# Aotearoa New Zealand Code of Practice for Online Safety and Harms | TikTok First Report

| | |
|---|---|
| **Signatory:** | TikTok New Zealand Limited |

| | |
|---|---|
| **Relevant Products / Services:** | TikTok |

## 4.1 Reduce the prevalence of harmful content online

Signatories commit to implementing policies, processes, products and/or programs that would promote safety and mitigate risks that may arise from the propagation of harmful content online, as it relates to the themes identified in section 1.4.

**Outcome 1.** Provide safeguards to reduce the risk of harm arising from online **child sexual exploitation & abuse (CSEA)**

Measure 1. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent known child sexual abuse material from being made available to users or accessible on their platforms and services. [Opted in]

Measure 2. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent search results from surfacing child sexual abuse material. [Opted in]

Measure 3. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to adopt enhanced safety measures to protect children online from peers or adults seeking to engage in harmful sexual activity with children (e.g. online grooming and predatory behaviour). [Opted in]

Measure 4. Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to reduce new and ongoing opportunities for the sexual abuse or exploitation of children. [Opted in]

Measure 5. Work to collaborate across industry and with other relevant stakeholders to respond to evolving threats. [Opted in]

*TikTok response:*

- Our platform is designed with the safety of minors front of mind, and we have a range of policies, processes and enhanced safety measures in place to protect the safety of minors on TikTok. These include measures to detect, prevent and report the sexual exploitation of minors and grooming behaviour, policies prohibiting content containing nudity and sexual activity involving minors, and minimum age requirements to use TikTok, as stipulated in our Terms of Service.

- Our Community Guidelines prohibit activities that enable or perpetuate the abuse, harm, endangerment, or exploitation of minors on TikTok.
- TikTok will take action on any content or accounts involving child sexual abuse material (CSAM) or sexual exploitation of a minor. Any content, including animation or digitally created or manipulated media, that depicts abuse, exploitation, or endangerment of minors is a violation of our Community Guidelines and will be removed when detected. We report CSAM and supporting evidence to the National Center for Missing & Exploited Children (**NCMEC**) and to any additional relevant legal authorities.
- We use a combination of AI and human moderation to detect, remove and, where appropriate, report any content that depicts, promotes, normalises, or glorifies grooming behaviours, as well as content that solicits real-world contact between a minor and an adult or between minors with a significant age difference.
- Users can also report all content (videos, comments, direct messages, hashes and sounds) if they believe it violates our Community Guidelines. Users can report content in-app and through our website by choosing a reason why they think the content might violate our Community Guidelines.
- We do not allow users who have been convicted of crimes against children to have an account on our platform. These crimes include: sexual assault, molestation, murder, physical abuse or neglect, abduction, international parental kidnapping, trafficking, exploitation of minors for prostitution, live online sexual abuse of a minor, sexual exploitation of minors in the context of travel and tourism, attempts to obtain or distribute CSAM, and the production, possession, or distribution of CSAM.
- Account holders who are under the age of 16 cannot use direct messaging or host a livestream (from November 23, users will need to be at least 18 years old to host a livestream) and their content is not eligible to appear in the For You feed (the age thresholds are higher in some regions). Account holders who are under the age of 18 cannot send or receive gifts via our virtual gifting features.
- For users of all ages, images and external videos cannot be shared via Direct Messaging.
- Our Family Pairing feature also allows give parents and guardians more control over their younger ones experience on TikTok, allowing them to adjust privacy and experience settings and implement features like our Screen Time Management tool and Private Mode.
- We partnered with Netsafe and Professor Amanda Third to host a webinar to provide an educational session for parents in NZ - this included a strong focus on protecting children online, including education about our family pairing feature, age restriction information and community controls. Amanda is an expert in user-centred, participatory research, and her work investigates children's and young people's technology practices.

