FEDERAL COURT OF AUSTRALIA

eSafety Commissioner v X Corp (Civil Penalty) [2026] FCA 629

File number(s):	VID 1092 of 2023
Judgment of:	**WHEELAHAN J**
Date of judgment:	21 May 2026
Catchwords:	**COMMUNICATIONS LAW** — Twitter, Inc operated a social media service – the applicant issued a non-periodic reporting notice to Twitter, Inc pursuant s 56(2) of the *Online Safety Act 2021* (Cth) – Twitter, Inc merged into the respondent – respondent liable to comply with the notice – respondent admits that it contravened s 57 of the Act by failing to prepare a report in the manner and form required by the notice – application for declaration and pecuniary penalty – whether Court should grant relief sought – orders appropriate – relief granted **PRACTICE AND PROCEDURE —**application for order that certain confidential material held on the Court file not be available for inspection on the ground that disclosure of the material would enable third parties to understand and circumvent the respondent's cyber defences – relief granted
Legislation:	*Federal Court of Australia Act 1976* (Cth) ss 37AE, 37AF and 37AG *Federal Court Rules* rr 2.23 and 2.32 *Online Safety Act 2021* (Cth) ss 3, 56, 57 and 162 *Online Safety (Basic Online Safety Expectations) Determination 2022* *Regulatory Powers (Standard Provisions) Act 2014* (Cth) s 82
Cases cited:	*Australian Building and Construction Commissioner v Pattinson* [2022] HCA 13; 274 CLR 450 *Australian Competition and Consumer Commission v MSY Technology Pty Ltd* [2012] FCAFC 56; 201 FCR 378 *Commonwealth v Director, Fair Work Building Industry Inspectorate* [2015] HCA 46; 258 CLR 482 *X Corp v eSafety Commissioner* [2025] FCAFC 99; 311 FCR 205
Division:	General Division

Registry:	Victoria
National Practice Area:	Commercial and Corporations
Sub-area:	Regulator and Consumer Protection
Number of paragraphs:	30
Date of hearing:	21 May 2026
Solicitor for the Applicant:	Australian Government Solicitor
Counsel for the Applicant:	Mr C Tran
Solicitor for the Respondent:	Thomson Geer
Counsel for the Respondent:	Mr P Herzfeld SC with Mr H Rogers

ORDERS

VID 1092 of 2023

BETWEEN:	**ESAFETY COMMISSIONER** Applicant
AND:	**X CORP.** Respondent

ORDER MADE BY: WHEELAHAN J

DATE OF ORDER: 21 MAY 2026

THE COURT DECLARES THAT:

1. The respondent contravened s 57 of the *Online Safety Act 2021* (Cth) between 29 March 2023 and 5 May 2023 by providing a report in response to a non-periodic reporting notice issued on 22 February 2023 by the eSafety Commissioner to Twitter, Inc under s 56 that was not prepared in the manner and form specified in the notice, in that the respondent:

   (a) did not respond to the following questions in the notice in circumstances where it was capable of doing so by the deadline of 29 March 2023: 14(b), 14(e), 19(a), 26, 26(a); and

   (b) did not respond to the following questions in the notice to the extent it was capable of doing so by the deadline of 29 March 2023: 1(b), 1(c), 3(b), 4(b), 6(c), 7(a)–(f), 10, 16(a), 17(b), 22(a)–(i), 25 and 27.

THE COURT ORDERS THAT:

2. Within 45 days of the making of this order, the respondent pay to the Commonwealth of Australia, a pecuniary penalty in the sum of $650,000 pursuant to s 162 of the *Online Safety Act 2021* (Cth) and s 82 of the *Regulatory Powers (Standard Provisions) Act 2014* (Cth).

3. Within 45 days of the making of this order, the respondent pay a contribution to the applicant's costs of the following proceedings:

   (a) *eSafety Commissioner v X Corp* (VID 1092 of 2023);

   (b) *X Corp v eSafety Commissioner* (VID 956 of 2023); and

(c) *X Corp v eSafety Commissioner* (VID 1186 of 2024).

in the agreed lump sum of $100,000.

4. Pursuant to r 2.23(3)(a) of the *Federal Court Rules 2011* (Cth), Annexures C and D to the statement of agreed facts and admissions filed on 1 May 2026 are ordered to be confidential with the consequence that until further order no person who is not a party be entitled to inspect those annexures on the Court file.

5. For the avoidance of doubt order 4 does not preclude any of the following persons from accessing the documents referred to in that order:

   (a) Judges of this Court;

   (b) necessary Court staff (including transcription service providers);

   (c) the parties to these proceedings;

   (d) legal representatives of the parties instructed in these proceedings; and

   (e) judicial officers and necessary staff of any court hearing an appeal from any decision made in the course of this proceeding.

6. For the further avoidance of doubt, order 4 does not prevent disclosure of the information referred to in the annexures by a Commonwealth officer acting in the course of his or her duties.

Note: Entry of orders is dealt with in Rule 39.32 of the *Federal Court Rules 2011*.

REASONS FOR JUDGMENT

WHEELAHAN J:

1 The applicant, the eSafety Commissioner, is responsible for the administration and enforcement of the *Online Safety Act 2021* (Cth). The respondent, X Corp, is the operator of the social media service X, formerly known as Twitter.

2 This civil penalty proceeding arises out of a report that the respondent submitted to the Commissioner in response to a notice given to Twitter, Inc under s 56(2) of the Act. The liability of the respondent to comply with the notice resulted from the merger of Twitter, Inc into X Corp, which was the subject of the decision of the Full Court in *X Corp v eSafety Commissioner* [2025] FCAFC 99; 311 FCR 205.

3 The respondent admits that its response to the Commissioner's notice was not prepared in the manner and form required by the notice and that it thereby contravened s 57 of the Act which is a civil penalty provision.

Background

4 Division 2 of Part 4 of the *Online Safety Act* creates a framework for imposing minimum safety standards on social media and other internet service providers. Subsection 45(1) of the Act empowers the Minister to set basic online safety expectations that give effect to the core expectations by legislative instrument. Subsection 46(1) sets out the core expectations of service providers that a determination under s 45 must specify. Broadly, the core expectations are that service providers will take reasonable steps to ensure safe use of their platforms, co-operate with the Commissioner, minimise and limit access to certain material, and maintain a complaints mechanism. The *Online Safety (Basic Online Safety Expectations) Determination 2022* set the basic online safety expectations during the relevant period.

5 Division 3 of Part 4 of the Act creates a mechanism by which service providers can be required to provide the Commissioner with reports about their compliance with the minimum standards imposed under Division 2. Subsection 56(2) grants the Commissioner power to give a written notice to a service provider requesting that it prepare a report about the extent to which it complied with the basic online safety expectations in the manner and form specified in the notice. Section 57 is a civil penalty provision, which requires service

providers to comply with a notice under s 56(2) to the extent that they are capable of doing so.

6 On 22 February 2023, the Commissioner issued a notice requiring Twitter, Inc to prepare by 29 March 2023 a report on its compliance with certain basic online safety expectations regarding child sexual exploitation and abuse material and activity during the period from 24 January 2022 to 31 January 2023. Schedule B to the notice prescribed the manner and form of the required report, which consisted of a list of 31 questions. For each question, there was typically one sub-question requiring an affirmative or negative response or a numerical figure followed by sub-questions requiring further explanation.

7 On 15 March 2023, Twitter, Inc merged into the respondent, X Corp.

8 On 29 March 2023, the respondent provided a report in response to the Commissioner's notice.

9 On 6 April 2023, the Commissioner wrote to the respondent identifying deficiencies in the report, including instances where no response was provided to a question or where the response failed to provide certain details required by the question.

10 On 5 May 2023, the respondent provided a supplementary response addressing the deficiencies raised in the Commissioner's letter.

