))

One of the significant safety risks and harms in relation to the use of anonymous accounts is the ability for individuals to engage in **repeated** activity or conduct that is unlawful or harmful (recidivism).

The Explanatory Statement to the Determination identifies a number of suggested steps to comply with section 9. Specifically, it suggests providers could have processes that uses web identifiers (such as cookies, IP addresses, browser fingerprinting), device or hardware identifiers, or other identifiers (such as account or behavioural analysis, metadata and traffic signals) to identify and stop re-registrations or alternative accounts in appropriate circumstances.

Services should not rely solely on user reports and complaints in identifying individuals who may have previously generated or shared material or engaged in activity that is unlawful or harmful. Relevantly, subsection 9(2)(a) emphasises the importance of proactive steps.

Other steps to **address recidivism** through the use of anonymous accounts, including proactive steps, may include:

- using other identifiers, in addition to those listed in the previous paragraph, to identify and stop re-registrations or alternative accounts, including personal information provided by the account holder (such as their name, address, date of birth, phone number, email, account photos, credit card details or other payment information), or behavioural indicators (such as their registration date, email alias, posting behaviour, usernames, or key phrases they use)
- using technology to detect previously banned end-users (for example, hash-matching that detects the profile pictures of banned end-users when an attempt is made to use them again)
- scanning for indicators of known or suspected offenders across all of the services operated by a provider, and implementing effective cross-service bans for offenders where appropriate
- providing end-users with clear communication advising if they are engaging in unlawful or harmful conduct, including conduct that violates terms of use, standards of conduct or other policies (for example, providing a warning via a pop-up)
- enabling end-users to block content from unverified or unauthenticated accounts
- imposing a strike system to determine appropriate action in response to repeated conducted (for example, warnings, penalties, bans, requiring identity verification to continue using the service)
- taking effective and appropriate enforcement action where necessary, such as implementing a device block to prevent an account from re-registering on the same device, or blocking an IP address.

Section 10 of the Determination – Consulting and cooperating with other service providers to ensure safe use

Determination, section 10:

Additional expectation

1. The provider of the service will take reasonable steps to:
    a. consult and cooperate with providers of other services; and
    b. ensure consultation and cooperation occurs between all relevant services provided by that provider, in order to promote the ability of end-users to use all of those services in a safe manner.

Examples of reasonable steps that could be taken

2. Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following:
    a. working with other service providers and between all relevant services provided by a service provider to detect high volume, cross-platform attacks (also known as volumetric or 'pile-on' attacks);
    b. sharing information with other service providers and between all relevant services provided by a service provider on material or activity on the service that is unlawful or harmful, for the purpose of preventing and dealing with such material or activity.

Providers are expected to take reasonable steps to cooperate with other members of industry to identify and respond to new harms, trends and issues that impact the safety of end-users. The intent of information sharing and cooperation is to allow providers to prevent and deal with unlawful and harmful material and activity in an effective manner, that suits their circumstances.

Providers are also expected to take reasonable steps to ensure there is information sharing and cooperation across **their own services**. The barriers to sharing information across a provider's own services will be lower than those sharing with other providers' services.

It is not expected that providers would cooperate in a way that puts a service's intellectual property at risk or involves the sharing of commercial-in-confidence information. The focus is on consultation and cooperation which aims to minimise unlawful and harmful material or activity that adversely impacts online safety for Australians.

High volume, cross-platform attacks

Subsection 10(2)(a) suggests that a reasonable step that a provider could take to cooperate with other providers or services is to detect and share information regarding high volume and/or cross-platform attacks (also known as volumetric or 'pile-on' attacks).

High volume attacks occur when a person is named in, tagged, or linked to an abusive post, which others 'like', share, re-post with additional commentary, and/or link to via other services. The volume of material can proliferate rapidly across services.

Cooperating to promote safe use in this way could include making other services aware of a volumetric attack by sharing information like URLs, hashtags or account names, as well as information on the people or groups being targeted, and insights on sources and trends. This information would assist a service to respond, subject to its own terms of use and policies.

Behaviours associated with volumetric attacks may include:

- **Direct harassment through public posts and/or private messages:** (including those that might meet the thresholds of adult cyber abuse, child cyberbullying or image-based abuse).
- **Mass commenting:** users may leave comments on a particular post, generally containing criticism, insults or slurs. In some instances, comments are harmless in isolation (for example, commenting a single word or name), but the large volumes of activity or content can cause harm.
- **Mass 'liking':** users may 'like' a targeted individual's posts en masse to flood their notifications. While an individual 'like' would not be considered abuse, the repetition and volume may be used for intimidation or harassment.
- **Doxing:** revealing personal information to deliberately make someone feel unsafe. Sharing this information publicly undermines the targeted individual's privacy, security, safety and/or reputation. Often those responsible for doxing urge others to use the information to harass the targeted individual.
- **Abuse of reporting functions:** users may simultaneously report non-violative content to trigger removal.
- **Vote manipulation/brigading:** users may abuse like/upvote and dislike/downvote features to promote negative content while obscuring authentic posts in the recommender algorithm. Where a group of outsiders do this in a targeted, coordinated manner, it is sometimes known as brigading.
- **'Sock puppeting':** fake accounts may be used to show manufactured support for a user's viewpoint, or to participate in staged arguments to drive polarisation. Fake accounts are also often used to increase the volume of an attack, or to post content that a user may not want to post on their main account, either for anonymity or for the risk of being banned.

- **Ratioing:** if a post receives more replies, comments or views than likes (a skewed ratio), it is often an indicator that the post has been poorly received. Ratioing can occur organically, however it can also be a coordinated behaviour, and is commonly celebrated as a victory amongst participants in an attack.

Technologies can also be leveraged to increase the reach and impact of a coordinated volumetric attack. For example:

- **Bots** can be automated to share and amplify material and target certain groups and communities.
- **Deepfakes** and image, video or audio editing can be used directly to harass, and to mispresent events or a person's actions in order to manipulate third parties into participating.
- **Generative AI** can allow for the automatic creation and dissemination of messages, emails, posts, and comments across a wide range of platforms. Paired with bots, AI can drastically increase the speed and scale of an attack and enable a single user to carry out a volumetric attack on their own. AI may also scrape personal information available about a target online to create highly specific content which can be more intimidating and linked to physical world harms.
- **Recommender algorithms** may pick up on trends in activity and promote related hashtags or posts. This can cause previously uninvolved users to join in, without necessarily realising the nature or scale of the attack.

Information sharing

Subsection 10(2)(b) suggests that a reasonable step could be to share information with other providers or services about material or activity that is unlawful or harmful with a view to preventing and dealing with it. For example, providers or services could share information about a section of the community that is being targeted with abuse due to an identifying characteristic (such as sexuality, ethnicity or disability), or linked to a specific event (such as a sporting or political event). Providers or services that receive this information could then take appropriate actions to prevent and deal with unlawful or harmful material or activity targeted at that group or event.