**Outcome 2:** Provide safeguards to reduce the risk of harm arising from online **bullying or harassment**

Measure 6. Implement, enforce and/or maintain policies and processes that seek to reduce the risk to individuals (both minors and adults) or groups from being the target of online bullying or harassment. [Opted in]

Measure 7. Implement and maintain products and/or tools that seek to mitigate the risk of individuals or groups from being the target of online bullying or harassment. [Opted in]

Measure 8. Implement, maintain and raise awareness of product or service related policies and tools for users to report online bullying or harassment content. [Opted in]

Measure 9. Support or maintain programs, initiatives or features that seek to educate and raise awareness on how to reduce or stop online bullying or harassment. [Opted in]

*TikTok response:*

- TikTok does not tolerate, and has implemented robust content policies and safeguards against, bullying, shaming and harassment (including sexual harassment).
- Our Community Guidelines prohibit content and behaviour that expresses abuse, including threats or degrading statements intended to mock, humiliate, embarrass, intimidate, or hurt an individual. This includes content that threatens to hack, 'dox' or blackmail another individual, as well as content that glorifies, normalises or promotes sexual harassment, regardless of the user's intent. These prohibitions extend to the use of all TikTok features and content.
- To enable good faith expression about matters of public interest, critical comments of public figures may be allowed; however, serious abusive behaviour against public figures is prohibited.
- TikTok proactively detects emerging cyberbullying trends through media monitoring, content analysis, and moderator feedback. We implement preventative and mitigative strategies to manage risk and curb harmful trends on the platform.
- Users can also report all content (videos, comments, direct messages, hashes and sounds) if they believe it violates our Community Guidelines. Users can report content in-app and through our website by choosing a reason why they think the content might violate our Community Guidelines.
- We have comment controls and filters that enable users to restrict who can comment on their content, bulk delete comments and automatically block specific keywords or “filter all” comments. In addition, our **Rethink** feature - provides automated prompts that encourage people to consider the impact of their words before posting a potentially unkind or violative comment.
- In terms of age appropriate settings - the 'allow comments on videos' setting for younger teens (13-15) is set to 'Friends' by default and the comment filter for spam and

offensive comments is always switched on for younger teens (13-15), while older teens (16-17) will have this setting on by default.

- TikTok's Youth Portal and Safety Centre also contains information and resources to help users identify, address and report bullying and harassment.
- In June 2021, TikTok engaged leading anti bullying NGO, Project Rockit, to create a series of bullying awareness video content about safety tools and how to identify and stand up to online hate and bullying. The 8 videos created helped users identify online hate and understand and confidently use TikTok's reporting and safety features to help protect them on the platform.
- TikTok is currently working with NGO Bully Zero to design a workshop for a TikTok Creator Safety Summit in November 2022. The summit will teach top creators, including from New Zealand, about online harassment and bullying prevention. The creators will then make videos highlighting these key messages to their followers on the app.

**Outcome 3:** Provide safeguards to reduce the risk of harm arising from online **hate speech**

Measure 10. Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of hate speech. [Opted in]

Measure 11. Implement and maintain products and tools that seek to prohibit or reduce the prevalence of hate speech. [Opted in]

Measure 12. Implement, maintain and raise awareness of product or service related policies and tools for users to report potential hate speech. [Opted in]

Measure 13. Support or maintain programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of online hate speech. [Opted in]

Measure 14. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online hate speech. [Opted in]

*TikTok response:*

- TikTok is a diverse and inclusive community that has no tolerance for discrimination. Our Community Guidelines do not permit content that contains hate speech or involves hateful behaviour, or which praises, promotes, glorifies or supports any hateful ideology (e.g., white supremacy, misogyny, anti-LGBTQ, antisemitism). We ban accounts and/or users that engage in severe or multiple hate speech violations or that are associated with hate speech off the TikTok platform.

