

Achieving digital platform public transparency in Australia

JUNE 2024

Reset.
AUSTRALIA

 Susan McKinnon
Foundation

Achieving effective public transparency requires new legislation. As this report documents, multiple voluntary attempts at transparency, in Australia and internationally, have not generated meaningful outcomes. *It is clear that, left to their own devices, digital platforms will not offer sufficient transparency to correct this informational asymmetry, and reform is needed.*

Reset.Tech Australia
2024

This research was funded by the



Cover image: Digital collage created using Midjourney in response to the prompt "imagine/ A large scale transparent glass sculpture as the focal point in an Australian landscape at dawn, H6D-100c, Sirui 50mm f/1.8 anamorphic 1.33x, UHD --style raw --stylize 600 --v 6.0 --ar 1:1"

Summary

As the momentum to regulate digital platforms in Australia intensifies, two words continue to circulate: transparency and accountability. These are vital and inseparable elements in a digital platforms regulatory model. One application of transparency into practice is a public reporting framework.

Transparency reporting can be a powerful tool for corporate accountability. To fully realise this potential, transparency reporting cannot be left as an opt-in exercise of broad, general and unassessable statements. This report explores the limitations of Australia's current public transparency framework for digital platforms and draws on international examples to recommend a path forward.

This report provides a thorough review of the public transparency frameworks under the *Digital Services Act* (DSA) and the UK's new *Online Safety Act*, augmented by recent polling which indicates the public's broad support for meaningful transparency measures. A model with five specific measures could improve public transparency in Australia is presented.

These include:

1. Risk assessments
2. Annual transparency reports (with prescribed metrics)
3. Annual independent audits
4. Data portals, including ad repositories and content moderation databases
5. Researcher access requirements

These measures could be applied in both online safety and misinformation and disinformation frameworks to create transparency around a broad range of digital risks.

Three recent case studies from the *Digital Services Act* highlight the sorts of insights public transparency models can generate that could be useful in the Australian context.

These include:

- › Information about how platforms prioritise and respond to both user and government requests
- › Information about how platforms' content moderation processes work
- › Information about how researcher access regimes work



Contents

Introduction	5
Public transparency in Australia: the current state of play.....	6
1. Public transparency requirements regarding misinformation and disinformation in regulation.....	6
2. Public transparency requirements regarding online safety in regulation.....	8
International models.....	10
1. The EU's <i>Digital Services Act</i>	10
EU Case Studies.....	12
2. The UK's <i>Online Safety Act</i>	16
Recommendations and proposals for a transparency framework for Australia.....	20
1. Risk assessments.....	20
2. Annual transparency reports	20
3. Annual independent audits.....	22
4. Data portals, including ad repository and content moderation	22
databases.....	22
5. Researcher access requirements.....	23
Appendices.....	24
Appendix 1.....	25
Appendix 2.....	26



Reset.Tech Australia is an independent, non-partisan policy research lab committed to driving public policy advocacy, research and civic engagement to strengthen our democracy within the context of technology. We are the Australian affiliate of Reset, a global initiative working to counter digital threats to democracy.

Introduction

Informational asymmetry underpins the relationship between digital platforms and the public they engage. Platforms in the business of digital content distribution – Facebook, YouTube, TikTok, X – operate with a granular level of detail about the operation of their products, their underlying technical systems, and the behaviours of their users. Unsurprisingly, these titans of information-based extractive industries are also information's powerful gatekeepers, invoking complexity traps and trade secrets to justify their deep operational secrecy.

Transparency measures are a path to invert the one-way glass of digital platform systems. At their most effective, they compel prescriptive information sharing from platforms about the function and impact of their technical systems and decision-making processes to external parties (including regulators, academics, and research organisations), on scheduled, real-time, and on-request bases. This information sharing can take many forms, from reportable metrics on activities, or access to direct platform data, such as application programming interfaces (APIs). At their least effective, they are unmeasurable and untestable gestures at corporate social responsibility written by public relations departments on the platform's terms and timelines.

Effective transparency means that the public has up-to-date and actionable information on the operation of various systems and elements and greater insights into how platforms make decisions and manage risks to the public. These outcomes are not easily gained: providing this level of meaningful information takes effort and resources that platforms are incentivised to direct into more commercially prosperous areas.

Achieving effective public transparency requires new legislation. As this report documents, multiple voluntary attempts at transparency, in Australia and internationally, have not generated meaningful outcomes. It is clear that, left to their own devices, digital platforms will not offer sufficient transparency to correct this informational asymmetry, and reform is needed. This report explores some of the public transparency measures that might be possible in Australia, in both the online safety and misinformation and disinformation regulatory frameworks. It explores the current state of play, international best practice examples and proposes some mechanisms that might improve transparency in Australia going forward.

Transparency is a critical part of a regulatory framework that is required to protect Australians from a broad range of digital harms. But it is not a silver bullet. Alongside transparency measures, we need regulation to make platforms responsible for user safety through a duty of care, mandating risk assessments and risk mitigations, and empowering regulators to effectively enforce safety online (see Appendix 1 for more details about this broader framework).

Public transparency in Australia: the current state of play

This section outlines current arrangements for digital platform public transparency with reference to misinformation and disinformation, as well as online safety.

01. Public transparency requirements regarding misinformation and disinformation in regulation

In Australia, requirements for public transparency around misinformation and disinformation are relegated to the *Australian Code of Practice on Disinformation and Misinformation* (the 'Code').¹ The Code requires that any platforms that choose to sign up publish an annual transparency report about the measures the platforms take to address misinformation and disinformation on their platforms. It does not require platforms to create or share data about the efficacy of these measures, just to clearly describe the measures taken.

The process is administered by Digi, an industry association, and appears to work as follows:

- › Platforms prepare a report on their misinformation and disinformation mitigation efforts for a January-December calendar year;
- › These reports are submitted to Digi in the first quarter of the following year;
- › An independent expert reviews the reports;
- › Reports are released publicly in May.

Digi have prepared templates and 'Best Practice Guidelines' for the reports which are set out in Appendix 2 of the Code.² Four rounds of transparency reports have been released by platforms who have signed onto the Code.

However, there are concerns about the quality of these transparency reports. In 2021,³ and again in 2023, the Australian Communications and Media Authority (ACMA) expressed concerns about the quality and comparability of the annual reports (ACMA) platforms released in compliance with the Code, going so far as to say that they "are not working to provide transparency."⁴ Likewise, academics have questioned their efficacy,⁵ the third sector has raised concerns (see page 7) and a complaint has been raised to Digi by Reset.Tech Australia regarding potentially misleading statements contained within a report. The complaint was dismissed, as only complaints regarding materially false information can be considered under the Code, rather than misleading information.⁶ The ACMA has no powers to request information from platforms regarding their efforts to mitigate misinformation and disinformation, nor powers to request data about the prevalence or impact of misinformation and disinformation in Australia. As the examples in this report indicate, regulatory powers to compel platforms to share more meaningful information on the efficacy of their policies and systems would both enhance public accountability and strengthen public trust.

Now is a good time to evaluate the state and efficacy of transparency reports regarding misinformation and disinformation in Australia. The ACMA has put out a tender for a consultant to help develop a set of 'metrics' for platform transparency reporting,⁷ indicating that improving the state of digital platform transparency is an achievable and relevant pursuit.



A summary of concerns about current public transparency reports emerging from a third sector discussion⁸

Lack of definitional clarity.

Transparency reports are unclear in at least two key ways.

First, when data about violative content is presented in reports, a lack of underlying data about the nature of the violation makes the data hard to interpret. For example, when figures about violative content are presented in transparency reports (such as statements like ‘10,000 posts were removed’), these figures are often ‘bundled’, and it is unclear which violations they account for (that is, the 10,000 posts could be posts arising from inauthentic behaviour, posts inciting violence, posts threatening electoral integrity or posts proven to be false after fact-checking). Platforms have a range of community guidelines and rules around what content is considered violative, and without specificity, it is hard to understand what the figures in the current transparency reports refer to. Compounding this lack of clarity, variations in guidelines across platforms make comparison difficult if not impossible. What violates guidelines on one platform may be fine on another.

Second, reports also talk about ‘acting’ on content, without providing clear descriptions about what these actions are. Platforms have a range of actions they claim to take against violative content. Platforms can remove it, label it, demote it and so on. Providing bundled figures about actions taken against violative content does not provide sufficient insight.

Lack of clarity about how the content moderation system operates.

Broad and overarching claims are often made about platforms’ content moderation systems but with insufficient detail for interrogation. For example, it is unclear what guidelines content moderators are given to enable the evaluation of Australian-specific content. Furthermore, platforms often do not disclose crucial information about their content moderators, and it is unclear how many content moderators are tasked with addressing Australian content. For instance, TikTok’s 2022 transparency report claimed that the company had 40K trust and safety experts,⁹ but industry data suggests their total workforce is only 38K.¹⁰ It is unclear where these ‘outsourced’ trust and safety experts – presumably content moderators – work, who they work for and how many of them review Australian-specific content. Additionally, it is unclear whether content moderators are stationed in Australia or abroad, making it difficult to decipher their knowledge of local contexts. Sometimes it is unclear what third-party fact-checkers platforms use.¹¹ It is often unclear how fact-checkers work, which content they prioritise for checking and how platforms, in turn, moderate their content against fact-checkers’ findings.

The figures used in transparency reports lack sufficient context to make them intelligible.

