

# Playing it safe: how online game services are (or are not) keeping kids safe online

**Transparency report: April 2026 non-periodic notices**

October 2026

## Contents

|                                                                        |           |
|------------------------------------------------------------------------|-----------|
| <b>Executive summary .....</b>                                         | <b>3</b>  |
| Significant variation in protection for users .....                    | 4         |
| <b>Key findings.....</b>                                               | <b>5</b>  |
| Furthering transparency.....                                           | 8         |
| <b>Definitions .....</b>                                               | <b>9</b>  |
| Harms .....                                                            | 9         |
| Other key terms.....                                                   | 11        |
| <b>Why eSafety is looking at online games .....</b>                    | <b>14</b> |
| Transparency .....                                                     | 15        |
| How features in online games shape online harm .....                   | 15        |
| Potential harms.....                                                   | 16        |
| How gaming features can contribute.....                                | 18        |
| Examples of how harms may play out .....                               | 20        |
| <b>What we examined .....</b>                                          | <b>21</b> |
| Who received the Notices? .....                                        | 21        |
| Online game services.....                                              | 21        |
| Gaming-adjacent services .....                                         | 24        |
| How many people are on these services?.....                            | 24        |
| What questions did eSafety ask? .....                                  | 26        |
| What information has been published, and what has been excluded? ..... | 27        |
| <b>Focus: Age assurance .....</b>                                      | <b>29</b> |
| Why were service providers asked about this?.....                      | 30        |
| Enforcing minimum age rules .....                                      | 30        |
| Applying default safety settings effectively .....                     | 31        |
| Restricting access to age-inappropriate content and experiences .....  | 31        |
| Limiting contact between adults and children .....                     | 31        |
| What we found.....                                                     | 32        |
| Age assurance tools and techniques.....                                | 32        |
| Identifying incorrect user age .....                                   | 35        |
| Parental controls .....                                                | 36        |
| Key insights and safety gaps.....                                      | 37        |
| <b>Focus: Proactive detection .....</b>                                | <b>39</b> |
| Why were service providers asked about this?.....                      | 39        |
| Safety measures in place.....                                          | 40        |

|                                                                                   |           |
|-----------------------------------------------------------------------------------|-----------|
| Sub-focus: Detecting online harms in chats and other communication features ..... | 41        |
| Why were service providers asked about this?.....                                 | 41        |
| What we found .....                                                               | 42        |
| Key insights and safety gaps.....                                                 | 53        |
| Sub-focus: Detecting online harms in user-generated and in-game content .....     | 54        |
| Why were service providers asked about this?.....                                 | 54        |
| What we found .....                                                               | 54        |
| Key insights and safety gaps.....                                                 | 59        |
| Sub-focus: Detecting high-risk games and harmful players.....                     | 59        |
| Why were service providers asked about this?.....                                 | 59        |
| What we found .....                                                               | 60        |
| Key insights and safety gaps.....                                                 | 64        |
| Sub-focus: Measuring the accuracy of proactive detection tools .....              | 65        |
| Why were service providers asked about this?.....                                 | 65        |
| What we found .....                                                               | 66        |
| Key insights and safety gaps.....                                                 | 73        |
| Sub-focus: Detecting transactions of virtual items and currencies.....            | 73        |
| Why were service providers asked about this?.....                                 | 73        |
| What we found .....                                                               | 74        |
| Key insights and safety gaps.....                                                 | 74        |
| <b>Focus: User reporting .....</b>                                                | <b>75</b> |
| Why were service providers asked about this?.....                                 | 75        |
| What we found.....                                                                | 75        |
| <b>Focus: Staff capacity and capability.....</b>                                  | <b>79</b> |
| Why were service providers asked about this?.....                                 | 79        |
| What we found.....                                                                | 80        |
| Staff numbers .....                                                               | 80        |
| Time taken to reach an outcome.....                                               | 81        |
| Moderator languages .....                                                         | 82        |
| Key insights and safety gaps.....                                                 | 83        |
| <b>Focus: Industry collaboration and information sharing .....</b>                | <b>84</b> |
| Why were service providers asked about this?.....                                 | 84        |
| What we found.....                                                                | 86        |
| Key insights and safety gaps.....                                                 | 87        |
| <b>After the report period .....</b>                                              | <b>88</b> |
| <b>Conclusion .....</b>                                                           | <b>89</b> |

# Executive summary

On 1 April 2026, eSafety gave non-periodic reporting notices to three online game service providers, and one gaming-adjacent service provider under section 56(2) of the *Online Safety Act 2021* (Cth) (the **Act**) (the **Notices**).

The Notices asked the service providers to explain how they were meeting the Australian Government's Basic Online Safety Expectations (the **Expectations**) during the report period from 1 October 2025 and 31 March 2026. They asked for detailed information about the systems, processes, technologies and resources each service provider used to help protect children from unlawful and harmful material and activity on its service.

The Notices were given to:

- Roblox Corporation – about Roblox
- Epic Games – about Fortnite
- Mojang AB – about Minecraft
- Valve Corporation – about Steam and Steam Chat.

Online game services present unique challenges due to the combination of features such as real-time player chat, anonymous play, immersive environments, and easy interaction across gaming communities. In light of ongoing concerns about serious harms such as child sexual exploitation and abuse (including grooming), sexual extortion, and radicalisation in gaming environments, transparency reporting notices were issued to the providers of Roblox, Minecraft, Fortnite and Steam. These required them to explain how they were identifying, preventing and responding to child sexual exploitation and abuse material and activity, to violence material and activity (including pro-terror and violent extremism material), to cyberbullying and to online hate for the reporting period.

The decision to send notices to these service providers did not, by itself, indicate eSafety's level of concern about their compliance with the Basic Online Safety Expectations compared to other services. Other service providers that are accessible in Australia may have been more, or less, compliant with the Expectations than those named in this report.

## Significant variation in protection for users

This report shows significant variation in the steps taken by service providers to protect Australian users under 18 during the report period. While the findings revealed some positive practices, they also revealed disparities and shortcomings which needed to be addressed to ensure Australian children can play safely online. Overall, there was no common baseline in terms of the safety protections in place between the service providers.

# Key findings

The following points outline some of the key findings from the Notices, including steps taken by the service providers to detect and address harms to users.

**The service providers took different approaches to preventing high-risk games from being accessible through their services. Left unchecked or unmoderated, there is a risk that these games could contain unlawful material or serve as high-risk environments for child sexual exploitation and abuse or activity which promotes, incites or instructs in matters of crime or violence.**

- Epic Games undertook a comprehensive review of every game ('Island') before it became accessible on Fortnite, whereas Roblox employed several tools to detect harmful, user-created 'Experiences' after they were published.
- In Minecraft, players self-declared as 18 or older could access and play in Minecraft 'worlds' (servers) which were hosted by third-parties and were not proactively moderated by Mojang Studios. Mojang Studios relied on proactive scanning and user reports to detect third-party servers which were high risk for violence material and activity and for child sexual exploitation and abuse.
- While Valve did have review processes in place to detect games containing unlawful material, including material related to CSEA or pro-terror material, prior to publication, it relied solely on user reports to identify published games which contained unlawful material.

**While most of the service providers employed tools to proactively detect unlawful and harmful material and activity within the services, the quality and coverage of these tools varied across each service. eSafety expects the online games industry to raise the bar to identify and combat serious harms such as child sexual exploitation and abuse, sexual extortion, and material and activity which promotes, incites, or instructs in matters of crime or violence.**

- Unlike the other service providers, Valve did not use any hash-matching tools on Steam and Steam Chat. Hash-matching tools allow service providers to identify and remove material previously confirmed to be child sexual exploitation and abuse material or violence material on a service without this material having to be repeatedly re-analysed and re-viewed across services and jurisdictions.
- Unlike the other service providers, Valve did not use tools to proactively detect harms in text or voice chat on Steam Chat, including Steam Community group chats, outside of a URL detection tool that 'mainly blocked scams'. This left all chats on Steam Chat

with limited protections from child sexual exploitation and abuse material and activity, and from violence material and activity.

- Roblox and Valve relied solely on internal data sources to train the language analysis models used on Roblox and Steam. Roblox did not frequently update several of its tools against this data and Valve did not update this data regularly after initial training. Mojang Studios used external keyword lists related to violence material and activity to train the language analysis models it used on Minecraft but did not use any external keyword lists related to child sexual exploitation and abuse. By contrast, on Fortnite, Epic Games used both internal data and synthetically generated data which used ‘seed’ keywords and phrases from external keyword lists for its language analysis tools monitoring player chats.
- Roblox’s ‘Image Auto-mod’ tool, which it used to detect violence material and child sexual exploitation and abuse material on all images on its service, had a recall rate of only 51%. Low recall can indicate under-enforcement by a tool, which can be a safety risk.

**There was significant variation between the service providers in their participation in cross-industry information-sharing initiatives, including well-established initiatives to counter terrorism and child sexual exploitation and abuse. Online games and gaming-adjacent services are deeply interconnected, and moving between them is an ordinary part of the user experience. Information-sharing initiatives enable providers to identify new risks that arise from this and mitigate them more quickly than if they were acting alone.**

- For Minecraft, Mojang Studios participated in a number of information-sharing initiatives with industry groups, including the Thriving in Games Group, Lantern and The Global Internet Forum to Counter Terrorism (GIFCT).
- Roblox shared signals with Lantern and worked directly with Discord to share information on trends in child safety and violent extremism.
- For Fortnite, while Epic Games received URLs from the Internet Watch Foundation and Tech Against Terrorism, it did not participate in the ‘Lantern’ initiative, and did not share signals about breaches related to child sexual exploitation and abuse with other service providers through similar initiatives.
- For Steam and Steam Chat, Valve did not participate in any information sharing initiatives.

**Some of the service providers used fewer indicators than others to detect players who had been banned or suspended for offences related to child sexual exploitation and abuse or to violence material or activity. The fewer indicators used, the more likely it is that perpetrators can return and continue to carry out harms such as child sexual exploitation and abuse, sexual extortion, and activity which promotes, incites or instructs in matters of crime or violence.**

- Where Mojang Studios, Roblox and Epic Games used multiple indicators to detect users who had re-registered (or attempted to) with new details after being banned or suspended for child sexual exploitation and abuse or violence material and activity related offences, Valve used minimal (a small number of) indicators for detecting such users.

**The level of age assurance was highly variable across the different services. If services do not know the age of their users, they cannot apply appropriate age-based protections that help guard against harms such as child sexual exploitation and abuse, sexual extortion, and activity which promotes, incites, or instructs in matters of crime or violence.**

- Despite having a range of protections in place for child accounts, Epic Games and Mojang Studios relied on users to self-declare their age at sign up to play Fortnite and Minecraft. Players who self-declared as 18 or older were able to access some high-risk features without additional age assurance checks, including:
  - chats with players self-declared as 18 or older which would not be subject to Epic Games' default suite of safety tools for private text chats that include children
  - access to chat with players of all ages (other than child accounts) which had communications features off by default on Minecraft
  - access to unmoderated Minecraft servers hosted by third-parties, which can be higher risk for unlawful and harmful activity.
- During the report period, Valve relied on users to self-declare their age at sign up for both Steam and Steam Chat and did not collect any age-related information at all.<sup>1</sup>
- From 3 December 2025, Roblox started using more robust age assurance measures, including limiting access to certain high-risk features on the service to Australian users identified as 18 years and older (irrespective of the self-declared age of the

---

<sup>1</sup> The Age-Restricted Material Codes create new obligations which primarily focus on preventing children's access or exposure to age-inappropriate content, including high-impact violence, across the 8 sections of the online industry. Under the App Distribution Services Online Safety Code (Class 1C and Class 2 Material), companies including Valve are required to ensure that from 9 September 2026, games on their stores that are classified R18+ are not able to be purchased or downloaded by Australian children. As of 9 September 2026, Valve started to require users in Australia to provide credit card verification to access R18+ games.

user). These measures followed [eSafety's engagement with Roblox](#) on its obligations under Australia's [Online Safety Codes and Standards](#).

**Some of the service providers took longer than others to respond to user and automated reports of specific harms. This can leave users exposed to serious harms for longer when platforms are aware that users are at risk.**

- On average, Mojang Studios had longer median times to respond to user reports of unlawful and harmful activity on Minecraft compared to other providers.
- Across all the service providers, Epic Games took the longest time (18 hours and 53 minutes) after receiving automated reports of unconfirmed child sexual exploitation and abuse material and activity other than grooming and sexual extortion to assess whether these reports were violative.

**The service providers varied in their approach to resourcing trust and safety teams. Having teams with sufficient capacity and expertise to address a range of online harms – and a working knowledge of various languages and cultural specificities – is essential to maintain a safe environment for users.**

- Epic Games had the highest ratio of trust and safety staff to end-users, and reported having a specialist investigative team of five full-time investigators and three contracted investigators globally, who addressed all reports of high-harm risk, including sexual extortion and grooming.
- Mojang Studios had a limited number of human moderators, even taking into account its reliance on Xbox and Microsoft trust and safety teams to address issues. Mojang Studios' moderators also did not operate in any language commonly spoken in Australia apart from English.
- Unlike the other service providers, Valve had no dedicated trust and safety staff other than content moderators and engineers.

## Furthering transparency

Where the responses to Notices have identified any potential failure to implement the Expectations, eSafety will engage further with those service providers to understand their plans to address these safety shortcomings and any obstacles to compliance. The information received and published in this report represents a significant step towards greater transparency and understanding of what service providers are (and are not) doing to protect Australians online.

# Definitions

These definitions explain the harms and other key terms we use in this report. Some terms have specific legal or technical meanings. We include plain language explanations to help readers understand the report, but they do not replace the formal definitions we use in the transparency notices.

## Harms

**Child sexual exploitation and abuse (CSEA)** is material or activity which includes both ‘child sexual exploitation’ (a broad category that encompasses material and activity that sexualises and is exploitative to the child, but that does not necessarily show the child's sexual abuse) and ‘child sexual abuse’ (which involves a sexual assault against a child, a narrower category that can be considered a sub-set of child sexual exploitation). CSEA material is a subset of class 1 material, as defined under section 106 of the Online Safety Act 2021 (the Act) by reference to the National Classification Scheme, and includes images, videos and texts. CSEA activity refers to other illegal activity, including grooming and sexual extortion of children. See also the definitions in Australia’s Unlawful Material Standards for [Designated Internet Services regarding class 1A and class 1B material](#). **Note:** Australia’s definition of CSEA under both criminal and classification laws includes AI-generated material.

- **Known CSEA material:** Material that has been previously confirmed to contain content depicting child sexual exploitation or abuse, and which has been hashed and stored in a hash database.
- **New CSEA material:** Material that has not been previously confirmed to contain content depicting child sexual exploitation or abuse, and which has not previously been hashed and stored in a hash database. Also known as ‘first generation CSEA material’.

**Cyberbullying:** Section 6 of the Act defines cyberbullying material as material provided on an online service which an ordinary reasonable person would conclude is likely to have the intended effect of seriously threatening, intimidating, harassing or humiliating a particular Australian child.

**Grooming:** For the purposes of this report, ‘grooming’ refers to the process of building an emotional connection with a child in order to sexually abuse or exploit them. This increasingly involves children being tricked or coerced into sexual activity in video calls or into sending sexual images through online services. Grooming for the purposes of sexual exploitation is a type of CSEA activity.

**Online hate:** Material which falls under each service provider's terms relating to hate. These were identified in consultation with service providers:

- Roblox Corporation: Policies relating to Roblox's Community Standards on 'Discrimination, Slurs and Hate Speech'.<sup>2</sup>
- Epic Games: Fortnite policies relating to the Provider's Community Rules on 'Intolerance and Discrimination',<sup>3</sup> Content Guidelines on 'Discriminatory Content'<sup>4</sup> and Fortnite Developer Rules on 'Intolerance and Discrimination'.<sup>5</sup>
- Mojang AB: Policies relating to either Minecraft's EULA provisions on hate or extreme bias, read together with the 'Keep your content clean' section of Xbox's Community Standards and/or s3(a)(vii) of the Microsoft Services Agreement.<sup>6</sup>
- Valve Corporation: Policies relating to:
  - rules and guidelines for 'Steam discussions, reviews and user generated content', specifically the portions of the guidelines relating to disrespectful content including discrimination, encouraging or facilitating real-world violence, directing abuse or insults at others or using content to harass, abuse or otherwise disparage others, or uploading inappropriate content
  - portions of Steam's Online Code of Conduct which relate to engagement in unlawful activity, uploading or posting illegal or inappropriate content, violating others' personal rights, or harassing other users or Steam personnel.<sup>7</sup>

**Sexual extortion:** A crime involving online blackmail, where a victim is tricked into sending an intimate image or video of themselves to someone who then threatens to share the image or video unless demands are met, usually for payment. Sexual extortion targeting children is a type of CSEA activity.

**Violence material and activity:** Material or activity which promotes, incites or instructs in matters of crime or violence. Violence material is material which would be, or would likely be, defined as class 2 material for the purposes of section 107 of the Act, or class 1 material for the purposes of section 106 of the Act. Violence material includes pro-terror material, crime and violence material, and extreme crime and violence material as defined by

---

<sup>2</sup> Roblox (2026), [Roblox Community Standards](#), Roblox website. Accessed 25 February 2026.

<sup>3</sup> Epic Games (2026), [Epic Games Community Guidelines](#), Epic Games website. Accessed 30 January 2026; Epic Games (2026), [Epic Games Community Guidelines](#), Epic Games website. Accessed 30 January 2026.

<sup>4</sup> Epic Games (2024), [Epic Games – Content Guidelines](#), Epic Games website. Accessed 30 January 2026.

<sup>5</sup> Epic Games (2026), [Fortnite Developer Rules](#), Epic Games website. Accessed 2 February 2026.

<sup>6</sup> Mojang Studios (2020), [Minecraft End\(er\) User Licence Agreement](#), Minecraft website. Accessed 24 February 2026; Xbox (2024), [Xbox Community Standards](#), Xbox website. Accessed 2 March 2026; Microsoft (2025), [Microsoft Services Agreement](#), Microsoft website. Accessed 3 March 2026.

<sup>7</sup> Valve Corporation (n.d.), [Rules and guidelines for Steam discussions, reviews and user generated content](#), Steam website. Accessed 2 February 2026; Valve Corporation (2025), [Steam Online Conduct](#), Steam website. Accessed 26 March 2026.

Australia's [Unlawful Material Standards for Relevant Electronic Services](#). For the purposes of this notice, violence activity may include attempts to migrate children to an external platform for the purpose of sharing violence material (for example, sharing links to a specific forum or community associated with violence material or activity).

## Other key terms

**Age assurance:** An umbrella term which includes:

- age verification (determining a person's age to a high level of accuracy, such as through the use of identity documents)
- age estimation (providing an approximate age through analysis of biological or behavioural human features that vary with age, for example through the use of facial or voice analysis)
- age inference (using information other than a date of birth which indirectly implies a certain age or age range).

**AI (artificial intelligence):** Refers to an engineered system that generates predictive outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives or parameters without explicit programming. AI systems are designed to operate with varying levels of automation.<sup>8</sup>

**Child:** A person under 18 years of age.

**Classifier:** A classifier is an algorithm that automatically orders or categorises data into one or more of a set of 'classes.' They are rules that map input data into predefined categories.

**Gaming-adjacent service:** A broad term covering services used to facilitate, or used in conjunction with, the playing of games – for example, services that people use to buy games, chat with other players, join forums or share game-related content. For the purposes of this report, Steam and Steam Chat are classified as gaming-adjacent services.

**Global Internet Forum for Countering Terrorism (GIFCT):** A forum that brings together the technology industry, government, civil society organisations and academia to foster collaboration and information-sharing to counter terrorist and violent extremist activity online.

---

<sup>8</sup> Australian Government (n.d.), [Strategy glossary](#), Data and Digital Government Strategy website. Accessed 6 July 2026.

**Hash-matching tool:** Digital technology that is used to create a hash (digital signature) of an image or video which is then compared against hashes of other images or videos to find copies.

**Internet Watch Foundation (IWF):** IWF is an international not-for-profit organisation working with the global technology sector to disrupt, prevent and remove images and videos showing the sexual abuse and exploitation of children from digital spaces.

**Language analysis tool:** Technology that assigns a probability that text or conversations involve certain behaviour, such as grooming a child for sexual abuse. This may involve AI (artificial intelligence).

**Lantern:** A cross-platform information sharing program led by the Tech Coalition, a community of technology companies working together to protect children online.

**Material:** Defined in section 5 of the Act to mean content in the form of text, data, speech, music or other sounds, visual images (moving or otherwise), or any other form or any combination of forms of content.

**Mod:** A modified version of a game, often created by a third-party.

**Monthly active users (MAUs):** Under section 20(5) of the Online Safety (Basic Online Safety Expectations) Determination 2022, service providers are expected, at eSafety's request, to provide a report on the number of 'active end-users of the service in Australia (disaggregated into active end-users who are children and those who are adult end-users) during a specified period'. The Determination does not define 'monthly active end-users'. For the purposes of this report, service providers reported on 'monthly active users' but their methodologies for calculating those figures varied.

**National Center for Missing & Exploited Children (NCMEC):** A U.S.-based nonprofit organisation that serves as a national resource and clearinghouse on issues relating to missing and exploited children. NCMEC operates the CyberTipline and works with technology companies and other partners to support the detection, disruption and prevention of online child sexual exploitation and abuse worldwide. NCMEC also makes several hash lists available to participating technology companies, including CSAM, Exploitative, Generative AI and Take It Down hash lists.

**Notice:** For the purposes of this report, a Notice is a non-periodic notice given to a service provider under section 56(2) of the Act on 1 April 2026.

**Online game:** A video game played online. Video games (also known as computer games) are interactive games that can be played on a variety of devices, including PCs (desktop

computers or laptops), consoles (such as Xbox and Playstation), mobile devices, or virtual reality (VR) devices.

**Online game service:** A service that facilitates the playing of online games.

**Off-platforming:** For the purposes of this report, this refers to the process of migrating a child to another online service, which is usually encrypted or private. Off-platforming can occur for a variety of purposes, including radicalisation or grooming for the purposes of sexual exploitation or abuse.<sup>9</sup>

**Recidivism:** In the context of this report, banned or suspended users re-registering to a service with new details to continue perpetrating online abuse. Multiple fake or imposter accounts may be created by a human user or by an automated bot programmed by a human.

**Report period:** Service providers were asked to provide information for the period between 1 October 2025 and 31 March 2026. Information provided reflects this period, unless stated otherwise.

**User-generated content:** A broad term that can include the creation of usernames, third-party modifications, custom objects (for example, skins), playthrough videos, comments and other content. While user-generated content can also refer to games created within an online game service (for example, 'Experiences' within Roblox, 'Islands' within Fortnite), eSafety has addressed games created within a service separately from other user-generated content for the purposes of this report.

---

<sup>9</sup> The Bracket Foundation (2022), [\*Gaming and the Metaverse: The Alarming Rise of Online Sexual Exploitation and Abuse of Children Within the New Digital Frontier\*](#), United Nations Interregional Crime and Justice Research Institute (UNICRI) website. Accessed 19 September 2026; NYU Stern Center for Business and Human Rights (2023), [\*Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter Them\*](#), NYU website. Accessed 12 February 2026.

# Why eSafety is looking at online games

## About the Basic Online Safety Expectations

The Australian Government developed the Basic Online Safety Expectations (known as ‘the Expectations’ or ‘BOSE’) to help make sure online services take reasonable steps to keep Australians safe. This means they should be protected from specific types of unlawful and harmful material and activity.

BOSE transparency notices require online services to explain how they are complying with these expectations.

Information about the Expectations, including the responses to all transparency notices and the regulatory guidance which sets out the material and activity that is considered ‘unlawful and harmful’, can be found on the [eSafety website](#).