- We collaborate with government, industry partners and other relevant stakeholders across New Zealand to proactively prevent and respond to online hate speech. In 2022 this work included new outreach initiatives with Oranga Tamariki, proactive moderation efforts in the lead up to Matariki, as well as ongoing engagement with New Zealand Police.
- We have comment controls and filters that enable users to restrict who can comment on their content, bulk delete comments and automatically block specific keywords or “filter all” comments. In addition, our Rethink feature - provides automated prompts that encourage people to consider the impact of their words before posting a potentially unkind or violative comment.
- Users can also report all content (videos, comments, direct messages, hashes and sounds) if they believe it violates our Community Guidelines. Users can report content in-app and through our website by choosing a reason why they think the content might violate our Community Guidelines.
- TikTok is currently working with Multicultural NSW (who created the "remove hate from the debate" web tool) to design a workshop for a TikTok Creator Safety Summit in November 2022. The summit will train top creators, including from New Zealand, in online the prevention of Hateful Behaviour and Hate Speech. The creators will then make videos highlighting these key messages to their followers on the app.

**Outcome 4:** Provide safeguards to reduce the risk of harm arising from online **incitement of violence**

Measure 15. Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of content that potentially incites violence. [Opted in]

Measure 16. Implement and maintain products and tools that seek to prohibit or reduce the prevalence of content that potentially incites violence. [Opted in]

Measure 17. Implement, maintain and raise awareness of product or service related policies and tools for users to report content that potentially incites violence. [Opted in]

Measure 18. Support or maintain programs and initiatives that seek to educate users on how to reduce or stop the spread of online content that incites violence. [Opted in]

Measure 19. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online content that incites violence. [Opted in]

*TikTok response:*

- TikTok maintains a zero-tolerance policy on violent extremism. Our Community Guidelines clearly outline that we do not allow people to use our platform to threaten or incite violence, or to promote violent extremist organisations, individuals, or acts. When there is a threat to public safety or an account is used to promote or glorify off-platform

violence, we ban the account. When warranted, we will report threats to relevant legal authorities.

- To effectively protect our community, we may consider off-platform behaviour to identify violent extremist organisations and individuals on our platform. We do not allow organisations or individuals on our platform who promote or engage in violence, including terrorist organisations, organised hate groups, criminal organisations, and other non-state armed groups that target civilians. If we find such organisations or individuals, we will ban their accounts.
- TikTok takes an uncompromising stance against enabling violent extremism on or off our platform. To further strengthen our commitment to user safety and human rights, we have partnered with Tech Against Terrorism, which brings together technology companies, civil society, and academics over the shared goal of countering violent extremism.
- In addition, our trust and safety teams partner with local experts and civil society organisations to understand the unique cultures and experiences of communities affected by violent extremism. We take into account publicly available information from experts, including the United Nations Security Council and Southern Poverty Law Centre, to designate dangerous or hateful individuals and organisations.
- Users can also report all content (videos, comments, direct messages, hashes and sounds) if they believe it violates our Community Guidelines. Users can report content in-app and through the web-app by choosing a reason why they think the content might violate our Community Guidelines.

**Outcome 5:** Provide safeguards to reduce the risk of harm arising from online **violent or graphic content**

Measure 20. Implement, enforce and/or maintain policies and processes that seek to prohibit and/or reduce the spread of violent or graphic content online. [Opted in]

Measure 21. Implement and maintain products and tools that seek to and/or reduce the spread of violent or graphic content. [Opted in]

Measure 22. Implement, maintain and raise awareness of product or service related policies and tools for users to report potential violent and graphic content. [Opted in]

*TikTok response:*

- Our Community Guidelines do not allow content that is gratuitously shocking, graphic, sadistic or gruesome, or content that promotes, normalizes, or glorifies extreme violence or suffering on our platform. TikTok uses a combination of AI and human moderation to proactively identify and prevent the spread of such content on our platform. When it is a threat to public safety, we ban the account and, when warranted, we will report it to relevant legal authorities.

- At the same time, we recognise that some content that would normally be removed per our Community Guidelines may be in the public interest. Therefore, we may allow exceptions under certain limited circumstances, such as educational, documentary, scientific, artistic, or satirical content, content in fictional or professional settings, counterspeech, or content that otherwise enables individual expression on topics of social importance. To minimise the potentially negative impact of graphic content, we may first include safety measures such as an “opt-in” screen or warning.
- Users can report all content (videos, comments, direct messages, hashes and sounds) if they believe it violates our Community Guidelines. Users can report content in-app and through the web-app by choosing a reason why they think the content might violate our Community Guidelines.