11 The respondent admits that it contravened section 57 of the Act by failing to prepare a report in the manner and form specified in the notice to the extent it was capable of doing so by 29 March 2023. Specifically, the respondent admits that it –

(a) did not respond at all to questions 14(b), 14(e), 19(a), 26 and 26(a) notwithstanding that it was capable of doing so; and

(b) did not respond to the extent that it was capable of doing so to questions 1(b), 1(c), 3(b), 4(b), 6(c), 7(a)-(f), 10, 16(a), 17(b), 22(a)-(i), 25 and 27.

12 The parties agree that the respondent's contravention ended on 5 May 2023 when it provided its response to the deficiencies identified in the Commissioner's follow up letter.

13 The respondent's admissions of contravention are contained in a statement of agreed facts and admissions which was received into evidence. A copy of the statement of agreed facts

and admissions, with the redaction of two confidential annexures, is attached to these reasons at Schedule A.

The remedies sought by the Commissioner in this proceeding

14 The remedies sought by the Commissioner against the respondent are a declaration of contravention, a pecuniary penalty of $650,000, and an order that the respondent pay an agreed contribution of $100,000 towards the Commissioner's costs in this and the related proceedings in which the respondent unsuccessfully challenged its liability to comply with the notice that was given to Twitter, Inc.

General considerations

15 The declaration, penalty, and costs order are sought in the exercise of judicial power, and they are sought by consent. Although the relief is sought by consent the Court must still be persuaded that what is proposed is appropriate: *Commonwealth v Director, Fair Work Building Industry Inspectorate* [2015] HCA 46; 258 CLR 482 (***Agreed Penalties Case***) at [57] to [58] (French CJ, Kiefel, Bell, Nettle and Gordon JJ).

16 Because this is an adversarial proceeding of a civil nature there is considerable scope for the parties to agree on facts and consequences, including upon the remedies provided that they are appropriate: *Agreed Penalties Case* at [57]. The question whether remedies are appropriate acknowledges that there might be a range of appropriate remedies.

17 There is a public interest in the settlement of the controversy between the Commissioner and the respondent in relation to the respondent's contraventions of s 57 of the Act. In this respect, being a civil proceeding, it is consistent with the nature of civil proceedings for a court to make orders by consent and to approve a compromise of proceedings on terms proposed by the parties, provided the Court is persuaded that what is proposed is appropriate: *Agreed Penalties Case* at [57].

18 The parties' agreement as to liability and relief has facilitated the resolution of this dispute and thereby saved time and resources of the applicant and the Court: *Agreed Penalties* at [38]. The respondent's compliance with s 57 was capable of being an evaluative issue that might have been the subject of extensive argument. That arises from the qualification in s 57 that a person must comply with a notice under subsection 56(2) "to the extent that the person is capable of doing so". The time and resources saved by the parties reaching agreement in relation to the fact of contravention favours giving effect to the proposed settlement.

19 It is also important that the Court take into consideration the effect on other cases when considering whether to give effect to an agreement to compromise in a civil penalty proceeding. Other parties in other instances might be more willing to acknowledge contraventions and agree to submissions on penalty if there is a degree of predictability of outcome and some level of assurance that the Court will give effect to compromises where it is appropriate to do so: *Agreed Penalties Case* at [46].

The declaration is appropriate

20 In relation to the appropriateness of the declaration that is sought, in *Forster v Jododex Australia Pty Ltd* [1972] HCA 61; 127 CLR 421 at 437 to 438, Gibbs J cited the speech of Lord Dunedin in *Russian Commercial and Industrial Bank v British Bank for Foreign Trade Ltd* [1921] 2 AC 438 at 448 who summarised the principles applied by Scottish courts to the action of declarator and identified the following requirements –

> The question must be a real and not a theoretical question; the person raising it must have a real interest to raise it; he must be able to secure a proper contradictor, that is to say, some one presently existing who has a true interest to oppose the declaration sought.

21 It is in the public interest for the Commissioner, as the regulator, to seek a declaration and for a declaration to be made. The objects of the *Online Safety Act* are to improve and promote the safety of Australians online: s 3. In furtherance of those objects, the basic online safety expectations are directed to minimising the prevalence of abusive, violent or otherwise sensitive material. The reporting requirements under the Act are an essential aspect of enforcing those expectations. Accordingly, where the operator of a large social media platform has failed to comply with those reporting requirements, the public has an interest in the Commissioner seeking and obtaining a public declaration of contravention, which will contribute to a deterrent effect.

22 I am satisfied that the statement of agreed facts and admissions provides a sufficient factual foundation for making the declaration. I am satisfied that the terms of the proposed declaration relate to conduct that contravened the *Online Safety Act* and that the matters in issue have been identified with sufficient precision: *Australian Competition and Consumer Commission v MSY Technology Pty Ltd* [2012] FCAFC 56; 201 FCR 378 at [35] (Greenwood, Logan and Yates JJ).

The penalty is appropriate

23 In deciding whether to order an agreed penalty, the Court is not bound by the parties' proposal and must satisfy itself that the penalty is appropriate: *Agreed Penalties Case* at [48]. If the Court is persuaded that the agreed proposed penalty is appropriate in all the circumstances, it is desirable for the Court to accept the parties' proposal, owing to the importance of promoting the predictability of outcome in civil penalty proceedings: *Agreed Penalties Case* at [46], [58]. Rather than there being a single appropriate penalty, there is a permissible range of penalties in which "a particular figure cannot necessarily be said to be more appropriate than another": *Agreed Penalties Case* at [47].

24 Under s 162 of the *Online Safety Act* a civil penalty provision in the Act is enforceable under Part 4 of the *Regulatory Powers (Standard Provisions) Act 2014* (Cth) (the **Regulatory Powers Act**). As with other civil penalty regimes the principal if not the sole purpose of imposing a penalty in this instance is the promotion of the public interest in compliance with the *Online Safety Act* by general and specific deterrence of further contraventions: *Australian Building and Construction Commissioner v Pattinson* [2022] HCA 13; 274 CLR 450 (***Pattinson***) at [9] (Kiefel CJ, Gageler, Keane, Gordon, Steward and Gleeson JJ). In assessing whether a penalty is appropriate the Court must also consider the matters mandated by s 82(6) of the Regulatory Powers Act, namely –

(a) the nature and extent of the contravention;

(b) the nature and extent of any loss or damage suffered because of the contravention;

(c) the circumstances in which the contravention took place; and

(d) whether the person has previously been found by a court (including a court in a foreign country) to have engaged in any similar conduct.

25 The above matters inform the significance that should be attributed to the need for specific deterrence in this case. The parties jointly submitted that the Court should take into account the relatively short period of time over which the contravention occurred, the fact that the notice was the first non-periodic reporting notice issued to Twitter, Inc pursuant to the Act, and the significant time and resources required to respond to the notice having regard to the sensitive information requested and the manner in which the respondent's data were stored. Those matters provide some explanation as to why the breach occurred and suggest that the likelihood of further breaches is lower than it would otherwise be. Accordingly, I accept that they reduce the need for specific deterrence in this case.

26 While it is necessary to have regard to the maximum penalty, the maximum penalty is not to be reserved for the worst category of contravening conduct: *Pattinson* at [49] to [50]. The parties agreed and I accept that the maximum penalty for the respondent's contravention of s 57 of the Act was $687,500. The agreed penalty is near to the maximum penalty but reduced to some extent to account for the mitigating factors described above. I accept that the penalty proposed is appropriate. A penalty near the maximum is appropriate in the case of the respondent, which is a substantial corporation so that it operates as a real deterrent and is not simply a cost of doing business.

27 I also accept that it is appropriate that the parties have compromised the question of costs on a global basis, thereby saving the resources of the parties and the Court.

Access to confidential information

28 The respondent made an interlocutory application for an order pursuant to r 2.32(3)(a) of the *Federal Court Rules 2011* (Cth) that Annexures C and D to the statement of agreed facts and admissions be the subject of a confidentiality order with the consequence that inspection is restricted. The application was supported by an affidavit of the respondent's solicitor, Justin Healy Quill sworn 18 May 2026. The Commissioner did not oppose the application.