Online games (video games played online) are a major part of childhood in Australia. eSafety’s 2024 research found that 89% of children aged 8 to 17 had played online games in the past year.<sup>10</sup> The Interactive Games & Entertainment Association’s ‘Australia Plays 2025’ report found that 82% of 1- to 17-year-olds played video games, and that most Australian households had two or more gaming devices.<sup>11</sup> eSafety’s ‘Early days, early insights’ report on the social media age restrictions found a statistically significant increase in children playing video games online every day, three months after the commencement of Australia’s social media age restrictions.<sup>12</sup>

Children have a human right to recreation and play.<sup>13</sup> Online games are an important source of play, creativity and connection for many children. They can help young players relax, socialise, learn and develop skills that support their wellbeing and digital literacy. Children should be able to enjoy these benefits without being exposed to harms such as exploitation, violence or abuse.<sup>14</sup>

---

<sup>10</sup> eSafety Commissioner (2024), [Leveling up to stay safe: young people’s experiences navigating the joys and risks of online gaming](#), Australian Government, accessed 12 February 2026.

<sup>11</sup> eSafety Commissioner (2024), [The risks and benefits of online gaming for children and young people](#), Australian Government; Interactive Games & Entertainment Association (2025), [Australia Plays 2025](#), IGEA website.

<sup>12</sup> eSafety Commissioner (2026), [Early days early insights – Social Media Minimum Age restrictions Report](#), Australian Government, accessed 12 February 2026.

<sup>13</sup> United Nations Convention on the Rights of the Child (1989), [Article 31](#), OHCHR website.

<sup>14</sup> UN Committee on the Rights of the Child (2021), [General comment No. 25 \(2021\) on children’s rights in relation to the digital environment](#) (Doc. CRC/C/GC/25), OHCHR website.

Services that have the sole or primary purpose of enabling end-users to play online games with other end-users are excluded from the Australian Government's age restrictions for under-16s that apply to social media platforms. However, the Expectations require service providers to take reasonable steps to address the safety risks that can arise in online environments, including while gaming. There are also enforceable obligations on online games and gaming-adjacent services in the Act's industry codes and standards in relation to unlawful material such as child sexual exploitation material and pro-terror material, as well as age-restricted material such as online pornography and self-harm related material.

In August 2026, eSafety [announced](#) it had accepted a court-enforceable undertaking from Roblox in relation to its compliance with the relevant industry standard.<sup>15</sup> The undertaking commits Roblox to taking additional measures to ensure:

- adults cannot contact children without parental consent
- children's accounts are private by default
- user reporting options are accessible and easy to use.

## Transparency

This transparency report examines what four major service providers told eSafety about the steps they are taking to protect children on their online game services and gaming-adjacent services.

Transparency is not just about disclosure – it is about driving real safety improvements. In publishing this report, eSafety expects industry members to learn from each other and make concerted efforts to uplift their safety, taking a best practice approach to meeting the Expectations.

## How features in online games shape online harm

Online games are not just games on a screen. Many combine open worlds, real-time player chat, anonymous play, immersive environments, virtual currencies, user-generated content, and easy interaction with wider gaming communities. These features can increase risks, especially for children. They can also influence how harms occur, spread on a platform and move between online environments. This means safety measures need to account for the design of the service, not only the content or conduct that occurs on it.

---

<sup>15</sup> eSafety Commissioner (2026), [eSafety secures court-enforceable undertaking from Roblox to do more to protect Australian kids from grooming and sexual exploitation](#), eSafety website. Accessed 20 August 2026.

The risks are compounded by a service's popularity with children. Analysis of reports made to eSafety indicates that people who want to exploit or abuse children often intentionally gravitate towards platforms that attract children.

## Potential harms

### Child sexual exploitation and abuse (CSEA), including grooming and extortion

- Perpetrators of CSEA may take advantage of a game's communication features to form relationships with children with the intention of sexually exploiting or abusing them.<sup>16</sup>
- Research has found perpetrators often initiate contact on games before luring a child to a gaming-adjacent service with private communication options,<sup>17</sup> such as Discord. This is known as 'off-platforming'.

### Cyberbullying

- Cyberbullying is not inherent to games, but anonymous play, group interaction and normalised bad behaviour can make it more common and harder to interrupt.<sup>18</sup> Some forms of bullying are specific to online games, such as 'griefing', where players irritate or harass other players by deliberately misusing the game in unintended ways.<sup>19</sup>
- eSafety's 2024 'Leveling up to stay safe' report found that 25% of children aged 8 to 17 had experienced griefing or trolling, while almost a third (32%) had experienced other bullying-type behaviours.<sup>20</sup>

### Online hate

- Online hate is when a person is targeted because of race, ethnicity, disability, religion, caste, sexual orientation, sex, gender identity, disease, immigration status, asylum seeker or refugee status, or age.<sup>21</sup>

---

<sup>16</sup> UNICEF (2023), [Child Sexual Exploitation in Online Gaming: Risks and Realities](#), UNICEF website.

<sup>17</sup> The Bracket Foundation (2022), [Gaming and the Metaverse: The Alarming Rise of Online Sexual Exploitation and Abuse of Children Within the New Digital Frontier](#), United Nations Interregional Crime and Justice Research Institute (UNICRI) website. Accessed 19 September 2026

<sup>18</sup> Yunhao, H., Sophie, E., Elizabeth M., C., & Bianca, K. (2025), 'Behind the screen: How gaming habits, motives, and content shape cyberbullying behaviour'. *Computers in Human Behavior Reports*, 18, 100675. doi: <https://doi.org/10.1016/j.chbr.2025.100675>; Kowert, R. (2020), 'Dark Participation in Games'. *Frontiers in Psychology*, 11. doi: <https://doi.org/10.3389/fpsyg.2020.598947>.

<sup>19</sup> Kowert, R. (2020), 'Dark Participation in Games'. *Frontiers in Psychology*, 11. doi: <https://doi.org/10.3389/fpsyg.2020.598947>.

<sup>20</sup> eSafety Commissioner (2024), [Leveling up to stay safe: young people's experiences navigating the joys and risks of online gaming](#), Australian Government.

<sup>21</sup> eSafety Commissioner (2026), [Basic Online Safety Expectations regulatory guidance](#) (p.7), eSafety website.

- Online hate may be targeted at particular players or at a group of people more generally. The ‘Levelling up to stay safe’ report found that 1 in 5 teenagers (aged 13 to 17) had seen or heard other players share or use online hate.<sup>22</sup>

## Violence material and activity

- There is evidence online games have in some instances been used to promote, incite or instruct players in crime and violence. This has included violent extremists exploiting gaming spaces to spread ideas, recruit users or build communities.<sup>23</sup>
- While online hate does not always lead to violence or radicalisation, extremist groups often spread online hate as part of their rhetoric, including threatening violence against a particular group. A 2024 study of Australia and six other countries found that 38% of respondents saw endorsements of violence against a particular group on online game services and gaming-adjacent services.<sup>24</sup> The report also noted that communities associated with gaming ‘can serve as sites of misogynistic and gendered harassment of women and members of the LGBTQI+ community – the normalisation of which can lower resilience to radicalisation and have offline consequences’.
- Similar to off-platforming following sexual grooming, children are known to be moved from games that promote extremist ideas to more private spaces that allow indoctrination and mobilisation.<sup>25</sup>

## Sadistic online exploitation

- Sadistic online exploitation, sometimes known as ‘nihilistic violent extremism’ is characterised by the deliberate use of the internet to harm, manipulate or exert control over others for personal pleasure, gratification, a sense of power, ideological beliefs or to corrupt victims and inflict ongoing harm.<sup>26</sup>
- Perpetrators have been known to groom their victims, as well as coercing them through sexual extortion and other kinds of online blackmail. Certain groups

<sup>22</sup> eSafety Commissioner (2024), [Leveling up to stay safe: young people's experiences navigating the joys and risks of online gaming](#), eSafety website.

<sup>23</sup> NYU Stern Center for Business and Human Rights (2023), [Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter Them](#), NYU website. Accessed 12 February 2026; UNICEF (2025), [Protecting Children in Online Gaming: Mitigating Risks from Organized Violence](#), UNICEF website; Radicalisation Awareness Network (2021), [Digital Grooming Tactics on Video Gaming & Video Gaming Adjacent Platforms: Threats and Opportunities](#), EU Migration and Home Affairs website.

<sup>24</sup> White, J., Wallner, C., Lamphere-Englund, G., Frankie, L., Kowert, R., Schlegel, L., Kingdon, A., Phelan, A., Newhouse, A., Saiz Erasquin, G., & Regeni, P. (2024), *Radicalisation through Gaming: The Role of Gendered Social Identity* (p.22), RUSI website

<sup>25</sup> NYU Stern Center for Business and Human Rights (2023), [Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter Them](#), NYU website, Accessed 12 February 2026.

<sup>26</sup> Australian Centre to Counter Child Exploitation (2025), [Sadistic Online Exploitation](#), ACCCE website.

associated with sadistic online exploitation have been known to use Neo-Nazi, white supremacist, or other extremist language.<sup>27</sup>

## How gaming features can contribute

Gaming features can contribute to risks and harms both separately and in conjunction with each other. This list is not exhaustive.

### Open worlds

Without adequate privacy and safety settings being implemented, a child can play with people of any age, skill level and behaviour standards, including adults (and other children) who may harm them deliberately or unintentionally.

### Player chat

Without adequate privacy and safety settings being implemented, other players can directly communicate with a child without the knowledge of their parents and carers.

### Anonymous play

Harms can be amplified when players are operating under real or perceived anonymity. This tends to make bad behaviour more common and harder to interrupt because the players are difficult to identify, hold to account, remove and/or prevent from returning to the environment.<sup>28</sup>

### Immersive environments

Games with immersive, metaverse-style elements can heighten risks. Through avatars and shared worlds, experiences can feel more intense, immediate, social and difficult to separate from play.<sup>29</sup> Harms in immersive environments can also be harder to detect.

### Virtual currencies

Online game services sometimes offer players virtual currencies to buy in-game items and other game-related content. Perpetrators of CSEA have been known to offer children

---

<sup>27</sup> Resolver (2026), [Weaponized loneliness: Understanding a hybrid threat to young people globally](#), Resolver website.

<sup>28</sup> Yunhao, H., Sophie, E., Elizabeth M., C., & Bianca, K. (2025), 'Behind the screen: How gaming habits, motives, and content shape cyberbullying behaviour'. *Computers in Human Behavior Reports*, 18, 100675. doi: <https://doi.org/10.1016/j.chbr.2025.100675>; Kowert, R. (2020), 'Dark Participation in Games'. *Frontiers in Psychology*, 11. doi: <https://doi.org/10.3389/fpsyg.2020.598947>.

<sup>29</sup> The Bracket Foundation (2022), [Gaming and the Metaverse: The Alarming Rise of Online Sexual Exploitation and Abuse of Children Within the New Digital Frontier](#), United Nations Interregional Crime and Justice Research Institute (UNICRI) website. Accessed 19 September 2026; eSafety Commissioner (2025), [Immersive technologies – position statement](#), eSafety website.

rewards such as in-game currencies or other virtual items in exchange for sexual interactions (sometimes known as ‘commodified sexual interaction’).<sup>30</sup>

A 2025 study by Thorn on people who had experienced commodified sexual interactions as a minor found that 1 in 11 had received some form of gaming currency as part of the exchange.<sup>31</sup>

## User-generated content

Without adequate detection and moderation, high-risk user-generated environments can thrive, such as ‘public bathroom’ games where players simulate sexual activity with children, or games which recreate mass shootings or incorporate terrorist propaganda.<sup>32</sup>

## Easy interaction with wider gaming communities

The layered ecosystem encourages movement between online games and related services. This means it seems natural for interaction between players to occur or continue off the main platform of a game, so ‘off-platforming’ by someone with ill intent is relatively easy. For example, an abuser may first contact a child in a game, before moving them to a private channel such as Discord.<sup>33</sup> Reports to eSafety show off-platforming can also occur as part of radicalisation. To obtain information about off-platforming for this report eSafety asked service providers about ‘attempts to migrate children to another service’.

---

<sup>30</sup> The Bracket Foundation (2022), [Gaming and the Metaverse: The Alarming Rise of Online Sexual Exploitation and Abuse of Children Within the New Digital Frontier](#), United Nations Interregional Crime and Justice Research Institute (UNICRI) website. Accessed 19 September 2026.

<sup>31</sup> Thorn (2025), [Commodified Sexual Interactions Involving Minors: New data on evolving dynamics in technology-facilitated child sexual exploitation](#) (p.23), Thorn website.

<sup>32</sup> Martin, S. (2025). [My chilling week on Roblox: sexually assaulted and shat on as a child avatar roaming the online world](#), The Guardian website, accessed 12 February 2026; Revealing Reality (2025). [A digital playground: The real guide to Roblox](#), Revealing Reality website, accessed 12 February 2026; Revealing Reality (2025), [Four months on: is Roblox any safer?](#), Revealing Reality website, accessed 12 February 2026; ADL Center on Extremism (2021). [The Dark Side of Roblox: ‘Active Shooter Studios’ Create Maps Based on Real-Life Mass Shootings](#), ADL website, accessed 16 February 2026; Hancorn, G. (2025). [Mosque shootings and far-right skins: Teens playing Roblox exposed to extremist content](#), accessed 16 February 2026; Global Network on Extremism and Technology (2025), [Playing with Hate: How Far-Right Extremists Use Minecraft to Gamify Radicalisation](#), accessed 12 February 2026; Anti-Defamation League (2022). [Breaking the Building Blocks of Hate: A Case Study of Minecraft Servers](#), ADL website; [Global Project Against Hate and Extremism](#) (2024) [Fortnite’s Political Content Encourages Violence and Allows Players to Stage a Capitol Insurrection](#), 3 May 2024, accessed 12 February 2026; [Global Project Against Hate and Extremism](#) (2024), accessed 12 February 2026, [Despite Previous Reporting, Fortnite Continues to Platform Violent Political and Antisemitic Content](#).

<sup>33</sup> Suojellaan Lapsia (2024), [Tech Platforms Used by Online Child Sexual Abuse Offenders](#) (p.16-17), Suojellaan Lapsia website; NYU Stern Center for Business and Human Rights (2023), [Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter Them](#), NYU website. For information on what measures Discord is taking to prevent CSEA material and activity on its service, please see eSafety’s periodic reporting series on child sexual exploitation and abuse, grooming and sexual extortion.

## Examples of how harms may play out

These examples are created from an amalgamation of real experiences. They are not about specific incidents on the services that were given Notices but serve to illustrate how features on games and gaming-adjacent services may add to risks and harms.

**Liam** is contacted on a gaming platform by an unknown stranger who he believes to be his age. The stranger asks for Liam's username on a particular social media platform, bypassing the game's detection efforts by using symbols and numbers instead of the name of the social media platform. After Liam and the stranger connect on social media, the stranger offers Liam virtual game items in exchange for a nude photo of himself.

While searching for communities associated with her favourite games, **Emily** comes across a community on a gaming-adjacent service which organises and advocates for harassment and in-game 'raids' of Muslim and LGBTIQ+ players. This community also provides a link to a website with white supremacist content.

**Olivia** is sexually groomed by an adult user on a social media platform. As part of the grooming process, Olivia's abuser provides a link to a metaverse-style game. Once in that game, the abuser pressures her to simulate sexual acts with him on the platform.

After becoming radicalised through prolonged contact with extremist communities online, **Matteo** uses a metaverse gaming platform to recreate the Christchurch terrorist attack, complete with a 'skin' which allows his in-game avatar to look like the Christchurch shooter.

# What we examined

## Who received the Notices?

eSafety gave notices to three providers of online game services and one gaming-adjacent service:

- Roblox Corporation – about Roblox
- Epic Games – about Fortnite
- Mojang AB – about Minecraft
- Valve Corporation – about Steam and Steam Chat.

The following information specifies service features, access and interaction that were current during the report period.

## Online game services

### Roblox

[The eSafety Guide profile](#)

Roblox is a gaming platform where users can design their own games, known as ‘Experiences’, and play games created by other users. Experiences range from competitive games where players compete in rounds of gameplay, such as ‘Dress to Impress’ and ‘Steal a Brainrot’, to free-form ‘hangout’ games where players can roleplay or socialise, such as ‘Brookhaven RP’. In many Experiences, players use their own custom Roblox avatar. Roblox also has a virtual currency, Robux, which players can use to buy items such as custom clothing for avatars and items for particular Experiences. Globally, users reportedly spend more time on Roblox than on Steam, PlayStation and Fortnite combined.<sup>34</sup> Roblox also has a ‘Roblox Communities’ feature, for groups of Roblox developers (or ‘Creators’) and users who share similar interests.

**Accessible on:** PC (desktop), mobile, gaming consoles such as Xbox and PlayStation, and virtual reality (VR) devices.

---

<sup>34</sup> Carpenter, L. (2026), [Players spent roughly 10,250,000,000 monthly hours in Roblox in 2025, analyst says—more than Steam, PlayStation, and Fortnite combined](#), PC Gamer website.

**User interaction and common terms:** Depending on their age and whether they have completed an age assurance process, players can interact via a range of chat functions, including Experience chat (voice and text chat between all players in the same Experience), Direct Experience chat (one-to-one text chat that can only be seen between two specific players in the same Experience) and Party chat (text and voice chat between a fixed group of players who play together on Roblox). Players' avatars can also usually interact with each other within Experiences. Members of a Roblox Community can also post and interact with each other using text on Roblox Community forums.

**Note:** From 7 January 2026, Roblox required users to verify their age to access chat features. From 3 December 2025, voice chat was not available for Australian children under the age of 16.

**Notice issued to:** Roblox Corporation

## Fortnite

### [The eSafety Guide profile](#)

Fortnite is a gaming platform made up of multiple games, known as 'Islands'. Some Islands are made by Epic Games, the provider of Fortnite, while others are made by Fortnite users either using Epic Games assets (Fortnite Creative) or by using the in-house game development tool, Unreal Editor for Fortnite (UEFN). One of the best-known Islands is 'Fortnite: Battle Royale', where players use guns and other weapons to shoot their enemies and compete against each other to be the last player or team standing. Players use 'skins' to change the appearance of their avatar, which may look like a regular person, an animal, a fictional character or a celebrity. Players can use Fortnite's virtual currency, V-Bucks, to buy skins and items for particular Islands. Developers of Islands can also talk to players on the platform Fortnite Communities.

**Accessible on:** PC (desktop), mobile and gaming consoles such as Xbox and PlayStation.

**User interaction and common terms:** Players can interact through a range of chat functions, including Game Chat (voice or text chats between all players on the same Island), Party Chat (voice or text chats between a fixed group of players), Direct Messaging (one-to-one messages between two Fortnite friends for up to 30 days) and Groups (conversations between Fortnite friends and Fortnite friends of friends for up to 30 days). Fortnite also has forums as part of its Fortnite Communities feature, which allow for

communication between Fortnite players and developers of Islands through posts and replies in channels made by developers.

**Note:** During the report period, chat features were disabled by default for players who declared they were under 18. Text and voice chat could be activated for players classified as under the age of 16 with parental permission.

**Notice issued to:** Epic Games

## Minecraft

### [The eSafety Guide profile](#)

Minecraft is a game where players explore and build in 3D worlds made up of blocks. These worlds are hosted on servers run by the provider of Minecraft, official partners, players, or external third-parties. Players can also pay for Minecraft Realms, a Microsoft-hosted subscription service, to create and manage their own online Minecraft worlds. Players use custom avatars and can add third-party content to Minecraft, including modified versions of the game known as 'mods'. They can also buy items and game content on Minecraft Marketplace using Minecraft's virtual currency, Minecoin.

**Accessible on:** PC (desktop), mobile, gaming consoles such as Xbox and PlayStation, and virtual reality (VR) devices. The original version of Minecraft, 'Minecraft: Java Edition', is only available on PC. Users must use a Microsoft Services Account ('MSA') to access online multiplayer experiences regardless of the gaming console used.

**User interaction:** Players can interact with each other on multiplayer servers using whatever text chat functions are available on that server.

**Note:** During the report period users were required to self-declare their age to access chat features.

**Notice issued to:** Mojang AB

## Gaming-adjacent services

### Steam and Steam Chat

[The eSafety Guide profile](#)

Steam is a digital game store, library and community platform used to download, launch and play games developed by other providers as well as by its own provider, Valve. Users with a Steam account can access Steam Chat, a direct messaging service available through Steam Community.

**Accessible on:** Web browser, mobile, desktop app and Steam Deck.

**User interaction and common terms:** Steam users can participate in open online forums, known as ‘Steam Discussions’, either through the Discussions tab or by accessing a ‘Game Hub’ for a particular game. Game Hubs also allow users to share game-related content, including custom items, mods, artwork, videos and reviews. Steam also has a broadcasting function that allows users to livestream their gameplay to friends or other Steam users. Users with a Steam account can use Steam Chat to exchange text and voice messages with their Steam friends, primarily while playing online games together, or to participate in group chats linked to particular Game Hubs.

**Note:** During the report period users were required to self-declare their age to access chat features.

**Notice issued to:** Valve Corporation

## How many people are on these services?

eSafety chose these services due to their popularity among Australian children and users.

**Roblox, Fortnite and Minecraft** were selected because they were the three games played most by Australian children, according to the most recent eSafety research available at that time. eSafety’s 2024 ‘Leveling up to stay safe’ report on young people’s participation in online gaming asked children aged 8 to 17 which services they had played in the past year (even if they only played it once or twice) and found 53% of children in that age range had played Minecraft, followed by Roblox (50%) and Fortnite (46%).<sup>35</sup>

---

<sup>35</sup> eSafety Commissioner (2024), [Levelling up to stay safe](#), eSafety website.

**Steam** was selected because it is one of the world's largest online storefronts for PC games (games played on computers rather than consoles). Prior to eSafety sending notices to providers, Steam advertised itself to developers as having 132 million monthly active users globally.<sup>36</sup>

To gain greater insights on the number of child users on these services, eSafety sent written requests alongside the Notice requiring service providers to provide their Australian monthly active users (MAUs) during the report period, broken down by adults and children. Service providers are required to provide age-aggregated data on their active end-users on request from eSafety under s20 of the Online Safety (Basic Online Safety Expectations)

Determination 2022. For Roblox, eSafety asked for monthly active user report numbers in the months before and after the global rollout of its age-verified chats. These MAU numbers provided a snapshot of the total audience size and engagement level on each platform over the report period, in particular their popularity among Australian children. MAU numbers also provide important context to information about the volume of harm detected by proactive tools or user reports, and the effectiveness of service interventions.