**Outcome 6:** Provide safeguards to reduce the risk of harm arising from online **misinformation**

Measure 23. Implement, enforce and/or maintain policies, processes and/or products that seek to reduce the spread of online misinformation. [Opted in]

Measure 24. Implement, enforce and/or maintain policies and processes that seek to penalise users who repeatedly post or share misinformation that violates related policies. [Opted in]

Measure 25. Support or maintain media literacy programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of misinformation. [Opted in]

Measure 26. Support or maintain programs and/or initiatives that seek to support civil society, fact-checking bodies and/or other relevant organisations working to combat misinformation. [Opted in]

Measure 27. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from misinformation. [Opted in]

*TikTok response:*

- We do not allow users to post content containing misinformation that causes significant harm to individuals, our community, or the larger public regardless of intent. This includes inaccurate or false content that may cause serious physical injury, illness, or death; severe psychological trauma; large-scale property damage, and the undermining of public trust in civic institutions and processes such as governments, elections, and scientific bodies.
- TikTok's Community Guidelines also prohibit the following:
    - Content containing misinformation that incites hate or prejudice

    - Misinformation related to emergencies that induces panic
    - Medical misinformation that can cause harm to an individual's physical health
    - Conspiratorial content including content that attacks a specific person or a protected group, includes a violent call to action, or denies a violent or tragic event occurred
    - Digital Forgeries (Synthetic Media or Manipulated Media) that mislead users by distorting the truth of events and cause significant harm to the subject of the video, other persons, or society
- In addition to removing content that is inaccurate and harms our users or community, we also remove accounts that seek to mislead people or use TikTok to deceptively sway public opinion. These activities range from inauthentic or fake account creation, to more sophisticated efforts to undermine public trust. These actors never stop evolving their tactics, and we continually seek to strengthen our policies as we detect new types of content and behaviours.
- We empower users to make informed decisions about the content and information they consume on TikTok, providing publicly available account information (no. of accounts following, followers, and likes) on TikTok profiles and supporting blue-tick verification for public figures and organisations, to help users identify authentic sources of information.
- While TikTok uses a combination of technology and thousands of safety professionals to enforce our Community Guidelines, we also rely on specialized misinformation moderators who have enhanced training, expertise, and tools to take action on misinformation.
- Content which may be potentially misleading is flagged by our moderators and may be removed from TikTok's "For You page" (FYP) while the information is sent to third party fact-checkers for review. Subsequently, content which is deemed false and misleading will be removed in accordance to our Community Guidelines.
- TikTok partners with third party fact checking organisations Agence France Presse and Australian Associated Press to detect and mitigate the spread of misinformation and disinformation in Oceania markets; leveraging the insights of IFCN-certified fact-checking experts.
- TikTok engaged the Australian Associated Press to create and deliver targeted digital media literacy education to a small group of content creators, including from New Zealand. The aim was to empower the participants to confidently source reliable and factual information, understand how to interrogate that information effectively, and encourage them to use those skills to create factual, reliable and educative content. The creators selected have large followings, with the project aiming to have a wide reach and impact in educating audiences on how to identify mis- and disinformation.
- TikTok also has a range of online resources, in-app PSA's and hubs developed in partnership with reputable third-parties focusing on a range of important topics

including COVID-19, elections, mental health, and first nations issues that provide users access to reliable information.

- Users can also report all content (videos, comments, direct messages, hashes and sounds) if they believe it violates our Community Guidelines. Users can report content in-app and through our website by choosing a reason why they think the content might violate our Community Guidelines.