29 Under r 2.23(2A) a person who is not a party to a proceeding may inspect a statement of agreed facts. However, under r.2.23(3) a person is not entitled to inspect a document that the Court has ordered be confidential or restricted from publication. The note to r 2.32(3) refers to the powers of the Court to prohibit the publication of evidence in s 37AF of the *Federal Court of Australia Act 1976* (Cth).

30 I am satisfied having regard to the nature of the information that the respondent wishes to have treated as confidential that the order is necessary to prevent prejudice to the proper administration of justice, and the Commissioner did not submit otherwise. Annexures C and D comprise the respondent's responses to the Commissioner's notice. They contain detailed confidential information about the systems used by the respondent to detect and prevent the transmission of child sexual exploitation and abuse material on the X platform and were provided on a confidential basis under threat of penalty. Permitting inspection of this information might risk enabling actors to understand and therefore circumvent these systems. That would be contrary to the purpose of the *Online Safety Act*, which is to protect Australians online: s 3. The proper administration of justice would be prejudiced if the Court

did not preclude the inspection of material which would undermine the purpose of the Act being enforced in these proceedings. It is therefore necessary to make the order.

I certify that the preceding thirty (30) numbered paragraphs are a true copy of the Reasons for Judgment of the Honourable Justice Wheelahan.

Associate:

Dated: 21 May 2026

SCHEDULE A

No. VID 1092 of 2026

Federal Court of Australia
District Registry: Victoria
Division: General

ESAFETY COMMISSIONER

Applicant

X CORP.

Respondent

STATEMENT OF AGREED FACTS AND ADMISSIONS

A. INTRODUCTION

1. This Statement of Agreed Facts and Admissions (**SAFA**) is jointly filed by the Applicant, the eSafety Commissioner, and the Respondent, X Corp., pursuant to s 191 of the *Evidence Act 1995* (Cth).

2. The Applicant and the Respondent jointly seek the declarations and orders set out in **Annexure A** to this SAFA.

B. THE PARTIES

3. The Applicant is and was at all material times the statutory appointee with the powers and functions set out in the Act. The Applicant is and was at all relevant times responsible for the civil enforcement of the Act.

4. Twitter, Inc. (**Twitter, Inc.**) was a publicly traded company incorporated in the State of Delaware in the United States of America.

5. The Respondent is a large private company incorporated in the State of Nevada in the United States of America.

6. On 15 March 2023, Twitter, Inc. merged with and into the Respondent and ceased to exist.

1

C. BACKGROUND

7. At all material times, 'Twitter' was a social media service within the meaning of s 13(1) of the Act, as it:

   (a) is an "electronic service" within the meaning of s 5 of the Act:

      (i) the sole or primary purpose of which is to enable online social interaction between two or more end-users;

      (ii) which allows end-users to link to, or interact with, some of the other end-users; and

      (iii) which allows end-users to post material on the service; and

   (b) is not an exempt service.

8. The Twitter social media service will be referred to as the '**Twitter service**' in this SAFA.

9. At all material times, the Twitter service was provided through the domain https://twitter.com.

10. Twitter, Inc. was the provider of the Twitter service until 15 March 2023. On and from 15 March 2023, the Respondent was the provider of the Twitter service.

11. In or about March to May 2023, there were approximately 2 million Australian users of the Twitter service.

The notice issued to Twitter, Inc.

12. On 22 February 2023, the Applicant gave to Twitter, Inc. the Notice pursuant to s 56(2) of the Act.

13. The Notice required Twitter, Inc. to prepare and give to the Applicant a report in the manner and form specified in the Notice about the extent to which Twitter, Inc. complied with one or more specified applicable basic online safety expectations from 24 January 2022 to 31 January 2023 (**specified period**). The Notice asked about the measures that Twitter had in place during the specified period to combat child sexual exploitation and abuse material and activity. The Notice required a response by 29 March 2023, or a later date if any extensions were granted. A copy of the Notice is at **Annexure B** to this SAFA.

14. On 29 March 2023, the Respondent provided a response to the Notice. A copy of the 29 March 2023 response is at **Annexure C** to this SAFA.

15. On 6 April 2023, the Applicant wrote to the Respondent to ask clarifying questions about the 29 March 2023 response. The Applicant sought a response to these clarifying questions by 24 April 2023. The Applicant then agreed to extend that time to 5 May 2023.

16. On 5 May 2023, the Respondent provided a response to the clarifying questions. A copy of the 5 May 2023 response is at **Annexure D** to this SAFA.

D. FACTS RELEVANT TO PENALTY

17. Section 57 of the Act requires that a person must comply with a notice given under s 56(2) of the Act to the extent that the person is capable of doing so.

18. Section 57 is a civil penalty provision. At the relevant time, the value of a penalty unit was $275. The maximum penalty for a single contravention by a body corporate was $687,500.

19. The Applicant alleges, and the Respondent admits, that:

    (a) The Respondent contravened s 57 of the Online Safety Act 2021 (Cth) (**Act**) between 29 March 2023 and 5 May 2023 by providing a report in response to a non-periodic reporting notice issued on 22 February 2023 by the eSafety Commissioner to Twitter, Inc. (**Notice**) under s 56 that was not prepared in the manner and form specified in the Notice, in that the Respondent:

        (i) did not respond to the following questions in the Notice in circumstances where it was capable of doing so by the deadline of 29 March 2023: 14(b), 14(e), 19(a), 26, 26(a), and

        (ii) did not respond to the following questions in the Notice to the extent it was capable of doing so by the deadline of 29 March 2023: 1(b), 1(c), 3(b), 4(b), 6(c), 7(a) – (f), 10, 16(a), 17(b), 22(a)–(i), 25 and 27.

20. The Applicant does not allege that the contravening conduct continued after 5 May 2023.

21. The parties agree that the agreed penalty is the appropriate sum to resolve this proceeding.

22. The facts underpinning [19] are contained in **Annexures B, C, D and E** to this SAFA.

E. OTHER FACTS RELEVANT TO PENALTY

23. The Applicant regularly requires information from the regulated community of providers of social media services and other service providers about their compliance with the Basic Online Safety Expectations or 'BOSE' as determined by the Minister. The purpose of obtaining information about the extent to which providers have complied with the BOSE is to inform the Applicant in the discharge of its statutory functions, and typically involves publishing information to the public about the extent to which those providers have complied with the BOSE.

24. The Applicant considers that compliance by the providers of social media services, including the Respondent, with requirements to respond to non-periodic reporting notices assists regulators to put in place effective and informed mechanisms targeted towards minimising and eliminating Class 1 material and related activities online. Transparency also enables other service providers to understand the measures available to them, supporting more robust compliance.

25. No quantifiable loss resulted from the contravening conduct. The Applicant says that there was damage caused by the contravening conduct, in that it impeded the discharge of the Applicant's statutory function.

Cooperation by the Respondent

26. The Respondent has cooperated with the Applicant in the making of this SAFA. If the Respondent had not cooperated in this way and had instead contested this proceeding, the time and cost required by the parties and this Court would have been significantly greater.

Other relevant matters

27. The Respondent has not previously been found to have engaged in any similar conduct before. The Respondent also has not been previously found by a court to have contravened the Act.

28. On 3 October 2023, the Applicant publicly issued a service provider notification under s 48(2) of the Act, stating that the Respondent failed to comply with the Notice. The publication of that notification is likely to have reputational consequences for the Respondent. The publication of the service provider notification is a relevant consideration in the assessment of deterrence. It publicly signalled the importance of

compliance with reporting notices issued under the Act and the potential consequences of non-compliance.