Of the four service providers that received Notices, Roblox was by far the most popular service. In the period between October and December 2025, Roblox reported Australian MAUs of around 3.36 million and between January and March 2026, Roblox reported 3.14 million monthly active users.<sup>37</sup> Roblox's global rollout of age-verified chats occurred on 7 January 2026. Approximately 53–54% (1.7 million) of Roblox's Australian MAUs were reportedly under the age of 18. Fortnite was the next most popular, with 1.5 million Australian users on the main Fortnite service. Of these, 19% (288,526) were reportedly under 18.<sup>38</sup>

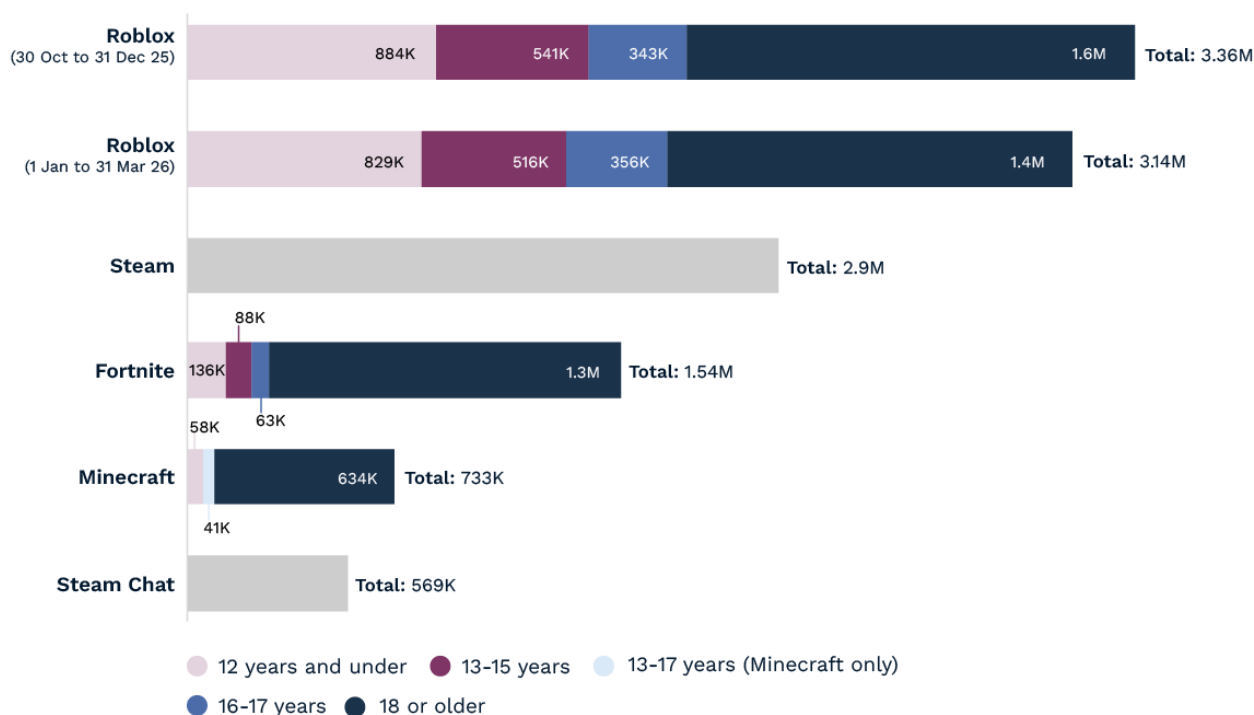
Valve Corporation reported that it had 2.09 million Australian MAUs on Steam overall but stated it did not track the age of its users, and that it could only provide numbers of users of Steam Chat for the last two weeks of the report period (even though the Expectations apply to its services).

---

<sup>36</sup> Valve Corporation. (n.g.). [Steamworks](#). Steamworks website.

<sup>37</sup> Roblox (2026), [A New Era of Safety: Facial Age Checks Now Required to Chat on Roblox](#), Roblox website.

<sup>38</sup> Epic Games reported that it relied on self-declaration of age to determine if a user is 18 or older. Self-declaration is considered a less reliable age assurance measure as it presents no barrier to a user declaring a false age; for example, see eSafety Commissioner (2023), [Roadmap for age verification and complementary measures to prevent and mitigate harms to children from online pornography – background report](#) (p.146–147), eSafety website.

**Figure 1: Australian monthly active users (MAU) by service**

**Note:** Averages have been calculated at the age segment level for some service providers so figures may not add to the totals displayed. The number of monthly active users aged 18 or older was calculated by deducting the number of monthly active users under 18 from the total monthly active users. Valve did not track the age of its users. Data on Steam Chat's MAUs is restricted to the last two-week data 'retention period' before 31 March 2026. Minecraft did not provide a further age breakdown for users between 13 to 17 years old. Fortnite provided Australian MAU numbers for Fortnite, Fortnite.com and Fortnite communities stating that each metric was calculated from different underlying data systems and could not be aggregated. Only the figures for Fortnite have been provided, as these are relevant to users of the main part of the Fortnite service.

## What questions did eSafety ask?

The Notice questions focussed on the systems, processes, technologies and resources service providers had in place to comply with the Expectations and keep children safe. We asked service providers about:

- the age assurance measures they used to reduce the risk of adults contacting children, and the availability of parental controls to further reduce this risk (Expectations 6(2), 6(2A), 11 and 12)
- the tools they used to proactively detect CSEA and violence material and activity and how they assessed the effectiveness of those tools (Expectations 6(2), 6(2A), 11 and 12), as well as how they proactively detected users with a history of harmful behaviour (Expectation 14(1A))

- the number of user reports they received about CSEA (including grooming) and sexual extortion, violence material and activity, cyberbullying and online hate on their services (Expectation 14(1A))
- the human resources they dedicated to the safety of their services (Expectations 6(2), 6(2A), 11 and 12)
- how they collaborated with other members of the online game service industry and wider gaming ecosystem to prevent harm across different platforms (Expectations 10(1) and 14(1A)).

The responses have been set out in separate focus areas later in this report.

Where service providers reported on information that addressed the same or similar issues, eSafety has compiled that information in summary tables. This enables greater transparency by setting out the actions that service providers were taking side-by-side, so they can be compared. It also provides a fuller understanding of how different service providers operate.

eSafety recognises each service provider is different – with different functionalities, architectures, business models and user bases. This means an intervention or tool may be proportionate and appropriate on one service, but not on another. Accordingly, when reviewing the tables, the nature of the service, the context in which it operates, and the risk of online harms associated with it should be taken into account.

## What information has been published, and what has been excluded?

This report is not a statement about the appropriateness of actions taken by service providers, or a conclusion regarding their compliance with the Expectations. eSafety intends this report to be a useful transparency and accountability tool, providing information about the actions that service providers were taking to keep Australians safe online at a particular point in time. Their policies and processes may have changed since, and may continue to change.

The report is a summary and does not reflect services providers' responses to eSafety's questions in their entirety.

In line with eSafety's regulatory guidance, certain information has been withheld where eSafety considered it was not appropriate to disclose, for example because it contained commercial-in-confidence information or because publication of the information would not serve the public interest.

The following points should also be noted:

- The information provided in responses to the Notices has not been verified by eSafety, although service providers are required to respond truthfully and accurately.
- Where data has been provided it covers the report period 1 October 2025 to 31 March 2026 inclusive, unless otherwise stated.
- The information summarised in this report is based on the responses eSafety received, which reflect the report period or other periods within this timeframe as specified. Service providers may have implemented changes to policies and processes since this information was provided.
- All data is global, unless otherwise stated.
- Terms used are defined in the Definitions section of this report, unless otherwise stated
- Where eSafety refers to actions or measures taken by a service provider, this refers to actions or measures taken by the provider with respect to the services covered by the Notice and should not be interpreted as applying to all products and services of a provider.

## Focus: Age assurance

Service providers are expected to take reasonable steps to make sure their users, including children, can use their services safely.<sup>39</sup> Implementing reasonable age assurance allows service providers to more effectively take steps that support this aim,<sup>40</sup> especially where safety settings are tied to a service provider's knowledge of a user's age, and/or where a service is intended for children.

Service providers were asked questions about the types of age assurance measures they used, when they assessed or reassessed a user's age, and how the effectiveness of their age assurance was evaluated.

Age assurance is an umbrella term referring to the measures that service providers can take to determine the age of the users on their services.

Providers can use a range of age assurance methods to determine whether a user is a child. These include:

- **self-declaration**, which requires users to state their age or date of birth<sup>41</sup>
- **parental vouching**, where a parent or guardian can create an account for a child by attesting to their age
- **age verification**, where a user's age is verified using genuine identity documents
- **facial age estimation**, where an image of a user's face is analysed for physical or behavioural characteristics using AI models to predict age without the need for formal identity documents
- **age inference**, where a user's likely age or age range is determined based on contextual, behavioural, transactional or environmental signals.

Service providers may adopt a 'waterfall' approach using multiple age assurance methods in combination, starting with the least onerous. Extra layers may be used to tighten access to high-risk features or to look for signals of contraventions of age rules.<sup>42</sup>

---

<sup>39</sup> Australian Government (2022), [Online Safety \(Basic Online Safety Expectations\) Determination 2022 s6\(2\), 6\(2A\), s11](#), Federal Register of Legislation website.

<sup>40</sup> Australian Government (2022), [Online Safety \(Basic Online Safety Expectations\) Determination 2022 s6\(2\), s11, s12](#), Federal Register of Legislation website.

<sup>41</sup> For the purpose of other schemes administered by eSafety requiring online services to implement appropriate age assurance measures, including the Social Media Age Restrictions and the Age-Restricted Material Codes, self-declaration alone is not considered an appropriate age assurance measure.

<sup>42</sup> Also known as successive validation; see part F of the final report of the [Age Assurance Technology Trial](#) which was commissioned by the Department of Infrastructure, Transport, Regional Development, Communications, Sport and the Arts.

Where less onerous methods of age assurance are used at sign-up, applying default settings which limit high-risk features is a best practice approach. Higher risk features should only be enabled following more rigorous age assurance measures.<sup>43</sup> Service providers should consider the proportionality of the measures that they adopt, including the impact on a user's privacy.

## Why were service providers asked about this?

Knowing whether individual players are under 18, and the number or proportion of child users on a service, allows the provider to identify and manage risks.

The service provider can use the information to:

- enforce minimum age rules
- apply default safety settings effectively and make parental controls available where relevant
- restrict access to age-inappropriate content and experiences
- limit contact between adults and children.

## Enforcing minimum age rules

Age assurance is necessary to enable service providers to enforce their own terms of service where those terms specify a minimum age. It can also be required for some services in order to comply with obligations, such as those in the Age-Restricted Material Codes.

Valve's terms of service prohibited children under 13 from using Steam services.<sup>44</sup>

Roblox did not have a minimum age for players to use the service, but parental consent and account management was required for users under the age of 13.<sup>45</sup>

Mojang did not have a minimum age on Minecraft but users under the age of 18 are required to have parental consent to create an account.<sup>46</sup>

Epic Games required Fortnite users to be 16 years or older to create an account without parental restrictions.<sup>47</sup>

---

<sup>43</sup> eSafety Commissioner (2025), [Social Media Minimum Age Regulatory Guidance](#) (p.38–39), eSafety website.

<sup>44</sup> Valve Corporation (n.d.), [Steam Subscriber Agreement](#), Steam website.

<sup>45</sup> Roblox (2026), [Roblox Terms of Use](#), Roblox website.

<sup>46</sup> Mojang (2023), [Minecraft Parents' Guide](#), Minecraft website.

<sup>47</sup> If a player indicates they are under 16 years, they are placed in a 'Cabined Account' which disables features like voice or text chat until a parent provides consent; Epic Games (n.d.), [Cabined Accounts](#), Epic Games website.

## Applying default safety settings effectively

The Expectations require that safety settings are robust and that they initially default to the most restrictive level for children's accounts.<sup>48</sup> Age assurance helps ensure these settings are applied correctly – and that parental controls are made available where appropriate.

Parental controls are software tools that allow a parent or guardian to monitor and limit what their child sees and does once they have access to an online service. Common parental controls include mechanisms to:

- block access to specific websites, apps or functions
- filter different kinds of restricted content
- limit who can communicate with a child and how
- monitor a child's use of a particular app or game, including setting time limits for use.

## Restricting access to age-inappropriate content and experiences

Under the Expectations, service providers are expected to take reasonable steps to prevent children from accessing content that is, or would likely be, classified as R18+. This can include implementing appropriate age assurance measures.<sup>49</sup> Some service providers also have mandatory obligations under the Age-Restricted Material Codes to implement 'appropriate age assurance' measures.

## Limiting contact between adults and children

All four services are popular with Australian adults as well as children. Roblox, Fortnite and Minecraft explicitly market their services to under-18s. While Valve did not market Steam or Steam Chat as a platform for use by children, its terms of service allow users who are 13 years or older and it provides parental control options for users under 18 years. The presence of children on these services, combined with features like open worlds and normalised anonymous chat, can make them attractive to adults who intend to exploit, coerce or abuse children.<sup>50</sup> Age assurance measures that distinguish children from adults can help service providers, limit contact between them.

<sup>48</sup> Australian Government (2022), [Online Safety \(Basic Online Safety Expectations\) Determination 2022 s6](#), Federal Register of Legislation website.

<sup>49</sup> Australian Government (2022), [Online Safety \(Basic Online Safety Expectations\) Determination 2022 s12](#), Federal Register of Legislation website; eSafety Commissioner (2026), [Basic Online Safety Expectations Regulatory Guidance](#) p55–57, eSafety website.

<sup>50</sup> Suojellaan Lapsia (2024), [Tech Platforms Used by Online Child Sexual Abuse Offenders](#) (p.17–18), Suojellaan Lapsia website; UNICEF (2025), [Protecting Children in Online Gaming: Mitigating Risks from Organized Violence](#) (p.12–14), UNICEF website; Resolver (2026), [Weaponized loneliness: Understanding a hybrid threat to young people globally](#) (p.33–34), Resolver website.

Game services with communications features also have mandatory obligations under the Online Safety Codes and Standards to provide tools and settings that prevent adults from using the service to contact Australian children under 16 without the consent of a parent or caregiver.

## What we found

### Age assurance tools and techniques

eSafety found disparities in the tools and techniques used by service providers to assess a user's age.

Roblox initially relied on self-declaration or parent-set age at sign-up to verify a user's age during the report period. However, from 3 December 2025, Roblox introduced age-based chat in Australia, which required users to verify their age using facial age estimation technology or ID verification in order to use chat functions. Users who did not verify their age could access Roblox Communities but could not access any form of text or voice chat. Users verified as under 18 could only chat within their own or similar age groups, with one-on-one and voice chats turned off by default for 9 to 12-year-olds and all in-game text chat turned off by default for under 9s.<sup>51</sup> Roblox rolled out age-based chat globally on 7 January 2026.<sup>52</sup> Roblox stated voice chat was unavailable for Australian users under 13 for the whole reporting period, and that from December 2025 it was also unavailable for Australian users under 16. Roblox stated the third-party provider of its age assurance tools, Persona, assessed the effectiveness of these mechanisms.

Mojang Studios stated it relied solely on self-declaration to assess a user's age as part of the Microsoft Service Account ('MSA') Authentication process. Parents or guardians were required to create an account on a child's behalf and had access to parental controls for Minecraft, if it was played on Xbox or PC, but were not required to verify their own age or their status as a parent. Mojang Studios stated players must also authenticate their age on the platform on which they accessed the game.

Epic Games stated players of Fortnite who self-declared as under 16 in Australia were initially put into a 'Cabined Account', which required the player's parent to provide 'Verified Parental Consent' for in-game features, such as spending money or communicating with other players via voice or text chat.<sup>53</sup> Epic Games stated the age assurance services provided

---

<sup>51</sup> Roblox (2025), [Roblox Requires Age Checks for Communication, Ushering in New Safety Standard](#), Roblox website.

<sup>52</sup> Roblox (2026), [A New Era of Safety: Facial Age Checks Now Required to Chat on Roblox](#), Roblox website.

<sup>53</sup> Epic Games (2026), [Cabined Accounts](#), Epic Games website.

by Kids Web Services for Fortnite in Australia included facial age estimation technology developed by Yoti and ID verification technology developed by Veriff. If a parent had undergone age assurance through Kids Web Services before, they could opt to verify their identity by providing their email address, which was hashed against a database of previous verifications, instead of going through the full verification process.<sup>54</sup>

‘Fortnite Developer’ accounts were required to confirm via Kids Web Services that they were at least 18 years old to join teams of ‘Creators’ for specific Unreal Editor for Fortnite projects.<sup>55</sup> Additional age assurance was required to publish or monetise an Island.

Valve stated it relied on self-declaration at sign-up to age assure Australian end-users. Valve stated it did not track end-users’ birth dates and could not provide monthly active users statistics broken down by the number of child and adult accounts. If a child self-declared they were 13 years or older at sign up, they would be able to use the Steam and Steam Chat services, as well as access certain age-restricted games.

Roblox, Minecraft and Fortnite had safety settings in place by default when a child account was created.

Roblox and Valve also provided some games on their services that were only for users 18 or older. Roblox stated it did not permit R18+ content even in 18+ games.<sup>56</sup>

Table 1 on the following page sets out the types of age assurance used by each of the service providers during the report period. Epic Games stated it used six of the eight types of age assurance tools identified in the table, as well as other techniques, including moderators flagging users. Mojang Studios only relied on self-declaration at sign up. Roblox was the only service provider to use an age inference model, however this was only launched in January 2026. None of the providers used an age key.

---

<sup>54</sup> Kids Web Services (n.d.), [Age Verification](#), Epic Games website.

<sup>55</sup> Epic Games (n.d.), [Creating Teams in Creator Portal in Unreal Editor for Fortnite](#), Epic Games website.

<sup>56</sup> Roblox (n.d.), [Restricted Content Policy – Roblox Support](#), Roblox website.

Table 1: Types of age assurance tools

|                                                         | Epic Games | Mojang Studios | Roblox | Valve |
|---------------------------------------------------------|------------|----------------|--------|-------|
| Self-declaration                                        | ●          | ●              | ●      | ●     |
| Facial age estimation                                   | ●          |                | ●      |       |
| ID document verification (government or non-government) | ●          |                | ●      |       |
| Credit card verification                                | ●          |                |        |       |
| Age inference model*                                    |            |                | ●      |       |
| Age key                                                 |            |                |        |       |
| Vouching (parental or other institution)                | ●          |                | ●      |       |
| Moderators flagging users                               | ●          |                |        |       |

\*Roblox launched its age inference model in January 2026. This was part of a broader fraud model, rather than a standalone tool.

**Note:** Valve reported that it used credit card validation, but this was only used for age verification on UK users and therefore is not included in the table.

Table 2 on the following page shows the parts of the service where age assurance tools were used. eSafety notes a larger number of tools used to assess age of users does not necessarily equate to stronger age assurance measures being in place.

**Table 2: Number of age assurance tools and techniques used by service providers at different stages of accessing each service**

|                                                                                 | <b>Epic Games</b> | <b>Mojang Studios</b> | <b>Roblox</b> | <b>Valve</b> |
|---------------------------------------------------------------------------------|-------------------|-----------------------|---------------|--------------|
| <b>Total tools used</b>                                                         | <b>10</b>         | <b>2</b>              | <b>4</b>      | <b>1</b>     |
| At sign up                                                                      | 1                 | 1                     | 2             | 1            |
| During gameplay                                                                 | 0                 | 0                     | 3             | 0            |
| To access certain games*                                                        | 0                 | 0                     | 3             | 1            |
| To gain access to parental controls                                             | 4 <sup>^</sup>    | 1                     | 1             | 0            |
| Upon accessing certain game features or parts of the service                    | 9                 | 0                     | 3             | 0            |
| When an account has been flagged for review by automated tools/human moderators | 5                 | 0                     | 1             | 0            |
| When the service is accessed on a particular device                             | 0                 | 1                     | 0             | 0            |
| Other                                                                           | 2 <sup>+</sup>    |                       |               |              |

\*To access certain games includes: To access certain Experiences (Roblox), to access certain Islands (Fortnite), to access certain servers (Minecraft), to access certain games (Steam).

<sup>^</sup>To gain access to parental controls: includes parents undergoing parent verification in order to provide verified parental consent for a child.

<sup>+</sup>Other includes moderator identification of age indicators when assessing content for publication.

**Note:** Tools can be used on multiple parts or stages of a service, therefore the total number of tools used at different stages does not add up to the total number of tools.

## Identifying incorrect user age

It is important for service providers to take action when they identify they hold incorrect age information for a user. We asked providers about the steps they took when they became aware that a user's age was incorrectly assured, as well as how they assessed the effectiveness of their age assurance tools and techniques.

Service providers had different levels of organisational maturity in how they identified accounts with incorrect age settings. They were also inconsistent in how they evaluated the performance of their age assurance tools and techniques.

Roblox stated after it became aware the age information it held about an end-user may be incorrect, it would either direct the account to undergo age assurance or revoke the age checked status and remove access to all features that required a user to have undergone an age check. However, Roblox did not ban accounts in these circumstances.

In the period from 7 January 2026 to 31 March 2026, Roblox only identified 15,105 incorrectly age-assured accounts globally. In the previous three months (prior to implementing additional age assurance measures), it did not detect any. Roblox stated that, during the

report period, it was developing age inference capabilities through its Continuous Age Check fraud model, which would infer age using behavioural signals, including those related to gameplay. Roblox stated this Continuous Age Check model was implemented from January 2026 onwards.

Epic Games and Mojang Studios did not use tools to proactively detect incorrectly age-identified Fortnite and Minecraft users.

- Epic Games said it was able to identify 1,931 incorrectly age-assured Australian accounts during the report period, where players reached out to change their age or made statements about their age to player support or where moderators reviewed reported voice or text chat.<sup>57</sup>
- When asked how many end-users were found to be incorrectly age assured, Mojang Studios responded, ‘none to our knowledge’.

Epic Games stated Fortnite’s moderators could direct a player or a player’s parent to undergo age assurance if they identified signals indicating that a player was incorrectly age assured.

Valve stated it did not assess the effectiveness of self-declaration as an age assurance mechanism. Valve also did not proactively monitor whether child users had been incorrectly assessed as adults. However, Valve did identify five accounts that had been incorrectly assessed as adults during the report period and took enforcement action in response.<sup>58</sup>

## Parental controls

Parental controls provide an extra layer of protection for children playing online games, by allowing parents to monitor and limit what their child does online.<sup>59</sup> Service providers were asked about the proportion of Australian under 18 accounts on their services with parental supervision enabled, and what those parental controls were.

Roblox stated only 4% of Australian under-18 accounts had parental controls enabled. Parents could use Roblox’s parental controls to enable chat functions that were turned off by default for younger children.

Minecraft accounts tied to a Microsoft Services Account for under-18s in Australia were subject to default restrictions such as blocking multiplayer and

---

<sup>57</sup> Epic Games stated this number included cases where the individual reached out to player support to change their age, or if Epic Games had reason to believe that a player’s declared age was lower than their true age and required the player to confirm their age.

<sup>58</sup> Actions included terminating or suspending an account or transferring it to a parent.

<sup>59</sup> eSafety Commissioner (2026), [Parental controls](#), eSafety website.

communication features unless explicitly enabled by a parent. Parents could also elect to monitor or pause their children's gameplay or approve friend requests. Mojang Studios stated almost all (94%) of accounts in Australia classified as under 18 that logged into Xbox services during the report period were associated with a Microsoft/Xbox Family Group, but could not provide details on how many of these services had parental controls enabled outside of the default settings.<sup>60</sup>

Epic Games stated 74% of Australian accounts self-declared as under 18 had parental controls enabled, with an additional 11% playing in a Cabined Account where certain features were inaccessible without verified parental consent. For Australian accounts self-declared as under 16, 100% of players either had parental controls enabled or were playing in a Cabined Account.

Valve's parental controls enable parents to limit children's access to Steam Chat and Steam Community features, and could limit the ability of children to receive friend requests. Unlike Roblox, Minecraft and Fortnite, children's Steam accounts in Australia did not have default safety settings in place and were not automatically linked to parental controls.<sup>61</sup> During the report period, Valve stated 75,330 Australian user accounts had parental controls enabled., approximately 4% of its monthly active users for the same period.

## Key insights and safety gaps

- Players of Fortnite and Minecraft who self-declared as 18 or older could access some high-risk features without the benefit of default protections in place for child accounts, such as chat with players of all ages (other than with self-declared child accounts) and, in the case of Minecraft, unmoderated, unaffiliated servers hosted by third-parties. More information about these servers and player chat can be found in this report's Focus section on ['Proactive detection'](#).
- Valve stated it did not track end-users' birth dates and could not provide MAUs broken down by the number of child and adult accounts. If a child self-declared they were 13 years or older at sign up, they would be able to use the Steam and Steam Chat services, as well as access certain age-restricted games.
- Mojang Studios, Epic Games and Valve stated they did not assess the effectiveness of self-declaration as an age assurance method during the report period.