**Outcome 7:** Provide safeguards to reduce the risk of harm arising from online **disinformation**

Measure 28. Implement, enforce and/or maintain policies, processes and/or products that seek to suspend, remove, disable, or penalise the use of fake accounts that are misleading, deceptive and/or may cause harm.[Opted in]

Measure 29. Implement, enforce and/or maintain policies, processes and/or products that seek to remove accounts,  (including profiles, pages, handles, channels, etc) that repeatedly spread disinformation. [Opted in]

Measure 30. Implement, enforce and/or maintain policies, processes and/or products that seek to provide information on public accounts (including profiles, pages, handles, channels, etc) that empower users to make informed decisions (e.g. date a public profile was created, date of changes to primary account information, number of followers). [Opted in]

Measure 31. Implement, enforce and/or maintain policies, processes and/or products that seek to provide transparency on paid political content (e.g. advertising or sponsored content) and give users more context and information (e.g. paid political or electoral ad labels or who paid for the ad).

Measure 32. Implement, enforce and/or maintain policies, processes and/or products that seek to disrupt advertising and/or reduce economic incentives for users who profit from disinformation. [Opted in]

Measure 33. Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from disinformation. [Opted in]

*TikTok response:*

- We do not allow content or activities that facilitate the spread of disinformation or which may otherwise undermine the integrity of our platform or the authenticity of our users. In addition to our Community Guidelines prohibiting harmful misinformation, we also remove content or accounts that involve spam, fake engagement, impersonation, and coordinated inauthentic behaviour, such as the use of multiple accounts to exert influence and sway public opinion while misleading individuals, our community, or our systems about the account's identity, location, relationships, popularity, or purpose.

- Beyond content that is inaccurate and harms our users or community, we also remove accounts that seek to mislead people or use TikTok to deceptively sway public opinion. These activities range from inauthentic or fake account creation, to more sophisticated efforts to undermine public trust. These actors never stop evolving their tactics, and we continually seek to strengthen our policies as we detect new types of content and behaviours.
- Our strict advertising policies exceed industry standards and all ads must undergo a review process which involves vetting the products/services promoted, ad caption, text, images, audio, visuals, age/region targeting, and landing pages. We do not allow promotion, sale, or solicitation of or facilitation of access to products or services that might be or are considered deceptive, misleading, or unlawful such as: unwarranted claims, misinformation, including pricing/discount or promotion information inconsistency, missing T&C or privacy policy pages, or any of such. We also require 18+ targeting for certain categories of products and service and proper disclaimers must be included when applicable.
- While TikTok uses a combination of technology and thousands of safety professionals to enforce our Community Guidelines, we also rely on specialised misinformation moderators who have enhanced training, expertise, and tools to take action on misinformation.
- Content which may be potentially misleading is flagged by our moderators and may be removed from TikTok's "for-you-page" (FYP) while the information is sent to third party fact-checkers for review. Subsequently, content which is deemed false and misleading will be removed in accordance to our community guidelines.
- We empower users to make informed decisions about the content and information they consume on TikTok, providing publicly available account information (no. of accounts following, followers, and likes) on TikTok profiles and supporting blue-tick verification for public figures and organisations, to help users identify authentic sources of information.
- Users can also report all content (**videos, comments, direct messages, hashes and sounds**) if they believe it violates our Community Guidelines. Users can report content in-app and through our website by choosing a reason why they think the content might violate our Community Guidelines.
- TikTok partners with third party fact checking organisations Agence France Presse and Australian Associated Press to detect and mitigate the spread of misinformation and disinformation in Oceania markets; leveraging the insights of IFCN-certified fact-checking experts.
- TikTok engaged the Australian Associated Press to create and deliver targeted digital media literacy education to a small group of content creators, including from New Zealand. The aim was to empower the participants to confidently source reliable and factual information, understand how to interrogate that information effectively, and encourage them to use those skills to create factual, reliable and educative content.

The creators selected have large followings, with the project aiming to have a wide reach and impact in educating audiences on how to identify mis and disinformation.

- TikTok also has a range of online resources, in-app PSA's and hubs developed in partnership with reputable third-parties focussing on a range of important topics including COVID-19, elections, mental health, and first nations issues that provide users access to reliable information.