29. The following factors stand in mitigation of the contravening conduct:

(a) the Notice was issued at a time in which Twitter, Inc. underwent significant corporate changes and associated internal personnel changes,

(b) the Respondent's relevant Safety personnel[1] prioritised their primary work of securing and ensuring the safety of the Twitter service, at a time where there were increased levels of attempts by bad actors to test and attack the Respondent's systems,

(c) the Notice requested sensitive information, requiring the Respondent to allocate significant resources to respond to the Notice,

(d) the Respondent's systems made the aggregation of data required to respond to the Notice a substantial undertaking,

(e) the data was not stored by the Respondent in a way that easily corresponded to certain Notice questions,

(f) the Notice was the first ever non-periodic reporting notice that was issued to Twitter, Inc. or the Respondent pursuant to the Act, and

(g) the Respondent subsequently provided clarifying responses on 5 May 2023.

Date: 1 May 2026

Signed by Matthew Garey
AGS lawyer for and on behalf of the Australian Government Solicitor
Lawyer for the Applicant

Date: 1 May 2026

Signed by Justin Quill
Thomson Geer
Lawyer for the Respondent

[1] Twitter, Inc. underwent a reduction in personnel as a result of its merger with the Respondent.

ANNEXURE A

AGREED FORM OF DECLARATIONS AND ORDERS

FEDERAL COURT OF AUSTRALIA
DISTRICT REGISTRY: VICTORIA
DIVISION: GENERAL

NO VID **1092 OF** 2026

ESAFETY COMMISSIONER

Applicant

X Corp.

Respondent

ORDERS

JUDGE: **WHEELAHAN J**

DATE OF ORDER:

WHERE MADE: Melbourne

BY CONSENT, THE COURT DECLARES THAT:

1. The Respondent contravened s 57 of the Online Safety Act 2021 (Cth) (**Act**) between 29 March 2023 and 5 May 2023 by providing a report in response to a non-periodic reporting notice issued on 22 February 2023 by the eSafety Commissioner to Twitter, Inc. (**Notice**) under s 56 that was not prepared in the manner and form specified in the Notice, in that the Respondent:

   (a) did not respond to the following questions in the Notice in circumstances where it was capable of doing so by the deadline of 29 March 2023: 14(b), 14(e), 19(a), 26, 26(a), and

   (b) did not respond to the following questions in the Notice to the extent it was capable of doing so by the deadline of 29 March 2023: 1(b), 1(c), 3(b), 4(b), 6(c), 7(a) – (f), 10, 16(a), 17(b), 22(a)–(i), 25 and 27.

BY CONSENT, THE COURT ORDERS THAT:

2. Within 45 days of the making of this order, the Respondent pay to the Commonwealth of Australia a pecuniary penalty in the sum of $650,000 pursuant to s 162 of the Act and s 82 of the *Regulatory Powers (Standard Provisions) Act 2014* (Cth).

3. Within 45 days of the making of this order, the Respondent pay a contribution to the Applicant's costs of the following proceedings:

   (a) *eSafety Commissioner v X Corp.* (VID 1092 of 2026),

   (b) *X Corp v eSafety Commissioner* (VID 956 of 2023), and

   (c) *X Corp v eSafety Commissioner* (VID 1186 of 2024),

   in the agreed lump sum of $100,000.

ANNEXURE B

NOTICE

22 February 2023

Our reference: BOSE-2023-124

Twitter, Inc.
c/o Justin Quill, Partner, Thomson Geer Lawyers
By email: jquill@tglaw.com.au

Dear Mr Quill,

Non-periodic reporting notice requiring you to prepare a report about the extent to which you have complied with the basic online safety expectations

Please see enclosed a non-periodic reporting notice (**the Notice**) given to Twitter, Inc. (**Twitter**) under section 56(2) of the *Online Safety Act 2021* (Cth) (**the Act**). The Notice requires Twitter to prepare a report, in the manner and form specified, about the extent to which it complied with one or more specified applicable basic online safety expectations during the specified period in respect of the following service:

1. Twitter, being a 'social media service' within the definition of section 13 of the Act.

The Notice requires Twitter to give the report to the eSafety Commissioner by 17:00 Australian Eastern Daylight Time on 29 March 2023.

If you require additional time to comply, you may request this by sending an email to industryBOSE@esafety.gov.au.

Definitions of relevant terms are included in **Schedule B** to the Notice.

The specified applicable basic online safety expectations

The specified basic online safety expectations are contained in the *Online Safety (Basic Online Safety Expectations) Determination 2022* (Cth), which commenced on 24 January 2022. See **Schedule A** to the Notice for full details of the specific expectations.

Section 6	Provider will take reasonable steps to ensure safe use
Section 8	Provider will take reasonable steps regarding encrypted services
Section 9	Provider will take reasonable steps regarding anonymous accounts
Section 10	Provider will consult and cooperate with other service providers to promote safe use
Section 11	Provider will take reasonable steps to minimise provision of certain material

Section 13	Provider will ensure mechanisms to report and make complaints about certain material
Section 14	Provider will ensure service has terms of use, certain policies etc.
Section 15	Provider will ensure service has mechanisms to report and make complaints about breaches of terms of use

The decision to give the Notice

I have decided to give the Notice under section 56(2) of the Act for the purpose of my statutory functions set out in section 27 of the Act, including to promote online safety for Australians, to monitor compliance with this Act, and to collect, analyse, interpret and disseminate information relating to online safety for Australians. In deciding whether to give a notice, I am required by section 56(5) of the Act to have regard to specific considerations. I have specified those considerations immediately below:

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;

- In the previous 12 months, eSafety has received high numbers of complaints regarding Twitter, including regarding child sexual exploitation and abuse (CSEA) material. I therefore consider Twitter to be a priority for a section 56(2) notice.

b) whether the provider has previously contravened a civil penalty provision of Division 3 (Reporting) of the Act:

- Twitter has not previously contravened a civil penalty provision of Division 3 (Reporting) of the Act. I note that I have not previously exercised my powers under Division 3 Part 4 of the Act in relation to Twitter nor are there any relevant reporting determinations in effect.

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;

- I have had regard to reports that Twitter has recently made signifcant cuts to staffing, and in particular roles responsible for trust and safety, and reductions in its Australia based workforce responsible for engaging with eSafety and local experts on safety issues. I am concerned that these staffing cuts may have diminished Twitter's operational capacity to respond to cases of CSEA, and I consider this to be a potential deficiency.
- I have also had regard to media reporting and internal and external expert advice concerning the use of Twitter by offenders to disseminate, and promote the sale and trade of CSEA material. I note that evidence has previously been provided to Twitter by eSafety of these issues.

- I also note media and non-governmental organisation reporting that CSEA, including known imagery, is available on Twitter, which I consider may show deficiencies relating to the systems, tools, and processes Twitter has in place to detect and proactively minimise CSEA on its service.

d) whether there are deficiencies in the service's terms of use, so far as they relate to the capacity of end users to use the service in a safe manner;

- I have had regard to Twitter's 'child sexual exploitation policy' which specifies the types of CSEA material and behaviour prohibited on the service. I do not regard this consideration as a reason for the giving of the Notice.

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;

- Twitter has not agreed to give the Secretary of the Department of Infrastructure, Transport, Regional Development, Communications and the Arts regular reports relating to the capacity of end users to use the service in a safe manner. As such, I consider this notice to be of assistance to my statutory functions, including those referred to above.

f) such other matters (if any) as the Commissioner considers relevant.

- In addition to the considerations above, I have had regard to:
  - The widespread use of Twitter, which I consider to heighten the risk of any harm arising from a failure to implement relevant safety expectations.
  - Twitter's livestreaming feature, and its publicly discussed plans to introduce end-to-end encryption on direct messages, which I regard to be features that pose potential risks for CSEA if deployed without reasonable measures to ensure safe use.
  - Very limited transparency concerning the systems, tools, processes, and human resources that Twitter deploys to prevent, detect, and remove CSEA on all parts of its service.

Required action

The Notice requires you to prepare a report, in the manner and form specified, about the extent to which you complied with one or more specified applicable basic online safety expectations during the specified period and give the report to the Commissioner by 17:00 Australian Eastern Daylight Time on 29 March 2023.

If you require additional time to comply with the Notice, please email industryBOSE@esafety.gov.au before 17:00 on 8 March 2023 with specific reasons why additional time is needed.