---

<sup>60</sup> Microsoft and Xbox updated their approach to minor accounts in November 2025, adjusting the age of consent from 13 to 18. Mojang Studios attributed the remaining 5.6% of accounts under 18 that had been active on Microsoft services to reflect the 13 to 17 year olds that did not have minor accounts prior to this period.

<sup>61</sup> Valve Corporation (2024), [Steam Support: Steam Families User Guide & FAQ](#), Steam website.

- Roblox and Valve stated only 4% of under-18 accounts in Australia on either service had parental controls enabled.
- Epic Games and Mojang Studios did not use tools to proactively detect incorrectly age identified end-users. When asked how many end-users were found to be incorrectly age assured, Mojang Studios responded, 'none to our knowledge'.

# Focus: Proactive detection

## Why were service providers asked about this?

Proactive detection helps service providers identify unlawful and harmful material or activity before, or soon after, it appears or occurs on the service. It is an important way to reduce the spread of online harms and respond quickly when serious risks occur.<sup>62</sup>

This focus area is specifically about CSEA material and activity (including grooming), sexual extortion and violence material and activity. It examines tools and techniques used on online game and gaming-adjacent services to proactively detect:

- online harms in communication features, such as player chat
- online harms in user-generated content and in-game items
- games and users who are high risk for harmful activity, including players with a history of harmful behaviour
- high-risk transactions of virtual items and currencies.

A range of tools and techniques can be used to proactively detect online harms, including:

- **hash-matching tools**, which use digital fingerprints (known as ‘hashes’) to detect copies of known unlawful material, such as CSEA and pro-terror material
- **language analysis tools**, which can be used on usernames, text chat and other written content as well as voice chat to identify terms or patterns associated with CSEA or violence material or activity
- **monitoring signals and metadata**, such as suspicious patterns of virtual currency transactions between adults and children, which may help identify harmful behaviour or risk
- **AI classifiers**, which use artificial intelligence to sort information or content such as images or text into categories of harmful and benign content (AI classifiers can also be used as part of other tools, such as language analysis tools).

---

<sup>62</sup> NYU Stern Center for Business and Human Rights (2023), [\*Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter Them\*](#), NYU website; The Bracket Foundation (2022), [\*Gaming and the Metaverse: The Alarming Rise of Online Sexual Exploitation and Abuse of Children Within the New Digital Frontier\*](#), United Nations Interregional Crime and Justice Research Institute (UNICRI) website. Accessed 19 September 2026.

eSafety does not consider responding to user reports to be a proactive detection tool or technique.

The effectiveness of these tools can depend on several factors, including:

- whether the tool is used across the whole service or only on part of it, such as text chat
- whether the tool is used by default or only in particular circumstances
- how frequently the tool is updated.

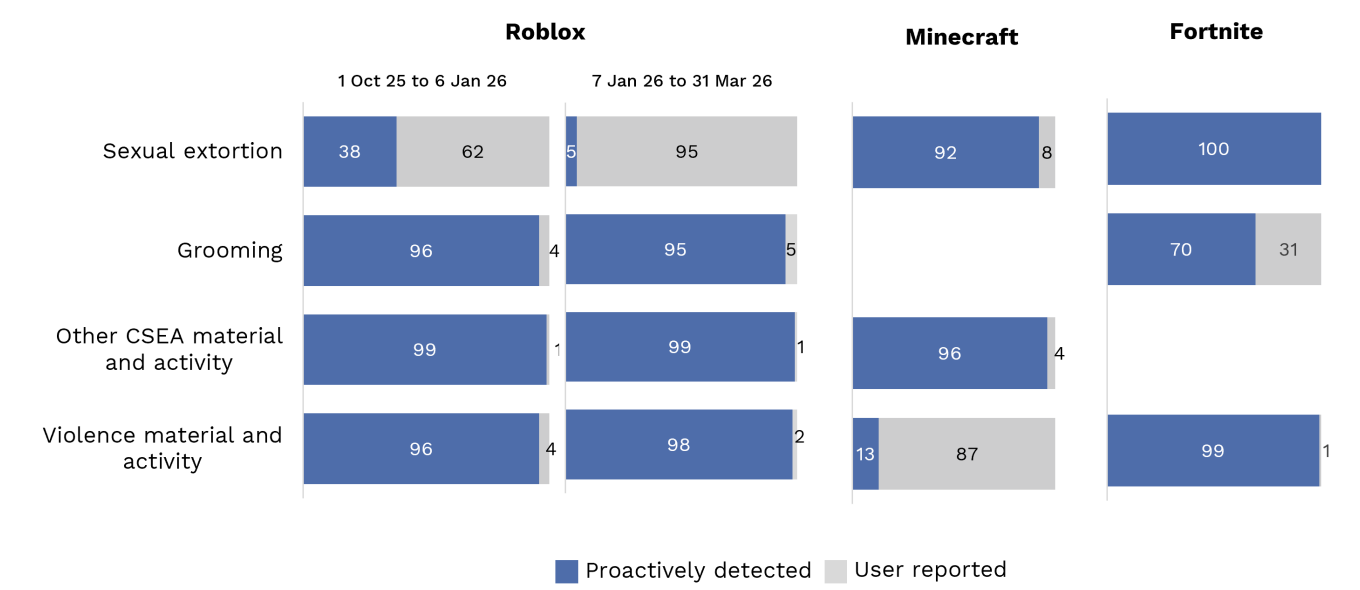
The source and quality of any keyword lists, hash databases or other reference data the tool relies on how frequently a tool is assessed or evaluated for accuracy, and whether any action is taken as a result of those assessments.

## Safety measures in place

eSafety found all service providers used multiple tools and techniques to proactively detect online harms across several different parts of their services.

Roblox, Epic Games and Mojang Studios reported that overall, nearly all CSEA material, and activity and violence material and activity (that service providers were aware of) was found using proactive tools and techniques. Unlike other providers, Mojang Studios provided the percentage of unlawful and harmful activity that was reported by users, rather than detected through user reports. This means that the statistics on Minecraft in Figure 2 may include inaccurate user reports. Steam and Steam Chat are not included in the figure as Valve did not provide a breakdown of reports by harms.

**Figure 2: Proportion of sexual extortion, grooming, other CSEA material and activity, and violence material and activity, that was detected proactively versus through user reporting, by type of harm**



**Note:** Steam and Steam Chat are not included in the graph as Valve did not provide a breakdown of reports by harms.

**Note:** Figure totals in the graph may not add to 100 due to rounding. For most providers, reporting categories did not directly align with eSafety harms categories. Roblox’s category for Sexual extortion includes all categories that relate to blackmail and extortion reports. Unlike other providers, Mojang Studios used unconfirmed user reports for this statistic. For Mojang Studios, Grooming is included in Other CSEA material and activity, and the figure for Sexual extortion represents Minecraft’s Sexually Inappropriate category, which includes harms other than sexual extortion. For Epic Games, confirmed reports of other CSEA material and activity other than child sexual abuse material is captured in Grooming and Sexual extortion. Epic Games stated there were no confirmed cases of child sexual abuse material detected during the reporting period.

## Sub-focus: Detecting online harms in chats and other communication features

### Why were service providers asked about this?

Communication between players forms a key part of many major online game and gaming-adjacent platforms, whether through community forums, text or voice chat. While these features can provide an opportunity for players to socialise and communicate, they can also enable exposure to contact harms such as grooming, sexual extortion, and promotion of or incitement to violence. Communication features can also be used to off-platform a child, moving them to a more private (sometimes end-to-end encrypted) service where unlawful or harmful material and activity may be easier to share undetected.<sup>63</sup>

<sup>63</sup> Australian Government (2022), [Online Safety \(Basic Online Safety Expectations\) Determination 2022 s6\(2\), s6\(3\) s11, s12](#), Federal Register of Legislation website.

Many online service providers use language analysis tools to detect keywords or patterns associated with unlawful or harmful material or activity. To remain effective, language analysis tools must be regularly updated to detect new slang and patterns of behaviour as well as attempts to circumvent chat filtering systems. It is also important that language analysis tools operate across the languages of the users on the service. In Australia, the top five languages spoken other than English are Mandarin, Arabic, Vietnamese, Cantonese, and Punjabi (see this report's Focus section on [‘Staff capacity and capability’](#) for more details on the languages covered by human moderators).

In addition to language analysis tools, URL detection tools can be used to detect links to websites associated with CSEA and violence material and activity (including terrorism or extremism), or to help detect attempts at off-platforming. Where communication features allow for images and videos to be shared, hash-matching tools and AI classifiers can also be used to detect known and new CSEA and violence material.

Organisations and initiatives such as The Tech Coalition, Thorn, and the Internet Watch Foundation (IWF) help provide lists of common terms, abbreviations and codes associated with CSEA (including grooming), and sexual extortion. Organisations such as Tech Against Terrorism and the Global Internet Forum to Counter Terrorism (GIFCT) also provide keywords relating to violent extremist groups. These same organisations also variously provide databases of hashes and URLs of known and new CSEA and terrorist and violent extremist material.

## What we found

### Roblox

During the report period, users could communicate in the following ways:

- **Experience Chat:** This enables users to interact with each other while playing a specific Experience. Depending on a user's age and whether they have completed an age assurance process, users can use text chat to communicate, which appears in a live scrolling feed during gameplay, or use voice chat, which activates when a user's avatar is in the proximity of other user avatars. Users can also communicate one-on-one with anyone else in the Experience via the live text feed ('Direct Experience chat').<sup>64</sup> Experience text and voice chats are available both in public Experiences, which any player can join, or user-owned versions of an Experience (known as 'private servers') which require permission from the owner of the server to join.

---

<sup>64</sup> Roblox (2026), [Safety by Age](#), Roblox website.

- **Party Chat:** This allows players to communicate in and outside of Experiences. Unlike Experience Chat, Party text chats are permanent and ongoing. Users are unable to send or receive images, videos, or links in chats. Players are also able to use voice chat with their Party.
- **Roblox Communities:** These are groups of Roblox developers (or ‘Creators’) and users who share similar interests. Communities can be set so that anyone can join them or can request approval to join from the creators or moderators of a Community. Members of a Community can post and interact with each other using text on Community forums. Members cannot upload images onto Community forums, but moderators of a forum may use images in their announcements.
- **Unofficial chat functions:** Roblox Creators were able to create their own ‘unofficial’ text chat functions in Experiences, in violation of Roblox’s Community Standards.<sup>65</sup> Roblox stated it was not possible for Creators to create unofficial voice chat functions.

Following its rollout of age-verified chat, Roblox reported all users could now only chat with other users in their respective age groups.<sup>66</sup> These age-verification techniques did not apply to Roblox Communities or unofficial chat functions. eSafety raised concerns with Roblox as part of its investigation into compliance with the relevant industry standard about ongoing communication features on Community forums available to children, as well as other concerns that Roblox did not have adequate safety measures in place to prevent attempts by adults to contact children and view information about those children. These concerns, and others, resulted in an [enforceable undertaking](#) committing Roblox to take additional measures.

Roblox stated it used a total of six language analysis tools on its service. One of these tools, Roblox’s Text Filter, ran on all material in text chats and Roblox Communities and reportedly detected grooming and other CSEA activity, sexual extortion, violence material and activity (including terrorism and violent extremism, suicide, self-harm, gore, and what Roblox called ‘real-world sensitive events’),<sup>67</sup> and attempts to migrate children to another service. Roblox also stated its Text Filter applied to unofficial chat functions. Roblox’s grooming detection tool, Sentinel, ran on Party text chat. In addition to grooming, Sentinel also reportedly detected other CSEA material and activity as well as attempts to migrate children to another service. All language analysis tools monitoring text chat ran on Experience chats within private servers.

---

<sup>65</sup> Roblox (2026), [Roblox Community Standards](#), Roblox website.

<sup>66</sup> Roblox (2026), [Roblox Requires Age Checks for Communication. Ushering in New Safety Standard](#), Roblox website.

<sup>67</sup> Roblox (2026), [Roblox Community Standards](#), Roblox website.

Two tools were run on Experience and Party voice chat to moderate voice chats for grooming and attempts to migrate children to another service. eSafety notes that Roblox's use of proactive detection tools on voice chats may have aided in detecting cases of sexualised voice chat or simulated sexual contact with children, such as that previously reported occurring on Roblox.<sup>68</sup> If a user was flagged as suspicious for 'child exploitation', two further language analysis tools (which are referred to as Additional language analysis tools A and B for the purposes of this report) were run on the user's voice chats as well as Party text chats and private server chats.

Only one of Roblox's language analysis tools, its Text Filter could reportedly interpret coded or indirect language (including the use of emojis or symbols in words). This is concerning noting the public information on Sentinel on Roblox's website, which stated that 'Bad actors often use subtle, indirect, or coded language' and that their detection systems 'continually evolve to keep pace with new ways bad actors attempt to evade our systems'.<sup>69</sup>

In addition to the tools Roblox mentioned in the Notice, Roblox has also published information on its website with respect to a Personally Identifiable Information (PII) classifier.<sup>70</sup> When asked whether this constituted a tool for the purposes of the Notice, Roblox stated, 'we were not using our open source personally identifiable information (PII) classifier during the reporting period'.

Roblox stated it used an internal team of subject matter experts to develop and update the keywords and training data for these tools to leverage what it called 'platform-specific expertise'. Of these tools, one was updated against source keywords/training data multiple times a day, one was updated every four months, and three were updated on an annual basis. Roblox stated it did not update its grooming classifier as it was developed by a third-party, despite also stating that the tool used its internal training data. This suggests the tool had not been updated since it was developed.

Roblox assessed five of these tools for accuracy during the report period. More information on these assessments can be found in this report's Sub-focus section on '[Measuring the accuracy of proactive detection tools](#)'.

In addition to its language analysis tools, Roblox also stated it had tools in place to detect unofficial text chat functions.

---

<sup>68</sup> Revealing Reality (2025), [A digital playground: The real guide to Roblox](#), Revealing Reality website.

<sup>69</sup> Koneru, N., & Vonck, E. (2025), [Open sourcing Roblox Sentinel: Our approach to preemptive risk detection](#). Roblox website. Accessed 15 May 2026; Roblox (2025), [Open Sourcing Roblox Sentinel: Our Approach to Preemptive Risk Detection](#), Roblox Website.

<sup>70</sup> Roblox (2025), [Open Sourcing Roblox PII Classifier: Our Approach to AI PII Detection in Chat](#), Roblox website.

**Table 3: Roblox tools used on communications surfaces**

|                                                         | Voice Grooming Detection Model | Text Filter | Sentinel | Additional language analysis tool A | Additional language analysis tool B | Specialised review techniques – A | Audio language analysis tool | Specialised review techniques – B |
|---------------------------------------------------------|--------------------------------|-------------|----------|-------------------------------------|-------------------------------------|-----------------------------------|------------------------------|-----------------------------------|
| <b>Communications surface</b>                           |                                |             |          |                                     |                                     |                                   |                              |                                   |
| Experiences                                             |                                | •           |          |                                     |                                     |                                   |                              | •                                 |
| Direct Experience chat                                  |                                | •           |          |                                     |                                     |                                   |                              |                                   |
| Experience text chat                                    |                                | •           |          |                                     |                                     |                                   |                              |                                   |
| Experience voice chat                                   | •                              |             |          | •                                   | •                                   | •                                 | •                            |                                   |
| Party text chat                                         |                                | •           | •        | •                                   | •                                   | •                                 |                              | •                                 |
| Party voice chat                                        | •                              |             |          | •                                   | •                                   | •                                 |                              |                                   |
| Private server chat                                     |                                | •           | •        | •                                   | •                                   | •                                 |                              |                                   |
| Unofficial text chat functions                          |                                | •           |          |                                     |                                     |                                   |                              |                                   |
| <b>Harm covered and if the tool ran on all material</b> |                                |             |          |                                     |                                     |                                   |                              |                                   |
| Grooming                                                | •                              | •           | •        | ○                                   | ○                                   | ○                                 |                              |                                   |
| Sexual extortion                                        |                                | •           |          |                                     |                                     |                                   |                              |                                   |
| Other CSEA material and activity                        |                                | •           | •        | ○                                   | ○                                   | ○                                 |                              |                                   |
| Violence material and activity                          |                                | •           |          |                                     |                                     |                                   |                              | ○                                 |
| Attempts to migrate children to another server          |                                | •           | •        | ○                                   | ○                                   | ○                                 | •                            |                                   |
| <b>Type of tool</b>                                     |                                |             |          |                                     |                                     |                                   |                              |                                   |
| AI classifier                                           |                                |             |          |                                     | •                                   |                                   |                              |                                   |
| Hash-matching tool                                      |                                |             |          |                                     |                                     |                                   |                              |                                   |
| Language analysis tool (audio)                          | •                              |             |          |                                     |                                     |                                   | •                            |                                   |
| Language analysis tool (text)                           |                                | •           | •        | •                                   | •                                   |                                   |                              |                                   |
| Signals-based detection                                 |                                |             |          |                                     |                                     |                                   |                              |                                   |
| Other                                                   |                                |             |          |                                     |                                     | •                                 |                              | •                                 |
| <b>Languages</b>                                        |                                |             |          |                                     |                                     |                                   |                              |                                   |

|                                                                                      | Voice Grooming Detection Model | Text Filter                  | Sentinel       | Additional language analysis tool A | Additional language analysis tool B | Specialised review techniques – A | Audio language analysis tool | Specialised review techniques – B |
|--------------------------------------------------------------------------------------|--------------------------------|------------------------------|----------------|-------------------------------------|-------------------------------------|-----------------------------------|------------------------------|-----------------------------------|
| English                                                                              | ●                              | ●                            |                | ●                                   | ●                                   |                                   | ●                            |                                   |
| Mandarin                                                                             |                                | ●                            |                |                                     |                                     |                                   | ●                            |                                   |
| Arabic                                                                               |                                | ●                            |                |                                     |                                     |                                   | ●                            |                                   |
| Vietnamese                                                                           |                                | ●                            |                |                                     |                                     |                                   |                              |                                   |
| Cantonese                                                                            |                                |                              |                |                                     |                                     |                                   | ●                            |                                   |
| Punjabi                                                                              |                                |                              |                |                                     |                                     |                                   |                              |                                   |
| Other                                                                                | ●                              | ●                            |                |                                     |                                     |                                   | ●                            |                                   |
| Language agnostic                                                                    |                                |                              | ●              |                                     |                                     |                                   |                              |                                   |
| Frequency of updating and source of keyword/training data of language analysis tools |                                |                              |                |                                     |                                     |                                   |                              |                                   |
| Source of keyword/training data                                                      | Internal lists                 | Internal lists               | Internal lists | Internal lists                      | Internal lists                      |                                   | Internal lists               |                                   |
| Frequency of update                                                                  | Annually                       | Hourly/ Multiple times a day | Annually       | Annually                            | Tool updated by a third-party       |                                   | Every 4 months               |                                   |

●

 Used on material by default.

○

 Used on material in some circumstances. Not applicable

Fortnite

During the report period, Epic Games said Fortnite users could communicate in the following ways:

- **Game chat:** This enables temporary text and voice chat functions during gameplay.
- **Party chat:** This allows text and voice chat functions for groups of players playing in different Islands together.
- **Direct messages/Groups:** This enables end-users to send messages directly to players on their Fortnite friends lists or groups of friends and ‘connections’ (friends of friends), which expire after 30 days.<sup>71</sup>
- **Fortnite Communities:** This allows Fortnite developers to communicate with other players on a series of forums. Users can post text and images to Fortnite Community discussion boards.

Epic Games stated it used two language analysis tools proactively on all text in Fortnite Game chats as well as all text chats involving players classified as under 18. It stated one of these language analysis tools detected child endangerment (including grooming) and sexual extortion, self-harm, threats of real-world violence, and attempts to migrate children to other services. Epic Games also used a ‘Personal Information Filter’ on all Game chats to

<sup>71</sup> Epic Games (2024), [Text Chat in Fortnite](#), Epic Games website.

identify and block any mentions of personal information as email addresses, ID information or phone number. The Personal Information Filter relied on pattern matching and AI machine learning and did not use keyword lists. Both of these tools were off by default but could be turned on for users classified as 18 or older. In chats other than Game Chat, the Personal Information Filter could be turned off by users classified as 13 or older, and was off by default but could be turned on for users classified as 18 or older.

In addition to the tools Epic Games used proactively, Epic Games said it used language analysis tools on player reports for all text and voice chats to detect grooming, sexual extortion, imminent threats of real-world violence or self-harm, communications promoting terrorism or violent extremism, and attempts to migrate children to other services.

Epic Games reportedly used a separate language analysis tool on text in Fortnite Community posts to detect references to terrorism and violent extremism, violence, gore, real-world deaths and physical violence, extreme crimes, and attempts to migrate children to another service. Epic Games also used the Personal Information Filter on Fortnite Communities.

Epic Games stated it used Microsoft's PhotoDNA to detect CSEA material and an AI classifier to detect terrorism, violent extremism and graphic or violent content on images uploaded to Fortnite Community.

Epic Games stated all language analysis tools were updated every two to three months with additional training data, with the base large language model used by all of the tools being updated 'at least yearly'. For its source materials, Epic Games stated it used both internal data and synthetically generated data which used 'seed' keywords and phrases from datasets shared by the Tech Coalition, Thorn keyword lists, and publicly available datasets.<sup>72</sup> Epic Games stated it trained its language analysis tools to detect indirect, coded or coercive language as well as difficult-to-analyse cases such as what it called 'game banter'.

In addition to its language analysis tools, Epic Games ran a URL blocking tool on all material by default on all chats and Fortnite Community posts, which detected and removed URLs sourced from Internet Watch Foundation and Tech Against Terrorism block lists. Epic Games said that this tool was updated daily. This tool was also used to block all URLs (regardless of their source) in game text chat.

Epic Games assessed all of these tools for accuracy during the report period. More information on these assessments can be found at this report's Sub-focus section on ['Measuring the accuracy of proactive detection tools'](#).

---

<sup>72</sup> GitLab (2026), [Early Sexual Predator Detection / eSPD Datasets](#), GitLab website; Inches, G., & Crestani, F. (2012), [PAN12 Deception Detection: Sexual Predator Identification](#), Zenodo website.