There are a number of additional reasonable steps that could be taken.

- Wherever possible, providers should take part in regular forums organised or facilitated by an industry association to discuss and evaluate effectiveness of safety tools and features that promote and ensure compliance with the Expectations and any other applicable safety laws.
- Providers could consider the off-platform behaviour of end-users of their services when making internal decisions affecting end-users. For example, when considering whether an end-user or account has violated terms of use, community guidelines or other policies, or whether an end-user poses an unacceptable safety risk to a service,

services could take into account credible information (such as that published, provided or validated by another service or provider) about significant threats related to that end-user, such as those related to child sexual exploitation and abuse or terrorism.

- Providers could consider collaborating or partnering with organisations that seek to work with industry to address particular online harms.

Relevant industry code and industry standard measures

Certain providers are required under the SMS Code to take part in an annual forum to discuss online safety and evaluate the effectiveness of measures implemented under the code and share best practice with other industry participants. Additionally, certain providers are required under the SMS Code to collaborate with expert groups that tackle child sexual exploitation and abuse and pro-terror material. The RES and DIS Standards require large services to establish and implement development programs, which can include activities such as joining relevant industry organisations or collaborating with relevant non-government organisations.

Chapter 2: Expectations regarding certain material

Division 3 of the Determination sets out expectations regarding certain material and activity, including that reasonable steps will be taken to minimise the extent to which the following material is provided on a service.

- Section 11: child cyberbullying material, adult cyber abuse material, non-consensual intimate images, class 1 material, and material that promotes, incites, instructs and depicts abhorrent violent conduct.
- Section 12: class 2 material.

Section 11 of the Determination – Minimising provision of certain material

Determination, section 11:

The provider of the service will take reasonable steps to minimise the extent to which the following material is provided on the service:

a. cyber-bullying material targeted at an Australian child;

b. cyber-abuse material targeted at an Australian adult;

c. a non-consensual intimate image of a person;

d. class 1 material;

e. material that promotes abhorrent violent conduct;

f. material that incites abhorrent violent conduct;

g. material that instructs in abhorrent violent conduct;

h. material that depicts abhorrent violent conduct.

Section 11 relates specifically to material set out in subsections 11(a)-(h) (**certain material**). eSafety has published regulatory guidance on eSafety's powers in relation to these categories of material, separate to the Expectations. For more detail on the nature of each category of material, see eSafety's other regulatory guidance documents.[48]

The reasonable steps taken to minimise the extent to which certain material is provided on a service may differ, depending on each category of material and the way in which this material is provided, or able to be provided, on a service.

Providers should assess the risks of this certain material being provided on their service, and tailor their steps to address the risks.

[48] See eSafety's website for regulatory guidance: Regulatory schemes | eSafety Commissioner.

For example, the risks for a service may include:

- end-users storing certain material on a service
- end-users generating certain material of themselves or others
- facilitating the creation of certain material (for example, through generative AI)
- end-users sharing certain material with other users, or sharing links to certain material
- end-users advertising the sale of, or access to, certain material.
- end-users encouraging other users to produce, share, store or otherwise access certain material
- end-users finding and connecting with victims or potential victims (including children) to obtain certain material
- repeated harassment, threatening, bullying, intimidating or abuse of a person, including through anonymous accounts or by creating multiple accounts to continue the behaviour.

Reasonable steps to minimise the provision of certain material should include both organisational and technical measures, to ensure that this material is:

- communicated to end-users as material that is not permitted on a service, or is subject to moderation (for more detail, see guidance on section 14 regarding terms of use and certain policies regarding reports, complaints and conduct)
- proactively detected by the provider, where appropriate (see examples in the following paragraph)
- able to be reported to the provider by end-users and trusted flaggers (see guidance on user reporting in section 13 for more detail)
- prioritised for review and action expeditiously by the provider.

A key step to minimising provision of certain material is the ability to detect it – either before it is uploaded or shared on a service, or immediately after it is provided on the service. A number of steps may be used to proactively detect certain material, including the following options.

- Hash matching technology to detect known images and videos of unlawful material such as CSEA and terrorism material.
- Hash matching technology to detect non-consensual intimate images shared on a service (see, for example, the National Center for Missing and Exploited Children's (NCMEC) Take It Down hash list for images of under 18 year-olds and StopNCII hash data base for images of people 18 years and older). Additionally, providers could use hash matching technology internally to hash content or material that is reported to them from end-users or otherwise detected by the provider, and scan for these internal hashes across their service.

- AI classifiers to identify new material that is likely to be unlawful (such as CSEA and terrorism material) or harmful, and prioritise for human review, including where this material is livestreamed on a service (for example, broadcast to a wide audience or occurring in a private video chat or call).
- Technologies such as language or text analysis which can identify a wide range of unlawful or harmful activity occurring on online services. These technologies and processes should be regularly evaluated and updated to respond to evolving use of language by end-users, including deliberate attempts to avoid detection through the use of new words, phrases, symbols and text.

Where content is unlawful it should be removed and reported to appropriate authorities. It may also be appropriate to ban the account holder and prevent them from re-registering on the service.

Providers could also use proactive nudges or prompts to end-users that the material they are attempting to upload, save, send or otherwise share may be unlawful or harmful, including whether such material is prohibited in terms of use or other policies. For more serious content, end-users should also be notified that the material may be unlawful.

Additionally, providers are expected to exercise vigilance in detecting ongoing patterns of abuse against end-users once abuse has been reported to the service. Material set out in section 11 may be provided on a service by end-users in a manner that demonstrates repeated abuse of other users, and providers should ensure they are taking reasonable steps to minimise the repeated provision of material.

It is important that tools are used on all appropriate parts of a service in order to detect certain material. Subject to technical or other constraints, eSafety considers that a provider is unlikely to be meeting the section 11 expectation (and section 6) if a service is only using relevant tools on one part of its service, but leaves other at-risk parts of a service without any intervention.

Additionally, eSafety will have regard to the extent to which these tools are implemented and relevant processes are updated. For example, it is unlikely to be sufficient to deploy a hash matching tool to detect CSEA, but only update the list of available hashes once a year.

Relevant industry code and industry standard measures

eSafety notes that certain providers are required under the SMS Code and RES and DIS Standards to proactively detect known CSAM and terrorism material.

eSafety notes that some compliance measures apply only to child sexual abuse material, whereas the Expectations apply to all class 1 material, including material that shows the sexual exploitation of child, but does not show their abuse.

The use of technological tools to proactively detect certain material should be supported by human moderators who review content flagged in appropriate circumstances and take steps to remove and report or otherwise deal with the material. Appropriately resourcing systems and processes to ensure that user reports of unlawful and harmful content are responded to, and actioned, in a timely manner support compliance with this expectation.