The absence of contextual information in transparency reports raises questions about the significance of reported numbers in three ways. First, the ‘denominator issue’ makes reported figures impossible to understand. For example, platforms will report the amount of violative content they addressed but without providing estimates of the total number of violative posts on a platform or indeed the total number of posts on a platform, making it very difficult to comprehend the significance of a claim like ‘we removed 80K videos that violated community guidelines’, as it is unclear if this is a significant proportion of violative content or a drop in the ocean. Second, there is a lack of international comparison. If, for example, 80K videos were removed in Australia, but 500K were removed in New Zealand, we might reach different conclusions about the efficacy of platform actions in Australia. Transparency reports in Australia do not currently require this information, making it impossible to know what proportion of violative content is detected and removed. Third, there is no consistency over time, making it difficult to tell if the situation is getting worse or better on a platform.

02. Public transparency requirements regarding online safety in regulation

Requirements for public transparency around online safety emerge from the *Online Safety Act*, specifically in the requirements around the Basic Online Safety Expectations (BOSE). Under the expectations, the Office of the eSafety Commissioner has the powers to request a range of information from platforms via 'transparency notices'.¹² These powers include:

- › Requiring platforms to provide periodic reports, ranging from 6 month to 24 month cadences about compliance with the BOSE. The reports would need to provide information about the extent to which the platform complied with the BOSE in general or specific elements of the BOSE,¹³ in a manner and form that is specified by the Commissioner.¹⁴ In requiring platforms to prepare periodic reports, the Office of the eSafety Commissioner needs to consider a range of things, including the number of complaints made to the Commissioner against a platform in the last year, previous contraventions of the *Online Safety Act*, and any safety deficiencies on the platforms.¹⁵ To date, we are unaware of any platform that has been required to produce periodic reports.
- › Requiring non-periodic reporting about compliance, where each platform provides a particular type of service (like search engines, or online messaging services) about compliance with one or more elements of the BOSE, in a manner and form that is specified by the Commissioner.¹⁶

While responses to these notices are sent directly to the Office of the eSafety Commissioner, the Commissioner is empowered to then publish a statement regarding the platform's periodic and non-periodic reports on their website; this is the *public* transparency element of the process.¹⁷

To date, three non-periodic transparency report statements have been published by the Office of the eSafety Commissioner.¹⁸

These include:

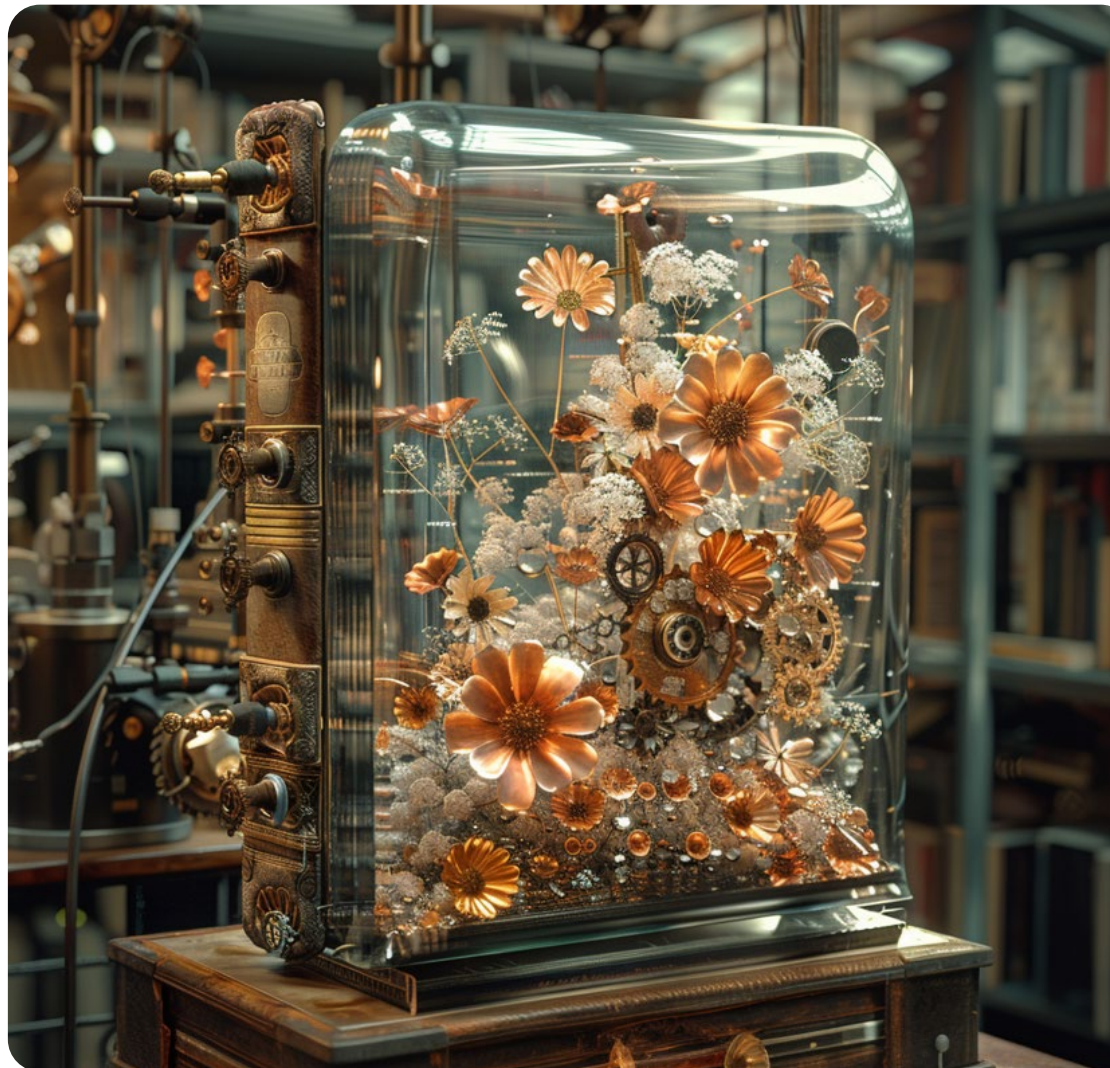
- › Statements summarising the notice given to X Corp in relation to online hate on X;
- › Statements summarising the notices given to Google, X, TikTok, Twitch and Discord regarding the steps being taken to detect and address child sexual exploitation and abuse on their services, and;
- › Statements summarising notices given to Apple, Meta, WhatsApp, Microsoft, Skype, Snap and Omegle regarding the steps being taken to detect and address child sexual exploitation and abuse on their services.

There are concerns about the quality and availability of information provided to the Commissioner in these non-periodic reports, which in turn hampered the public transparency of the published notices. The Office of the eSafety Commissioner has so far found that:

- › X did not comply with the non-periodic reporting notice regarding online hate by providing responses that were incorrect, significantly incomplete or irrelevant;¹⁹
- › Google failed to answer a number of questions in response to the notice regarding child sexual exploitation and abuse;²⁰
- › X failed to provide any response to some questions, such as by leaving the boxes entirely blank, and in other instances, provided a response that was otherwise incomplete and/or inaccurate in response to the notice regarding child sexual exploitation and abuse.²¹ The Office of the eSafety Commissioner issued X with an infringement notice for \$610,500 in October 2023, which X challenged by seeking judicial review of the transparency notice, the service provider notification, and the infringement notice.²²

The next section explores potential refinements and additions to the current model in detail.

Artwork created using Midjourney in response to the prompt *"imagine/a surreal glass book sculpture, futuristic antique style, hasselblad H6D-100c, Sirui 50mm f/1.8 anamorphic 1.33x, --v 5.2*



International models

There are a number of international models for regulating transparency that Australia can draw from. These include the EU's *Digital Services Act* (DSA) which covers a range of risks including misinformation, disinformation and user safety, and the UK's *Online Safety Act*, which places specific duties of care on platforms regarding user safety. Below, we summarise the key points in each model.

01. The EU's *Digital Services Act*

The DSA²³ places broad obligations on digital platforms to address four categories of risk.

These include:

- i. the dissemination of illegal content, as defined by the national laws of member states;
- ii. negative effects on fundamental rights, as defined in the EU Charter of Fundamental Rights;
- iii. negative effects on civic discourse and electoral processes, and public security;
- iv. negative effects in relation to gender-based violence, public health and minors and negative consequences to physical and mental well-being.

The DSA also creates five different public transparency measures for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs):²⁴

1. Annual risk assessments.

Platforms are required to produce detailed risk assessments that identify actual and reasonably foreseeable systemic risks, stemming from the design or functioning of their services. This includes the design of recommender systems, content moderation systems, terms and conditions, ads review and ads manager systems, and also the data related practices of the platforms. These risk assessments need to align with the four broad risks covered by the DSA above.

Requirements for risk assessments and mitigations are a central pillar of accountability under the DSA and also serve as a public transparency tool. While risk assessments are submitted directly to regulators for review and analysis, a summary of the risk assessment is also published alongside an independent audit (see next page).

2. Annual, public transparency reports.

Platforms are required to produce annual, public transparency reports on their content moderation practices. Unlike Australia's current transparency reports, these reports are heavily prescriptive, so there is a clear 'template' of information platforms must provide and there is little room for interpretation from platforms. The templates were designed after a public consultation by the European Commission, asking for feedback on the guidance on the form, content and other details to meet the reporting requirements. There has been one 'round' of transparency reports released in October 2023 (before the publication of the templates), and this has brought visibility to a range of systemic issues. For example, the first round of transparency reports released by VLOPs revealed a trove of information, such as the low number of moderators employed in non-English speaking markets (see page 14) for more information). An implemented act on transparency reports is expected for later in 2024, which will bring further detail to reporting rules and templates.²⁵

3. Annual, independent audits.

Platforms must commission and publish an independent audit of their compliance with all the DSA obligations, including the risk assessment and transparency obligations. The audits are eventually published.