**Table 4: Epic Games tools used on communications surfaces**

|                                                         | Image<br>moderations<br>tool for Fortnite<br>communities | Photo<br>DNA | General-<br>purpose<br>text<br>moderation<br>tool | Personal<br>Information<br>Filter | URL<br>blocking<br>tool | Text chat<br>moderation<br>tool | Automated<br>text<br>reporting<br>tool | Voice chat<br>moderation<br>tool |
|---------------------------------------------------------|----------------------------------------------------------|--------------|---------------------------------------------------|-----------------------------------|-------------------------|---------------------------------|----------------------------------------|----------------------------------|
| <b>Communications surface</b>                           |                                                          |              |                                                   |                                   |                         |                                 |                                        |                                  |
| Party text chat                                         |                                                          |              |                                                   | •                                 | •                       | •                               | •                                      |                                  |
| Game text chat                                          |                                                          |              |                                                   | •                                 | •                       | •                               | •                                      |                                  |
| Direct messages                                         |                                                          |              |                                                   | •                                 | •                       | •                               | •                                      |                                  |
| Groups                                                  |                                                          |              |                                                   | •                                 | •                       | •                               | •                                      |                                  |
| Fortnite community posts                                | •                                                        | •            | •                                                 | •                                 | •                       |                                 |                                        |                                  |
| Voice chat                                              |                                                          |              |                                                   |                                   |                         |                                 |                                        | •                                |
| <b>Harm covered and if the tool ran on all material</b> |                                                          |              |                                                   |                                   |                         |                                 |                                        |                                  |
| Grooming                                                |                                                          |              |                                                   |                                   |                         | ○                               | ○                                      | ○                                |
| Sexual extortion                                        |                                                          |              |                                                   |                                   |                         | ○                               | ○                                      | ○                                |
| Other CSEA material and activity                        | •                                                        | •            |                                                   |                                   | •                       |                                 |                                        |                                  |
| Violence material and activity                          | •                                                        |              | •                                                 |                                   | •                       | ○                               | ○                                      | ○                                |
| Attempts to migrate children to another server          |                                                          |              | •                                                 | ○                                 | •                       | ○                               | ○                                      | ○                                |
| <b>Type of tool</b>                                     |                                                          |              |                                                   |                                   |                         |                                 |                                        |                                  |
| AI classifier                                           | •                                                        |              | •                                                 | •                                 |                         | •                               | •                                      | •                                |
| Hash-matching tool                                      |                                                          | •            |                                                   |                                   | •                       |                                 |                                        |                                  |
| Language analysis tool (audio)                          |                                                          |              |                                                   |                                   |                         |                                 |                                        | •                                |
| Language analysis tool (text)                           |                                                          |              | •                                                 | •                                 |                         | •                               | •                                      | •                                |
| Signals-based detection                                 |                                                          |              |                                                   |                                   |                         |                                 |                                        |                                  |
| Other                                                   |                                                          |              |                                                   | •                                 | •                       |                                 |                                        |                                  |
| <b>Languages</b>                                        |                                                          |              |                                                   |                                   |                         |                                 |                                        |                                  |
| English                                                 |                                                          |              | •                                                 |                                   |                         | •                               | •                                      | •                                |
| Mandarin                                                |                                                          |              | •                                                 |                                   |                         | •                               | •                                      | •                                |
| Arabic                                                  |                                                          |              | •                                                 |                                   |                         | •                               | •                                      | •                                |
| Vietnamese                                              |                                                          |              | •                                                 |                                   |                         | •                               | •                                      | •                                |

|                                                                                                      | Image<br>moderations<br>tool for Fortnite<br>communities | Photo<br>DNA | General-<br>purpose<br>text<br>moderation<br>tool | Personal<br>Information<br>Filter | URL<br>blocking<br>tool             | Text chat<br>moderation<br>tool                             | Automated<br>text<br>reporting<br>tool                      | Voice chat<br>moderation<br>tool                           |
|------------------------------------------------------------------------------------------------------|----------------------------------------------------------|--------------|---------------------------------------------------|-----------------------------------|-------------------------------------|-------------------------------------------------------------|-------------------------------------------------------------|------------------------------------------------------------|
| Cantonese                                                                                            |                                                          |              | ●                                                 |                                   |                                     |                                                             |                                                             |                                                            |
| Punjabi                                                                                              |                                                          |              | ●                                                 |                                   |                                     |                                                             |                                                             |                                                            |
| Other                                                                                                |                                                          |              | ●                                                 |                                   |                                     | ●                                                           | ●                                                           | ●                                                          |
| Language<br>agnostic                                                                                 |                                                          |              |                                                   | ●                                 |                                     |                                                             |                                                             |                                                            |
| Frequency of update and source of keyword/training data of language analysis and hash-matching tools |                                                          |              |                                                   |                                   |                                     |                                                             |                                                             |                                                            |
| Source of<br>keyword/<br>training<br>data                                                            |                                                          | NCMEC        | Fortnite<br>data                                  | Relies on<br>pattern<br>matching  | IWF<br>Tech<br>Against<br>Terrorism | Fortnite<br>data and<br>syntheticall<br>y generated<br>data | Fortnite<br>data and<br>syntheticall<br>y generated<br>data | Fortnite data<br>and<br>synthetically<br>generated<br>data |
| Frequency<br>of update                                                                               |                                                          | Daily        | Weekly                                            | 2–3 Months<br>or as<br>required   | Daily                               | 2–3 Months<br>or as<br>required                             | 2–3 Months<br>or as<br>required                             | 2–3 Months<br>or as<br>required                            |
| ● Used on material by default.<br>○ Used on material in some circumstances.<br>■ Not applicable      |                                                          |              |                                                   |                                   |                                     |                                                             |                                                             |                                                            |

Minecraft

During the report period, Mojang Studios reported Minecraft had one in-game text chat function which allowed users in the same server to temporarily communicate with each other.

Mojang Studios stated on text chats in Minecraft Realms and Featured Bedrock Servers it used one language analysis tool, Microsoft’s Community Sift, to detect grooming, sexual extortion, other CSEA material and activity, violence material and activity, and attempts to migrate children to another service.

Mojang Studios also used a tool (labelled in this report as ‘Internal Proprietary Tool A’) to aid its investigative team in detecting nihilistic violent extremism using data from an open-source investigative tool to identify nihilistic violent extremism trends and organisations. This tool was used in conjunction with Community Sift on Featured Bedrock servers, which were hosted by official partners of Mojang Studios, and on Realms servers, which were hosted by Mojang Studios’ parent company Microsoft. eSafety finds this particularly encouraging, noting media reports of Minecraft being used as a first point of contact for sadistic online exploitation groups<sup>73</sup>. Training data for both tools was reportedly updated weekly to monthly. While Community Sift had been trained to interpret code or indirect language as well as sentence level context, Mojang Studios stated Internal Proprietary Tool A solely relied on keyword detection.

<sup>73</sup> The Fifth Estate, [This violent extremist network targets kids online](#), *CBC News*, 21 March 2025, accessed 12 February 2026; NL Times, [Dutch kids active in online chat groups grooming children to commit violence, sex abuse](#), 25 August 2025, accessed 12 February 2026.

Mojang Studios ran its language analysis tools in both text chats and other texts which players could use to communicate with others in-game, such as text written in player-crafted books.

Mojang Studios assessed all of its language analysis tools for accuracy during the report period. More information on these assessments can be found in this report's Sub-focus section on [‘Measuring the accuracy of proactive detection tools’](#).

**Table 5: Minecraft tools used on communications surfaces**

|                                                         | Community Sift | Internal proprietary tool A |
|---------------------------------------------------------|----------------|-----------------------------|
| Featured Bedrock servers                                | •              | •                           |
| Third-party unaffiliated servers                        |                |                             |
| Minecraft Realms servers                                | •              | •                           |
| User-hosted servers                                     |                |                             |
| Text chat                                               | •              | •                           |
| Text on books                                           | •              | •                           |
| Bedrock chat/commands                                   | •              | •                           |
| <b>Harm covered and if the tool ran on all material</b> |                |                             |
| Grooming                                                | •              | •                           |
| Sexual extortion                                        | •              | •                           |
| Other CSEA material and activity                        | •              | •                           |
| Violence material and activity                          | •              | •                           |
| Attempts to migrate children to another server          | •              |                             |
| <b>Type of tool</b>                                     |                |                             |
| AI classifier                                           | •              |                             |
| Hash-matching tool                                      |                |                             |
| Language analysis tool (audio)                          |                |                             |
| Language analysis tool (text)                           | •              | •                           |
| Signals-based detection                                 |                |                             |
| Other                                                   | •              |                             |
| <b>Languages</b>                                        |                |                             |
| English                                                 | •              | •                           |
| Mandarin*                                               | •              |                             |
| Arabic                                                  | •              |                             |
| Vietnamese                                              | •              |                             |
| Cantonese*                                              | •              |                             |
| Punjabi                                                 |                |                             |

|                                 | Community Sift                                                                           | Internal proprietary tool A                                                              |
|---------------------------------|------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| Other                           | ●                                                                                        |                                                                                          |
| Language agnostic               |                                                                                          |                                                                                          |
| Source of keyword/training data | GIFCT,<br>Tool that helps identify nihilistic violent extremism trends and organizations | GIFCT,<br>Tool that helps identify nihilistic violent extremism trends and organizations |
| Frequency of update             | Weekly                                                                                   | Monthly                                                                                  |

- Used on material by default
- Used on material in some circumstances
- Not applicable

**Note:** Community Sift is also a tool that can identify and enable/disable list of URLs that are placed in text. Known Asset Database is also a tool that can string and character match.  
\*Mojang Studios stated its Community Sift tool operated in simplified Chinese.

### Steam and Steam Chat

During the report period, Valve said users on Steam and Steam Chat could communicate in the following ways:

- **Steam community forums (known as Steam Discussions):** This enables users to discuss content relating to games and other topics on public forums. Steam users can participate in Steam Discussions as a standalone feature of the Steam service, through accessing a Steam Community ‘Game Hub’, which compiles all of Steam’s Community features for a particular game, or as part of a Community group (which is not necessarily tied to a particular game).
- **Steam Community group text chats:** This enables group chats associated with a particular game’s Community Hub. On mobile, these chats are accessed via the Steam Chat app.
- **Reviews:** This allows users to leave reviews on a Community Hub. Steam Community Groups can create curated lists of these games.
- **Steam broadcast chat:** Users watching Steam broadcasts can participate in a live chat feed which disappears when users are no longer broadcasting.
- **Steam Chat:** This allows users to direct message or voice chat with friends. Images and videos can be uploaded to Steam Chat.

Valve stated it used a language analysis tool to detect CSEA material and activity (other than grooming), glorification, celebration or encouragement of real-world violence or terrorism, and attempts to migrate children to another service on all written content outside of chats. Valve stated it did not use proactive detection tools to specifically detect grooming or sexual extortion. Valve stated the tool did pick up indirect or coded language and was based on data tagged by moderators as violative of Steam’s policies. Valve stated the model was

updated based on the frequency of encountered content and its confidence in the results, and while it was initially updated daily during the design and implementation stage, it was now updated ‘much less frequently’ due to Valve’s confidence in the model’s performance. Valve also stated it used a URL detection tool to detect CSEA material and activity and attempts to migrate children to another service on Steam Chat, but this tool ‘mainly blocked scams’.

In addition to using proactive detection tools, Valve stated its human moderators ‘manually search for specific terms on Steam using a detection tool, and then fully moderate the results as if they were reported, manually removing community content (usually user-names and profile pictures)’.

**Table 6: Valve’s tools used on communications surfaces**

|                                                         | Steam internal text analysis tool | Steam URL detection tool |
|---------------------------------------------------------|-----------------------------------|--------------------------|
| <b>Communication surface</b>                            |                                   |                          |
| Reviews                                                 | ●                                 |                          |
| Broadcast chats                                         |                                   |                          |
| Steam Chat (text chat)                                  |                                   | ●                        |
| Steam Chat (Community group chat)                       |                                   | ●                        |
| Steam Discussions                                       | ●                                 |                          |
| Other                                                   |                                   |                          |
| <b>Harm covered and if the tool ran on all material</b> |                                   |                          |
| Grooming                                                |                                   |                          |
| Sexual extortion                                        |                                   |                          |
| Other CSEA material and activity                        | ○                                 | ○                        |
| Violence material and activity                          | ○                                 |                          |
| Attempts to migrate children to another server          | ○                                 | ○                        |
| <b>Type of tool</b>                                     |                                   |                          |
| AI classifier                                           | ●                                 |                          |
| Hash-matching tool                                      |                                   |                          |
| Language analysis tool (audio)                          |                                   |                          |
| Language analysis tool (text)                           | ●                                 |                          |
| Signals-based detection                                 |                                   |                          |
| URL detection and blocking                              |                                   | ●                        |
| <b>Languages</b>                                        |                                   |                          |
| English                                                 |                                   |                          |
| Mandarin                                                |                                   |                          |
| Arabic                                                  |                                   |                          |
| Vietnamese                                              |                                   |                          |
| Cantonese                                               |                                   |                          |

|                                                                | Steam internal text analysis tool | Steam URL detection tool |
|----------------------------------------------------------------|-----------------------------------|--------------------------|
| Punjabi                                                        |                                   |                          |
| Other                                                          |                                   |                          |
| Language agnostic                                              | ●                                 |                          |
| <b>Frequency of update and source of keyword/training data</b> |                                   |                          |
| Source of keyword/training data                                | Internal lists                    |                          |
| Frequency of update                                            | Ad hoc                            |                          |

- Used on material by default
- Used on material in some circumstances
- Not applicable

## Key insights and safety gaps

- Valve stated it did not use any tools to proactively detect unlawful and harmful activity in Steam Chat, including Steam Community group chats, other than detection of prohibited URLs, nor did it use tools to proactively detect harms on Steam Broadcast chats.
- The coverage of Roblox's language analysis tools varied greatly depending on which part of the service a player used. Notably, Roblox did not report using proactive detection tools for violence material and activity on voice chats. Additionally, only one of Roblox's language analysis tools, its Text Filter, could interpret coded or indirect language, including use of emojis or symbols in words.
- Epic Games stated it did not proactively detect unlawful or harmful material and activity in voice chats for users who self-declared as 18 and older, and instead it only detected it in user-reported voice chats. Epic Games also stated that it did not use tools to proactively detect CSEA (including grooming), sexual extortion, or violence material and activity in private text chats for users with a self-declared age of 18 years and older.
- Both Roblox and Valve relied exclusively on internally developed keyword lists and training data to update their language analysis tools. Despite engaging more broadly with Lantern, Mojang stated that it did not use external keyword lists or training data to update its language detection tools for CSEA on its services.
- Roblox reported that three of its six language analysis tools were only updated against their source keywords or training data on an annual basis. Valve stated during the design implementation stage it assessed its language analysis tool on a daily basis but assessed it much less frequently now, due to the tool 'working well'.
- Roblox, Epic Games and Mojang Studios stated that their language analysis tools running on text and voice chats did not all operate in all the major languages spoken in Australia.

## Sub-focus: Detecting online harms in user-generated and in-game content

### Why were service providers asked about this?

Many online game and gaming-adjacent services allow players to create and share their own game-related content, including items for use in-game. This is often central to gameplay, creativity and community, but it can also create safety risks if services do not have effective safeguards.

User-generated content across the services examined for this report included the following examples:

- Roblox allowed users to create custom items on its Marketplace for player avatars, such as character skins, accessories and ‘emotes’ (virtual gestures an avatar can make).
- Minecraft allowed players to create 3D structures made from blocks, from basic shapes and statues to elaborate buildings and landscapes.
- Steam users could share artwork, videos, reviews and screenshots of games on Steam Community.

While not user-generated content, Fortnite also had a marketplace of items (known as Fortnite’s Item Shop) that, similar to Roblox, included character skins, accessories and emotes.

Information on whether service providers tested these tools for accuracy and details of these assessments can be found in this report’s Sub-focus section on [‘Measuring the accuracy of proactive detection tools’](#).

## What we found

### Roblox

Roblox reported it used its Image Auto-mod tool on all images on the platform, including catalogue/library items. Roblox stated if content was flagged by Roblox’s Image Auto-mod tool as sexual or containing personally identifiable information, the content was run through an additional image detection tool to determine if it could be CSAM. It stated if content was confirmed as containing material related to terrorism and violent extremism or ‘real-world sensitive events’, an additional image detection tool was run to detect similar images on the platform. For all images that were not AI-generated, Roblox ran Microsoft’s PhotoDNA hash-matching tool to detect CSEA material and activity.

Roblox assessed its Image Auto-mod tool accuracy during the report period. More information on this assessment can be found at the [‘Measuring the accuracy of proactive detection tools’](#) section of the report.

**Table 7: Roblox tools used on user generated content**

|                                                                                         | Image Auto-mod | PhotoDNA             | Hive’s Image Classifier | Search Party | Specialised review technique – B |
|-----------------------------------------------------------------------------------------|----------------|----------------------|-------------------------|--------------|----------------------------------|
| <b>User generated content</b>                                                           |                |                      |                         |              |                                  |
| Catalogue/library items                                                                 | •              | •                    | •                       | •            | •                                |
| <b>Harm</b>                                                                             |                |                      |                         |              |                                  |
| Grooming                                                                                |                |                      |                         |              |                                  |
| Sexual extortion                                                                        |                |                      |                         |              |                                  |
| Other CSEA material and activity                                                        |                | •                    | •                       |              |                                  |
| Violence material and activity                                                          | •              |                      |                         | •            | •                                |
| Attempts to migrate children to another server                                          | •              |                      |                         |              |                                  |
| <b>Type of tool</b>                                                                     |                |                      |                         |              |                                  |
| AI classifier                                                                           |                |                      | •                       |              |                                  |
| Hash-matching tool                                                                      |                | •                    |                         |              |                                  |
| Language analysis tool (audio)                                                          |                |                      |                         |              |                                  |
| Language analysis tool (text)                                                           |                |                      |                         |              |                                  |
| Signals-based detection                                                                 |                |                      |                         |              |                                  |
| Technique using a combination of tools                                                  |                |                      |                         |              | •                                |
| Image detection                                                                         | •              |                      | •                       |              |                                  |
| Image similarity detection with Roblox assets                                           |                |                      |                         | •            |                                  |
| <b>Frequency of updating and source of keyword/training data of hash-matching tools</b> |                |                      |                         |              |                                  |
| Source of keyword/training data                                                         |                | NCMEC                |                         |              |                                  |
| Frequency of update                                                                     |                | Updated by Microsoft |                         |              |                                  |

## Fortnite

Although developers cannot make Item Shop items, Epic Games stated it restricts the way players use emotes from its Items Shop to communicate with each other. Epic Games stated players could not perform certain emote combinations in-game that may be perceived as explicit in close proximity to other players, and players can elect not to see other players doing certain emotes. Epic Games also reported using language analysis tools and AI classifiers to assess the content of user generated names.

**Table 8: Epic Games tools used on user generated content**

|                                                                                                | General purpose text moderation tool | Video highlighting tool | Personal Information Filter |
|------------------------------------------------------------------------------------------------|--------------------------------------|-------------------------|-----------------------------|
| <b>User generated content</b>                                                                  |                                      |                         |                             |
| User generated names (Epic Games display names, Fortnite sidekick names, developer team names) | •                                    |                         | •                           |
| <b>Harm</b>                                                                                    |                                      |                         |                             |
| Grooming                                                                                       |                                      |                         |                             |
| Sexual extortion                                                                               |                                      |                         |                             |
| Other CSEA material and activity                                                               |                                      | •                       |                             |
| Violence material and activity                                                                 | •                                    | •                       |                             |
| Attempts to migrate children to another server                                                 | •                                    |                         | •                           |
| <b>Type of tool</b>                                                                            |                                      |                         |                             |
| AI classifier                                                                                  | •                                    | •                       | •                           |
| Hash-matching tool                                                                             |                                      |                         |                             |
| Language analysis tool (audio)                                                                 |                                      |                         |                             |
| Language analysis tool (text)                                                                  | •                                    |                         | •                           |
| <b>Frequency of updating and source of keyword/training data of language analysis tools</b>    |                                      |                         |                             |
| Source of keyword/training data                                                                | Internal data                        |                         |                             |
| Frequency of update                                                                            | Weekly                               |                         |                             |

## Minecraft

Before a user entered a server, Mojang Studios said it employed a hash-matching tool (known in this report as ‘Internal Proprietary Tool B’) to detect any player skins and known usernames which violate Minecraft’s community standards, including usernames associated with CSEA, violent content and gore. Mojang Studios also reported it used Microsoft’s hash-matching tool, PhotoDNA, on resource packs and gallery screenshots in Minecraft Realms servers. Mojang Studios also reported using Community Sift and Internal Proprietary Tool A to detect harms contained in user generated names.

Mojang Studios assessed both of these for accuracy during the report period. More information on these assessments can be found at the [‘Measuring the accuracy of proactive detection tools’](#) section of the report.

**Table 9: Mojang Studios tools used on user generated content**

|                                                                                                                                                      | Community Sift | Internal Proprietary Tool A | PhotoDNA | Internal proprietary tool B |
|------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|-----------------------------|----------|-----------------------------|
| <b>User generated content</b>                                                                                                                        |                |                             |          |                             |
| User generated names (Bedrock nametags, Java profile names, gamertags, entity names, resource pack names and description, real name and description) | •              | •                           |          |                             |
| Usernames and player skins                                                                                                                           |                |                             |          | •                           |
| Resources packs                                                                                                                                      |                |                             | •        |                             |
| <b>Harm</b>                                                                                                                                          |                |                             |          |                             |
| Grooming                                                                                                                                             | •              | •                           |          | •                           |
| Sexual extortion                                                                                                                                     | •              | •                           | •        | •                           |
| Other CSEA material and activity                                                                                                                     | •              | •                           | •        | •                           |
| Violence material and activity                                                                                                                       | •              | •                           | •        | •                           |
| Attempts to migrate children to another server                                                                                                       | •              |                             |          |                             |
| <b>Type of tool</b>                                                                                                                                  |                |                             |          |                             |
| AI classifier                                                                                                                                        | •              |                             |          |                             |
| Hash-matching tool                                                                                                                                   |                |                             | •        | •                           |
| Language analysis tool (audio)                                                                                                                       |                |                             |          |                             |
| Language analysis tool (text)                                                                                                                        | •              | •                           |          |                             |
| Signals-based detection                                                                                                                              |                |                             |          |                             |
| Other*                                                                                                                                               | •              |                             |          | •                           |

|                                                                                                               | Community Sift                                                                        | Internal Proprietary Tool A                                                           | PhotoDNA                                                      | Internal proprietary tool B                        |
|---------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------|----------------------------------------------------|
| <b>Frequency of updating and source of keyword/training data of language analysis and hash-matching tools</b> |                                                                                       |                                                                                       |                                                               |                                                    |
| Source of keyword/training data                                                                               | GIFCT, Tool that helps identify nihilistic violent extremism trends and organizations | GIFCT, Tool that helps identify nihilistic violent extremism trends and organizations | GIFCT, IWF, NCMEC, Tech Coalition, Microsoft partner networks | Internal database of Minecraft skins and usernames |
| Frequency of update                                                                                           | Weekly                                                                                | Monthly                                                                               | At least daily                                                | Hourly/Multiple times a day                        |

\* Other includes tools enabling/disabling lists of URLs in text, and tools that apply string and character matching.

## Steam and Steam Chat

Valve stated it used two AI classifier tools (Google cloud vision image scoring and Amazon Rekognition image scoring) and one language analysis tool on its service to proactively detect CSEA (excluding grooming), violence material and activity and attempts to migrate children to another service on Workshop Items, Artworks, and Screenshots in Steam Community.

Valve stated it also conducted hourly audits of screenshots from all active broadcasts to detect CSEA, sexual extortion, attempts to migrate children to another service and/or any real-world violence.

When asked whether these tools ran on all material on these surfaces by default, Valve stated it ‘employs different moderation tools and techniques on different parts of the platform’.

**Table 10: Valve’s tools used on user generated content**

|                                | Automated text analysis | Google’s cloud vision image scoring | Amazon Rekognition image scoring |
|--------------------------------|-------------------------|-------------------------------------|----------------------------------|
| <b>User generated content</b>  |                         |                                     |                                  |
| Screenshots                    | •                       | •                                   | •                                |
| Artwork                        | •                       | •                                   | •                                |
| Workshop items                 | •                       | •                                   | •                                |
| Videos                         |                         |                                     |                                  |
| Broadcasts                     |                         |                                     |                                  |
| User profile names and avatars |                         |                                     |                                  |

|                                                | Automated text analysis | Google's cloud vision image scoring | Amazon Rekognition image scoring |
|------------------------------------------------|-------------------------|-------------------------------------|----------------------------------|
| <b>Harm</b>                                    |                         |                                     |                                  |
| Grooming                                       |                         |                                     |                                  |
| Sexual extortion                               |                         |                                     |                                  |
| Other CSEA material and activity               | •                       | •                                   | •                                |
| Violence material and activity                 | •                       | •                                   | •                                |
| Attempts to migrate children to another server | •                       | •                                   | •                                |
| <b>Type of tool</b>                            |                         |                                     |                                  |
| AI classifier                                  |                         | •                                   | •                                |
| Hash-matching tool                             |                         |                                     |                                  |
| Language analysis tool (audio)                 |                         |                                     |                                  |
| Language analysis tool (text)                  | •                       |                                     |                                  |
| Signals-based detection                        |                         |                                     |                                  |
| Technique using a combination of tools.        |                         |                                     |                                  |

**Note:** Valve stated it relied on user reports on videos that were available on its service and proactively reviewed every broadcast before publication.