It is particularly important that end-users are provided with clear and readily identifiable mechanisms to report certain material and make complaints. For more detailed guidance on reporting and complaint mechanisms, see Chapter 3.

Section 12 of the Determination – Preventing children's access to class 2 material

Determination, section 12:

Core expectation

1. The provider of the service will take reasonable steps to ensure that technological or other measures are in effect to prevent access by children to class 2 material provided on the service.

Examples of reasonable steps that could be taken

2. Without limiting subsection (1) of this section, reasonable steps for the purposes of that subsection could include the following:
   a. Implementing appropriate age assurance mechanisms;
   b. conducting child safety risk assessments;
   c. Continually seeking to develop, support or source, and implement improved technologies and processes for preventing access by children to class 2 material.

What is class 2 material?

Class 2 material is defined earlier in this guidance on page 6.

Why should children be prevented from accessing this material?

There are risks for children and young people under the age of 18[49] as a result of intended, unintended, non-consensual or coerced access to class 2 material. Therefore, a range of interventions should be adopted by providers to suit the evolving developmental needs of children and young people.

More information on the risks and harms related to children and young people's access to pornography can be found in eSafety's Age Verification Roadmap and background report.[50]

[49] References to 'children and young people' generally means children and young people under the age of 18.
[50] See eSafety's website: Age verification | eSafety Commissioner.

This guidance will be updated in the future to address any overlap between the section 12 expectation and industry codes or industry standards relating to class 2 material.

Technological and other measures that may be used to prevent access by children to class 2 material

In determining what reasonable steps should be taken to prevent access by children and young people to class 2 material, it is important to consider the extent to which class 2 material is provided on a service. For example, providers may operate services that:

1. deliberately host or provide access to class 2 material for end-users (for example, porn sites),
2. permit class 2 material, or do not actively enforce prohibition of this material, but it is not a core aspect of the service (for example, end-users can share material or distribute links to class 2 material, advertisements may be placed that contain or link to class 2 material), or
3. prohibit class 2 material.

Subsection 12(2) of the Determination provides three examples of reasonable steps that can be taken to ensure compliance with section 12:

(a) implementing appropriate age assurance mechanisms

(b) conducting child safety risk assessments

(c) continually seeking to develop, support or source, and implement improved technologies and processes for preventing access by children to class 2 material.

Age assurance is not defined in the Determination, and is an umbrella term which includes both age verification and age estimation solutions.

- Age verification measures determine a person's age to a high level of accuracy and can involve the use of physical or digital government identity documents to establish a person's age.
- Age estimation technologies provide an approximate age to allow or deny access to age-restricted online content or services. Age estimation can involve the use of biometric data, such as a facial scan or voice recording, to infer a person's age or age range.

By identifying 'appropriate age assurance mechanisms' as an example of a reasonable step, providers have a degree of flexibility as to how they protect children and young people from access to class 2 material. The Explanatory Statement to the 2024 Amendment Determination notes that whether an age assurance mechanism is 'appropriate' will depend on relevant factors such as:

- the effectiveness of the age assurance mechanisms

- the extent to which class 2 material is provided on the service
- the likelihood of children accessing the material on the service.

Age assurance mechanisms may also support compliance with other applicable expectations, for example by:

- ensuring that underage or prohibited end-users are not able to access services (for example, many services do not permit children who are under 13 – which relates to the section 6 expectation on ensuring safe use of a service)
- assisting providers in enforcing their minimum age requirements and terms of use (also relevant to section 14)
- providing an indication to a service that an end-user is of a certain (or approximate) age, which enables high privacy and safety settings to be implemented by default for that end-user, including preventing access or exposure to certain content on a service (also relevant to section 6).

Providers can consider the elements of a Restricted Access System,[51] as set out in the Online Safety (Restricted Access Systems) Declaration 2022[52] in terms of measures that may be adopted to prevent children and young people from accessing class 2 material on their service, although additional steps may be required, depending on the nature of the service. These elements include:

- requiring an end-user to apply for access to relevant class 2 material, with a declaration that they are at least 18 years old
- giving warnings and safety information for class 2 material
- incorporating reasonable steps to confirm the age of applicants.

Measures to prevent children and young people from accessing this material should not unduly restrict the rights of adults to create, access and share lawful content, and it is important that steps to achieve this be balanced against the need to preserve age-appropriate access to sexual health and wellbeing information and support.

For services that **deliberately permit class 2 material as a core part of the service,** it is important that robust measures are in place to prevent children and young people under 18 from accessing the service.

This may include:

- clearly communicating to end-users that the service contains class 2 material and is

[51] A restricted access system is a means of limiting access to material that is inappropriate to children and young people under 18. The Commissioner may give remedial notices to certain providers requiring the recipient to take all reasonable steps to remove class 2B material from a service, or place the material behind a restricted access system. See eSafety's Online Content Scheme.

[52] Online Safety (Restricted Access Systems) Declaration 2022: Online Safety (Restricted Access Systems) Declaration 2022.

intended for adult access (over 18 years old)

- applying meta-tags to the site, such as the Restricted to Adults label, to ensure the service or platform is blocked by any filters that may be in place for children on accounts or devices
- implementing age assurance or age verification mechanisms to prevent access to the service, and to prevent account registration if accounts are required.
- ensuring that landing pages or first point of contact with a service do not contain class 2 material and that this material is placed behind an age-gate.

For services that do not have class 2 material as a core part of their service but **permit class 2 material**, steps should be taken to prevent access to that material by children and young people under 18. For example, the service may:

- take the same steps listed for services that intentionally permit class 2 material (communicating to end-users that the service may contain class 2 material, using meta-tags to ensure the service is blocked by filters in place for children, and using age assurance mechanisms where appropriate)
- limit the searchability or discoverability of class 2 content by children and young people under 18, for example by preventing autocomplete or predictive entries for searching for terms that are known to be associated with class 2 material, and filtering out search responses for children and young people under 18
- blur class 2 material by default for all end-users to prevent unintentional access or exposure
- deploy technology or other tools to minimise the risk that class 2 material is provided, promoted or otherwise accessible to children and young people via the service, either as content or in advertisements
- deploy technology or tools to ensure that any permitted class 2 material, and any accounts dedicated to or commonly providing class 2 material, are appropriately tagged and that end-users are provided with appropriate warnings and options not to view the tagged content
- provide support to children and young people where they are specifically seeking out class 2 material – for example, pop up messages, tools or resources that explain why this material is not available to them (or is otherwise inappropriate for their age) or direct them to appropriate resources or support
- provide clear and accessible guidelines for end-users about access to class 2 material on the service and what safety measures are in place for children and young people under 18
- provide clear and readily identifiable reporting tools for children and young people (or their parents or carers) to flag class 2 material that they encounter, and ensure that flagged or reported material is not provided to the child or young person again

- provide strong parental controls, filtering and other supervision tools to support parents in ensuring that class 2 material is not accessible to a child or young person.