4. Data portals, including ad repositories and content moderation data.

The DSA requires at least two different data portals be produced for public transparency;

- a. Searchable ad repositories: Particular attention is given in the EU to paid-for advertising and making visible and searchable what paid-for advertising is being run and by whom. To enable this, article 39 of the DSA requires services to create up-to-date searchable ad repositories. These are openly searchable databases of all ads presented on platforms including targeting options and data about advertisers;
- b. Data about content moderation decisions: Articles 17(1) and 24(5) mandate that platforms submit 'statements of reasons', which are 'clear and specific' explanations about content moderation decisions and restrictions imposed on users who post illegal content or content that is otherwise incompatible with their terms and conditions. These statements must be sent to the European Commission, for their inclusion in a publicly accessible machine-readable database. The Commission launched the DSA Transparency Database²⁶ in September 2023 and since its launch, the Database has been constantly updated in real time by VLOPs. The Commission has provided a template for the statements, which are expected to be submitted via its designated API. From February 2024, all online platforms in the EU (and not just VLOPs) are required to start reporting regarding content moderation decisions.

5. Researcher access to public interest data.

Underpinning all of these public transparency requirements is a right for vetted researchers to analyse and explore platform data themselves. Article 40 places requirements on platforms to provide vetted researchers with '*access to data ... for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union... and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures*'.²⁷ There are stringent requirements for who is considered a vetted researcher (largely ensuring that they are non-commercial researchers, i.e. academics and civil society) and the nature of the data that can be requested (it must relate to systemic risks on the platform) to ensure this scheme works in the public interest. See page 15 for more information.

These five public transparency measures exist alongside extensive information gathering powers afforded to the European Commission under articles 67-75 of the DSA, such as the power to request information, to take interviews and statements, and to conduct inspections.

The DSA provides the European Commission with strong enforcement tools should any of these public transparency measures not be met.

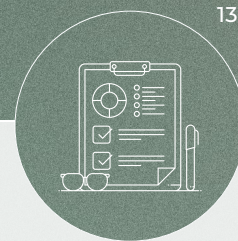
The Commission is able to issue interim measures and fines of up to 6% of global annual turnover to platforms who fail to deliver on their public transparency requirements. In case of systemic non-compliance in relation to a criminal offence, Member States' courts are also empowered to order a temporary restriction of access to the platform concerned.

EU Case Studies

The following three case studies provide some real-life examples for the insights available under the EU's public transparency regime. Note that the EU's transparency reports will be further refined with a delegated act expected later in 2024. See Appendix 2 for a comparison of EU transparency report data against current Australian misinformation and disinformation reporting frameworks.



Photo credit: Maximalfocus on Unsplash

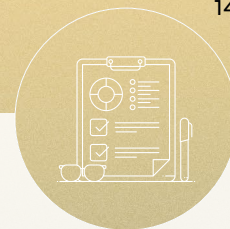


What sorts of information can meaningful transparency reports highlight?

An analysis of transparency reports submitted under the DSA highlights some of the useful information that effective transparency reports can showcase. The Slovak Council for Media Services (the national regulator and Designated Services Coordinator) analysed the first tranche of transparency reports published by the VLOPs and VLOSEs in October 2023²⁸ and reached a number of conclusions about how content moderation practices were prioritised in Europe.

We have arranged the insights into the following themes:

- › **Government orders:** The reports show that almost all platforms treat governmental orders as a top priority in their content moderation efforts and seek to resolve them as soon as possible. The median time needed to give effect to an order is generally below 7 days, with orders concerning product safety getting resolved within 2 days, while orders concerning harassment require 7 days.
- › **User notices:** Most VLOPs manage to process user notices within hours of receiving them (TikTok 13 hours; Instagram 20.4 hours; LinkedIn 23 minutes). However, most VLOPs actioned a relatively small number of notices. For example, the Apple App Store actioned only 2% and X (formerly Twitter) 19% of all submitted notices.
- › **Complaints handling:** The quantitative data indicates that a complaint has a fairly high chance of overturning the original decision once submitted.
- › **Own-initiative content moderation:** All VLOPSEs prefer limiting the visibility of violative content rather than removing it. They also target primarily hate speech, adult nudity and sexual content, and incitement to violence. Marketplaces, such as Amazon or AliExpress, focus heavily on scanning their respective services, while social media platforms refrain from scanning and instead focus on a wide range of detection methods, emphasising algorithms and external contractors that flag content.
- › **Automated content moderation:**
 - **Reliance:** The reports showed an extensive reliance on automated means of content moderation. For example, automation is responsible for over 90% of all content moderation actions taken by Meta's social media platforms (Facebook 93%; Instagram 98%). All social media platforms utilise a combination of both automated and manual review, with automation being used for the most graphically discernible categories of violative content (e.g. child pornography or CSAM). YouTube and Meta's Facebook and Instagram use 'hashing' to prevent the re-upload of violative content while TikTok manually reviews all pieces of content once they reach a particular virality.
 - **Accuracy:** All companies display an adequate compliance rate (average 84%), with the lowest (72%) being the description of accuracy indicators for each applicable Member State language. Platforms adopted varying techniques to calculate the accuracy of their automated tools. Many used successful appeals against automated decisions as a proxy for accuracy. A different methodology applied by Google involved comparing human evaluations of a random sample of user content against decisions made by its automated systems on that same content, with the human evaluation assumed as the ground truth. Alternatively, Meta described a methodology for measuring random samples of automated decisions against pre-established expectations for policy enforcement.²⁹
- › **Suspensions:** There were mixed approaches to suspensions. Several VLOPSEs did not suspend any user for either manifestly illegal content or submission of manifestly unfounded notices or complaints. On the other hand, a few VLOPSEs took a deliberate suspension approach, amounting to hundreds or even thousands of cases.
- › **Human resources:** Facebook and Instagram, each with more than 255 million users in the EU, have together 1362 content moderators, while TikTok, with more than 135 million users, has 5827. Google services (excluding Google Maps), Facebook, Instagram and TikTok span some, but not all national languages. The remaining VLOPSEs focus only on a few major EU languages while neglecting the rest.



What sort of information could content moderation data tell us?

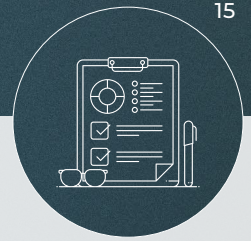
Under the DSA, platforms are required to submit regular data about content moderation decisions made on their platform to the European Commission. The Commission then collates this data into a single, interactive database for public transparency. In April 2024, this database—called ‘The Transparency Database’—contained 16 billion entries. Since its launch in September 2023, several studies have been conducted on the Database, with different temporal scope and different datasets, but they all aim to provide more information about the nature of content moderation decisions—and their impact on freedom of expression—across the Union. A comparable Australian database could be compiled by an Australian regulator or by an academic or third sector partner, provided that regular data is provided by platforms in prescribed ways.

In Europe, three major studies (Trujillo *et al.*,³⁰ Drolsbach and Pröllochs,³¹ and Kaushal *et al.*³² found that:

- › **Moderation decisions:** All studies reveal that content moderation decisions by platforms, for the overwhelming majority (more than 99%) are made based on violation of terms of services and not because the content is illegal under EU member state laws. This suggests that platforms rely more on their community guidelines to drive moderation actions, rather than EU or member state law.
- › **Grounds for decisions:** Trujillo *et al.* provided comparative analysis of the platforms based on the different infringement categories and identified three groups of similar platforms:
 - (1) those in which ‘scope of platform service’ was the most common infringement (YouTube, Facebook, LinkedIn, TikTok, and Instagram);
 - (2) those in which ‘pornography or sexualized content’ was the most common infringement (Pinterest and X);
 - (3) a single-platform group (Snapchat), for which the previous categories are the most common, along with ‘scams and fraud’.

Platforms sometimes used additional optional keywords: Snapchat used ‘sexual abuse material’ and ‘online bullying intimidation’. Pinterest predominantly used ‘adult sexual material’. X used additional keywords for each of its entries, using in total 17 distinct keywords, more than any other social media. However, in almost half (43%) of its entries, the keyword is ‘other’ (detailed as a policy violation), while in 49% it is adult sexual material. In the remaining entries, the keywords are primarily ‘child sexual abuse material’, ‘nudity’, and ‘goods/services not permitted’.

- › **Moderation actions:** Kaushal *et al.* showed that most of the restrictions are done at the level of content visibility (93.4%), followed by account (6.5%), service provision (1.1%), and monetary limitations (0.015%).
 - In terms of subcategories, restriction by visibility is further distributed among values of disabled content (66%), removed content (24%), and other (8%). Very rarely, content has been demoted (0.08%) and age restricted (0.03%).
 - Content is mainly disabled by Google Shopping (84%), followed by Amazon (14%). Most of the content removal is done by TikTok (56%), Facebook (15%), Google Maps (12%), Amazon (8%), YouTube (7%), and Pinterest (4%).
 - Demotion of content is only being done by Facebook (98.2%) and Instagram (1.8%) whereas age restriction of content is only being done by TikTok (90.3%) and YouTube (9.7%).



How does the DSA's 'Article 40' researcher access regime work?

Who can request data?

- › Only 'vetted' researchers can request data. To become a vetted researcher, a researcher must meet the following requirements:
 - They must be affiliated to a research organisation. The definition potentially covers a wide range of non-university organisations, including civil society organisations;
 - The applicant, or the lead applicant, needs to be affiliated with a European research organisation;
 - They must be independent from commercial interests;
 - They must have an appropriate research proposal, including demonstrating that:
 - The proposal will contribute to the understanding of systemic risks as laid out in Article 34(1) and/or to assess risk mitigation measures as outlined in Article 35;
 - They have disclosed the funding of the research;
 - The proposal will fulfil data security and confidentiality requirements, as well as to protect personal data, and explain how they intend to do so;
 - Access to the data within the indicated timeline is necessary and proportionate to the purposes of the research;
- › They must commit themselves to making their research results publicly available free of charge, within a reasonable period after the completion of the research.