Failure to comply with the Notice may result in enforcement or compliance action being taken without further notice.

esafety.gov.au

Failure to comply

Under section 57 of the Act, you must comply with the requirement under the Notice to the extent that you are capable of doing so.

Failure to comply with the Notice is a contravention of section 57 of the Act and could result in the commencement of civil penalty proceedings for a civil penalty order of up to AUD$687,500 in respect of a body corporate or AUD$137,500 in respect of an individual for a single contravention, or other action as determined by the eSafety Commissioner.

False or misleading information or documents

It is a serious offence under the Criminal Code (Cth) to give false or misleading information or produce false or misleading documents to the eSafety Commissioner or the Office of the eSafety Commissioner.

Review rights

You have a right to seek an internal or external review of the decision to give you a notice under section 56(2) of the Act.

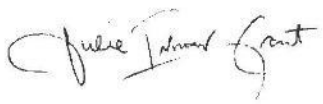
An internal review is a review conducted by the eSafety Commissioner under eSafety's Internal Review Scheme. There is no fee associated with a request for an internal review.

An external review is a review conducted by the Administrative Appeals Tribunal.

The enclosed information sheet sets out your rights in regard to the different review options available to you, as well as other options if you do not agree that the Notice should have been given to you.

Please note that you are required to comply with the Notice even if you have made an application for internal or external review, unless otherwise notified by a reviewer.

Regards



Julie Inman Grant
eSafety Commissioner

Enclosed: *Notice under section 56(2) of the Act*
Information Sheet: Review Rights

NON-PERIODIC REPORTING NOTICE TO PREPARE A REPORT ABOUT THE EXTENT TO WHICH THE PROVIDER OF A SOCIAL MEDIA SERVICE AND RELEVANT ELECTRONIC SERVICE COMPLIED WITH BASIC ONLINE SAFETY EXPECTATIONS

Under section 56(2) of the *Online Safety Act 2021* (Cth)

To: Twitter, Inc.

Attention: Justin Quill, Partner, Thomson Geer Lawyers

By: Email

This non-periodic reporting notice (**Notice**) is given to you under section 56(2) of the *Online Safety Act 2021* (Cth) (**the Act**) in respect of the following service that you provide in Australia:

1. Twitter, being a 'social media service' within the definition of section 13 of the Act.

You are required to prepare a report about the extent to which you complied with the applicable basic online safety expectations specified in **Schedule A** from 24 January 2022 to 31 January 2023 inclusive (**Report Period**).

The report must include answers to the questions specified in **Schedule B,** and use the template provided.

You are required to give the report to the Commissioner within 35 days of being given this notice (being no later than 17:00 Australian Eastern Daylight Time on 29 March 2023) or within such longer period as I allow if I am contacted by you with a request for extension.

The report is to be given to the Commissioner by email to industryBOSE@esafety.gov.au

Section 57 of the Act provides that a person must comply with a notice to the extent that the person is capable of doing so. Failure to comply with a notice may result in enforcement action including attracting a civil penalty of up to AUD$687,500 in respect of a body corporate or AUD$137,500 for an individual, for a single contravention. It is a serious offence under the Criminal Code (Cth) to give false or misleading information or produce false or misleading documents to the eSafety Commissioner or the Office of the eSafety Commissioner.

Date: 22 February 2023

Julie Inman Grant

eSafety Commissioner

E: industryBOSE@esafety.gov.au

esafety.gov.au

Schedule A – specified applicable basic online safety expectations

The Online Safety (Basic Online Safety Expectations) Determination 2022 (**Determination**) commenced on 24 January 2022, which specifies the basic online safety expectations for the Twitter Services that are the subject of this Notice (**Expectations**). The following table sets out the specified applicable basic online safety expectations to be reported on by Twitter for the purpose of this Notice:

6 Expectations—provider will take reasonable steps to ensure safe use

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.

8 Additional expectation—provider will take reasonable steps regarding encrypted services

(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.

9 Additional expectation—provider will take reasonable steps regarding anonymous accounts

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.

10 Additional expectation—provider will consult and cooperate with other service providers to promote safe use

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.

11 Core expectation—provider will take reasonable steps to minimise provision of certain material

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.

13 Expectations—provider will ensure mechanisms to report and make complaints about certain material

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;
(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.]

14 Additional expectations—provider will ensure service has terms of use, certain policies etc

(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 and 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.

15 Expectations—provider will ensure service has mechanisms to report about make complaints about breaches of terms of use

Core expectation

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

Note: Definitions of terms, such as 'class 1 material', 'abhorrent violent conduct', and 'cyber-abuse material' are provided in the Act.

Australian Government | eSafetyCommissioner

Information sheet: Right of Review

Internal review by the eSafety Commissioner

You have a right to seek an internal review of this decision under the Internal Review Scheme. An internal review is an impartial review of the merits of a decision. The purpose of an internal review is to consider whether the original decision made was the correct one.

You must make an application for an internal review **within 30 days** of receiving the notice of this decision. If you are unable to make your application within 30 days, please email internalreview@eSafety.gov.au.

There are no fees associated with an application for internal review.

To request an internal review, you will need to download and complete the **Request for internal review form** available on eSafety's website: www.esafety.gov.au/about-us/corporate-documents/internal-review.

Please fill the form out and email it or post a hard copy to eSafety.

Email: internalreview@esafety.gov.au

Post: **Attention:** Internal Review
eSafety Commissioner
PO Box Q500
Queen Victoria Building
NSW 1230

For additional information on eSafety's Internal Review Scheme, including the **eSafety Internal Review Procedure** and the **Online Safety (Internal Review Scheme) Instrument 2022**, please visit eSafety's website: www.esafety.gov.au/about-us/corporate-documents/internal-review.

External review by the Administrative Appeals Tribunal

You have a right to seek review of this decision by the Administrative Appeals Tribunal (AAT). You can also request that the AAT review a decision that has been made under the Internal Review Scheme.

It is recommended that you seek an internal review prior to seeking a review by the AAT however, there is no requirement to do so. You can choose to apply directly to the AAT.

The AAT is an independent body that can, among other things:

- confirm the eSafety Commissioner's decision

esafety.gov.au

- vary the eSafety Commissioner's decision; or
- set the eSafety Commissioner's decision aside and replace it with its own decision.

You must apply to the AAT for review in writing. The AAT has a form available on its website which you can use.

Applications for review should be made **within 28 days** of being told about the decision.

You must enclose the $1,011 application fee with your application. If you want to apply for the application fee to be waived, you can obtain the application form for this from the AAT.

The AAT website (www.aat.gov.au) has more information. If you have any questions about the AAT's procedures and requirements, please contact the AAT. Information about how to contact the AAT is available at www.aat.gov.au/contact-us.

Requesting a statement of reasons for decision

If we have not provided the reasons for this decision, you may request a statement of reasons under section 28 of the *Administrative Appeals Tribunal Act 1975* (Cth). Your request needs to be made in writing **within 28 days** of being told of this decision. To request a statement of reasons, please email internalreview@esafety.gov.au.

Access to documents

You have a right to seek access to documents held by the eSafety Commissioner under the *Freedom of Information Act 1982* (**FOI Act**).

You must apply to the eSafety Commissioner in writing through one of the following options:

Online: using the **Contact Us** form on the eSafety Commissioner's website

Post: **Attention:** The FOI Coordinator
eSafety Commissioner
PO Box Q500
Queen Victoria Building
NSW 1230

Email: enquiries@esafety.gov.au

When you make your application, you should:

- state that the request is an application for the purpose of the FOI Act;

- provide information about each document to which you are seeking access to enable us to process your request, and
- provide a postal, email or fax address for us to reply to and which we can use to communicate with you about your application.

The eSafety Commissioner's website has more information on how to make an FOI application: www.eSafety.gov.au/about-us-corporate-documents/freedom-of-information

Complaints

If you are dissatisfied with the way that the eSafety Commissioner has handled this matter, we ask that you contact us using the **Contact Us** form on the eSafety Commissioner's website so that we can try to help resolve any issues.