For services that **do not permit class 2 material**, steps should be taken to ensure that this policy is known to end-users and enforced. For example, the service may:

- set out this prohibition clearly in terms of use, community guidelines and/or other relevant policies
- take steps to enforce these terms of use – for example by warning, suspending or banning end-users who breach the terms of use, or preventing them from re-registering where appropriate
- enable end-users to report class 2 material to the service, and respond to these reports
- provide proactive detection of class 2 material
- provide strong parental controls, filtering and other supervision tools to support parents in ensuring that class 2 material is not accessible by children and young people
- use AI classifiers to detect nudity, combined with human moderation.

Importantly, technologies and tools continue to develop and improve in relation to the prevention of access to certain material, including class 2 material. Subsection 12(2)(c) recognises this and provides an example of continually seeking to develop, support or source, and implement improved technologies and processes. The Explanatory Statement to the 2024 Amendment Determination notes that this could be done by developing a service's own, improved technologies or through supporting or sourcing external technologies.

Relevant industry standard measures

Providers subject to the DIS Standard should note that the 'high impact DIS' category includes websites with the predominant purpose of enabling access to high impact materials (R18+, X18+ or RC) posted by end-users, such as pornography sites.[53] Key obligations of Tier 1 services apply in relation to class 1A and 1B material. However, some obligations needed to protect children from this material, have the same impacts for children accessing class 1C and class 2 material, for example preventing end-users known to be under 18 from using high impact services, and requiring that only account holders can post or distribute material on the services.

[53] See section 6 of the DIS standard on eSafety's register of industry codes and industry standards: Register of industry codes and industry standards for online safety | eSafety Commissioner.

Chapter 3: Expectations regarding reports and complaints

Division 4 of the Determination sets out expectations in relation to:

- Section 13: mechanisms to report and make complaints
- Section 14: terms of use, certain policies etc.
- Section 15: mechanisms to report and make complaints about breaches of terms of use
- Section 16: accessible information on how to complain to the Commissioner

Section 13 of the Determination – Providing mechanisms to report and make complaints about certain material

Determination, section 13:

Core expectation

1. The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable end users to report, and make complaints about, any of the following material provided on the service:
   a. cyber-bullying material targeted at an Australian child;
   b. cyber-abuse material targeted at an Australian adult;
   c. a non-consensual intimate image of a person;
   d. class 1 material;
   e. class 2 material;
   f. material that promotes abhorrent violent conduct;
   g. material that incites abhorrent violent conduct;
   h. material that instructs in abhorrent violent conduct;
   i. material that depicts abhorrent violent conduct.

Additional expectation

2. The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable any person ordinarily resident in Australia to report, and make complaints about, any of the following material provided on the service:
   a. cyber-bullying material targeted at an Australian child;
   b. cyber-abuse material targeted at an Australian adult;
   c. a non consensual intimate image of a person;
   d. class 1 material;
   e. class 2 material;
   f. material that promotes abhorrent violent conduct;
   g. material that incites abhorrent violent conduct;
   h. material that instructs in abhorrent violent conduct;
   i. material that depicts abhorrent violent conduct.

The intention of this section is to ensure that services have appropriate complaints processes for all Australians to report certain material regulated under the Act to a service, without the requirement to have an account with that service.[54]

Reporting and complaint mechanisms should be clear and readily identifiable to end-users and others at all relevant points in time when they are engaging with material, activity or other users.

Providers should conduct a safety risk and impact assessment of what harms and risks end-users and individuals ordinarily resident in Australia are likely to experience on their services, and design intuitive reporting options for end-users accordingly. This assessment should include accessibility requirements to ensure all end-users are able to effectively use the reporting options.

Additionally, providers should ensure that report and complaint mechanisms on their services are designed in a way that enables for the prioritisation of reports for escalation and rapid response – for example, reports that are likely to relate to unlawful material or activity or present a serious threat to life, health or safety.

Clear and readily identifiable reporting and complaint mechanisms are particularly critical as a safety intervention where providers are limited in their ability to deploy technologies on their services that proactively detect unlawful and harmful material and activity.

What is a 'clear' mechanism for reporting and making a complaint?

A reporting or complaint mechanism is more likely to be 'clear' if individuals are presented with a menu which contains an appropriate category or description of the issue that they want to report.

Issue-specific reporting options are important to empower individuals to clearly identify the reason they are concerned with the content, and to enable the provider to respond appropriately including by prioritising certain reports. For example, a specific CSEA reporting option is critical to ensuring that this extremely harmful, unlawful material is reported and able to be prioritised for review and action (such as banning the account and referral to law enforcement). This might be provided alongside a 'general' reporting category to ensure those who want to make a report that is not harm-specific also have the opportunity to do so.

Providers should offer a clear mechanism for individuals who do not have an account with the service to report material or other end-users, without the need to create an account themselves. This is important where material or activity may be impacting an individual who is not an end-user of the service – for example, cyberbullying or other abusive material where the victim is not an end-user of the service where the material is being shared.

---

[54] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 17: Federal Register of Legislation - Online Safety (Basic Online Safety Expectations) Determination 2022.

If a service is known, or likely, to be used in a way that facilitates extremely harmful, unlawful activity such as CSEA and the promotion of terrorism, it is unlikely that the provider can demonstrate compliance with section 13 if they do not have a specific reporting option for these categories (for example, if a service requires individuals to rely on broad reporting options such as 'inappropriate content' or 'sexual activity' to report this unlawful content).

Individuals should also be provided with relevant information, at the time of reporting, about how their personal information will be used (if at all) as a result of making a report or a complaint, to ensure individuals feel comfortable, informed and empowered to make a genuine report or complaint without fear of consequences. Providers should consider eliminating barriers to reporting and complaints, such as requirements to provide personal information or to follow multiple steps to locate reporting options.

What is a 'readily identifiable' mechanism for reporting and making a complaint?

A reporting option is 'readily identifiable' if it can be quickly and easily accessed and used by an individual without barriers, at every part of the user experience. For example, reporting and complaint mechanisms should:

- be provided on all aspects of a service so that an individual can report all relevant material and activity – including material they have seen in a post, a livestream, a video chat or direct communication, or activity by another end-user or by a group or forum
- enable individuals to report and complain about material that an individual has knowledge of but does not have direct access to (for example, an intimate image that they know has been shared on a service, but the individual does not know where on the service, or who has access to it)
- be accessible in-service at the point in which the individual wishes to flag material, meaning they can report content without needing to navigate to a separate part of the service or exit the service to report via email or complaint form
- be available to all end-users of a service, regardless of whether they have an account, or are logged in or not
- be consistently accessible for individuals where a service may be accessed via an app or browser or via desktop
- ensure a seamless process for material of concern to be identified to the provider (for example, report and complaint mechanisms should be designed so they automatically flag and preserve the material in question for review by the service)
- not require individuals to take screenshots, save links or otherwise create their own copy of the material in order to make a report or complaint to the service, although this additional functionality may be useful to individuals.