Regarding **Measure 31:**

- Paid political advertising, including sponsored political content, is not allowed on TikTok. For more information regarding TikTok's advertising policies, please see here.

## 4.2 Empower users to have more control and make informed choices

Signatories recognise that users have different needs, tolerances, and sensitivities that inform their experiences and interactions online. Content or behavior that may be appropriate for some will not be appropriate for others, and a single baseline may not adequately satisfy or protect all users. Signatories will therefore empower users to have control and to make informed choices over the content they see and/or their experiences and interactions online. Signatories will also provide tools, programs, resources and/or services that will help users stay safe online.

**Outcome 8.** Users are empowered to **make informed decisions** about the content they see on the platform

Measure 34. Implement, enforce and/or maintain policies, processes, products and/or programs that helps users make more informed decisions on the content they see. [Opted in]

Measure 35. Implement, enforce and/or maintain policies, processes, products and/or programs that seek to promote accurate and credible information about highly significant issues of societal importance and of relevance to the digital platform's user community (e.g. public health, climate change, elections). [Opted in]

Measure 36. Launch programs and/or initiatives that educate or raise awareness on disinformation, misinformation and other harms, such as via media/digital literacy campaigns. [Opted in]

*TikTok response:*

- We present users with a stream of videos on our 'For You' feed curated to their interests, making it easy to find content and creators they love. It is powered by a recommendation system that delivers content to each user that is likely to be of interest to that particular user, but also works to intersperse recommendations that might fall outside people's expressed preferences, offering an opportunity to discover new categories of content.
- Our recommendation system on TikTok is also designed with safety as a key consideration. Reviewed content found to depict things like graphic medical procedures or legal consumption of regulated goods, for example – which may be shocking if surfaced as a recommended video to a general audience that hasn't opted in to such content – may not be eligible for recommendation.
- The feed can also be curated by users. If a video is not quite to a user's taste, we empower our users to long-press on a video and tap "Not Interested" to indicate that they don't care for a particular video. We are also working on a feature that would let people choose words or hashtags associated with content they don't want to see in their For You feed, to offer another way to help people customise their feed – e.g. a vegetarian who wants to see fewer meat recipes.
- We empower users to make informed decisions about the content and information they consume on TikTok, providing publicly available account information (no. of accounts following, followers, and likes) on TikTok profiles and supporting blue-tick verification for public figures and organisations, to help users identify authentic sources of information.
- Users can also report all content (videos, comments, direct messages, hashes and sounds) if they believe it violates our Community Guidelines. Users can report content in-app and through our website by choosing a reason why they think the content might violate our Community Guidelines. All these actions contribute to future recommendations in the For You feed.
- TikTok also has a range of online resources, in-app PSA's and hubs developed in partnership with reputable third-parties focussing on a range of important topics including COVID-19, elections, mental health, and first nations issues that provide users access to reliable information.
- TikTok partners with third party fact checking organisations Agence France Presse and Australian Associated Press to detect and mitigate the spread of misinformation and disinformation in Oceania markets; leveraging the insights of IFCN-certified fact-checking experts.
- TikTok engaged the Australian Associated Press to create and deliver targeted digital media literacy education to a small group of content creators, including from New Zealand. The aim was to empower the participants to confidently source reliable and factual information, understand how to interrogate that information effectively, and encourage them to use those skills to create factual, reliable and educative content. The creators selected have large followings, with

the project aiming to have a wide reach and impact in educating audiences on how to identify mis- and disinformation.

**Outcome 9.** Users are **empowered with control** over the content they see and/or their experiences and interactions online

Measure 37. Implement, enforce and/or maintain policies, processes, products and/or programs that seek to provide users with appropriate control over the content they see, the character of their feed and/or their community online. [Opted in]

Measure 38. Launch and maintain products that provide users with controls over the appropriateness of the ads they see.