## Key insights and safety gaps

- Mojang Studios did not have any tools or techniques to proactively detect uses of Minecraft blocks to recreate certain shapes or structures, instead relying on user reports. This is of concern noting media reports of extremists using Minecraft blocks to recreate extremist imagery and environments such as swastikas and recreations of mass shooting events.
- Valve did not use any hash-matching tools on user-generated content on Steam Community.

## Sub-focus: Detecting high-risk games and harmful players

### Why were service providers asked about this?

All four service providers allowed users to play different games as part of their services. Roblox and Epic Games allowed users to develop their own games within Roblox and Fortnite for other users to play. In Minecraft, users played on different servers which each housed their own Minecraft 'worlds'. Steam sold nearly 30,000 games through its games

store,<sup>74</sup> mostly developed by third-party providers. Without effective safeguards, it is possible for games which have a higher risk of unlawful and harmful material and activity to be made available on these services.

Roblox, Epic Games and Mojang Studios were asked about the tools and techniques they used to detect games that were higher risk for unlawful and harmful material and activity.<sup>75</sup> As Valve was a seller of third-party games rather than a platform for games developed by its users, eSafety asked Valve about its process for screening games from third-party developers.

Service providers were also asked questions about off-platform monitoring. Off-platform monitoring is distinct from off-platforming and refers to the monitoring of activity on other services either provided internally or by third-party services to identify potentially harmful accounts or communities.

Service providers were also asked about indicators they used to detect banned or suspended users re-registering with new details. This is known as ‘recidivism’.

As with previous transparency reports, eSafety has chosen not to publish the specific indicators reported by service providers to prevent recidivism, to avoid this information being misused. Instead, eSafety has indicated the range of indicators used.

Information on whether service providers tested these tools for accuracy and details of these assessments can be found in this report’s Sub-focus section on [‘Measuring the accuracy of proactive detection tools’](#).

## What we found

### Roblox

On Roblox, Creators could create games (‘Experiences’). Experiences could be played in a public or private server. Roblox reported using two specialised review techniques to detect Experiences and users associated with CSEA or violence material and activity. Roblox stated AI classifiers were run on all Experiences (including private servers) at all times to monitor for harmful signals and behaviours of a user or Experience to flag them for further review. Roblox stated these signals included signals relating to terrorism and violent extremism, real-world sensitive events, grooming and other CSEA material and activity. These signals

---

<sup>74</sup> Valve Corporation (2019), [Steam. The Ultimate Online Game Platform](#), Steam website.

<sup>75</sup> Australian Government (2022), [Online Safety \(Basic Online Safety Expectations\) Determination 2022 s6\(2\), s6\(3\)\(a\) s11, s12](#), Federal Register of Legislation website.

could reportedly be picked up by its other moderation tools and then ‘escalated into dedicated review workflows’.

As part of its specialised review process on violence material and activity, Roblox said it undertook ‘targeted monitoring’ for content associated with terrorism and violent extremism and mass-violence related themes. Roblox said its additional internal language analysis tools formed part of its review pipeline for child exploitation. Roblox also stated its image detection and hash-matching tools detected harmful activity before it was published, and that assets imported by creators in Roblox Studio ‘must undergo an asset moderation process pre-publication, and are only visible in a published Experience if approved by Roblox.’ More information on Roblox’s image detection and hash-matching tools can be found in this report’s Sub-focus section on [‘User-generated and in-game content’](#).

Roblox stated it used multiple indicators to detect recidivism and had a dedicated team of engineers focused on improving Roblox’s alternative account detection capabilities. Roblox also reported using off-platform monitoring to detect accounts and communities associated with grooming, sexual extortion and violence material and activity.

## Fortnite

Epic Games said it allowed developers to create their own games (or ‘Islands’) on Fortnite. Fortnite Developers could either use Fortnite Creative or Unreal Editor for Fortnite (UEFN) to create games on Fortnite. Fortnite Creative limited developers to using Epic Games’ pre-existing game assets and mechanics, whereas UEFN allowed developers more freedom to create custom games and in-game assets.

Epic Games stated it proactively reviewed all of Fortnite’s Islands before they were published. For games developed with UEFN, every asset was reportedly automatically screened by moderation tools, which included the following:

- An AI classifier which detected images associated with CSEA, terrorism and violent extremism, gore, and/or extreme crimes.
- A separate AI classifier which detected CSEA and violence material and activity in video trailers for Islands.
- A language analysis tool which was used to detect references to terrorism and violent extremism, violence, gore, real-world deaths and physical violence in text in developer-made content, including display names.
- Microsoft’s hash-matching tool PhotoDNA, which was used to detect CSAM.
- Epic Games’ Personal Information Filter, which ran on all text in Fortnite Islands and filtered out personal information such as email addresses, which could be used as a means to off-platform a potential victim.

Epic Games assessed its image detection AI classifier and its language analysis tool for developer-made content for accuracy during the report period. More information on these assessments can be found at the [‘Measuring the accuracy of proactive detection tools’](#) section of the report.

Epic Games stated these tools were also used on any items Developers created for in-game use, including items sold in-game by the developer for V-Bucks (known as ‘in-island transactions’). Every Island submitted for publishing reportedly underwent ‘holistic human review’, irrespective of whether tools had flagged the presence of violative content. Epic Games stated this meant that moderators flagged any potentially hidden rule-breaking content and conducted a ‘playthrough’ of the Island in a dedicated moderator mode. If the Island contained any content which violated the Fortnite Developer Rules,<sup>76</sup> the Island was rejected from publishing with the developer being able to submit the Island again with the content removed. In the case of ‘high-harm violations’, Epic Games stated Islands were permanently blocked from being published and the developer was banned. Post-publication, the Epic Games Investigation team would reportedly conduct ‘periodic reviews’ of island titles and other indicators, including manually reviewing island and assets.

Epic Games stated it used multiple indicators to detect recidivism. Epic Games also stated it used third-parties to identify on-platform abuse including grooming, sexual extortion and violence material and activity.

## Minecraft

In Minecraft, members of the Minecraft community other than Microsoft or its official partners could make their own servers by configuring their computer as a public internet server. These were labelled by Mojang Studios as ‘third-party unaffiliated servers’.<sup>77</sup>

Mojang Studios has stated, ‘Over ~90% of Minecraft Australian players play solo, offline with family or friends, or online on Realms with known family or friends.’ However, eSafety notes the CEO of Mojang Studios has previously stated third-party unaffiliated servers were among the more likely servers where online harms may occur, due to ‘inconsistent safety approaches taken by independent server operators’. Mojang Studios stated third-party unaffiliated servers ‘do not have connected end points integrated into Mojang Studios safety systems’, and that content moderation on these servers is determined by the operators of these servers. Mojang Studios clarified that this means that it did not use any tools that proactively detected harms in third-party unaffiliated servers, which are independently created and operated by members of the Minecraft community. eSafety also notes these

---

<sup>76</sup> Epic Games (2026), [Fortnite Developer Rules](#), Epic Games website.

<sup>77</sup> Kayleen Walters (2025), ‘16.3’ [Witness Statement for UK Covid-19 Inquiry](#), 17 May 2025.

types of servers can be modded to include voice chat functions, which can carry additional risk.

Mojang Studios stated users had to complete Microsoft Service Account Authentication and pass ‘applicable parental control checks’ to access both of these types of experiences. Mojang Studios stated it relied on in-game reporting, which was built into the client (the player)’s end of the game rather than that of the server host, and collaboration with third-party server-hosting service providers to identify unlawful and harmful material and activity. In addition to in-game reporting, Mojang Studios stated it had a dedicated enforcement team that reviewed any server reported through a formal ‘Report a Server’ form or that came through other help and moderation channels.

Mojang Studios stated it required operators of third-party unaffiliated servers to link and integrate Minecraft’s guidelines and terms and conditions in the hosts’ terms and conditions. Mojang Studios stated it had partnered with GamerSafer to create an ‘Official Server List’ of third-party unaffiliated servers. Mojang Studios stated servers featured on this list must comply with Minecraft Usage Guidelines and demonstrate certain requirements including providing the purpose of the server, intended audience, and foundational community management practices that set the tone, values, and principles of each server. However, these servers did not need to have proactive moderation efforts in place to be featured on this list.<sup>78</sup>

Mojang Studios stated it used a ‘nightly scanner’, which used on-platform and off-platform indicators to identify potentially high-risk servers. Mojang Studios also reported using off-platform monitoring to detect accounts and communities associated with grooming, sexual extortion and violence material and activity.

Mojang Studios stated it used multiple indicators to detect recidivism.

## Steam and Steam Chat

Valve stated it required game developers to complete a survey on the content in their game as part of its onboarding process and it reviews each game and its Steam Store page, along with the completed content survey during the onboarding process. Valve stated its review process included human and AI review. Under Valve’s Steam Distribution Agreement (the agreement third-party game providers enter into in order to distribute online games on Steam), third-party game providers on Steam are responsible for ensuring their games comply with any territorial restrictions and laws. As part of the on-boarding process, Valve stated it checked:

---

<sup>78</sup> GameSafer (n.d.), [Introducing Server Badges: Rewarding servers for leveling up gameplay and safety](#), FindMCServer website.

- game titles against keyword lists associated with CSEA or violence material and activity
- whether the game provider has had previous games removed from Steam for distributing class 1 material
- whether the game provider had any patches available off-platform which would allow the game to contain class 1 material after purchase.

Valve stated it would retest a game based on factors such as the number and content of user reports and any known history of the game and developer.

Valve stated it received a number of reports of potential violations of Valve's policies which could include Class 1 material. These reports included 2,424 reports from Australian Steam end-users and 196,626 reports from global end-users, covering categories of AI content, exploitation of children, non-labelled adult content, scams and law violations.

Valve stated during the report period it permanently removed 142 games as part of its review process for violations of its terms of service policies which include content that potentially contained class 1 material (such as CSEA or terrorism and violent extremism material) and reinstated 20 games after the breach was addressed. This process resulted in 10 developers being banned and one being suspended.

Valve used minimal (a small number of) indicators to detect recidivism. Valve did not use off-platform monitoring to identify accounts or communities present on its services that were known to have been engaging in sexual grooming, sexual extortion, or violence material and activity.

## Key insights and safety gaps

- While Roblox, Mojang and Epic Games, all used multiple indicators to detect recidivism, Valve used minimal indicators to detect recidivism. In general, service providers looking for a wider range of indicators to detect recidivism will have a better chance of preventing the re-registration of banned or suspended users.
- While Valve did have a process in place to review games, it did not provide details of any measures in place to proactively detect games that included unlawful material after they had been published on its platform.
- Valve was the only service provider not using off-platform monitoring to identify accounts or communities present on its services that were known to have been engaging in sexual grooming, sexual extortion, or violence material and activity.
- Mojang's reliance on self-declaration of age risked children being able to access third-party unaffiliated servers through Minecraft with little to no moderation.

## Sub-focus: Measuring the accuracy of proactive detection tools

### Why were service providers asked about this?

eSafety's regulatory guidance on the Expectations states service providers should continually assess and evaluate the effectiveness of online safety measures on their services and update, refine and adjust these safety measures to ensure safe use of the service.<sup>79</sup> For the purposes of this report, eSafety asked service providers whether they had assessed their proactive moderation tools for accuracy and, if so, what the results of these assessments were. eSafety also asked service providers what metrics they used as part of these assessments, and whether service providers had taken any action after assessing a tool for accuracy.

eSafety asked service providers, where applicable, to provide the precision and recall rates for a tool.

- The **precision rate** of a tool measures how many inputs it correctly identifies as meeting certain criteria. It is focused on the quality (correctness) of the results. For example, out of a set of items labelled as harmful content, how many were actually harmful content? High precision can indicate fewer false positives. Low precision can indicate over-enforcement by a tool, which could be considered a reputational risk.
- A **recall rate** measures how many times a tool correctly identifies inputs as meeting certain criteria compared to the actual number of inputs in the dataset that meet that criteria. It is focused on the coverage (completeness) of the results. For example, out of all harmful content a tool could have picked up, how much did it detect? High recall can indicate that most or all actual violations are detected. Low recall can indicate under-enforcement by a tool, which is a safety risk.

These are some other metrics that service providers reported on:

- **False positive rate** – this measures how many times a tool incorrectly identifies input as meeting certain criteria compared to the actual number of inputs in the dataset that do not meet that criteria. For example, how much benign content does a tool incorrectly flag as harmful? Unlike precision rates, false positive rates for proactive detection tools measure incorrectly flagged benign content as a percentage of all benign content a tool could have detected (not as a percentage of all content labelled

---

<sup>79</sup> eSafety Commissioner (2026), [Basic Online Safety Expectations Regulatory Guidance](#) (p.30), eSafety website.

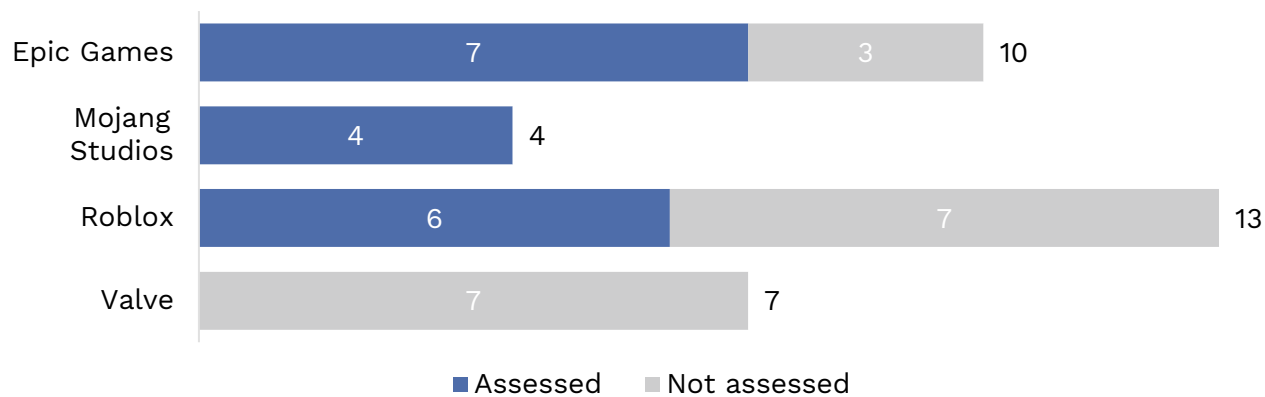
by the tool as harmful). High false positive rates can indicate over-enforcement by a tool.

- **Accuracy** – this measures the percentage of correct decisions made by the tool out of all moderation decisions. Unlike recall, accuracy measures how many correct decisions a tool made out of **all content** (not as a percentage of how much **harmful content** a tool could have picked up).
- **Actionability rates** – this measures the proportion of all items escalated for action that were escalated as a result of content flagged by the tool.

## What we found

All providers except for Valve assessed at least one of their proactive detection tools for accuracy during the report period. Mojang Studios was the only provider to report assessing all their tools for accuracy during the report period.

**Figure 3:      Number of tools assessed for accuracy by service provider**



**Note:** Valve stated two of its tools were assessed by the third-party provider.

### Roblox

Roblox stated it assessed the accuracy of six out of its 13 tools and techniques during the report period. Five of the tools Roblox assessed were language analysis tools. Roblox also assessed its Image Auto-mod tool, which it ran on all images uploaded to the service. Roblox used a variety of methods to assess its six tools which are set out in the following table.

Roblox stated it preferred to use false positive rates instead of precision rates when measuring the accuracy of its Text Filter tool, as precision rates ‘can be significantly influenced by the underlying content of the platform’.

Roblox stated it upgraded two of its tools to newer models and ‘implemented improvements’ to three of its tools’ models following assessment. Notably, Roblox stated it

had enhanced its human reviewer training as a result of assessments of its Image Auto-mod and Text Filter tools. Roblox also stated it increased the training data following assessment of its Image Auto-mod tool and its audio language analysis tool.

**Table 11: Roblox's assessment of its tools**

| Name of tool        | Precision rate                            | Recall rate                                                                                  | Other metrics                                                                                                                                                                                                                                   | Description of methodology                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|---------------------|-------------------------------------------|----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Text Filter         | Not provided                              | Approximately <b>95%</b> for English-language content and <b>88%</b> for non-English content | Block rates: Text Filter blocked <b>4.5%</b> of English-language content and <b>7.57%</b> of non-English content                                                                                                                                | <p>Roblox stated it used a combination of continuous human labelling, where human moderators consistently review, tag and evaluate tools outputs, as well as AI-assisted red-teaming where dedicated individuals or a team use a separate AI to develop input that is likely to generate undesirable behaviour in a tool and then use this input to probe a tool for weaknesses.</p> <p>No specific details were provided on the methodologies used for human labelling, or details on red-teaming and adversarial prompting.</p> |
| Image Auto-mod tool | 61% (compared to 68% for human reviewers) | 51% recall (compared to 68% recall for human reviewers)                                      | Roblox stated automated decisions were made by the tool for approximately 98.5% of images uploaded to the platform, with the remaining images escalated for human review. Of the images assessed by the tool, approximately 4.2% were rejected. | Roblox stated it used continuous human labelling to assess the efficacy of this tool. To assess the performance of the tool, Roblox stated it took a random subset of images labelled by human experts and compared the content that the tool flagged as harmful with the content that had been flagged as harmful by human review. Roblox then measures the precision and recall of both the tool and human review, based on the human-labelled data.                                                                            |
| Sentinel            | Not provided                              | Not provided                                                                                 | Newer model identified 28% more candidates than original model. Overall, 15% more 'sufficient child exploitation evidence' was identified per candidate.                                                                                        | Roblox used 'curated' datasets to compare the performance of newer models against its original models.                                                                                                                                                                                                                                                                                                                                                                                                                            |

| Name of tool                        | Precision rate | Recall rate                                                                                                                                                              | Other metrics                                                                                | Description of methodology                                                                                                                                                                                                                       |
|-------------------------------------|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Additional language analysis tool A | Not provided   | Final rate not provided. Roblox stated testing at different thresholds produced 70% recall with a false positive rate of 10%, 99% recall with false positive rate of 60% | 8% increase in 'the automation of child exploitation detection rates'.                       | Roblox used 'curated' datasets to evaluate model performance across different thresholds. Roblox also stated it measured 'child exploitation detection rates' through both human moderator identification and model-based (ML Judge) assessment. |
| Additional language analysis tool B | Not provided   | Not provided                                                                                                                                                             | Approximately 80% of identified candidates surfaced relevant evidence for moderation tickets | Roblox stated it used 'a combination of indirect performance evaluation and system-level analysis', focusing on how often the tool found relevant evidence to support the work of human moderators.                                              |
| Audio language analysis tool        | Not provided   | Not provided                                                                                                                                                             | Approximate false positive rate of 0.7%                                                      | Roblox stated it assessed the tool by monitoring overall 'model trigger rates' and the volume of user-reported violations as an additional indicator of detection effectiveness.                                                                 |

## Fortnite

Epic Games stated it assessed 7 out of 10 of its proactive moderation tools and techniques during the report period. This included all of Epic Games' language analysis and image detection tools, as well as its URL blocking tool. Epic Games did not assess the accuracy of its video highlighting tool, its Personal Information Filter, or its use of Microsoft's PhotoDNA tool.

Epic Games stated it updated four of its tools to newer models following assessment. Notably, the new version of its automated text reporting tool resulted in an increased precision rate of 25.3%, while its newer version of its voice chat moderation tool resulted in an increased recall rate of 75.7% at the same level of precision. Epic Games also stated it 'regularly improved' and updated its general-purpose text moderation tool.

**Table 12: Epic Games' assessment of its tools**

| <b>Name of tool</b>                              | <b>Precision rate</b>                                                                                                                                                        | <b>Recall rate</b>                                                                                                                                 | <b>Other metrics</b> | <b>Description of methodology</b>                                                                                                                                                                                                                                                                                                                                                                          |
|--------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Image moderation tool for developer-made content | Not provided, on the basis that decisions on this content were subject to human review.                                                                                      | 95.38%                                                                                                                                             | Accuracy of 99%      | Epic Games took samples of Islands previously submitted for publishing and compared the content that the tool flagged as a violation of Epic Games policies with the content that had been previously flagged as a violation by human moderators.                                                                                                                                                          |
| Image moderation tool for Fortnite Communities   | Not provided, as 'users receive immediate feedback on images that the tool detects as violative, and have the opportunity to replace the images, with no sanctions applied'. | 93%                                                                                                                                                |                      | Epic Games took samples of image assets previously uploaded by developers and compared the content that the tool flagged as a violation of Epic Games policies with the content that had been previously flagged as a violation by human moderators.                                                                                                                                                       |
| General-purpose text moderation tool             | 68%                                                                                                                                                                          | 64%, due to the relatively few examples of violative content in the sample dataset being 'more likely to involve ambiguity in the interpretation'. | Accuracy of 99%      | Epic Games stated it took a random subset of live data labelled by human experts and compared the content that the tool flagged as a violation of Epic Games policies with the content that had been flagged as a violation by the experts<br><br>Epic Games also conducted manual reviews on display names to 'collect qualitative insights and prioritise model improvements' as part of its assessment. |
| URL blocking tool                                | 99.81%                                                                                                                                                                       | Improvement of 22.9%. Epic Games stated absolute recall for its current model 'had not been computed' and that it instead measured                 |                      | Epic Games applied the URL blocking tool to 'historical' data and compared the URLs that the newer model flagged as matching relevant blocklists with the URLs                                                                                                                                                                                                                                             |

| Name of tool                  | Precision rate                                                                          | Recall rate                                                                                                                                                                                                                                                                       | Other metrics     | Description of methodology                                                                                                                                                                                                       |
|-------------------------------|-----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                               |                                                                                         | improvements in recall from its previous model                                                                                                                                                                                                                                    |                   | that the original model had previously flagged.                                                                                                                                                                                  |
| Text chat moderation tool     | Not provided, on the basis that decisions on this content were subject to human review. | 81.5%                                                                                                                                                                                                                                                                             |                   | Epic Games stated it took a random subset of live user reports labelled by human experts and compared the content the tool flagged as a violation of Epic Games policies with the content flagged as a violation by the experts. |
| Automated text reporting tool | 25.3%                                                                                   | Not provided<br><br>Epic Games stated the content that was not made into a report by this tool was automatically deleted, however the tool's low precision rate reflected a 'deliberate tradeoff' in order to ensure that a greater proportion of violative content was detected. |                   | Epic Games stated it measured the proportion of all automatically created reports with at least one high-harm violation confirmed by human review.                                                                               |
| Voice chat moderation tool    | 97.8%                                                                                   | 75.7%                                                                                                                                                                                                                                                                             | Accuracy of 91.8% | Epic Games stated it took a random subset of live user reports labelled by human experts and compared the content the tool flagged as a violation of Epic Games policies with the content flagged as a violation by the experts. |

## Minecraft

Mojang Studios stated it assessed all four of its tools and techniques during the report period. For two of these tools, Mojang Studios stated it updated the rules for the tool based on their assessments. For the assessment of PhotoDNA, Mojang Studios stated similar hashes were grouped together by Microsoft to ensure they provided the same result, and poorly performing hashes could be removed from the underlying hash set altogether.