Relevant industry code and industry standard measures

There are requirements under the SMS Code and RES and DIS Standards in relation to reporting mechanisms and class 1 material. For example, some providers are required to provide user reporting tools which are 'on-platform' or 'in-service' or which allow users to specify the harm associated with the material the user wishes to report.

Section 14 of the Determination – Providing terms of use and certain policies and procedures regarding reports, complaints and conduct

Determination, section 14:

1. The provider of the service will ensure that the service has:
   a. terms of use; and
   b. policies and procedures in relation to the safety of end-users; and
   c. policies and procedures for dealing with reports and complaints mentioned in section 13 or 15; and
   d. standards of conduct for end-users (including in relation to material that may be posted using the service by end-users, if applicable), and policies and procedures in relation to the moderation of conduct and enforcement of those standards.

Note 1: see section 17 in relation to making this information accessible to end-users.

Note 2: for paragraph (b), the policies and procedures might deal with the protection, use and selling (if applicable) of end-users' personal information.

1A. The provider of the service will take reasonable steps (including proactive steps) to detect breaches of its terms of use and, where applicable, breaches of policies and procedures in relation to the safety of end-users, and standards of conduct for end-users.

2. The provider of the service will take reasonable steps (including proactive steps) to ensure that any penalties specified for breaches of its terms of use, policies and procedures in relation to the safety of end-users, and standards of conduct for end-users, are enforced against all accounts held or created by the end-user who breached the terms of use and, where applicable, breached the policies and procedures, and standards of conduct, of the service.

3. The provider of the service will, within a reasonable period of time:
   a. review and respond to reports and complaints mentioned in sections 13 and 15; and
   b. take reasonable steps to provide feedback on the action taken.

4. For the purposes of subsection (3), in determining 'a reasonable period of time', the provider must have regard to:
   a. the nature and impact of the harm that is the subject of the report or complaint;
   b. the complexity of investigating the report or complaint; and
   c. any other relevant matters.
5. For the purposes of paragraph (3)(a):
   a. review means considering a report or complaint from when it is first made; and
   b. respond means taking and implementing a decision to have content removed and reported, have an end-user banned, or other content moderation decisions, or a decision to take no action.

Terms of use, standards of conduct, policies and procedures in relation to online safety

Terms of use, standards of conduct, policies and procedures are key mechanisms for providers to communicate what is and is not allowed on their service (in terms of both material and activity). They are also important mechanisms for providing a clear and transparent rationale for action a provider may take to address unlawful and harmful material and activity on the service.

Some providers refer to the relevant parts of terms of use, standards of conduct, policies or procedures as community guidelines, community standards or rules. eSafety considers these important mechanisms to be interrelated and core to ensuring safe use of a service. A provider should set out clearly how standards of conduct and/or relevant policies are linked to terms of use of a service.

It is expected that terms of use, policies and procedures and standards of conduct will be clear, explicit and easy to understand. One of the factors eSafety is required by the Act to consider in determining whether to give a reporting notice is 'whether there are deficiencies in a service's terms of use, so far as they relate to the capacity of end-users to use the service in a safe manner'.[55]

Relevant industry code and industry standard measures

Certain social media service providers, relevant electronic services and designated internet service providers are also required to ensure their service's policies and terms of use regarding treatment of class 1A and 1B material, meets the requirements set out in the applicable industry code or industry standard.

[55] See sections 56(5)(d) (non-periodic reporting notice) and 49(5)(d) (periodic reporting notice) of the Act.

What online safety harms should be covered in terms of use, policies and procedures and standards of conduct?

Terms of use should prohibit activity and material that is unlawful and harmful. At a minimum, providers should ensure that their terms of use and other policies align generally with the unlawful and harmful matters dealt with under the Act (the matters specified in section 13 of the Determination). Additional harms suggested in the Explanatory Statement to be covered by terms of use and other policies include, but are not limited to:

- hate against a person or group of people on the basis of race, ethnicity, disability, religious affiliation, caste, sexual orientation, sex, gender identity, disease, immigrant status, asylum seeker or refugee status, or age
- promotion of suicide and self-harm, such as pro-anorexia content, that does not meet the threshold of class 1 or class 2 material
- high volume, cross-platform attacks that have a cumulative effect that is damaging but does not meet the threshold of adult cyber abuse when reported as singular comments or posts
- promotion of dangerous 'viral' activities that have the potential to result in real injury or death.

Providers should consider whether their terms of use, policies, procedures and/or standards of conduct effectively address the range of harms and risks that currently do, or may, arise on their service. Providers are best placed to identify emerging forms of harmful end-user conduct or material, and are afforded flexibility by the Determination to choose the best and most responsive way to address them on their service. Providers should update their terms of use, standards of conduct and other policies and procedures as new risks and harms emerge over time.

Where a provider provides multiple services, there should be service-specific terms of use, policies and procedures that are tailored to the service and any particular safety risks or harms posed by, or to, end-users of that service. It may not be sufficient for a service to rely on high-level, broad terms of use that do not clearly and explicitly set out what material and activity is prohibited or restricted, and how the service enforces these rules.

Detecting breaches of terms of use, policies and procedures and standards of conduct (subsection 14(1A))

Providers are expected to take reasonable steps, including proactive steps, to detect breaches of its terms of use and, where applicable, breaches of relevant policies and procedures and standards of conduct.

It is unlikely to be sufficient for providers of services with risks of online harm occurring to rely on end-users or individuals to make reports or complaints as the primary means of detecting relevant breaches.

Proactive steps may include the use of tools and technology that detect material and activity either before it is created, uploaded or shared, or immediately after it is provided on a service, such as:

- hash matching technology – including the use of verified hash databases to detect known unlawful material such as CSEA and terrorism, as well as hash matching used by a service to detect material already known to the service
- AI classifiers to detect new material that is likely to breach terms of use, policies and procedures or standards of conduct
- language, text and audio analysis technology which can identify a wide range of material and activity
- remaining alert and detecting ongoing patterns of unlawful and harmful behaviour in breach of terms of use, policies and procedures or standards of conduct.

Dealing with reports and complaints (subsections 14(3), (4) and (5))

Section 14 sets out expectations in relation to how providers will deal with reports and complaints, including that providers:

- will have policies and procedures for dealing with reports and complaints (subsection 14(1)(c))
- will, within a reasonable period of time, review and respond to reports and complaints and provide feedback on the action taken (subsection 14(3)).