*TikTok response:*

- We present users with a stream of videos on our 'For You' feed curated to their interests, making it easy to find content and creators they love. It is powered by a recommendation system that delivers content to each user that is likely to be of interest to that particular user, but also works to intersperse recommendations that might fall outside people's expressed preferences, offering an opportunity to discover new categories of content.
- The feed can be curated by users. If a video is not quite to a user's taste, we empower our users to long-press on a video and tap "Not Interested" to indicate that they don't care for a particular video. We are also working on a feature that would let people choose words or hashtags associated with content they don't want to see in their For You feed, to offer another way to help people customise their feed – e.g. a vegetarian who wants to see fewer meat recipes.
- Users can also report all content (videos, comments, direct messages, hashes and sounds) if they believe it violates our Community Guidelines. Users can report content in-app and through our website by choosing a reason why they think the content might violate our Community Guidelines. All these actions contribute to future recommendations in the For You feed.
- We have comment controls and filters that enable users to restrict who can comment on their content, bulk delete comments and automatically block specific keywords or “filter all” comments. In addition, our Rethink feature - provides automated prompts that encourage people to consider the impact of their words before posting a potentially unkind or violative comment.
- We also empower TikTok users with other various privacy and safety features to control their interactions on the app by:
    - Blocking the accounts they don’t want to interact with

    - Setting their account private, where only approved users can follow them and watch their content
    - Choosing if they wish to receive Direct Messages (DMs) or not
    - Choosing who can comment on their content or even turning off comments on their content altogether
    - Filtering comments containing any keywords they don't approve of
    - Hiding all the comments on their content until they review and approve them
    - Choosing who can tag them in their content or even prohibiting anyone from tagging them in their content
    - Choosing who can mention them in their content or even prohibiting anyone from mentioning them in their content
    - Choosing if anyone can duet with their videos or not
    - Choosing who can see their following list i.e. people they follow
- Our Family Pairing feature also allows give parents and guardians more control over their younger ones experience on TikTok, allowing them to adjust privacy and experience settings and implement features like our Screen Time Management tool and Private Mode.

Regarding **Measure 38**:

- TikTok enforces strict advertising policies that go beyond industry standards to protect all users on our platform.
- All ads must undergo a review process which involves vetting the products/services promoted, ad caption, text, images, audio, visuals, age/region targeting, and landing pages.
- We have an extensive list of prohibited products and services that cannot be advertised on the platform including: gambling, tobacco, alcohol, drugs, adult services, weight loss management/supplements, and political advertising.
- We take a special level of care and caution when it comes to advertising that our minors see - and craft our policies to ensure that any ads that could be shown to younger audiences are safe for those viewers.
- The broadest category of protections for minors in TikTok ads starts with the company's outright ban of any advertising that appeals directly to children, either by influencing children directly/indirectly or appealing to children to get their parents to buy a product. Ads for children's toys and clothing are allowed provided they are targeted and appeal to adults.
- Beyond this broad protection, we moderate ads for any products or services that could potentially pose a higher physical, emotional or financial risk to minors and

restrict them to an 18+ audience. Examples include dating apps, restricting financial services and health products.

- Additionally, we understand that our platform's products encourage mimicry, so we've taken minor safety into consideration and limit the behaviours that can be shown in ads well beyond what is legally required (banning ads showing unsafe driving, dangerous stunts, etc).
- We regularly review our advertising policies, and our Community Guidelines, to ensure they are keeping up with the development of new products and services and cultural trends, and keep our users safe on TikTok.

## 4.3 Enhance transparency of policies, processes and systems

Transparency helps build trust and facilitates accountability. Signatories will provide transparency of their policies, processes and systems for online safety and content moderation and their effectiveness to mitigate risks to users. Signatories, however, recognise that there is a need to balance public transparency of measures taken under the Code with risks that may outweigh the benefit of transparency, such as protecting people's privacy, protecting trade secrets and not providing threat actors with information that may expose how they may circumvent or bypass enforcement protocols or systems.