If you are still dissatisfied, you may make a complaint to the Commonwealth Ombudsman. The Ombudsman usually prefers that your concerns are raised with the eSafety Commissioner first.

There is a Commonwealth Ombudsman office in each capital city. Further information may be obtained at www.ombudsman.gov.au.

Attachment A
Transparency: Opportunity to Make Submissions

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

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 non-periodic reporting notice given to you.

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

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

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

eSafety will carefully consider your response in line with the guidance below. As set out in the Regulatory Guidance on the Basic Online Safety Expectations, eSafety assesses that the objective of the Act of improving and promoting online safety for Australians will be met most effectively through transparency in the exercise of the eSafety Commissioner's functions and through publishing information received from industry.

In determining what is appropriate to publish, eSafety will take into account the objectives of the Act and the functions of the eSafety Commissioner pursuant to section 27 of the Act:

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

Please note that any submissions are voluntary and do not form part of your obligation to response to the notice.

You are invited to provide submissions in the section below, via separate correspondence, or within your Schedule B response. Please provide any submissions to the same deadline as your response to the notice.

Please note that any submissions you make as part of this process will not be published, unless required by another legal process. These submissions are solely for the purpose of eSafety's consideration of information it will publish.

Categories of information that eSafety will consider not publishing

Category of information	Includes	Relevant Factors
Commercial in confidence	Trade secrets. Information with commercial value, where that value would be diminished if the information were published.	Matters eSafety will consider include: - the extent to which information is already publicly known - measures taken to guard secrecy - the value of the information to its owner and competitors - the effort and money spent by the owner in developing the information - the ease or difficulty with which others might acquire or duplicate the information - the commercial harm that could occur from publication - other relevant information or submissions raised.
Other business information that would be unreasonable to publish	Information about an individual's business or professional affairs, or information about the business, commercial or financial affairs of an organisation or undertaking (business information) that would unreasonably affect that person adversely in respect of their lawful business or professional affairs or that organisation or undertaking in respect of its lawful business, commercial or financial affairs.	Matters eSafety will consider include: - whether the information is business information - how the publication of the information could have an unreasonable adverse impact on the individual or business.
Law enforcement and public safety	Information that could affect law enforcement or public safety, including disclosing methods or procedures for preventing, detecting, investigating, or dealing with matters arising out of, breaches or evasions of the law. Information that could assist individuals and groups from deliberately contravening or circumventing safety and security measures.	Matters eSafety will consider include: - whether the methods or procedures are publicly known or prevalent across industry. - the level of detriment that is likely to occur from the disclosure of any lawful methods or procedures for investigating, preventing, detecting or dealing with breaches of the law. - how information could assist individuals in contravening company safety policies and interventions and the level of detriment that is likely to occur. Information that eSafety will not normally publish: - Specific indicators, for example behavioural indicators (new account contacting multiple children) that might companies use, or technical indicators (e.g. device and IP addresses); - Language/terms searched for; - Detailed explanations or information on how technologies work and their weaknesses/vulnerabilities; - "New technology" that is not currently in the public domain.
Personal information	Information about a natural person.	Matters eSafety will consider include:

		- whether the information is about an identified individual or an individual who is reasonably identifiable from the summary or other sources - whether the information in the summary could be de-identified so that is no longer about an identifiable individual or individual who is reasonably identifiable.

Submission by provider

Please note there is no limit to the comments you can provide, and this box should automatically resize.

Schedule B: Manner and form of the required report.

This document relates to the non-periodic reporting notice (the Notice) issued to Twitter, Inc. (Twitter) under section 56(2) of the *Online Safety Act 2021* (Cth) (the Act).

22 February 2023

Instructions

Please respond to the questions below related to your implementation of the Basic Online Safety Expectations. The relevant expectations are stated at the beginning of each question. Please also note that:

- You are required to respond to the questions in the manner and form specified, including using this template for your responses.
- You are required to respond truthfully and accurately.
- Where information, including data, is requested, please provide it from 24 January 2022 to 31 January 2023 (the **Report Period**).
- If you do not collect the relevant information, you should endeavour to provide alternative relevant information. For example, where Australian data cannot be disaggregated from global data, the broader dataset may be acceptable.
- The questions relate to the specific Expectations in the *Online Safety (Basic Online Safety Expectations) Determination 2022 (the Determination)* (Cth) (**Determination**) referenced below.
- If you think a question is not applicable, please check with eSafety before finalising your response, and be prepared to explain why you do not consider a question to be applicable. The questions have been drafted to be specific to your services so should be relevant.
- The size of text boxes for responses should not be taken as an indication of any text limit. Please include as much information as you consider necessary to provide a fulsome response to the question. The boxes should automatically resize as needed.
- If you require additional time, eSafety will consider accordingly. However, we expect that most of the information requested should be able to be provided within the time permitted.
- Further information can also be found in the Basic Online Safety Expectations Regulatory Guidance[1] and on the relevant page[2] on eSafety's website.
- You can contact industryBOSE@safety.gov.au regarding any of the above points, or with any other questions regarding your response.

[1] Office of the eSafety Commissioner, 'Basic Online Safety Regulatory Guidance', July 2022, accessed 3 February 2023, URL: https://www.esafety.gov.au/sites/default/files/2022-07/Basic%20Online%20Safety%20Expectations%20regulatory%20guidance.pdf
[2] Office of the eSafety Commissioner, accessed 3 February. 22, URL: https://www.esafety.gov.au/industry/basic-online-safety-expectations

Questions

1. **Twitter has previously stated[3] that it uses PhotoDNA and internal, proprietary tools to detect child sexual exploitation and abuse[4] (CSEA). Does Twitter use hash matching[5] tools to scan for known[6] child sexual exploitation and abuse CSEA images on the following parts of its service?**

Determination, Expectations: s6; s11

a. Please answer yes/no for the following parts of the service: b. If yes, name the tools used.

	Yes/No	Please name all the hash matching tools used:
Public posts ('tweets')	Choose an item.	
Twitter direct messages	Choose an item.	

c. If the answer to question 1(a) is yes for any part of the service, where does Twitter source hashes of CSEA images from? Please list all the databases used. If these differ for the parts of the service listed in question 1(a), please specify the sources for each.

[3] Twitter, 15th Transparency Report: Increase in proactive enforcement on accounts, 31 October 2019, accessed 8 February 2023, URL: 15th Transparency Report: Increase in proactive enforcement on accounts (twitter.com)

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

[5] This type of technology creates a unique digital signature (known as a "hash") of an image which is then compared against signatures (hashes) of other photos to find copies of the same image. See for example: Microsoft, 'PhotoDNA', accessed 3 February 2023, URL: https://www.microsoft.com/en-us/photodna

[6] Material that has been previously confirmed to contain child sexual exploitation and abuse.

d. Does Twitter take all available hashes of CSEA images from these databases, or a subset? If a subset, please clarify what this is, either by name and/or by description.

e. How often does Twitter update the list of hashes of CSEA images it scans against from these databases?

f. If the answer to question 1(a) is no for any part of the service, i. please specify why hash matching tools are not used
ii. Explain what alternative reasonable steps Twitter is taking to detect known CSEA images on those services.

2. Does Twitter use any tools to detect known CSEA video?

Determination, Expectations: s6; s11

a. Please answer *yes/no to question 2* for the following parts of the service:		b. Please name all the tools used, if applicable
Twitter public posts ('tweets')	Choose an item.	
Twitter direct messages	Choose an item.	

c. If the answer to question 2a is no for any part of the service, please: i. specify why hash matching tools for known video are not used
ii. explain what alternative reasonable steps Twitter is taking to detect known CSEA video on those services.