Providers should have clear policies and procedures for dealing with reports and complaints and should take steps to communicate these to individuals. For example:

- Users should be provided with confirmation that their report or complaint has been received, and an indication of when they will receive a response from the provider – this could include providing the user with a receipt, reference or report number in relation to the report or complaint.
- Policies and procedures should include clear guidance on when reporting to external bodies is required – for example, to law enforcement bodies or in response to a request from eSafety.
- Providers are also expected to provide feedback on the action taken within a reasonable period of time. It may be the case that it is not reasonable to provide feedback in relation to every report or complaint made (for example, reports or complaints that are vexatious or without merit) although it is expected that for

legitimate reports and complaints, services will do this.

Providers should also have internal policies and procedures for prioritising and responding to reports or complaints that are likely to relate to unlawful material or activity, or present a serious threat to life, health or safety. Prioritisation of reports is important to ensure that the review and response time from the point at which the report was made is 'reasonable'.

Relevant factors in determining a 'reasonable period of time' in which to review and respond to reports and complaints include as per subsection 14(4):

- the nature and impact of the harm that is the subject of the report or complaint
- the complexity of investigating the report or complaint
- any other relevant matters.

The Explanatory Statement to the 2024 Amendment Determination emphasises that it is important that services address reports of harm as quickly as is reasonably possible to minimise the potential harm and provide feedback to the complainant on the outcome of their report. Timely action and informing users of decisions taken is important for effectiveness and transparency, but also assists users who may wish to subsequently report material to the Commissioner.

It is expected that providers will respond promptly to the most severe harms reported or complained about and that providers have the resources commensurate with the size and risk of their service to enable this outcome.

Relevant industry code and industry standard measures

'Other relevant matters' for the purposes of determining a reasonable period of time in which to review and respond to reports and complaints may include any time frames in industry codes or industry standards for responding to reports or complaints of class 1 material. For example, if a provider is compliant with a stipulated time frame or requirement in terms of responding to reports or complaints under an industry code or standard, eSafety is likely to consider that this constitutes a 'reasonable period of time'.

- The SMS Code requires providers of Tier 1 and Tier 2 social media services to take appropriate steps to respond 'promptly'.
- The RES and DIS Standards require that all relevant electronic services and identified designated internet services:
    - respond 'promptly' to the end-user to acknowledge their complaint or report
    - take appropriate and timely action to investigate the complaint and inform them of the outcome.

Reasonable steps to enforce penalties for relevant breaches (subsection 14(2))

In addition to setting out clear and comprehensive terms of use and policies relating to the safety of end-users, it is expected that providers will also have in place effective systems to enforce terms of use. It is also expected that providers will enforce any standards of conduct and policies included or incorporated in the terms of use. This would include providers making appropriate enquiries into any suspected breaches of terms of use, standards of conduct or other relevant policies.

Providers should consider a range of enforcement options and apply these in a manner that is proportionate to the nature of the breach. Enforcement against breaches should also have regard to issues such as minimising the risk of material or activity occurring again, including by banning accounts where there are severe breaches of the terms of use. More serious breaches are likely to require a more significant response.

Providers should also be able to explain these steps to eSafety in relation to an investigation or other escalation.

Options to enforce breaches of terms of use may include:

- warnings and strikes, nudges and prompts to end-users
- requiring an end-user to review certain safety information
- removing certain privileges or functionality for an end-user (such as the ability to monetise or livestream content, or removal of a 'credibility' or similar badge)
- account blocking or account limiting (or blocking or limiting content)
- removal of an account, or content
- requiring an end-user to apologise, in appropriate circumstances
- account suspension – accounts may be de-activated or suspended for a temporary period of time, and alerts may be sent to give the end-user time to address the issue
- disabling an account – accounts may be permanently disabled so they are no longer visible or active
- down-rank content – demote content visibility for some or all content posted by an end-user
- geo-blocking or geo-IP-blocking.

Reasonable steps which support the effective and consistent enforcement of penalties for breaches of terms of use may include:

- ensuring content moderation staff – including community or volunteer moderators - are trained to apply these terms of use, content guidelines and other internal guidelines consistently and objectively

- ensuring transparency regarding these enforcement processes and outcomes, and publish relevant information in a regular transparency report or other safety report
- ensuring terms of use and policies and procedures are regularly reviewed and updated as needed – this could be done as part of regular safety risk assessments
- ensuring effective measures are in place to detect end-users who attempt to re-register or regain access to a service when they have been banned, or had other enforcement action taken against them, and to prevent this recidivism (see chapter 1, section 9 on anonymous accounts for examples of steps that may be taken to address recidivism)
- appropriately resourcing trust and safety teams and content moderation teams.

It is unlikely to be sufficient for a service to only refer individuals who make reports or complaints about breaches of terms of use to external sources of support and to take no further steps to address the material and/or account that is the subject of the report or complaint, including to prevent future harm on the service. For example, for a severe or repeated breach of terms of use, policies and procedures or standards of conduct, the service should also take appropriate action such as banning the account.

Relevant industry code and industry standard measures

Certain social media service, relevant electronic service and designated internet service providers are required under industry codes and industry standards to enforce their terms of use by taking appropriate action, including by removing class 1A material on the service and ensuring that related breaches cease.

Section 15 of the Determination – Providing mechanisms to report and make complaints about breaches of terms of use

Determination, section 15:

Core expectation

1. The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable end users to report, and make complaints about, breaches of the service's terms of use.
2. The provider of the service will ensure that the service has clear and readily identifiable mechanisms that enable any person ordinarily residing in Australia to report, and make complaints about, breaches of the service's terms of use and, where applicable, breaches of the service's policies and procedures and standards of conduct mentioned in section 14.

The purpose of this section is to provide an avenue for all Australians to have material or activity that breaches a service's terms of use removed or otherwise dealt with in an appropriate manner by the service without requiring them to have an account with a particular service.[56]

For example, an individual may be aware that harmful material relating to them, which breaches the terms of use of a service, is accessible on a service that they do not have an account with or otherwise engage with. Providers should ensure that individuals (and, in certain circumstances, their parent or guardian) are not prevented from reporting or complaining about a breach of a service's terms of use because they do not have an account.

Providers should consider the list of steps set out in this guidance in relation to section 13 (reporting and complaints about certain material) as these are also relevant to providing reporting and complaint mechanisms in relation to breaches of terms of use, policies and procedures and standards of conduct.

Section 16 of the Determination – Providing access to information on how to complain to the eSafety Commissioner

Determination, section 16:

The provider of the service will ensure that information and guidance on how to make a complaint to the Commissioner, in accordance with the Act, about any of the material mentioned in section 13 provided on the service, is readily accessible to end-users.

The purpose of this expectation is to make end-users in Australia aware that they can make complaints to the Commissioner regarding material included in section 13 of the Determination.