**Table 13: Mojang Studios' assessment of its tools**

| Name of tool                                         | Precision rate | Recall rate                             | Other metrics                                                      | Description of methodology                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|------------------------------------------------------|----------------|-----------------------------------------|--------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Community Sift                                       | Not provided   | Not provided                            | Actionability rate of 24%                                          | <p>Mojang Studios stated it used a team of 'language and cultural specialists' to continuously review and update the application of text rules against relevant data and social media trends in their respective languages.</p> <p>Staff in Mojang's parent company, Microsoft, also reportedly regularly audited Community Sift and could change its ruleset for Mojang Studios specifically or for all games using Community Sift on Xbox. Mojang Studios also stated they used 'weekly safety validation tests' which 'test[s] all of the core safety functions (including chat filtering, automated escalations, and player abuse reporting) to make sure that all safety systems are working as expected before releases go live.'</p> |
| Internal tool detecting nihilistic violent extremism | Not provided   | Not provided                            |                                                                    | <p>Mojang Studios stated it did not assess its internal tool detecting nihilistic violent extremism in the same way as it assessed its other tools due to its use as an aid for investigators rather than as a tool which actively returns prohibited content. Instead, Mojang Studios said it reviewed internally created text rules for the tool daily and updated them to make sure they were working as intended. Mojang Studios stated it also continuously reviewed trends online and social media platforms to determine if text used in the tool's text rules remained current and was reflective of terminology being used on Minecraft.</p>                                                                                       |
| Internal proprietary tool B                          | Not provided   | '100% for a given collection of hashes' | 18 items were overturned on player appeal during the report period | <p>Mojang regularly updated the underlying hashes based on player appeals. Mojang also stated it conducted a biweekly safety validation test to identify false positives and false negatives in the underlying hashes, but did not provide eSafety with details of this test.</p> <p>Mojang Studios stated the nature of its hash-matching mechanisms meant the recall for this tool would in theory be '100% for a given collection of hashes'. It therefore focused its evaluation methods on assessing the quality of the</p>                                                                                                                                                                                                            |

| Name of tool | Precision rate | Recall rate                             | Other metrics                                                                                                                                                                                                                                                                                                                                                              | Description of methodology                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|--------------|----------------|-----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              |                |                                         |                                                                                                                                                                                                                                                                                                                                                                            | hash datasets used by its hash-matching tools.                                                                                                                                                                                                                                                                                                                                                                                                                       |
| PhotoDNA     |                | '100% for a given collection of hashes' | In the period between 1 April 2025 to 31 March 2026, 96.536% of content that matched hashes previously confirmed by Microsoft as reportable CSEA was determined to be violative of Microsoft's service terms. 81.684% of content that matched hashes whose hits had not previously been reviewed by Microsoft was determined to be violative of Microsoft's service terms. | As with its other hash-matching tool, Mojang Studios focused its evaluation methods on assessing the quality of the hash datasets rather than recall rates. Mojang Studios stated its parent company, Microsoft, measured the rate of confirmed hash 'hits' to ensure that content of hashes met Microsoft's guidelines on CSEA and violent extremist and terrorist content, as well as the rate at which content was confirmed to be violative or 'benign' content. |

Steam and Steam Chat

Valve stated content flagged by its internal language analysis tool was subject to human review but did not give details on any assessments made to ensure the tool's accuracy in detecting content.

Valve also stated support agents evaluated Steam accounts that were blocked as a result of their prohibited URL chat detection but did not provide details on whether this tool was assessed for accuracy.

Valve stated Google and Amazon assessed the images detected by their respective image detection tools, but did not provide details on whether Google, Amazon or Valve assessed these tools for accuracy during the report period.

## Key insights and safety gaps

- Based on its response, Valve appears not to have assessed any of its tools for accuracy after deployment. While it stated it ‘regularly reviews all of its tools for performance and potential improvements’, Valve appeared to mainly rely on user feedback and reporting. For the third-party tools used on its service, Valve stated it relied on the third-party provider testing.
- Roblox stated its Image Auto-mod tool makes the majority of its decisions without human review, despite only having a recall rate of approximately 51% and precision of approximately 61%. This is concerning as Roblox’s Image Auto-mod tool is the only tool used to detect violence material and activity on all images on the platform, including catalogue and library items, by default. Roblox also only runs Hive’s Image Classifier on images flagged by this Image Auto-mod tool. Roblox noted that for human reviewers, the recall rate was approximately 68% with a 68% precision rate, based on human-labelled data. Roblox stated ‘both automated systems and human reviewers may encounter challenges in assessing complex or borderline content’.

## Sub-focus: Detecting transactions of virtual items and currencies

### Why were service providers asked about this?

eSafety asked service providers whether they had any tools or other measures in place to detect the use of virtual currencies or items linked to grooming and/or sexual extortion. This is important because perpetrators of CSEA and have been known to offer children rewards such as in-game currencies or other virtual items in order to build trust or draw children into private chats, or in exchange for nude images or other forms of sexual interactions (this is sometimes known as ‘commodified sexual interaction’).<sup>80</sup> Communities linked to sadistic online exploitation have also been known to use in-game currencies to groom potential victims or to get children to share sexual images or videos of themselves.<sup>81</sup> Children who are involved in these transactions may not always realise that the request is exploitative.

---

<sup>80</sup> The Bracket Foundation (2022), [Gaming and the Metaverse: The Alarming Rise of Online Sexual Exploitation and Abuse of Children Within the New Digital Frontier](#), United Nations Interregional Crime and Justice Research Institute (UNICRI) website. Accessed 19 September 2026; WeProtect Global Alliance (2026), [Online gaming and risks to children](#), WeProtect Global Alliance website.

<sup>81</sup> Resolver (2026), [Weaponized loneliness: Understanding a hybrid threat to young people globally](#), Resolver website.

## What we found

Roblox stated it used ‘targeted investigative measures’ to detect transactions of Robux and Robux gift cards, in which investigators reviewed users’ payments, trade and gift card activity within Roblox’s internal customer support to identify anomalies or concerning patterns. Roblox did not state the circumstances in which investigators reviewed these payments or the frequency of these reviews.

Epic Games stated V-Bucks could not be gifted through Fortnite and that its automated text reporting tool, which ran on all conversations with a player under 18, could detect conversations that included transferring V-Bucks or V-Bucks gift cards.

Mojang and Valve did not use any tools to detect transactions of Minecoin or Minecoin gift cards, or trades of items on Steam Community Marketplace during the report period. Mojang Studios stated Minecoins could not be gifted to other users in-game or outside of the game, but that Minecoin gift cards may be purchased and provided outside the service through redemption codes rather than in-service.

## Key insights and safety gaps

Mojang and Valve did not use any tools to detect high-risk transactions of Minecoin gift cards or transfers of items on Steam Community Marketplace during the report period. While eSafety is not aware of any media reports specifying risks with these particular currencies or items, eSafety notes the emerging nature and complexity of this harm, and considers it important for service providers to have measures in place to detect high-risk transactions of virtual currencies and items related to sexual grooming and sexual extortion.

## Focus: User reporting

Under the Expectations, providers of online services are expected to take reasonable steps to enforce their terms of service in relation to the safety of their users.<sup>82</sup> User reporting, along with flagging by automated tools, helps service providers detect unlawful and harmful material and activity that breaches their terms of service.

### Why were service providers asked about this?

User reporting data helps show how users experience a service. The number of reports a service provider receives can vary depending on the size of the service, the nature of the harm, how easy it is to report, and whether users trust that reporting will make a difference. This is relevant especially because eSafety research shows children tend to underreport harms and other negative experiences on gaming platforms, often because they believe reporting will be ineffective.<sup>83</sup>

When considering the volume of user reports and effectiveness of responses on each service it is important to take into account how many children actively use it. A high number of user reports does not necessarily indicate there was a higher proportion of unlawful or harmful material or activity on these services in comparison to others. For example, it may have been proportionate given the overall number of users on a service, or it may have indicated user reporting options were more easily accessible than on other services.

### What we found

User reporting volumes for each harm category varied significantly across the services. These were not broken down by age of user.

Of the harms asked about under the Notices, cyberbullying was the complaint category with highest reporting on Roblox. Roblox stated, following implementation of age-verified chats on 7 January 2026, the number of cyberbullying reports globally dropped dramatically from 1,159,483 reports down to 274,601 reports. By contrast, Roblox stated reports of grooming in Australia rose from 84 to 337 cases after implementation of age-verified chats. These changes in user reporting numbers do not necessarily indicate a decrease in cyberbullying or

---

<sup>82</sup> Australian Government (2022), [Online Safety \(Basic Online Safety Expectations\) Determination 2022 s14\(1A\)](#), Federal Register of Legislation website.

<sup>83</sup> eSafety Commissioner (2024), [Leveling up to stay safe: Young people's experiences navigating the joys and risks of online gaming](#) (p.53–56), eSafety website.

an increase in grooming on Roblox during these periods, nor that age-verification was the cause of any change.

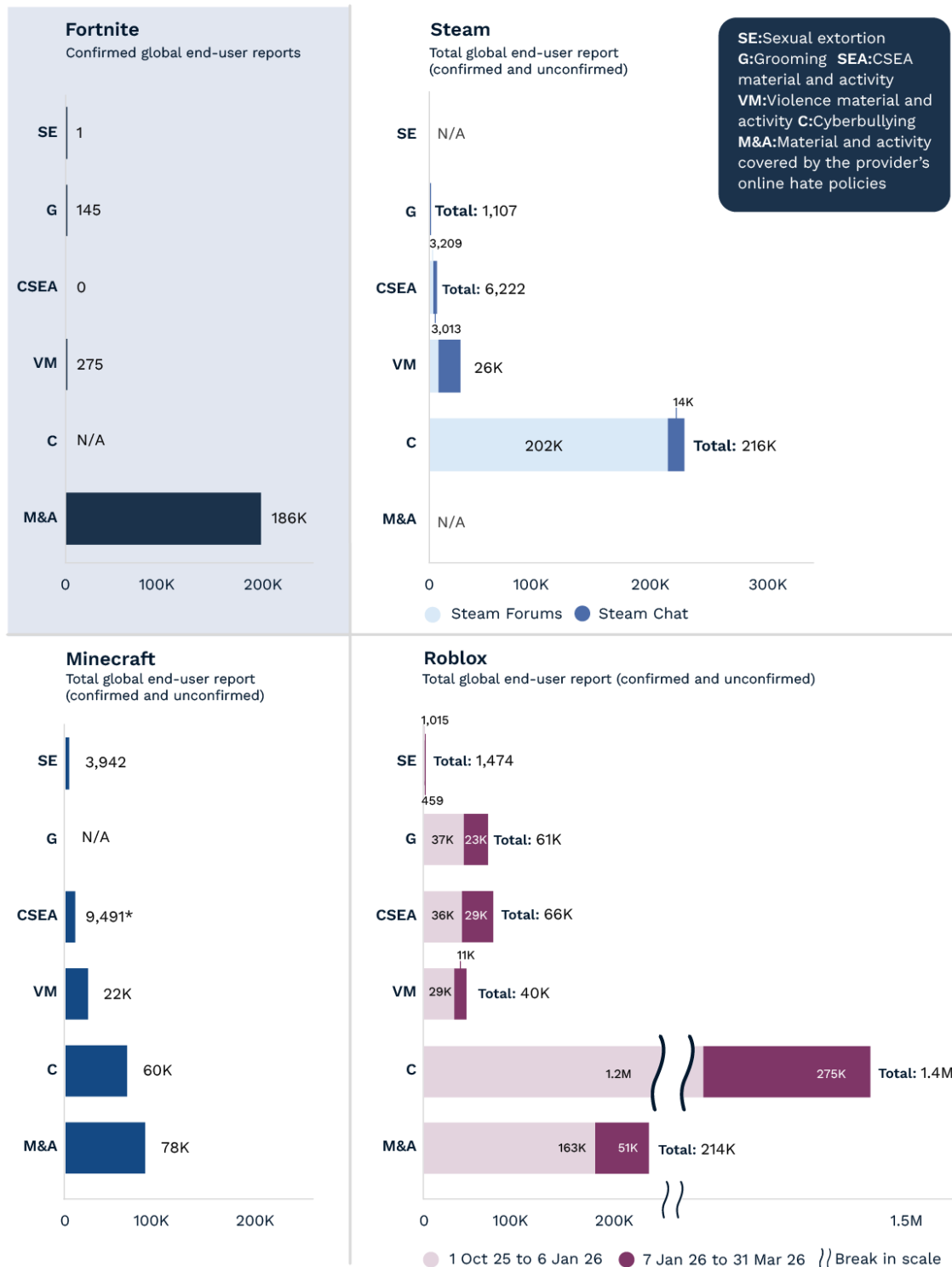
For Minecraft, Fortnite and Steam and Steam Chat, complaints relating to material or activity covered by their online hate policies received the highest number of user reports.

All services reported a significant proportion of the user reports could not be substantiated.

- Valve reported, in some cases on Steam and Steam Chat, ‘end users specifically used the most harmful descriptors of content in their reports, but when reviewed, that content is not found’.
- Mojang reported on Minecraft ‘player-selected harm categories are often a poor proxy for the true nature of a report. This can reflect bad-faith reporting (for example, to punish another player), selecting a more severe category to gain priority in moderation queues, uncertainty about how to label misconduct, or misclassification driven by frustration or misunderstanding of gameplay features.’

Epic Games only provided eSafety with the number of reports by end-users that it had confirmed as containing harmful material and activity. It stated it did not categorise end-user reports where moderators concluded that no further action was required.

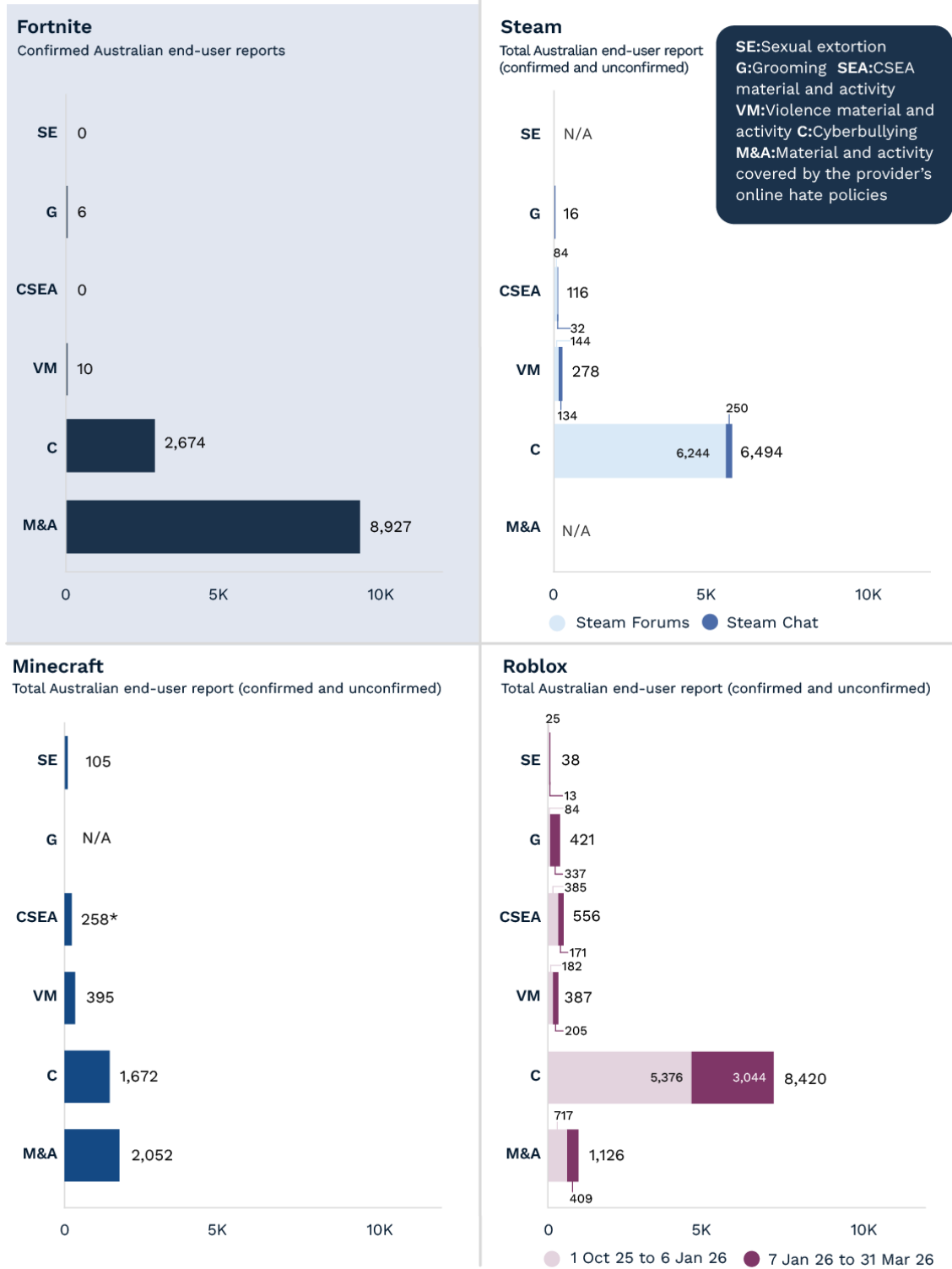
Figure 4 on the following page shows global end-user reports by type of unlawful and harmful material and activity.

**Figure 4: Global end-user reports by type of unlawful and harmful material and activity**

\* Includes reports on sexual grooming as well as reports on other CSEA material and activity.

**Note:** Number of end-user reports includes all reports made by end-users in these categories for all services except Fortnite. Epic Games stated it could only provide the number of end user reports that were for confirmed violations. For most service providers, reporting categories did not directly align with eSafety harms categories. Roblox's category for Sexual extortion includes all categories that relate to blackmail and extortion reports. For Steam and Steam Chat, where Valve could not confidently translate its reporting categories, it populated the aggregate number of reports for content violating its policies. For its Steam service, Valve provided user reporting figures broken down by harm type for its forums (Steam Discussions) but did not provide user reporting figures broken down by harm on user-generated content or reviews. For Steam Discussions, cyberbullying reports include reports for harassment and disruptive behaviour, CSEA material and activity reports include all reports on child safety and not specifically for CSEA material and activity and it did not have a category for grooming. Valve did not have a reporting category for sexual extortion across both Steam Chat and Steam Forums. Valve also did not provide details on the number of reports captured by Valve's policies on online hate. For Minecraft, reporting categories combine grooming and other CSEA material and activity together – this has been attributed to the Other CSEA material and activity category. Sexual Extortion is included in Mojang Studios' Sexually Inappropriate Behaviour category and therefore includes other harm types. For Fortnite, other CSEA is captured in Grooming and Sexual extortion. Epic Games stated it could not provide cyberbullying reports as it could not identify the nationality and age of an alleged victim in line with the Online Safety Act definition of cyberbullying.

Figure 5: Australian end user reports by type of harmful material and activity



\*Includes reports on grooming as well as reports on CSEA material and activity.

**Note:** Number of end-user reports includes all reports made by end-users in these categories for all services except Fortnite. Epic Games stated it could only provide the number of end user reports that were for confirmed violations. For most providers, reporting categories did not directly align with eSafety harms categories. Roblox’s category for Sexual extortion includes all categories that relate to blackmail and extortion reports. For Valve, where it could not confidently translate its reporting categories, it populated the aggregate number of reports for content violating its policies. For its Steam service, Valve provided user reporting figures broken down by harm type for its forums (Steam Discussions) but did not provide user reporting figures broken down by harm on user-generated content or reviews. For Steam Discussions, Cyberbullying reports include reports for harassment and disruptive behaviour, CSEA material and activity reports include all reports on child safety and not specifically for CSEA material and activity and it did not have a category for grooming. Valve did not have a reporting category for sexual extortion across both Steam Chat and Steam Forum. Valve also did not provide details on the number of reports captured by Valve’s policies on online hate. For Minecraft, reporting categories combine grooming and other CSEA material and activity together –this has been attributed to the Other CSEA material and activity category. Sexual Extortion is included in Mojang Studios’ Sexually Inappropriate Behaviour category and therefore includes other harm types. For Fortnite, other CSEA is captured in Grooming and Sexual extortion. Epic Games stated it could not provide cyberbullying reports as it could not identify the nationality and age of an alleged victim in line with the Online Safety Act definition of cyberbullying.

# Focus: Staff capacity and capability

## Why were service providers asked about this?

The number of Trust and Safety staff, along with the language skill set of moderators, can impact the ability of providers to address unlawful or harmful material and activity.

According to the 5Rights Foundation (a UK-based non-government organisation) '[Trust and Safety staff] plays a crucial role in defining and implementing safety standards'.<sup>84</sup>

Review by human moderators may be required to verify material flagged by tools or reported by users, particularly where the material has not been identified previously, or the unlawful or harmful activity is occurring in real time.

In the Notices, service providers were asked about their trust and safety teams, including their people and resources. These teams can include human content moderators, investigative and intelligence teams, policy staff, data analysts, teams implementing training for moderators and other staff, and engineers who develop and test tools used to detect unlawful and harmful material and activity.<sup>85</sup>

Service providers were also asked about the time taken to reach an outcome after unlawful or harmful material or activity was reported by a user or flagged by automated tools. This is an important indicator of how effectively service providers respond. Fast action is especially important for unlawful material and activity such as CSEA. In these cases, timely responses can help alert law enforcement agencies to urgent risks and reduce further harm to victim-survivors.

For this report, eSafety also focused on whether service providers used moderators who spoke the languages most commonly spoken in Australia: English, Mandarin, Arabic, Vietnamese, Cantonese and Punjabi.<sup>86</sup> Effective moderation depends on understanding the language and culture for context, so it's important that providers have human moderators operating in the languages of the communities that use their services. Some services may use translation services when reviewing material or activity involving languages that their human moderators do not cover. However, eSafety does not consider automated translation services to be a substitute for human moderation because relying on language translation tools risks losing important context and cultural or linguistic nuance.

---

<sup>84</sup> 5Rights Foundation (2025), [Advancing Trust & Safety: systems and standards for online safety professionals](#) (p 18), 5Rights website.

<sup>85</sup> 5Rights Foundation (2025), [Advancing Trust & Safety: systems and standards for online safety professionals](#), 5Rights website.

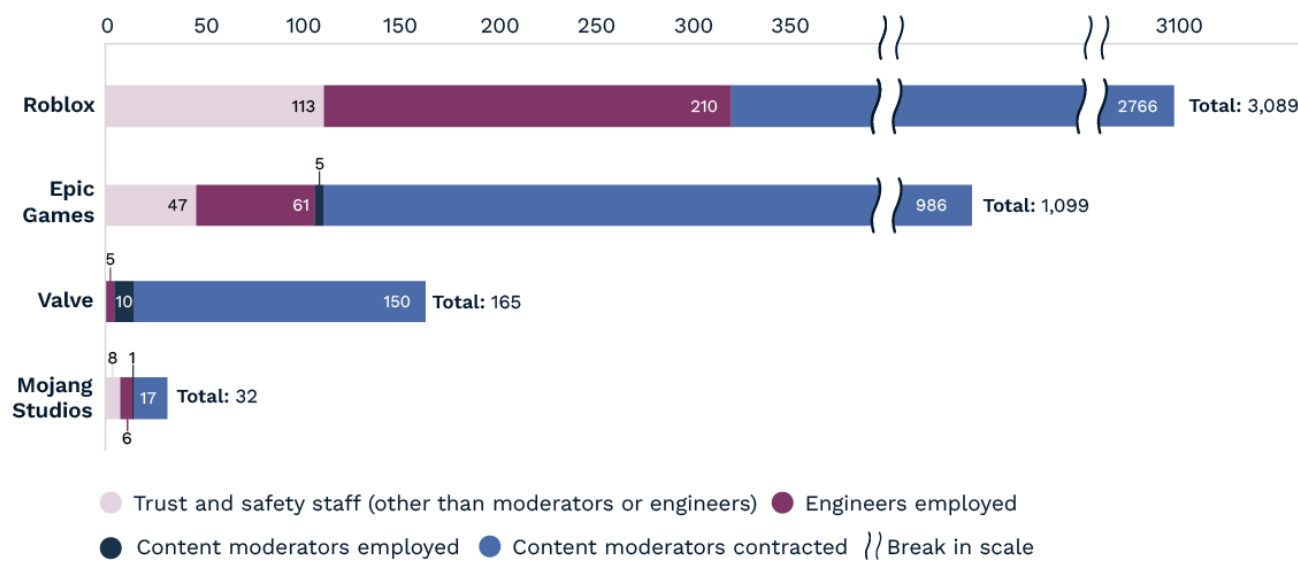
<sup>86</sup> Australian Bureau of Statistics (2022), [Cultural diversity: Census, 2021](#), ABS website.