3. Does Twitter block URLs[7] linking to known CSEA on the following parts of its service?
Determination, Expectations: s6; s11

a. Please answer yes/no to question 3 for the following parts of the service:	
Part of the service	**Yes/No**
Public posts ('tweets')	Choose an item.
Twitter direct messages	Choose an item.

b. If the answer to question 3(a) is 'yes' for any part of the service, where does Twitter source its URLs linking to known CSEA material? Please list all the databases/lists used.

c. If the answer to question 3(a) is 'no' for any part of the service, please:

[7] Uniform Resource Locator. An example of a URL list of known CSEA is the Internet Watch Foundations: URL List (iwf.org.uk)

i. specify why URLs to known CSEA are not blocked
ii. Explain what alternative reasonable steps Twitter is taking to prevent the sharing of links to known CSEA.

4. Does Twitter use tools to scan for new[8] CSEA material?

Determination, Expectations: s6; s11

a. Please answer yes/no to question 4 for the following parts of the service: b. If yes, name the tools used.

	Yes/No	Please name tools used:
Twitter public posts ('tweets')	Choose an item.	
Twitter direct messages	Choose an item.	

c. If the answer to question 4a is yes, do human moderators review all reports flagged by the technology as new CSEA material? If not, what proportion are reviewed by human moderators?
d. What steps are taken when new CSEA material is confirmed to have been shared by an account?

[8] New CSEA is material that has not been previously confirmed, hashed, and stored in a hash database.

e. If the answer to question 4a is no for any part of the service, please: i. specify why tools to detect new CSEA are not used.
ii. explain what alternative reasonable steps Twitter is taking on that service to detect new CSEA material.

5. What proportion of CSEA material is detected proactively, as compared to CSEA material reported by users/trusted flaggers for the following services? Please provide separate answers for the time periods specified.

Determination, Expectations: s6; s11

Time Period	Part of service	Proportion reported vs detected by Twitter (%s)
24 January 2022 – 27 October 2022	Twitter public posts ('tweets')	
	Twitter direct messages	
28 October 2022 – 31 January 2023	Twitter public posts ('tweets')	
	Twitter direct messages	

a. Please provide any further context or information

6. When Twitter confirms that an account has shared CSEA material, does it also review other material shared by that account?

Determination, Expectations: s6; s11

a. Please answer yes/no to question 6 for the following parts of the service:

Part of the Service	Yes/No
Twitter public posts ('tweets')	Choose an item.
Twitter direct messages	Choose an item.

b. If the answer to question 6a. is 'yes', does this process involve a review of all content or a subset? If a subset, please specify.
c. Does this process involve a human moderator review and/or use of technical tools? Please name any tools used.
d. If the answer to question 6a is 'no', please: i. specify why Twitter does not review all content.

ii. explain what alternative reasonable steps are being taken.

7. Does Twitter have measures in place to detect the livestreaming[9] of CSEA on its service?
Determination, Expectations: s6; s11

a. Please answer yes or no to question 7.
Choose an item.

b. If the answer to question 7a is yes, please confirm whether any of the measures below are used:					
Language analysis tools (text)	☐	Language analysis tools (audio)	☐	Video classifiers	☐
Age assessment tools	☐	Geographical indicators	☐	Behavioral indicators	☐
Traffic indicators	☐	Financial indicators	☐	Others (please specify):	☐
Provide any other information or context:					

c. If the answer to question 7a is yes, please detail the tool(s) used.
d. Do human moderators review all reports flagged by the technology to confirm livestreaming of CSEA? If not, what proportion are reviewed by human moderators?

[9] For the purpose of the Notice, livestreaming is the transmission or receipt of acts of sexual exploitation or abuse of children live via webcam or video to people anywhere in the world, whether or not in exchange for payment. Livestreaming includes one-on-one video calls and video calls where one or multiple people stream material to a group of any size.

e. What steps does Twitter take when an account is detected for livestreaming CSEA?

f. If the answer to question 7a is no, please: i. specify why no tools are used to detect livestreamed CSEA,
ii. explain what alternative reasonable steps Twitter is taking to ensure safe use of livestreams and prevent its use for CSEA?

8. What tools does Twitter use to detect underage users?

Determination, Expectations: s6; s14

9. How many underage users (under 13) has Twitter identified and removed on its service during the reporting period?

Determination, Expectations: s6; s11;

a. Please provide number of underage users identified and removed: i. Globally ii. In Australia:

10. What proportion of underage users detected were detected by Twitter proactively, as compared to underage users reported by users?

11. If reported by users what method does Twitter use to determine if the reported account was created by an underage user?

12. What measures does Twitter have in place to ensure users whose accounts are closed for being underage are prevented from opening a new account?

13. Does Twitter use any language analysis technology to detect likely online grooming[10]?
Determination, Expectations: s6; s11

a. Please answer yes/no to question 13 for the following parts of the service. If yes, name the tools used.

	Yes/No	Please name the tools used:
Twitter public posts ('tweets')	Choose an item.	
Twitter direct messages	Choose an item.	

b. If the answer to question 13a is yes, do human moderators review all reports flagged by the technology to confirm likely online grooming? If not, what proportion are reviewed by human moderators?
c. what steps does Twitter take when an account is detected to be involved in online grooming?

d. If the answer to question 13a is no for any part of the service, please:

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

i.	specify why no language analysis tools are used to detect grooming
ii.	explain what alternative reasonable steps Twitter is taking to ensure grooming does not occur on its service?

14. Does Twitter use any language analysis technology to detect terms, abbreviations, codes and hashtags indicating likely CSEA activity, in particular but not limited to, sexual extortion[11] and the trading and sale of CSEA material?
Determination, Expectations: s6; s11

a.	Please answer yes/no for the following parts of the service. If yes, name the tools used.

	Yes/No	Please name the tools used:
Twitter public posts ('tweets')	Choose an item.	
Twitter direct messages	Choose an item.	

b.	If the answer to question 14a is yes for any part of the service, please list all the specific organisations that Twitter uses to source these terms and abbreviations (for example name the specific law enforcement, civil society, academic and other organisations)

[11] 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.

c. Do human moderators review all reports flagged by these indicators? If not, what proportion are reviewed by human moderators?
d. What steps does Twitter take when such indicators are detected on an account?

e. If the answer to question 14a is 'no' for any part of the service, please: i. specify why language analysis tools are not used,
ii. explain what alternative reasonable steps Twitter is taking to proactively minimise CSEA activity such as sexual extortion on its services.

15. Are any of safety tools deployed by Twitter implemented on users' devices[1], rather[12] than on servers?

Determination, Expectations: s6

[12] Implemented on "devices" or "client" side, rather than on servers, "server side".

16. **a. Please list all user identifying information collected on sign up to Twitter. (Examples of information may be phone number, email address, device identifiers, IP address or others)**
Determination, Expectations: s6; s9; s14

b. Is there an authentication process for any of the information captured on sign up (for example, email or phone authentication) to ensure the account is controlled by the user signing up?
c. If the answer to question 16a is no, please specify why Twitter does not authenticate any of the identifying information it collects from users on sign up.

17. **What measures does Twitter have in place to prevent recidivism[13] of users for CSEA-related breaches on its service?**
Determination, Expectations: s6; s9; s11; s14

[13] Banned users re-registering new accounts.

a. Please confirm whether any of the indicators below are used to detect recividism:					
Device identifiers	☐	IP address	☐	Geographic indicators	☐
Phone number	☐	Name	☐	Financial indicators	☐
Email address	☐	Username	☐	Date of birth	☐
Others (please specify):					
Provide any other information or context:					

b. If any of the information collected on sign up listed in question 16a is not used to prevent recidivism, please specify why not.

18. Can users report instances of CSEA material or activity to Twitter within the following parts of the service? (i.e. without navigating to a separate webform or email address)

Determination, Expectations: s6; s11; s13; s14; s15

a. Please answer *yes/no* for the following parts of the service:	
Twitter public posts ('tweets')	Choose an item.
Twitter direct messages	Choose an item.

b. If the answer is no for any service, what alternative reasonable steps is Twitter taking to ensure users can provide clear and readily identifiable means of reporting CSEA material on those services?