Non-EU researchers can also apply, as long as they fulfil the criteria, are working with a EU research organisation or lead and their project is studying systemic risks in the Union (noting that risks can be transnational in nature).

What data can be requested?

Vetted researchers can only request access to data for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union, as set out pursuant to Article 34(1) DSA, and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures pursuant to Article 35.

What is the process for requesting data?

1. Researchers write an application for data access and send it to the national DSA regulator (called the 'Digital Services Coordinators');
2. The Digital Services Coordinator approves or rejects the access request;
3. If approved, the platform can ask for amendments to the data access request if:
 - The platform doesn't have the data;
 - The platform sees security risks or confidentiality issues, including trade secrets;
4. The platform provides the data to the researchers within a 'reasonable period'.

The practicalities of the process still remain open as a forthcoming delegated act works its way through the EU system, laying down the specific conditions under which VLOPs and VLOSEs are to provide data.

How's it working?

Researchers will be able to apply for access to data once Member States have appointed their respective Digital Services Coordinators and the delegated act is adopted, which is expected to occur later in 2024.

Publicly accessible data

Providers of VLOPs and VLOSEs must also give researchers access to publicly accessible data and if technically possible, real-time data. This is a different process that has been available to researchers since August 2023.

- › For this type of access, the researcher must fulfil a subset of conditions of "vetted researchers". They must be independent from commercial interests, disclose the funding of their research, be able to fulfil data security and confidentiality requirements, and explain how access to the data and the indicated timeline is necessary and proportionate to the purposes of the research.
- › The data must also be used solely for performing research that contributes to the detection, identification and understanding of systemic risks in the Union.
- › The process does not involve regulators and is between platforms and researchers.
- › This provision has been read as mandating tools such as CrowdTangle as well as to allow scraping activities.

02. The UK's Online Safety Act

Looking over to the United Kingdom, the new online safety regime provides for platform risk assessments and prescribed transparency reports. Note that the UK's scheme does not cover misinformation and disinformation. The framework as outlined here should be read as a reference for potential regulator-prescribed transparency reports and public-facing risk assessments.

The UK's *Online Safety Act* (UK OSA) creates a number of different duties that are applied to user-to-user and search services that are accessible to end users in the UK. The regulator—Ofcom—is currently undertaking a number of consultations to inform its implementation of the Act. Under the UK OSA, all regulated services have to comply with duties relating to illegal content and, if they are likely to be accessed by children, duties relating to child safety. The Act also has extraterritorial application, meaning it applies to companies established outside the UK. A number of additional duties are imposed on certain categories of services, broadly larger or riskier services, including duties relating to transparency (see Figure 1). The relevant Secretary of State has not yet laid regulations to set the criteria for inclusion in these categories.



All user-to-user systems have duties regarding:

- › Illegal content risk assessments
- › Illegal content
- › Content reporting
- › Complaints procedures
- › Freedom of expression and privacy
- › Record keeping and review.

All services likely to be accessed by children have duties regarding:

- › Children's risk assessments
- › Protecting children's online safety.

The largest online services also have additional duties regarding:

- › Adult risk assessment duties
- › Duties to protect adults' online safety
- › Duties to protect content of democratic importance
- › Duties to protect journalistic content.

Note, that the UK's 'plural duties' approach has reduced the breadth of industry's obligations, and has created gaps in protections for users. Reset.Tech advises that a 'singular' duty of care ensures protections can be truly systemic and preventative.

Figure 1: Overview of the duties of care present in the UK's Online Safety Act³³

The UK OSA has two key public transparency requirements:

1. Risk assessments.

Under the UK OSA platforms are required to produce risk assessments related to their duties. While these risk assessments serve as a key risk mitigation tool, they will also serve as a public accountability mechanism. Risk assessments will be submitted directly to Ofcom, but Category 1 and 2a (search) services will be required to publish summaries of their risk assessments, which will form a public transparency mechanism.

2. Annual transparency reports.

The UK OSA's public transparency provisions require services to provide annual transparency reports to Ofcom, who will specify the information, format, deadline for submission and publication date in a notice issued to the provider.³⁴ These metrics ensure comparability between annual reports and across platforms. Schedule 8 sets out a non-exhaustive list of the types of information that can be requested by Ofcom for inclusion in the transparency report.³⁵

Below are some of the types of information included in the list:

- › The formulation, development, scope and application of the terms of service.
- › The systems and processes for users to report content which they consider to be illegal content, content that is harmful to children or relevant content.
- › The systems and processes that a provider operates to deal with illegal content, content that is harmful to children and relevant content, including systems and processes for identifying such content and taking it down.
- › The design and operation of algorithms which affect the display, promotion, restriction or recommendation of illegal content, content that is harmful to children, or relevant content.³⁶

Providers are required by the OSA to ensure the material provided is “complete” and “accurate”.

These transparency reports will then become public, with some additional context from Ofcom. Ofcom is required³⁷ to publish annual reports with a “summary of conclusions drawn from the transparency reports” relating to patterns or trends, industry good practice or other information.

Ofcom will consult on its draft guidance on transparency reports in July 2024. The first transparency notices are likely to be served to regulated services in mid-2025.

These public transparency measures exist alongside strong information gathering powers for Ofcom. These information gathering powers came into force in October 2023³⁸ and Ofcom recently consulted on these powers,³⁹ but they broadly include:

- › Ofcom has powers to require interviews,⁴⁰ powers of entry, inspection and audit.⁴¹
- › Information notices: Ofcom can request “any information ... require(d) for the purpose of exercising, or deciding whether to exercise, any of their online safety functions” to assess providers’ compliance with their duties under the OSA and inform enforcement action.⁴²
- › Compelling the disclosure of information specifically about the use of service by deceased child users.⁴³ These powers came into force on 1 April 2024⁴⁴ and are the result of a half-decade long campaign and a landmark inquest into the suicide of Molly Russell.⁴⁵
- › Ad hoc, independent review: Ofcom can commission a ‘skilled persons report’ by a third-party to “identify and assess a failure, or possible failure by a provider of a regulated service to comply with a relevant requirement.” Ofcom does not envisage that it will request these reports regularly, however, and unlike in the EU these reports do not need to be made public.⁴⁶

The UK OSA includes strong enforcement tools should platforms fail to meet any of these public transparency measures. Ofcom may take enforcement action if services fail to respond to an information notice or provide incorrect or misleading information, which could result in a fine and a requirement to take steps to comply with the notice. Regulated services which fail to comply with their duties under the OSA, including transparency requirements, can be fined the greater of £18 million or 10 percent of their annual global turnover.⁴⁷ For extreme breaches like failing to follow information requests, senior managers can face criminal prosecution.⁴⁸ In the most extreme cases, Ofcom has the power to stop a service from being accessed in the UK.⁴⁹

Public support for stronger transparency measures

Vexed as the domain of platform regulation may be, regulating for enhanced transparency generally receives strong public support. In this section we present recent data demonstrating the Australian public's support for strengthened platform transparency measures.

In April 2024, Reset.Tech Australia worked together with YouGov to poll 1,514 people on their views about platform transparency. Specifically, we asked how people felt about measures for risk assessments, risk mitigations, transparency and accountability.

We started by asking the public if they thought digital platforms like social media companies should have a duty of care, and found strong support. Ninety-three percent of people agreed that social media companies should have a duty to take reasonable care of their users (see Figure 2).