It is at the discretion of providers to decide how they provide this information, and providers have flexibility to design their services in a way that best supports end-users with important safety information, including that a complaint can be made to eSafety in relation to certain material and activity. Providers may choose to make this information accessible at all points of the end-user experience, or at the point of account creation or first use, or at regular intervals, or in a sequence appropriate for that services' complaints process.

However, end-users should be provided with this information in a clear and readily accessible manner at the point when they report material to the service and when they complain to the service. This is important because complaining to a service is a necessary first step for end-

[56] See Explanatory Statement, Online Safety (Basic Online Safety Expectations) Determination 2022, page 19: Federal Register of Legislation - Online Safety (Basic Online Safety Expectations) Determination 2022.

users who are seeking removal of cyberbullying material directed at a child[57] or of adult cyber abuse, under eSafety's complaint schemes.

Additionally, this information (including direct links to information on the eSafety website about how to make a complaint) should be clearly set out in appropriate documents and links on a service, such as in the terms of use, community guidelines, safety centre or other safety resources.

Chapter 4: Expectations regarding accessible information

Section 17 of the Determination – Providing access to information on terms of use, policies, complaints and similar topics

Determination, section 17:

1. The provider of the service will ensure that the information specified in subsection (2) is:
   a. readily accessible to end-users; and
   b. in relation to the information mentioned in paragraph (2)(b)—accessible at all points in the end-user experience, including, but not limited to, point of purchase, registration, account creation, first use and at regular intervals (as applicable); and
   c. regularly reviewed and updated; and
   d. written in plain language.

2. For the purposes of subsection (1), the information is the following:
   a. the terms of use, policies and procedures and standards of conduct mentioned in section 14;
   b. information regarding online safety and parental control settings, including in relation to the availability of tools and resources published by the Commissioner.

This expectation relates to the provision, accessibility, review and presentation of information regarding a service's terms of use and information about online safety and parental control settings – including in relation to the availability of tools and resources published by eSafety.

[57] For more information, see eSafety's regulatory guidance on the Cyberbullying Scheme: Regulatory schemes | eSafety Commissioner.

eSafety has published a suite of tools and resources on the eSafety website[58] that providers could provide to end-users to supplement their own safety information such as terms of use, policies, procedures and standards of conduct.

The information specified in subsection 14(2) should be simple and as easy as possible for users to locate and to use, to make their (or their children's) user experience as safe and age-appropriate as possible. This is particularly important when a user is registering to use a service or using the service for the first time, but it is also important that the information is easy to find throughout a user's experience of the service.

Where a user has indicated to a service that they are seeking specific information, such as information for parents, services should provide relevant eSafety resources at that point in time to assist end users.

For the purpose of subsection 17(1)(b), provision of this information at 'regular intervals' may be satisfied through adhering to the section 18 expectation.

Information should be written in plain language and should be provided in multiple languages to ensure end-users are able to understand key safety information. Information should also be age-appropriate to suit the developmental needs of children if a service permits or has child-users.

Section 18 of the Determination – Providing updates about changes in policies, terms and conditions

Determination, section 18:

The provider of the service will ensure that end-users receive updates written in plain language in relation to changes in the information specified in subsection 17(2), including through targeted in-service communications.

Providers should ensure that end-users receive updates in plain language regarding changes to the terms of use, policies, procedures and standards of conduct and information available about online safety and parental control settings, including through targeted in-service communications.

[58] eSafety website: Online safety | eSafety Commissioner.

Depending on the nature of the update, end-users could be required to confirm that they understand the changes and how they will be impacted – for example, if terms of use are updated to prohibit certain activity or material, end-users should be required to confirm that they have read and understood this and agree to abide by this rule.

These updates should be provided in multiple languages to support end-users and should also be age-appropriate to suit the developmental needs of children and young people, if a service permits younger users.

Providers may choose how to best present these updates to end-users, including through age-appropriate means to young people and children. Infographics, videos, tiered notices and other measures to ensure end-users are able to understand the updates and how this impacts their safety experience may all be appropriate.

Chapter 5: Expectations regarding record keeping

Section 19 of the Determination – Keeping records regarding certain matters

> **Determination, section 19:**
>
> The provider of the service will keep records of reports and complaints about the material mentioned in section 13 provided on the service for 5 years after the making of the report or complaint to which the record relates.

The purpose of this expectation is to ensure providers can provide the Commissioner with information on complaints about the material in section 13 and how the provider actioned the complaints.

This information will help the Commissioner assess the effectiveness of complaint and moderation practices over time and point out areas where services are doing this well, as well as areas where improvements could be made.

Providers should retain an appropriate amount of detail in these records to assist the Commissioner in assessing the adequacy of a service's response to reports and complaints.

For example, where a report or complaint is made to the service about certain material, the service should include in its record:

- the mechanism by which the end-user made the report – such as through in-service reporting, or via a webform or email
- the specific category of material reported – both as reported by the end-user and as designated or established by the service

- the service's response to the report or complaint (such as any content moderation decision like material removed, report made to a law enforcement body, or enforcement action taken against the offending end-user)
- the time taken to respond to the report or complaint
    - this should include an overall indication of the time taken, from the point at which an end-user made a report to the point where action was completed by the service
    - this could also include more specific information such as the time taken for a report to be flagged to a specialist team for review and action, and any re-review required or other escalation.

Where certain enforcement action is taken against end-users as a result of a report or complaint, such as a permanent ban from the service, records could include details about offending end-users to ensure they are prevented from re-registering or accessing the service.

Where records of reports and complaints contain personal information, including sensitive information or information that is likely to be perceived as sensitive to end-users, providers are expected to ensure this information is subject to robust privacy protections.

Records should be kept for five years. Providers are not expected to have five years of records until at least five years following the making of the Determination.[59]

eSafety recognises that some jurisdictions may prevent providers from storing relevant data for this period of time, and will have regard to this when assessing compliance with this expectation.

Taking steps to ensure an appropriate level of detail is retained in records under this section is likely to support providers in responding to Commissioner information requests, as set out in section 20.

Relevant industry code and industry standard measures

Social media service providers are also required to keep records in relation to the measures they have adopted to comply with the SMS Code.

The RES and DIS Standards require that providers keep records that set out the actions that the provider has taken to comply with the relevant industry standard for at least two years.

[59] The Online Safety (Basic Online Safety Expectations) Determination 2022 was registered on 23 January 2023: Federal Register of Legislation - Online Safety (Basic Online Safety Expectations) Determination 2022.

Chapter 6: Dealings with the Commissioner

Section 20 of the Determination – Providing requested information to the Commissioner

Determination, section 20:

Core expectation

1. If the Commissioner, by written notice given to the provider of the service, requests the provider to give the Commissioner a statement that sets out the number of complaints made to the provider during a specified period (not shorter than 6 months) about breaches of the service's terms of use, the provider will comply with the request within 30 days after the notice of request is given.
2. If the Commissioner, by written notice given to the provider of the service, requests the provider to give the Commissioner a statement that sets out, for each removal notice given to the provider during a specified period (not shorter than 6 months), how long it took the provider to comply with the removal notice, the provider will comply with the request within 30 days after the notice of request is given.
3. If the Commissioner, by written notice given to a provider of the service, requests the provider to give the Commissioner specified information relating to the measures taken by the provider to ensure that end users are able to use the service in a safe manner, the provider will comply with the request within 30 days after the notice of request is given.