# What we found

## Staff numbers

Roblox had the largest number of Australian monthly active users (between 3.14 and 3.36 million) and also employed the highest number of staff working across trust and safety roles (3,089) during the report period.

**Figure 6: Number of staff employed or contracted for trust and safety.**



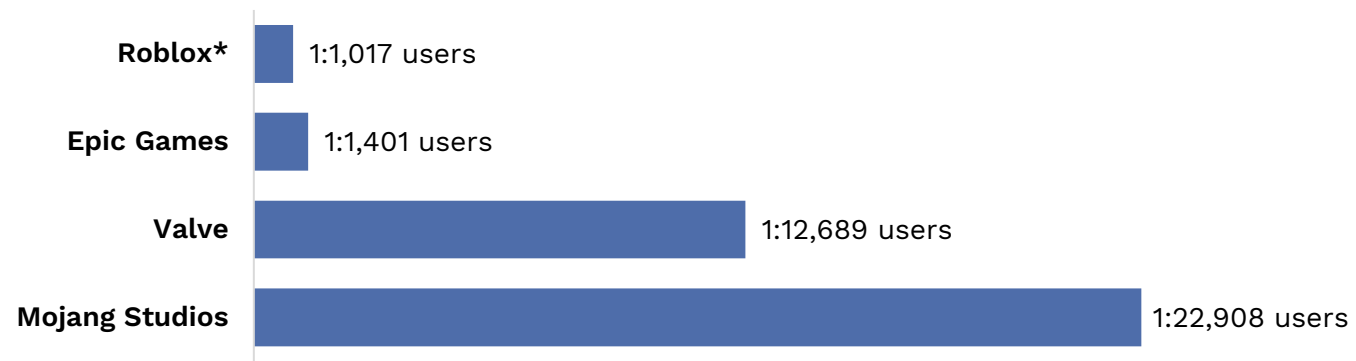
**Note:** In its response, Mojang Studios noted that it did not include resources employed or contracted by Microsoft and Xbox in its numbers.

Because Mojang Studios had only 32 staff covering trust and safety roles globally, they had a ratio of only one trust and safety staff member to approximately 22,908 Australian monthly active users. Mojang Studios stated it had access to and regularly relied on Xbox and Microsoft trust and safety teams, and the centralised Digital Safety team based on the types of issues presented. As part of eSafety’s latest periodic report on CSEA, Microsoft reported it had a total of 70 contractor moderators and 24 moderators employed directly by Microsoft during the period of 1 July to 31 December 2025. Questions on staffing in the CSEA report were not limited to CSEA specialists.

Trust and safety staff are expected and required to respond to online harms globally. A low ratio of trust and safety staff to users indicates sizeable risks of online harm, especially to children.

Valve stated it did not use titles within its organisation and therefore no employees were designated as trust and safety staff, and that many of Valve’s employees perform work related to content moderation functions.

**Figure 7: Ratio of trust and safety staff to the number of monthly active Australian users (1 T&S staff: number of MAUs)**



\* Ratio for Roblox is based off its monthly average user between 7 Jan 2026 and 31 March 2026

**Note:** Mojang Studios stated it has access to and relies on Xbox and Microsoft staff and resources where the issues require additional support. In its response, Mojang Studios noted that it did not include resources employed or contracted by Microsoft and Xbox in its numbers. As part of eSafety’s latest periodic report on CSEA, Microsoft reported it had a total of 70 contractor moderators and 24 moderators employed directly by Microsoft during the period of 1 July to 31 December 2025. Questions on staffing in the CSEA report were not limited to CSEA specialists.

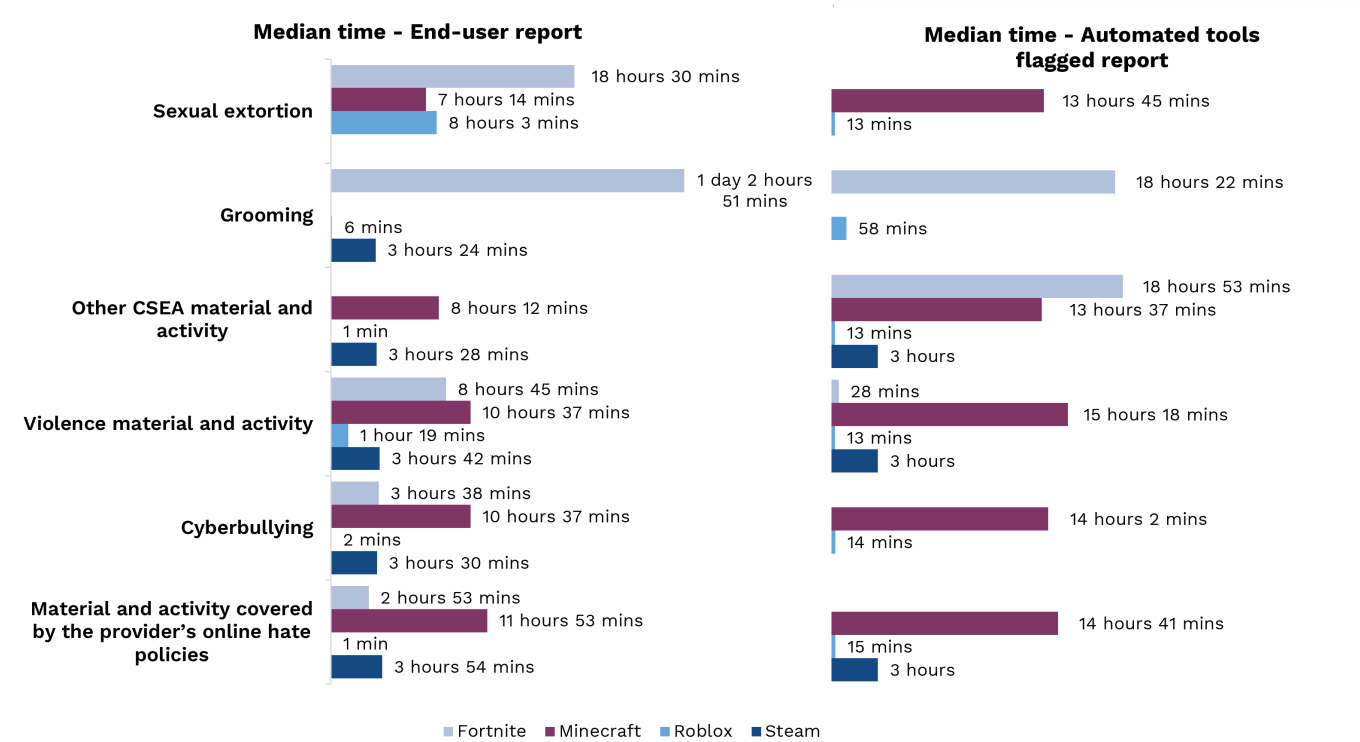
## Time taken to reach an outcome

The median time to review and action both user reports and issues detected by automated tools was significantly higher overall for Minecraft despite its lower overall report volume, highlighting the importance of maintaining a moderation system and workforce that is appropriately scaled to the size and activity of the user base.

Across all services, Epic Games took the longest time to reach an outcome on automated reports of other CSEA material and activity. Epic Games said that the reason for longer median times in reaching an outcome in response to all reports of grooming for the purposes of sexual exploitation and sexual extortion on Fortnite was that all reports of these high-harm categories were escalated to a specialist team of five FTE investigators and three contracted investigators globally. Epic Games stated that after they were reviewed none of these reports contained other CSEA material and activity. Epic Games stated that this content was not available to users while the automated reports were being reviewed.

For end-user reports, Epic Games stated its median time data only included reports where violations were confirmed, as it did not categorise reports where moderators concluded that no further action was required.

**Figure 8: Median time taken to action complaints by users and by automated tools where complaints subsequently reviewed by human moderators**



**Note:** For most providers, reporting categories did not directly align with eSafety harms categories. Roblox’s category for Sexual extortion includes all categories that relate to blackmail and extortion report. For Valve, where it could not confidently translate its reporting categories, it populated the aggregate number of reports for content violating its policies. Mojang Studios reported on Grooming and Other CSEA material and activity together and therefore is not split out in the graph. Epic Games stated reports of sexual extortion raised by its automated tool were categorised as ‘Grooming’. Epic Games also stated it ‘proactively rejects harmful content on [its] platform using tools that block and filter’ before material is published online and therefore it did not received reports on Cyberbullying and material and activity covered by its online hate policies during the report period. Epic Games’ stated it calculated median times for cases where an end-user report was confirmed by a human moderator to contain violative material or conduct. It stated the data excluded end-user reports where human moderators determined that no breach was identified. Epic Games stated it was not possible for it to calculate median time for all end-user reports, as non-violative end-user reports were not categorised. Epic Games stated by contrast the median times reported for automated tool flagged reports did include both violative and non-violative reports.

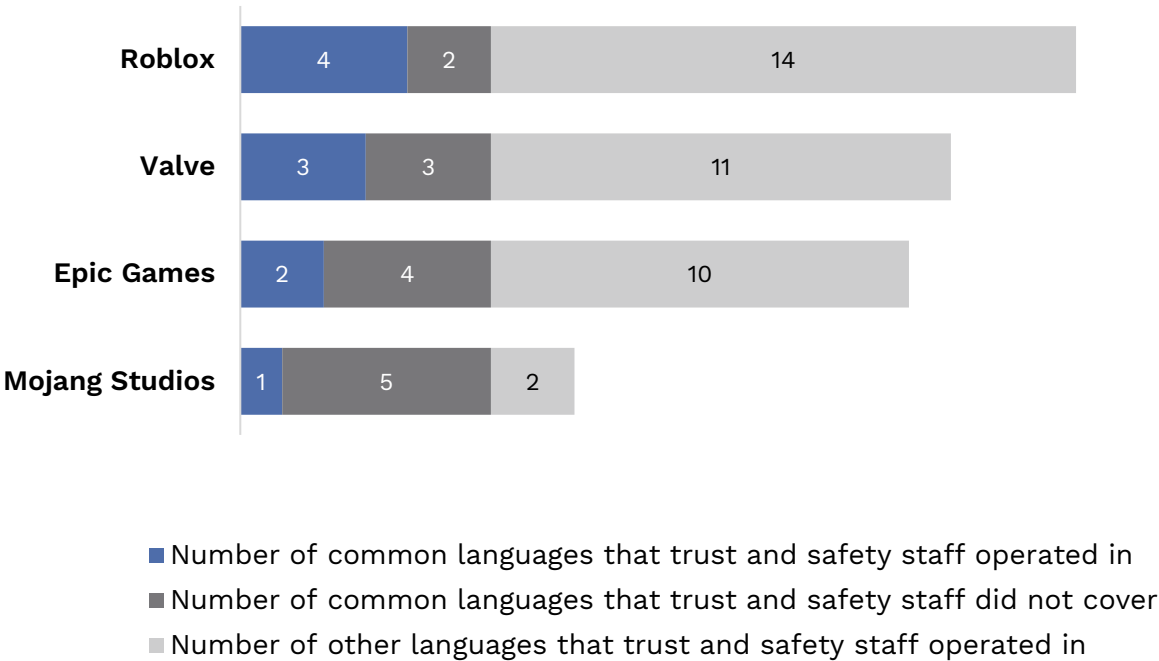
## Moderator languages

Human moderators often need to assess the style of language, context and intent behind online content or behaviour. eSafety focused on whether service providers used moderators who operated in the most common languages spoken in Australia: English, Mandarin, Arabic, Vietnamese, Cantonese and Punjabi.<sup>87</sup>

Mojang Studios did not operate in any language commonly spoken in Australia apart from English.

<sup>87</sup> Australian Bureau of Statistics (2022), [Cultural diversity: Census, 2021](#), ABS website.

Figure 9: Trust and safety staff that operate in the common languages spoken in Australia



**Note:** All providers reported that Trust and Safety staff utilise language translation software where needed. Unlike other providers, Epic Games interpreted operating in a language to include use of translation technology; accordingly, the graph has been amended to reflect how many languages Epic Games stated moderators were fluent in.

## Key insights and safety gaps

- Mojang Studios employed or contracted only 32 trust and safety staff globally.<sup>88</sup> This meant only one trust and safety staff member to every 22,908 monthly active users in Australia. Mojang Studios stated it had access to and regularly relied on Xbox and Microsoft trust and safety teams, and the centralised Digital Safety team based on the types of issues presented. Trust and safety staff are expected and required to respond to online harms globally.
- When Epic Games’ automated tools flagged other CSEA material and activity, it took a materially longer period of time for moderators to action those complaints compared to other providers, although in each case the flagged material was ultimately determined not to be violative.
- Mojang Studios stated its trust and safety staff operated in only two languages other than English, neither of which were among the most common spoken in Australia. Valve also stated none of the top five languages spoken in Australian homes (other than English) were covered by trust and safety staff.

<sup>88</sup> In its response, Mojang noted that they did not include resources employed or contracted by Microsoft and Xbox in their numbers. As part of eSafety’s latest periodic report on CSEA, Microsoft reported it had a total of 70 contractor moderators and 24 moderators employed directly by Microsoft during the period of 1 July to 31 December 2025. Questions on staffing in the CSEA report were not limited to CSEA specialists.

# Focus: Industry collaboration and information sharing

## Why were service providers asked about this?

Under the Expectations, service providers are expected to consult and cooperate with other service providers to promote the safe use of online services. Information and intelligence sharing relationships between service providers, government and non-government organisations also serve to reduce harms, through proactive identification of new risks, technologies, harms and solutions. eSafety's transparency reports reinforce these benefits by allowing service providers to identify any shortcomings in their own approach and to learn from other service providers' examples of best practice.

Information-sharing is particularly important where online game players may also use gaming-adjacent services to chat with friends, join communities, share content, or co-ordinate gameplay.<sup>89</sup> This means safety risks can easily move across platforms and services. Participating in timely information-sharing helps reduce the risks and harms of platform migration and off-platforming.<sup>90</sup>

Information sharing is also important between service providers and safety partners that are not necessarily specific to the gaming environment.

These are some notable information-sharing initiatives:

- **National Center for Missing & Exploited Children (NCMEC):** Shares emerging indicators, patterns and other relevant signals identified through its work with Electronic Service Providers to help inform proactive detection and disruption efforts. NCMEC also provides the Take It Down service, a free global service available in 38 different languages designed to help prevent the online sharing of nude, partially nude or sexually explicit images or videos taken when a person was under 18. Individuals can generate a unique digital fingerprint ('hash') of an image or video directly on their device without the image or video itself being uploaded to NCMEC. Participating online platforms can use those hashes to detect and limit the sharing of matching content on their services.<sup>91</sup>

---

<sup>89</sup> Englund, G., & White, J. (2023), [The Online Gaming Ecosystem: Assessing Digital Socialisation, Extremism Risks and Harms Mitigation Efforts](#), GNET website.

<sup>90</sup> WeProtect Global Alliance (2026), [Online gaming and risks to children](#), WeProtect Global Alliance website.

<sup>91</sup> NCMEC (2026), [Take It Down](#), NCMEC website.

- **Lantern:** An industry initiative run by the Tech Coalition that helps companies share indicators of CSEA and information about accounts that violate providers' CSEA policies. As of 2025, Lantern had reportedly facilitated the sharing of more than two million signals related to CSEA between companies, including signals related to sexual extortion. Lantern also launched a workstream for online game services and gaming-adjacent services to share intelligence related to grooming and off-platforming.<sup>92</sup>
- **Tech Against Terrorism:** Supports online platforms to identify, prevent and respond to terrorist and violent extremist activity online. Through its Knowledge Sharing Platform and wider suite of specialist tools and resources, it provides online service providers with actionable threat intelligence, datasets and practical guidance to strengthen the detection and mitigation of terrorist and violent extremist content and activity. Tech Against Terrorism also works directly with technology companies, providing tailored support and expertise to strengthen their capabilities and reduce the exploitation of their services by terrorist and violent extremist actors.<sup>93</sup>
- **The Global Internet Forum to Counter Terrorism (GIFCT):** Operates a hash-sharing database that gives its members access to hashes of images, videos, PDFs and URLs known to contain known terrorist and violent extremist content.<sup>94</sup>
- **Internet Watch Foundation (IWF):** Provides services including Hash Lists, Keyword Lists, and URL Lists containing known sexual abuse imagery to members to help block that material.<sup>95</sup>
- **Extremism + Gaming Research Network:** Facilitates information sharing through collaborative research activities, expert convenings, and partnerships that enable researchers, industry, and policymakers to exchange insights on extremist activity in gaming environments.
- **Christchurch Call Foundation:** Facilitates information sharing through coordinated crisis response mechanisms, multistakeholder working groups, and partnerships with initiatives like GIFCT, enabling governments and industry to collaboratively address terrorist and violent extremist content online.
- **Thriving in Games Group:** Allows members to confidentially share research and best practice on safety and inclusivity in gaming. Members include large 'AAA' game studios, smaller developers, and gaming-adjacent services such as Discord and Twitch.<sup>96</sup>

<sup>92</sup> Tech Coalition (2026), [2025 Transparency report](#), Tech Coalition website.

<sup>93</sup> Tech Against Terrorism (2026), [Knowledge Sharing Platform](#), Tech Against Terrorism website.

<sup>94</sup> GIFCT (n.d.), [GIFCT's Hash-Sharing Database](#), GIFCT website.

<sup>95</sup> IWF (n.d.), [Our services](#), IWF website.

<sup>96</sup> Thriving in Games Group (n.d.), [Join Us](#), Thriving in Games Group website.

eSafety encourages service providers to join information sharing initiatives (where they are not already involved) to further develop best practice for evaluating and improving trust and safety systems.

- Service providers should continue to explore new ways to proactively detect unlawful and harmful user generated content in games, such as exploring ways to hash games and game components and share these hashes. For example, Mojang Studios reports it has already developed in-house tools to hash user-generated skins.
- Service providers should explore ways to share signals to improve collaborative impact (including by working with the Tech Coalition's Project Lantern as the program expands its scope to cover threats with an Other CSEA nexus, including sadistic online exploitation).

## What we found

Roblox, Mojang Studios and Epic Games all reported participating in information-sharing initiatives in relation to CSEA and extremism.

Roblox and Mojang Studios reported they shared signals with and used flagged account lists from Technology Coalition's Lantern program.

Mojang Studios stated it shared results of newly identified extremist groups with Lantern, was a member of the Thriving in Games Group (formerly known as the Fair Play Alliance) and used its Disruption and Harms in Games Framework and Digital Thriving Playbook to help inform program strategy and understanding of safety risks. However, Mojang Studios did not provide any information on any information-sharing arrangements it had with service providers which host third-party mods for Minecraft, such as CurseForge.

Mojang Studios stated it participated in the GIFCT Incident Response framework. Roblox stated it joined GIFCT and is exploring participation in industry hash-sharing initiatives to address terrorist and violent extremist content.

Roblox stated it worked in partnership with Discord to 'share broader trends in key risk areas like child safety and violent extremism' and it engaged with collaborative research and development work on online safety with the Young and Resilient Centre at Western Sydney University, PROJECT ROCKIT, and the Prosocial Design Network.

Epic Games stated it received URLs from both Internet Watch Foundation and Tech Against Terrorism to populate its blocking tools for CSEA and violence material and activity.

## Key insights and safety gaps

Services that do not participate in signal-sharing initiatives may be at significant risk of failing to detect and remove unlawful and harmful material and activity.

- Valve stated it did not participate in information-sharing initiatives, noting it ‘follows legal process (for example, valid law enforcement requests) when required to share user data, but did not otherwise share user information with other third-parties as to do so would be a violation of its privacy policy’. eSafety notes some information sharing initiatives, including Project Lantern, are designed with explicit governance and privacy safeguards to support responsible signal sharing.<sup>97</sup>
- Despite being a member of the Tech Coalition, Epic Games stated it did not participate in Lantern and reported that it did not otherwise share signals with other service providers to prevent recidivism in relation to CSEA.
- Mojang Studios did not provide any information on any information-sharing arrangements it had with service providers which host third-party mods for Minecraft, such as CurseForge.

---

<sup>97</sup> Tech Coalition (2026), [Lantern](#), Tech Coalition website.

## After the report period

eSafety notes some service providers continued to develop their safeguards after this report period. These are some examples:

- Roblox rolled out a new approach to age-based accounts and parental controls in April 2026.<sup>98</sup>
- Mojang Studios has stated the Xbox Network continues to evaluate age assurance for voice and text communication and game invites, including for Minecraft (where players access Minecraft with their Microsoft Service Account as an Xbox gaming studio) following rollout in the United Kingdom in February 2026.<sup>99</sup>
- As of 9 September 2026, Valve started to require users in Australia to provide credit card verification to access R18+ games.
- Epic Games has stated it will launch automated voice reporting to detect high-harm interactions with child participants, such as child endangerment in voice chat, and automatically flag them for its review for all conversations with players under 18.
- Epic Games has stated it has implemented a new feature where if a frontline moderator escalates a report containing a suspected high-harm violation (including for grooming, sexual extortion and other CSEA material and activity), the allegedly violative account is immediately banned from using voice and text chat and sending new friend requests while the complaint is investigated. Epic Games has also stated the team responsible for investigating reports of potential high harm violations has grown to twelve investigators.
- Epic Games has stated that it is looking into requiring players to verify that they are at least 16 years old to access chat features, unless they have already received verified parental consent.

---

<sup>98</sup> Roblox (2026), [Roblox Introduces New Age-Based Accounts and Expanded Parental Controls for Users Under 16](#), Roblox website.

<sup>99</sup> Xbox Support (2025), [UK age verification on Xbox](#), Xbox website.

# Conclusion

This report sets out what four major online game and gaming-adjacent service providers told eSafety about the steps they are taking to help keep Australian children safe and to comply with the Expectations. It helps show where service providers are taking action, where they may need to improve and where more questions may need to be asked.

eSafety's regulatory guidance, first published in July 2022 and updated in September 2023, sets out eSafety's planned approach to the exercise of its powers in respect of the Expectations.

eSafety will use the information to build greater understanding of industry practices. eSafety will continue to check how these service providers are meeting the Expectations and addressing safety risks for children in online gaming environments. We will also continue to raise specific gaps and vulnerabilities with service providers that received Notices, to understand more about why certain aspects of the Expectations may not currently be complied with and any future steps they plan to take to ensure their services implement the Expectations.

Where findings are relevant to compliance with Australia's Online Safety Codes and Standards, eSafety may engage relevant services further and can commence investigations into compliance.

eSafety hopes the information will be used by all online service providers to address key challenges and incentivise greater transparency in the future – including through their own voluntary disclosures. eSafety also intends the information in this report to be used by researchers, academics, the media and the public to scrutinise the efforts of industry, to improve accountability and to encourage implementation of the Expectations.



[eSafety.gov.au](https://www.esafety.gov.au)