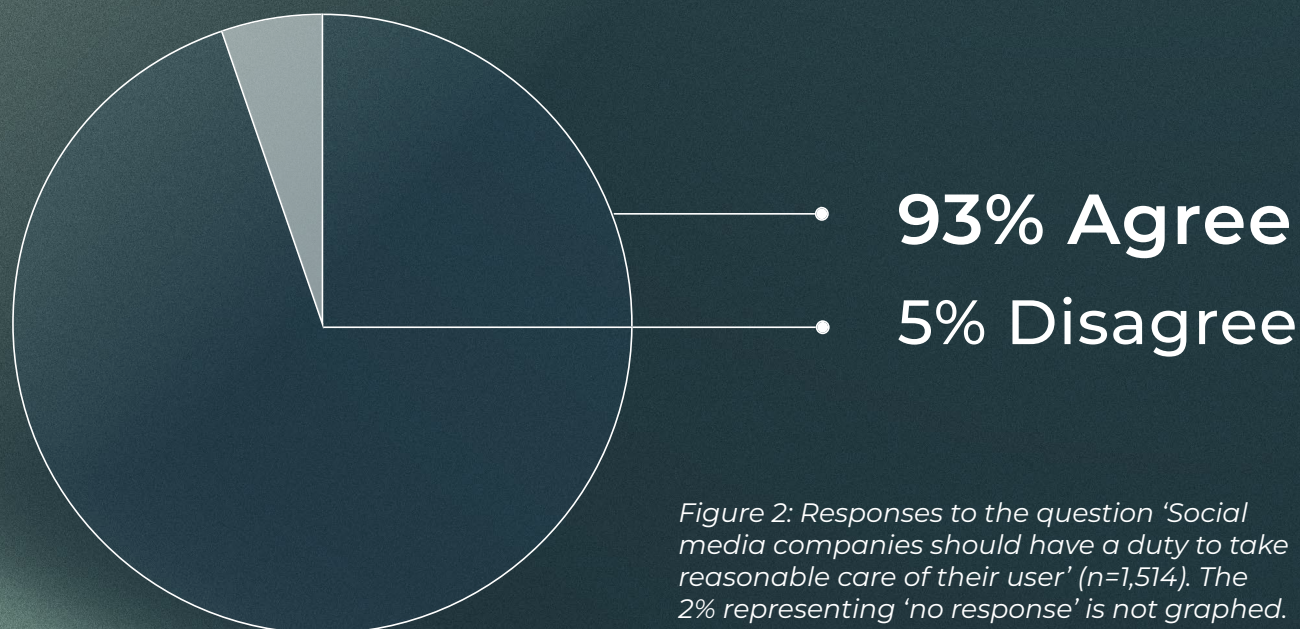
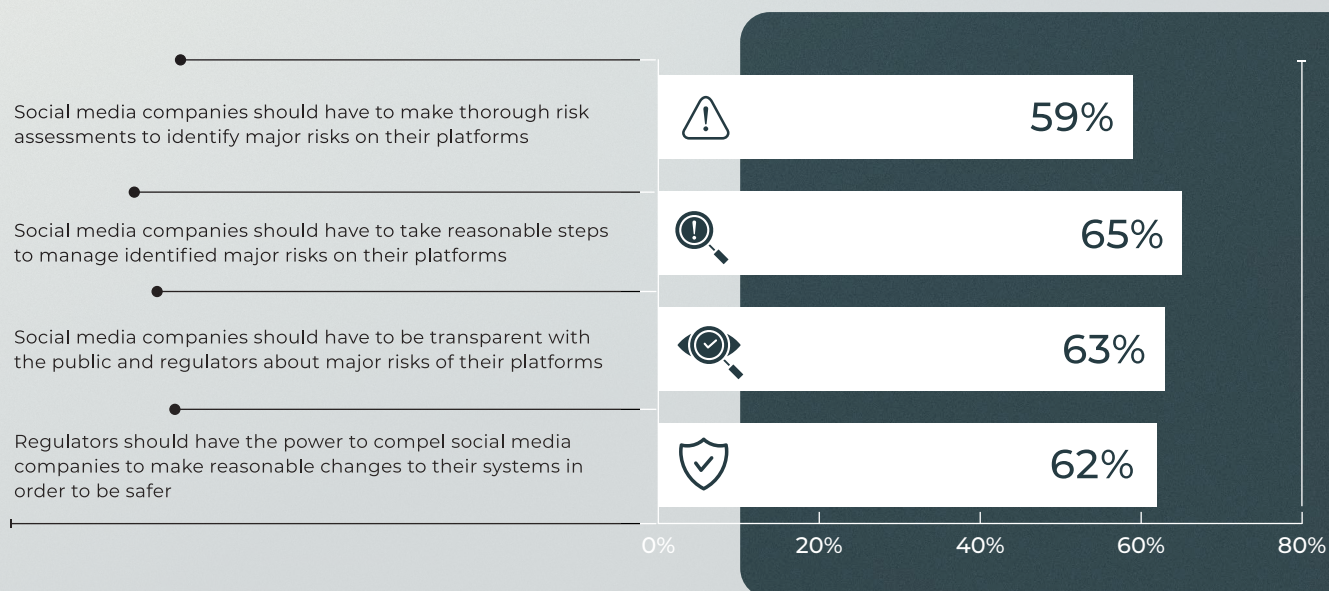


Figure 2: Responses to the question 'Social media companies should have a duty to take reasonable care of their user' (n=1,514). The 2% representing 'no response' is not graphed.

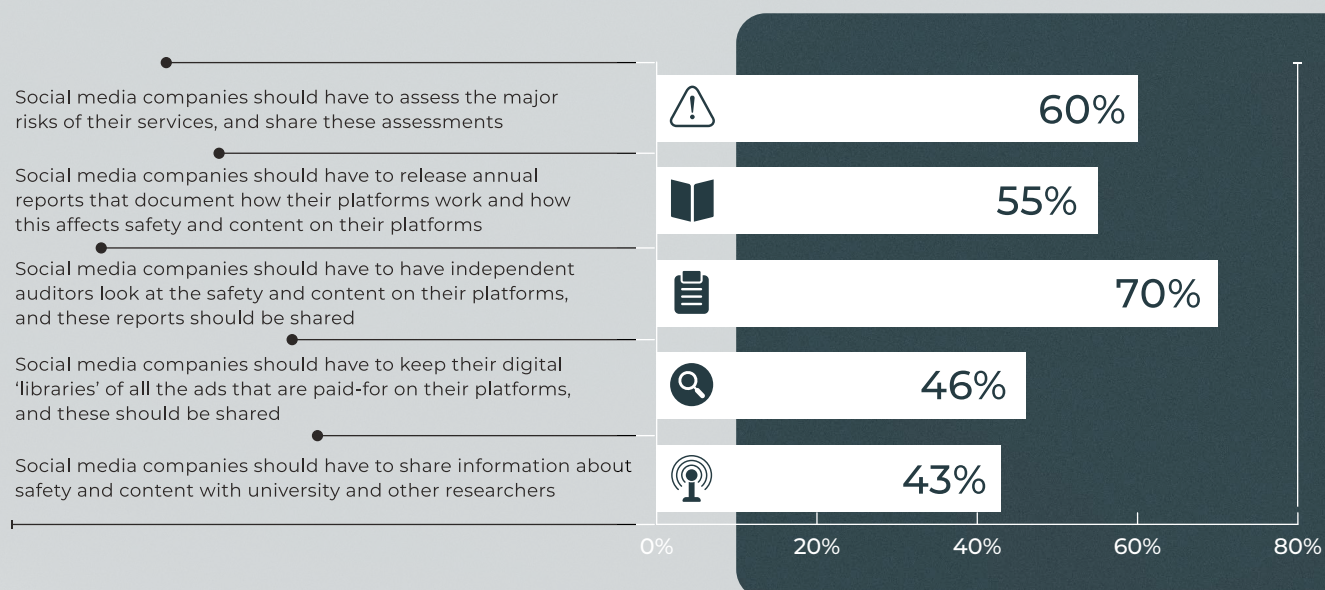
We also asked about measures for risk assessments and risk mitigations, transparency and accountability, and found strong support for systemic laws that increase accountability and transparency (see Figure 3).

Figure 3: Responses to the question 'It's not always clear if social media companies are responsible for the harms that happen on their platforms. There are some discussions that laws could be passed that make social media companies more responsible. Which, if any, of these responsibility measures would you support in law (select all you support)' (n=1,514)



We asked for more thoughts about transparency, using the ability to 'understand' how algorithms work as a case study. There was support for a battery of transparency measures (see Figure 4).

Figure 4: It's not always clear how social media companies build their systems and algorithms. There are some discussions that laws could be passed that make social media companies be more transparent about how platforms work and the consequences of this. Which, if any, of these transparency measures would you support in law? (select all you support)' (n=1,514)



Recommendations and proposals for a transparency framework for Australia

We have developed a template for public transparency measures in Australia for consideration.

These include five specific measures:

01. Risk Assessments

Risk assessments serve many purposes, but also have a public transparency purpose. We are not aiming to explore what makes an effective risk assessment—which requires detailed analysis beyond the scope of this transparency focused research—but rather to highlight their role in public transparency.

Australia's regulatory framework for digital risks, in both online safety regulations and misinformation and disinformation frameworks, should include requirements for detailed risk assessments. While the full risk assessments must go to regulators to review, summaries of these risk assessments should be made public. These public summaries need to include:

- › A detailed description of the risks identified on a platform;
- › Indications of the severity of the risks;
- › Measures of the scale of the risks in Australia, and;
- › Clearly outline what mitigation processes platforms have deployed in Australia.

02. Annual transparency reports

Annual transparency reports are necessary, that include clear prescriptions around information that must be shared to create meaningful transparency. It also helps create the conditions to compare services and track changes over time in meaningful ways.

These metrics could include, for example:

Metrics for misinformation and disinformation transparency reports

- › The volume and response to regulator orders and other legal requirements, e.g.:
 - Number of 'take down' orders issued by regulators, median, average and max time to respond to these, and final response;
 - Number of notices received regarding IP, defamation, privacy and illegal content notifications received from Australian end-users; median, average and max time to respond to these, and final response;

- Notices processed using automated means;
- Data about number of out of court settlement made;
- › Content moderation metrics, including impact on Australian businesses and pages. e.g.:
 - Number of organic content measures (i.e. how much content they proactively detected) that violated their community guidelines; by violation type (e.g. violated self harm policy, was fact-checked misinformation etc); amount detected by automated means; amount detected by human moderators; median, average and max time to detect these, and final response; business specific metrics, e.g.:
 - Number of organic business entity measures (i.e. how many Australian business accounts were removed and restricted as a result of organic content moderation);
 - Number of organic entity measures (i.e. how many Australian pages or products were removed and restricted as a result of organic content moderation);
 - Number of user-reported content measures (i.e. how much content was reported to the platform by Australian end-users) that violated their community guidelines; by violation type; median, average and max time to detect these; response; number of challenges against response; final outcomes; business specific metrics, e.g.:
 - Number of business entity measures following user-reporting (i.e. how many Australian business accounts were removed and restricted after user-reporting);
 - Number of entity measures following user reporting (i.e. how many Australian pages or products were removed and restricted after user-reporting);
 - Number of 'trusted-flagger' content measures (i.e. how much content was acted on by a platform as a result of Australian fact-checkers or trusted flaggers); amount reported to platform; by violation type; amount subsequently detected by automated means; median, average and max time to detect these; response; number of challenges against response; final outcome;
 - Indicators of accuracy and error rates for automated review processes; both for organic detection and following user reporting;

- Human resources dedicated to content moderation, including information about; number located within Australia; number dedicated to Australian content or addressing reports from Australian end-users; qualifications and training; support; volume of work (i.e. how much content per hour are they required to review); language addressed;
- Measures against misuse such as number of Australian end-users:
 - Number of accounts suspended or deleted and why; number of challenges and final outcome;
 - Number of Australian end-users monthly, including breakdowns by under 18 and over 18.