19. a. What is the median time taken to respond to a user report about CSEA[14] for the following services:

Determination, Expectations: s6; s11; s13; s14; s15

Service	Median time
Twitter public posts ('tweets')	
Twitter direct messages	

b. Please provide any other information or context you wish to provide in respect of each part of the service.

***20.* a. Twitter's CEO has indicated[15][16] that Twitter intends to implement end-to-end encryption[17] of direct messages. During the report period, was a decision taken that end-to-end encryption of direct messages should be implemented? Please answer yes or no.**

Determination, Expectations: s8

Choose an item.

[14] Content removed and reported, user banned, or other content moderation decision taken.

[15] Musk, Elon, November 2022. Accessed 6 February 2023, URL: https://twitter.com/elonmusk/status/1596718851097755648?s=20&t=z_6MuLY6BW45A1LdBCxiLw

[16] Forbes, 'Elon Musk Wanted Twitter to Encrypt Messages. His New Safety Chief Says It's On Hold', 7 December 2022, accessed 6 February 2023, URL: Elon Musk Wanted Twitter To Encrypt Messages. His New Safety Chief Says It's On Hold (forbes.com)

[17] End to end encryption (E2EE) is a is a specific method used to secure communications from one device, or "end point", to another. E2EE transforms standard text, imagery, and audio into an unreadable format while it is still on the sender's system or device and so that it can only be decrypted once it reaches the recipient's system or device.

b. If no decision has been made yet, did staff work on the development of end-to-end encryption of direct messages during the report period? Please answer yes or no.

Choose an item.

c. If either of the answers to question 20a or 20b are yes, please detail the measures considered, or intended to be implemented, to ensure that end-to-end encryption does not increase the risk of CSEA dissemination without detection.

21. What information does Twitter include in its NCMEC reports[18]? Please list all types of information included in a NCMEC report.

Determination, Expectations: s6; s10 s14

22. Please provide the numbers for the following categories of staff between the two specified periods:

Determination, Expectations: s6; s11

Category of staff	**Number of Staff Employed or Contracted by Twitter Between the Following Periods**

[18] Reports made to the National Centre for Missing and Exploited Children.

	Time period: 24 January 2022 - 27 October 2022	Time period: 28 October 2022 – 31 January 2023
a. Engineers focussed on trust and safety issues globally.		
b. Trust and safety staff dedicated to CSEA issues globally.		
c. Trust and Safety staff globally.		
d. Trust and safety staff in the APAC region.		
e. Trust and safety staff in Australia.		
f. Content moderators globally.		
g. Content moderators in the Asia Pacific region.		
h. Public policy staff globally.		
i. Public policy staff in the APAC region.		
j. Public policy staff in Australia.		

23. Please list all of the languages that Twitter human moderators operate across? Please specify for the period 24 January 2022 – 27 October 2022 and 28 October 2022 – 31 January 2023 respectively.

Determination, Expectations: s6; s11; s13; s14; s15

24. Describe the design objectives of Twitter's Recommender Systems, i.e. "For You". What are the top 5 objectives?

Determination, Expectations: s6; s11

25. Provide a list and description of the top 10 signals that Twitter uses in determining what content should be recommended to a user and provide details of their relative importance in meeting the objectives in question 24.

Determination, Expectations: s6; s11

26. Describe any measures Twitter employs to test and update recommender systems, in a way which contributes to the overall safety of Twitter's platform, and avoids the risk of amplifying harmful content? For example, this could include internal audits, external audits, risk and impact assessments, and a/b testing.

Determination, Expectations: s6; s11

a. If such tests and updates are performed, how often are they performed?

27. What criteria does Twitter have for evaluating content that is harmful to children but does not violate community guidelines for the purposes of not recommending, or recommending less often?

Determination, Expectations: s6; s11

28. What steps are taken to prevent the recommendation of accounts or content involved in the sale and trade of child sexual exploitation material?
Determination, Expectations: s6; s11

29. Does Twitter have internal metrics in place to assess the efficacy of its interventions to detect CSEA material?
Determination, Expectations: s6; s11

a. Please answer yes or no to question 29.

Choose an item.

b. If yes, please detail the metrics and how they enable Twitter to understand the effectiveness of its interventions to prevent and detect CSEA.

c. Please provide any further detail or context.

30. How many appeals have been made by users in the period of this reporting notice for accounts banned or content removed for CSEA material or activity?
Determination, Expectations: s6; s11

a. Please indicate number of appeals
b. How many of these appeals were successful?
c. Please provide any further information or context.

31. Does Twitter provide training to its staff on Safety by Design principles[19]?
Determination, Expectations: s6; s11

a. Please answer yes or no
Choose an item.
b. Please provide any other information or context.

[19] Principles and background | eSafety Commissioner

If you have any broader context or clarification to provide for any of the questions, please provide it below.

Please return to industryBOSE@esafety.gov.au

ANNEXURE C

ANNEXURE D

ANNEXURE E

Notice question	Deficiency
1(b)	While the Respondent said that it used Photo DNA and 'other proprietary technologies', it did not name those technologies in its 29 March 2023 response. The Respondent clarified that information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
1(c)	While the Respondent said that it used a number of sources, including NCMEC, the Internet Watch Foundation and other partners, it did not name all sources in its 29 March 2023 response. The Respondent clarified that information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
3(b)	While the Respondent said that it used a number of sources, including NCMEC, the IWF and other partners, it did not name all sources in its 29 March 2023 response. The Respondent clarified that information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
4(b)	While the Respondent answered 'Yes' to Question 4(a) in respect of 'Twitter public posts', it did not answer Question 4(b) in its 29 March 2023 response. The Respondent clarified that information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
6(c)	While the Respondent said that it used a combination of human moderator review and technical tools, it did not name all tools in its 29 March 2023 response. The Respondent clarified that information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
7(a) – (f)	While the Respondent said that it no longer had livestream video following the discontinuation of the Periscope app in its 29 March 2023 response, it subsequently identified that this response was inaccurate due to an inadvertent error. The Respondent also clarified this information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
10	While the Respondent provided a relevant answer to Question 10 in its 29 March 2023, it subsequently identified that this response was inaccurate because the Respondent did not track the relevant metric and that, at the time of the 29 March 2023 response, the Respondent was not capable of providing the information requested by this question.
14(b)	While the Respondent answered 'Yes' to Question 14(a) in respect of 'Twitter public posts', it did not answer Question 14(b) because "We are under confidentiality agreements with the providers we work with and are unable to disclose the information requested without violating our contractual agreements."
14(e)	While the Respondent answered 'No' to Question 14(a) in respect of 'Twitter direct messages', it did not answer Question 14(e) in its 29 March 2023 response. The Respondent clarified that information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
16(a)	While the Respondent provided examples in response to Question 16(a), it did not name all user identifying information collected on sign up to Twitter in its 29 March 2023 response. The Respondent clarified that information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
17(b)	While the Respondent provided a link to its privacy policy in response to Question 17(b) in its 29 March 2023 response, that privacy policy did not adequately answer

	Question 17(b). The Respondent clarified its answer in its 5 May 2023 response, in response to the Applicant's clarifying questions.
19(a)	The Respondent did not answer Question 19(a) in its 29 March 2023 response. It subsequently provided the information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
22(a) – (i)	The Respondent did not answer Question 22(a) – (i) in its 29 March 2023 response. It subsequently provided the information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
25	The Respondent did not answer Question 25 in its 29 March 2023 response. It subsequently provided the information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
26	The Respondent did not answer Question 26 in its 29 March 2023 response. It subsequently provided the information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
26(a)	The Respondent did not answer Question 26(a) in its 29 March 2023 response. It subsequently provided the information in its 5 May 2023 response, in response to the Applicant's clarifying questions.
27	While the Respondent said that it continued to put in place a 'range of settings' to evaluate content that is harmful to children, it did not name all criteria in its 29 March 2023 response. The Respondent clarified that information in its 5 May 2023 response, in response to the Applicant's clarifying questions.