Additional expectation

4. If the Commissioner, by written notice given to a provider of the service, requests the provider to give the Commissioner a report on the performance of online safety measures that relevant providers have announced publicly or reported to the Commissioner, the provider will comply with the request within 30 days after the notice of request is given.

Additional expectation

5. If the Commissioner, by written notice given to a provider of the service, requests the provider to give the Commissioner a report on the number of active end-users of the service in Australia (disaggregated into active end-users who are children and those who are adult end-users) during a specified period, the provider will comply with the request within 30 days after the notice of request is given.

The information which the Commissioner may request from a provider under section 20 can be directly relevant to how services are meeting the Expectations. For example, information requested under subsection 20(1) (number of complaints) can provide the Commissioner with

an indication of how effectively terms of use are communicated to users and enforced by a provider. It is also relevant to how a provider is ensuring safe use of a service, including by taking reasonable steps to proactively minimise the provision of certain material (section 6).

For more information on the various reporting powers and options, including a section 20 request for information, see page 13 of this guidance.

It is at the discretion of the provider to provide additional information regarding complaints (for example, how many were deemed vexatious, how many did not meet a threshold for action, how complaints were resolved), however providers should consider what additional information or context they could include in response to a section 20 request, as this would assist in better understanding and assessing how a provider is ensuring safe use of their service and meeting the Expectations.

For example, where complaints about breaches of terms of use indicate an increase in a specific type of harmful activity or trend, it is useful to provide additional, information (such as improved reporting options, updated terms of use or introduction of new safety features) which may be relevant to an increase in the number of reports.

Under subsection 20(2), the Commissioner may request a statement that, for each removal notice given to the provider during a specified period, sets out how long it took a provider to comply with the removal notice. This information will help the Commissioner assess how rapidly providers are complying with removal notices given under the Act's schemes.

Under subsection 20(3), the Commissioner may request information relating to the measures taken by the provider to ensure that end-users are able to use the service in a safe manner. The purpose of this expectation is to enable the Commissioner to request specified information concerning online safety measures being taken by a provider.

The Commissioner may also request a report on the performance of safety measures that it has publicly announced or reported to the Commissioner (subsection 20(4)). In practice, when a provider announces a significant new safety feature, that provider should expect to be asked by the Commissioner to report on the impact of that safety feature on the experience of end-users. The intention of this expectation is to address the scenario of a provider announcing a safety feature, but failing to disclose whether the feature was effective.

Providers should ensure they continually evaluate and assess safety features and collect relevant information about the performance of such measures, in order to comply with a subsection 20(4) request.

Under subsection 20(5), the Commissioner may request a report on the number of active end-users of the service in Australia, disaggregated into active end-users who are children and those who are adult. This information will assist the Commissioner in assessing the reach and prevalence of a service within Australia, and consequently the level of risk a service poses to Australian adults and children and whether certain steps taken to comply with the Expectations are reasonable. This will improve the Commissioner's capacity to

support Australians by identifying where Australians are most likely to need support and enable a more efficient deployment of resources.

The Commissioner may also consider information collected under subsection 20(5) notice as a relevant matter when deciding whether to issue a periodic or non-periodic reporting notice under Part 4, Division 3 of the Act. As noted on page 19, a service's reach and the profile of its end-users, including whether the service is used by children, is a factor that the Commissioner might consider relevant when deciding to give a reporting notice. The information may also be used by the Commissioner when assessing whether a particular safety tool, process, measure or policy, constitutes a 'reasonable step' towards implementing a particular expectation set out in the Determination.

eSafety may use section 20 requests for information as part of an escalation of regulatory engagement with providers. In the first instance eSafety may seek some of the information included in section 20 on an informal basis, including through regular engagement and specific queries. This reflects the regular and ongoing engagement that eSafety has with providers, and that some information can be shared through these mechanisms. This informal engagement helps inform eSafety regarding providers' practices, trends and specific risks.

However, where information is required for specific regulatory purposes, or if eSafety intends to publish relevant information for the purpose of improving transparency, eSafety may make a formal request through section 20.

A failure to respond or comply with a request through section 20 would provide the Commissioner with grounds to give and publish a statement to that effect. The Commissioner may also seek the information through a non-periodic or periodic reporting notice instead, which would carry civil penalties for non-compliance.

Section 21 of the Determination – Providing a designated contact point

Determination, section 21:

1. The provider of the service will ensure that there is an individual who is:
   a. an employee or agent of the provider; and
   b. designated as the service's contact point for the purposes of the Act.

Note: The provider of the service is expected to have a designated contact point regardless of whether the service has staff physically located in Australia.

2. The provider will ensure that the following: contact details of the contact point are notified to the Commissioner:
   a. an email address; and
   b. a phone number or voice chat address.
3. If there is a change to the identity or contact details of the individual designated as the service's contact point for the purposes of the Act, the provider will give the Commissioner written notice of the change within 14 days after the change.

Section 21 requires providers to notify eSafety of a designated contact point. Any changes must be notified to eSafety in writing within 14 days after the change. The designated contact point is not required to be physically located in Australia. Providers that choose to discontinue maintaining a physical presence in Australia, however, will still be expected to provide the Commissioner with a designated contact point.

In order to facilitate the sharing of contact details, and also to enable the sharing of other information, eSafety has established a webform for relevant providers. Providers are encouraged to use this webform. By completing and maintaining information via this form, eSafety will regard a provider as meeting the expectation under section 21. Contact details will not be made public without the consent of providers.

Contact details may be used for engagement on implementation of the Expectations, and on other online safety issues, as well as a point of contact for eSafety for communications related to the enforcement of the Act. Where eSafety has existing contacts, particularly those used for content removal notices and other engagement under the Act, these are likely to continue to be used. Providers may want to nominate these existing contacts for the purposes of section 21 to ensure consistency, or may choose alternative points.

The webform includes some voluntary questions that providers may answer (for example, details of terms of use and reporting processes). Where appropriate, this information may be published in the interests of transparency.

To access the webform link and share the relevant information, please contact: industrybose@esafety.gov.au.

Annex A

The following is an example of the guidance provided to previous notice recipients to assist in making submissions related to information which should not be published.

> 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 provided 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 following table; **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.
>
> eSafety will carefully consider your submissions in line with the guidance in Table 1. 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.

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 includes: ○ specific indicators that companies use

		○ 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		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.

eSafety.gov.au