Metrics for online safety transparency reports

Some of the metrics that provide transparency around misinformation and disinformation would also provide public transparency around safety, especially regarding data around content moderation and monthly average users. However, drawing on Article 35.1 of the DSA, which notes a range of mitigation measures platforms could adopt, a number of additional metrics could be developed to provide greater transparency around systemic safety features and their adoption. Below, we have listed some non-exhaustive options:

- › Metrics on the design, features, or functioning of services, e.g.:
 - Data around internal safety tests made of features and systems conducted, including a description of tests and outcomes, and nature of adaptations made as a result that affect Australian end users;
 - Changes to Community Guidelines and Terms of Service for Australian end users;
 - Human resources dedicated to trust and safety, including information about; number located within Australia; number dedicated to Australian safety issues and safety of Australian end-users; qualifications and training; and support.
- › Problematic use metrics, e.g.:
 - Number of adult users demonstrating problematic over-user, and data about average and median use times;
 - Number of push notifications sent to these users on average per day (in app and out of app);
 - Number of child users (under 18) demonstrating problematic over-user, and data about average and median use times;
 - Number of push notifications sent to these users on average per day (in app and out of app);
 - Number of child users (under 18) accessing the platforms between 10pm and 6am in their time zone, and data about average overnight usage;
 - Estimates of number of users under the minimum age of use according to the terms of service, and data about average detection and response to these accounts;
- › Child sexual exploitation and abuse metrics:
 - Numbers of adult users blocked for contact with minors, and data about response times and previous reportings of users;
 - Numbers of adult users reported by minors, and data about responses and response times;
 - Number of CSAM reports made, and data about response and response times;
- › Online scam metrics, e.g.:
 - Number of online scam and scam posts reported on the platform, including data about detection method (organic or user-report), average engagement and responses including average response time;
- › Content moderation metrics, including impact on Australian businesses and pages (as described in a potential misinformation and disinformation reporting framework) e.g.:
 - Number of organic content measures (i.e. how much content they proactively detected) that violated their community guidelines; by violation type (e.g. violated self harm policy, online scams, fact-checked misinformation etc); amount detected by automated means; amount detected by human moderators; median, average and max time to detect these, and final response; business specific metrics, e.g.:
 - Number of organic business entity measures (i.e. how many Australian business accounts were removed and restricted as a result of organic content moderation);
 - Number of organic entity measures (i.e. how many Australian pages or products were removed and restricted as a result of organic content moderation);
 - Number of user-reported content measures (i.e. how much content was reported to the platform by Australian end-users) that violated their community guidelines; by violation type; median, average and max time to detect these; response; number of challenges against response; final outcomes; business specific metrics, e.g.:
 - Number of business entity measures following user-reporting (i.e. how many Australian business accounts were removed and

restricted after user-reporting);

- Number of entity measures following user reporting (i.e. how many Australian pages or products were removed and restricted after user-reporting);
 - Number of ‘trusted-flagger’ content measures (i.e. how much content was acted on by a platform as a result of Australian fact-checkers or trusted flaggers); amount reported to platform; by violation type; amount subsequently detected by automated means; median, average and max time to detect these; response; number of challenges against response; final outcome;
 - Indicators of accuracy and error rates for automated review processes; both for organic detection and following user reporting;
 - Human resources dedicated to content moderation, including information about; number located within Australia; number dedicated to Australian content or addressing reports from Australian end-users; qualifications and training; support; volume of work (i.e. how much content per hour are they required to review); language addressed;
- › Number of Australian end-users monthly, including breakdowns by under 18 and over 18.

03. Annual independent audits

An independent expert or ‘skilled person’ should be required to review both the platforms’ risk assessment and transparency notice. More detailed, technical assessments could be offered to regulators, but at a minimum an independent auditor must issue an annual public assessment.

04. Data portals, including ad repository and content moderation databases

Following the EU model, platforms could make at least two data portals featuring Australian specific data. This includes:

- › **Searchable ad repositories.** A database of paid-for advertising posted by Australian users and / or targeting Australian users needs to be created. This should include details about the content of the ad, details about the advertiser, indications of advertising spend, details about targeting parameters, metrics of views and engagements and details about who received the ad (given that automated targeting may render targeting parameters void). Given platforms’ demonstrated weaknesses at differentiating between ‘political/ social affairs’ advertising and other advertising,⁵⁰ we would recommend an expansive view of what should be included in this database. These databases should be provided by platforms, but data could also be provided in standardised ways to allow the creation of a central database.
- › **Data about content moderation decisions.** Platforms should be required to provide data about what content moderation decisions have happened on their platforms that affect content posted by Australian users. This data should include reasons the content was moderated (violation of guideline X, illegal content, request from the Office of the eSafety Commissioner etc), the nature of the content moderation action taken (e.g. removal, demotion, labelling etc), data about if the moderation was user flagged or proactively detected by the platform, use of human vs automated moderation, data about the reach and engagement of the content before action was taken and data about challenges made to content moderation decisions and outcomes. This should be provided in a standardised way to allow the compilation of a searchable centralised database.

05. Researcher access requirements

Australia should have a researcher access scheme that is comparable in scope to the EU scheme. In general, we would support a system where:

- › Vetted researchers could request data. Like in the EU, requirements for an Australian vetted researcher could include:
 - Affiliation to a research organisation, including academic and third sector research organisations;
 - Researchers or at least the lead researcher should be an Australian resident;
 - Non-commercial purpose limitations;
- › Suitable research projects are provided with data. A suitable proposal would include information demonstrating that:
 - The research aligns with the ambitions of the underlying regulatory powers, and how it is broadly of public benefit. This does not include data about trade secrets;
 - Funding for the research is fully disclosed;
 - Access to the specific data requested, and the indicated timeline indicated, is necessary and proportionate to the purposes of the research;
 - Data security and confidentiality requirements, as well as personal data safety requirements, will be fulfilled, and;
 - The research results will be made publicly available free of charge within a reasonable period after the completion of the research.
- › The process for requesting data was managed by the ACMA or other independent organisation. The regulator or an appointed independent (i.e. non-industry body) should receive and regularly review requests. Once they have approved a request, the request should be sent to a platform to implement within a reasonable period. There needs to be a transparency process where researchers and platforms can challenge decisions made by the body.
- › Existing data and data tools should be made available to Australian researchers for free. For researchers who meet the threshold of vetted research and commit to using the data for advancing the ambitions of the *Online Safety Act*, access to APIs, scraping tools or research tools should be provided directly by the platforms. Platforms already build and maintain APIs and other tools which include data they decide is relevant to researchers. This provision is about ensuring that Australian researchers can access these. For example, the TikTok researchers' API is only available to European and American researchers currently. The regulator or independent body that oversees specific data access requests could also manage a mediation process for disputes about access to these tools, but we envisage that this process could be managed directly between platforms and researchers.

These five public transparency measures need to exist alongside information gathering powers afforded to the ACMA and Office of the eSafety Commissioner. Regulators need to be adequately resourced to exercise these powers.

Transparency also requires strong enforcement powers to be realised. We note the proposals in the exposure draft of the Combatting Misinformation and Disinformation Bill provide only modest enforcement measures, suggesting that ACMA be able to issue fines of up to 5%, and that the *Online Safety Act* has limited enforcement powers regarding transparency measures, and that maximum fines levels are inadequate. Other global digital regulators can issue fines of up to 10% of global turnover, and other Australian regulators also have these powers in other domains.⁵¹ This higher rate would be more appropriate. Further, following the UK's model, we recommend investigating the possibility of criminal penalties for executives where transparency requirements are blatantly disregarded.⁵² Other enforcement measures will be necessary for other aspects of digital regulation but are beyond the scope of this paper (e.g. enforcing a duty of care, etc).

Appendices

Appendix 1.....25

Appendix 2..... 26



Appendix 1

5 pillars of systemic regulation

1. A duty of care shaping platforms' actions	2. Requirements for risk assessments	3. Requirements for risk mitigations	4. Requirements for transparency measures	5. Requirements for accountability measures
An overarching duty of care would place broad obligations on platforms to ensure user safety in systemic ways. Specific responsibilities could be enumerated by focusing requirements for risk assessments. The UK OSA⁵³ introduces duties of care, and draft Canadian legislation⁵⁴ introduces duties on services; however, both are pluralised, which reduces the systemic focus. The EU's DSA⁵⁵ regulations have similar systemic obligations but are phrased as responsibilities to address particular risks, specifically risks posed by: • Illegal content • Negative effects for the exercise of fundamental rights, such as dignity and privacy and political freedoms, as outlined in the European Charter • Negative effects on civic discourse and electoral processes and public security • Negative effects on gender-based violence, public health, children's wellbeing and serious negative consequences to people's physical and mental wellbeing.⁵⁶	Requirements for platforms to assess <i>all their</i> systems and elements for risks would incentivise systemic change and help platforms realise their duty of care. Risk assessments could be focused on addressing the following: • Illegal materials (such as class 1A & 1B materials as already included in the <i>Online Safety Act</i>) • Harmful materials (such as image-based abuse, abhorrent violent material, cyberbullying of children and abuse of Australian adults as already included in the <i>Online Safety Act</i>) • Online scams • Risks to electoral processes and public security • Risks to human rights, such as political freedoms, hate speech and violence or serious harm to individuals • Risks to gender-based violence, children's best interests, public health and the environment. These risk assessments should go to regulators, and summaries could be made public after a period of time. Both the EU's DSA and the UK OSA require risk assessments.	As a corollary of risk assessments, platforms must be required to implement reasonable steps to mitigate against each risk identified. These measures must be included in the assessments sent to regulators. Both the EU's DSA and the UK OSA require risk mitigations against risks identified in assessments. Canada's Online Harms Bill also places obligations on platforms to mitigate risks aligning with their duties.	Five different measures could be introduced to enhance public transparency: • Annual risk assessments • Annual public transparency reports, which are heavily prescriptive • Annual independent audits of risk assessments and transparency reports • Data portals, including ad repository and content moderation databases • Researcher access to public interest data and requirements that vetted researchers can access public interest data. These need to exist alongside strong investigative powers for regulators. The EU's DSA requires this sort of public transparency regime.	To meaningfully drive change, regulations need to be enforceable. Specifically, regulators must be empowered and resourced to: • Compel redress and changes to platforms' systems and elements rather than just compel transparency or take-down • Issue penalties that match the scale of global profits of digital platforms • Have powers to 'turn off' services where failures are persistent and all other measures have been exhausted • Enhance the public-facing complaints mechanism to include complaints from individuals and consumer groups regarding systemic risks and breaches of their duty of care • Have strong investigative and information-gathering powers • Effective notice-and-take-down powers. The European Commission, the UK's Ofcom and the new Canadian Digital Safety Commissioner have combinations of these enforcement powers. The UK goes even further and includes criminal sanctions for executives regarding transparency.⁵⁷

Appendix 2

Transparency Reports Comparison

The table below compares data submitted by Meta to the *Australian Code of Practice on Disinformation and Misinformation* transparency reporting process and Facebook to the *Digital Services Act* transparency reporting process. The metrics selected are indicative and not exhaustive, and should be read in conjunction with the list of potential metrics from pages 20-22.

Notes:

Meta/Facebook’s selection is made without any qualitative assessment of good or bad performance, rather, it is offered as a significant and typical platform to examine in an Australian context.

The *Digital Services Act* asks for information on a different set of risks to the Australian categories of ‘misinformation and disinformation’. As a result, these are not neat

comparisons, but are offered to demonstrate what improvements could be made to Australia’s public reporting framework for misinformation and disinformation.

Meta provided two new metrics for Australia in its May 2024 report: the number of ads removed for not complying with misinformation policies, and the number of ads removed for not complying with policies on social issues, elections

or politics (‘SIEP’). While data on enforcement of advertising policies would be very welcome, these numbers are meaningless if not matched to denominators, such as the total number of advertising reports received. We also note that replicated, repeated evidentiary testing of Meta’s advertising reviews system throws doubt on the system’s robustness at detecting misinformation in ads.

Metric	Regulation (EU) 2022/2065 (Facebook)	<i>Australian Code of Practice on Disinformation and Misinformation</i> (Meta: Facebook and Instagram)	Commentary
Removal of content violating policies	Precise numbers provided of content removed, with reference to: - Whether removed of Facebook's own accord or in response to reporting, - If removed in response to reporting, a total volume is also given of removal requests, - The proportion removed by automated means, - The policy the content was removed under. For an example of this data for Facebook in the EU, please see over the page.	<i>"Between 1 January and 31 December 2023, we took action on over 9,700 pieces of content across Facebook and Instagram for violating our Misinformation policies. This is compared to 91,000 pieces of content for the 2022 reporting period and 180,000 (globally) in 2021."</i>	For Australia, the following points are unclear: - Allocation of the content across Meta's products, since the 9,700 is bundled across Facebook and Instagram, - What measures 'take action' precisely refers to (presumably removal, given the heading), - Whether this action was in response to user reports or Meta's proactive searches, - Meta's responsiveness to Australian user reports, - What policy the content was removed under (the policy cited appears to cover multiple areas, spanning elections to health), noting the drop from 91,000 to 9,700 is explained as due to policy changes to health misinformation.
Complaints handling system	Precise numbers provided for the complaints handling and appeals system, including: - Numbers of complaints received, by policy area, and number upheld, - Total number of reports received for content removal, and total number actually removed, - Number of complaints with a legal basis received and number restored, - Median time for decisions on complaints, - Median time for actioning appeals.	No data	The European data into complaints handling, while high-level, provides an indication of the volume of user complaints and appeals, and an indication of Facebook's responsiveness to this user input. Arguably, reporting these figures publicly puts the platform 'on notice' to resolve user complaints and appeals in a timely fashion. The data on volume of user reports and proportion of complaints leading to removal, or appeals upheld, would clearly be helpful for building public trust in the vexed area of misinformation and disinformation.
Average monthly active users	Provided in orders of 'millions', to one decimal place, updated every six months	No data	Australians lack official numbers for the number of users accessing Meta products (registered and logged in users browsing the platform through a website or mobile device at least once a month).
Volume of human resources dedicated to content moderation	Precise numbers provided per language group of (human) content moderators, broken down by official EU language	No data	Without this data, Australians have no way of understanding the human resources responsible handling Australian users' reports and assessing the appropriateness of the investment.

Appendix 2

EU Transparency Dashboard Data

- › Source of detection for terms of service violations: 22,000 in total
- › 99.94% detected by Facebook
- › Trusted flagger notices, 0%
- › Article 16 (illegal notice from Governments) less than 0.1%

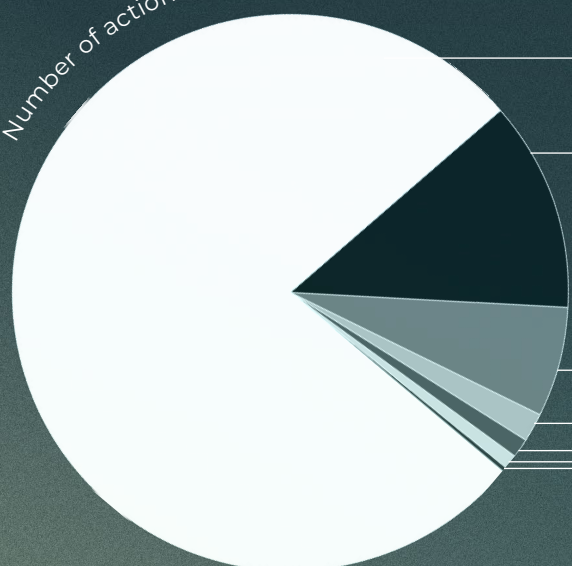
This figure draws upon the public data available in the EU's 'Transparency Dashboard' to show what possibilities exist in the EU for observing the volume and type of actions taken by platforms. In this example, data is taken for a 30 day period from May to June 2024 for Facebook specifically. Note, these figures can be calculated for any day range, for any platform in scope.

For the period of the data pull, Facebook took 35,869,477 'actions'.

The dashboard tool is available at:

<https://transparency.dsa.ec.europa.eu/dashboard>

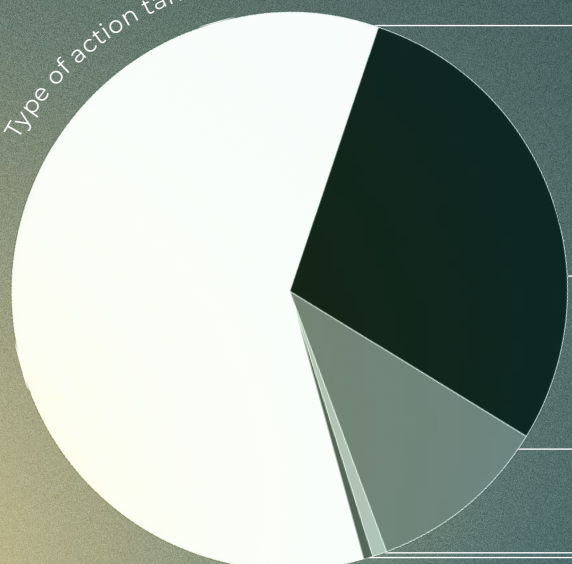
Number of actions taken



Number of actions taken:

- 27,696,000** for violations of scope of platform service (i.e. community guidelines)
- 4,341,886** for data protection and privacy violations
- 2,287,888** for violence
- 547,098** for pornography or sexualised content violations
- 457,534** for illegal or harmful speech
- 219,099** for scams or fraud
- 92,359** for illegal or unsafe products

Type of action taken



Type of action taken:

- 21,375,000** accounts suspended
- 10,267,000** content removed
- 3,753,000** content demoted
- 343,000** content labelled
- 131,000** accounts terminated

Endnotes

- 1 Digi 2022 *Australian Code of Practice on Disinformation and Misinformation* <https://digi.org.au/disinformation-code/>
- 2 See Appendix 2, Digi 2022 *Australian Code of Practice on Misinformation and Disinformation* <https://digi.org.au/wp-content/uploads/2022/12/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL--December-22-2022.docx.pdf>
- 3 Australian Communications and Media Authority 2021 *A report to government on the adequacy of digital platforms' disinformation and news quality measures* <https://www.acma.gov.au/sites/default/files/2021-11/Adequacy%20of%20digital%20platforms%20disinformation%20and%20news%20quality%20measures.pdf>
- 4 Australian Communications and Media Authority 2023 *Digital platforms' efforts under the Australian Code of Practice on Disinformation and Misinformation Second report to government*, p. 3 <https://www.acma.gov.au/sites/default/files/2023-07/Digital%20platforms%20efforts%20under%20Code%20of%20Practice%20on%20Disinformation%20and%20Misinformation.pdf>
- 5 Uri Gal 2022 'Transparency reports' from tech giants are vague on how they're combating misinformation. It's time for legislation' *The Conversation* <https://theconversation.com/transparency-reports-from-tech-giants-are-vague-on-how-theyre-combating-misinformation-its-time-for-legislation-184476>
- 6 Reset.Tech Australia 2024 *Functioning or Failing?* <https://au.reset.tech/uploads/Reset.Tech-Functioning-or-Failing-Report-V2C.pdf>
- 7 Aus Tender 2024 *Tender to establish Contract for Analysis of metrics to assess digital platform effectiveness under voluntary, self-regulatory* <https://www.tenders.gov.au/Atm/ShowClosed/32c0f214-5a50-4cf9-98e1-b2f72aa54ddc?PreviewMode=False>
- 8 See more at Reset.Tech Australia 2024 *Regulating for Transparency: Transparency Reports in Australia* <https://au.reset.tech/uploads/Regulating-for-Transparency-Reset.Tech.pdf>
- 9 TikTok 2022 *Annual Transparency Report* <https://digi.org.au/wp-content/uploads/2023/05/TikTok-2022-Annual-Transparency-Report.pdf>
- 10 Daniel 2024 'TikTok Users and Growth Statistics (2024)' *SignHouse* <https://www.usesignhouse.com/blog/tiktok-stats>
- 11 Note, some information about this is publicly available, for example:
 - Meta 2024 *Where We Have Fact Checking* <https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking/partner-map>
 - Arjun Narayan Bettadapur 2020 'TikTok partners with fact-checking experts to combat misinformation' *TikTok* <https://newsroom.tiktok.com/en-au/tiktok-partners-with-fact-checking-experts-to-combat-misinformation>

A full list of Australian operating fact-checkers can be found at IFCN 2024 *Verified signatories of the IFCN code of principles* <https://ifcncodeofprinciples.poynter.org/signatories>
- 12 eSafety Commissioner 2024 *Responses to transparency notices* <https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices>
- 13 Commonwealth of Australia *Online Safety Act 2021 Division 3(A) 49 3* <https://www.legislation.gov.au/C2021A00076/latest/text>
- 14 Commonwealth of Australia *Online Safety Act 2021 Division 3(A) 49 2 OSA* <https://www.legislation.gov.au/C2021A00076/latest/text>
- 15 Commonwealth of Australia *Online Safety Act 2021 Division 3(A) 49 5 OSA* <https://www.legislation.gov.au/C2021A00076/latest/text>
- 16 Commonwealth of Australia *Online Safety Act 2021 Division 3(A) 56 2 OSA* <https://www.legislation.gov.au/C2021A00076/latest/text>
- 17 Commonwealth of Australia *Online Safety Act 2021 Division 3(A) 59 2 OSA* <https://www.legislation.gov.au/C2021A00076/latest/text>
- 18 Available at eSafety Commissioner 2024 *Responses to transparency notices* <https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices>
- 19 eSafety Commissioner 2024 *Summary of response from X Corp. (Twitter) to eSafety's transparency notice on online hate* <https://www.esafety.gov.au/sites/default/files/2024-01/Full-Report-Basic-Online-Safety-Expectations-Summary-of-response-from-X-CorpTwitter-to-eSafetys-transparency-notice-on-online%20hate.pdf>, 19, note "Subsequent information was provided by X Corp. after the Notice by the deadline that did seek to rectify earlier omissions of information provided. eSafety took this into consideration into account in deciding upon the appropriate enforcement action."
- 20 eSafety Commissioner 2023 *Formal warning under section 58 of the Online Safety Act 2021 (Cth)* <https://www.esafety.gov.au/sites/default/files/2023-10/Formal-Warning-to-Google-LLC.pdf>
- 21 eSafety Commissioner 2023 *Service provider notification in relation to contravention of section 57 of the Online Safety Act 2021 (Cth)* <https://www.esafety.gov.au/sites/default/files/2023-10/Service-Provider-Notification-to-X-Corp.pdf>
- 22 eSafety Commissioner 2024 *Responses to transparency notices* <https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices>
- 23 See X Corp v eSafety Commissioner (VID956/2023), status available at: <https://www.comcourts.gov.au/file/Federal/P/VID956/2023/actions>
- 24 EU 2022 *Digital Services Act Regulation* <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065>
- 25 These are designated when a service reaches 45 million monthly active users in the EU (10% or more of the EU population). A full list of current VLOPs and VLOEs can be found at European Commission 2024 *DSA: Very large online platforms and search engines* <https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops>
- 26 European Commission 2024 *Published consultations and feedback: Digital Services Act – transparency reports (detailed rules and templates)* https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14027-Digital-Services-Act-transparency-reports-detailed-rules-and-templates_en
- 27 European Commission 2023 *DSA Transparency Database* <https://transparency.dsa.ec.europa.eu/>
- 28 See Article 40, EU 2022 *Digital Services Act* <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065>
- 29 Jakub Rybníkář and Katarína Drevená 2024 'DSA Transparency Reports - Briefing, 31 January 2024' *Council for Media Services* https://rpms.sk/sites/default/files/2024-02/CMS_DSA_TransparencyReports_briefing.pdf
- 30 Tremau 2023 *Peeling Back The Screen: Lessons From the First DSA Transparency Reports* <https://tremau.com/peeling-back-the-screen-lessons-from-the-first-dsa-transparency-reports/>
- 31 Amaury Trujillo, Tiziano Fagni, and Stefano Cresci 2023 *The DSA Transparency Database: Auditing Self-reported Moderation Actions by Social Media* <http://arxiv.org/abs/2312.10269>
- 32 Chiara Drolsbach and Nicolas Pröllochs 2023 *Content moderation on social media in the EU: Insights from the DSA Transparency Database* <https://arxiv.org/abs/2312.04431>

- 32 Rishabh Kaushal, Jacob van de Kerkhof, Catalina Goanta, et al. 2024 *Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database* <https://arxiv.org/abs/2404.02894>
- 33 HM Government 2023 *Online Safety Act* <https://www.legislation.gov.uk/ukpga/2023/50/enacted>
- 34 HM Government 2023 *Online Safety Act Chapter 5* <https://www.legislation.gov.uk/ukpga/2023/50/part/4/chapter/5/enacted>
- 35 HM Government 2023 *Online Safety Act Schedule 8* <https://www.legislation.gov.uk/ukpga/2023/50/schedule/8/enacted>
- 36 HM Government 2023 *Online Safety Act Schedule 8* <https://www.legislation.gov.uk/ukpga/2023/50/schedule/8/enacted>
- 37 HM Government 2023 *Online Safety Act Section 159* <https://www.legislation.gov.uk/ukpga/2023/50/section/159/enacted>
- 38 HM Government 2023 *Online Safety Act Section 100* <https://www.legislation.gov.uk/ukpga/2023/50/part/7/chapter/4/enacted>
- 39 Ofcom 2023 *Consultation: Protecting people from illegal harms online* <https://www.ofcom.org.uk/consultations-and-statements/category-1/protecting-people-from-illegal-content-online>
- 40 HM Government 2023 *Online Safety Act Sections 105 and 106* <https://www.legislation.gov.uk/ukpga/2023/50/part/7/chapter/4/crossheading/investigations-and-interviews/enacted>
- 41 HM Government 2023 *Online Safety Act Section 107* <https://www.legislation.gov.uk/ukpga/2023/50/part/7/chapter/4/crossheading/powers-of-entry-inspection-and-audit/enacted>
- 42 HM Government 2023 *Online Safety Act Section 100* <https://www.legislation.gov.uk/ukpga/2023/50/part/7/chapter/4/enacted>
- 43 HM Government 2023 *Online Safety Act Section 101 and 102* <https://www.legislation.gov.uk/uksi/2023/1420/regulation/3/made>
- 44 Ofcom 2024 *Ofcom's power to gather information relating to the death of a child* <https://www.ofcom.org.uk/online-safety/information-for-industry/information-gathering>
- 45 Adam Satariano 2022 'British Ruling Pins Blame on Social Media for Teenager's Suicide' *The New York Times* <https://www.nytimes.com/2022/10/01/business/instagram-suicide-ruling-britain.html>
- 46 HM Government 2023 *Online Safety Act Section 104* <https://www.legislation.gov.uk/ukpga/2023/50/part/7/chapter/4/crossheading/skilled-persons-reports/enacted>
- 47 HM Government 2023 *A guide to the Online Safety Bill* <https://www.gov.uk/guidance/a-guide-to-the-online-safety-bill>
- 48 HM Government 2023 *A guide to the Online Safety Bill* <https://www.gov.uk/guidance/a-guide-to-the-online-safety-bill>
- 49 HM Government 2023 *A guide to the Online Safety Bill* <https://www.gov.uk/guidance/a-guide-to-the-online-safety-bill>
- 50 Platforms routinely rely on advertiser self-disclosure of political advertisements, see e.g. Reset.Tech Australia 2023 *How do platforms handle electoral misinformation in paid-for advertising? An experimental evaluation using the Voice referendum* <https://au.reset.tech/uploads/Reset.Tech-Australia-Paid-For-Advertising.pdf>
- 51 Australian Competition and Consumer Commission n.d. *Fines and penalties* <https://www.accc.gov.au/business/compliance-and-enforcement/fines-and-penalties>
- 52 HM Government 2023 *A guide to the Online Safety Bill* <https://www.gov.uk/guidance/a-guide-to-the-online-safety-bill>
- 53 HM Government 2023 *Online Safety Act 2023* <https://www.legislation.gov.uk/ukpga/2023/50/enacted>
- 54 Canada 2024 *Online Harms Bill 2024* <https://www.parl.ca/LegisInfo/en/bill/44-1/c-63>
- 55 EU 2022 *Digital Services Act* <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
- 56 Article 34, EU 2022 *Digital Services Act* <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
- 57 UK Department for Science, Innovation & Technology 2024 *Online Safety Act: new criminal offences circular* <https://www.gov.uk/government/publications/online-safety-act-new-criminal-offences-circular/online-safety-act-new-criminal-offences-circular>



Reset.
AUSTRALIA

 **Susan McKinnon
Foundation**