**Outcome 10. Transparency of policies, systems, processes and programs** that aim to reduce the risk of online harms

Measure 39. Publish and make accessible for users Signatories' safety and harms-related policies and terms of service. [Opted in]

Measure 40. Publish and make accessible information (such as via blog posts, press releases and/or media articles) on relevant policies, processes, and products that aim to reduce the spread and prevalence of harmful content online. [Opted in]

*TikTok response:*

- TikTok's Community Guidelines, safety policies and terms of service are publicly available on the TikTok website. These policies cover a broad range of issue verticals including minor safety, bully and harassment, violent extremism and illegal activities and regulated goods.
- In consultation with relevant stakeholders, we update our Community Guidelines from time to time to evolve alongside new behaviours and risks, as part of our commitment to keeping TikTok a safe place for creativity and joy.

- TikTok also has a range of online resources, in-app PSA's and hubs developed in partnership with reputable third-parties focussing on a range of important topics including COVID-19, elections, mental health, and first nations issues that provide users access to reliable information.
- TikTok's Newsroom, a publicly accessible webpage, outlines information including media articles and other relevant publications which highlight the work being done by the platform to address the spread of harmful online content. This includes updates on our work to counter misinformation on the platform, and our efforts to prevent the spread of violent extremism and its associated ideologies.
- We also publish quarterly Community Guideline Enforcement Reports, with additional Transparency Reports every six months on Government and Law Enforcement requests.

**Outcome 11.** Publication of regular **transparency reports** on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics

Measure 41. Publish periodic transparency reports with KPIs/metrics showing actions taken based on policies, processes and products to reduce the spread or prevalence of harmful content (e.g. periodic transparency reports on removal of policy-violating content). [Opted in]

Measure 42. Submit to the Administrator an annual compliance report, as required in section 5.4, that set out the measures in place and progress made in relation to Signatories' commitments under the Code. [Opted in]

*TikTok response:*

- TikTok uses a combination of AI and human moderators to identify, review, and action content that violates our policies. We compile metrics on these actions through our Community Guideline Enforcement reports, which are publicly accessible on the TikTok Transparency Centre webpage, and provide quarterly insights into the volume and nature of content and accounts removed from our platform. We also publish reports on Government and Law Enforcement requests every six months.

- As founding signatories to the Aotearoa New Zealand Code of Practice for Online Safety and Harms, TikTok commits to submitting to the Administrator annual compliance reports that will be used to evaluate our compliance and progress made against relevant outcomes and measures in the Code. TikTok will prepare its first annual compliance report for 2023 accordingly.

## 4.4 Support independent research and evaluation

Independent local, regional or global research by academics and other experts to understand the impact of safety interventions and harmful content on society, as well as research on new content moderation and other technologies that may enhance safety and reduce harmful content online, are important for continuous improvement of safeguarding the digital ecosystem. Signatories will seek to support or participate in these research efforts.

Signatories may also seek to support independent evaluation of the systems, policies and processes they have implemented under the commitments of the Code. This may include broader initiatives undertaken at the regional or global level, such as independent evaluations of Signatories' systems.

**Outcome 12.** Independent research to understand the impact of safety interventions and harmful content on society and/or research on new technologies to enhance safety or reduce harmful content online.

Measure 43. Support or participate, where appropriate, in programs and initiatives undertaken by researchers, civil society and other relevant organisations (such as fact-checking bodies). This may include broader regional or global research initiatives undertaken by the Signatory which may also benefit Aotearoa New Zealand. [Opted in]

Measure 44. Support or convene at least one event per year to foster multi-stakeholder dialogue, particularly with the research community, regarding one of the key themes of online safety and harmful content, as outlined in section 4. This may include broader regional or global events undertaken by the Signatory which involve Aotearoa New Zealand. [Opted in]

*TikTok response:*

- We are continually assessing ways that we can create safer experiences and better resources for users. The decisions we make in this regard are informed by external research and engagement with experts. Examples of research, partnerships and multi-stakeholder programs we have supported include:
- Commissioning Praesidio Safeguarding, an independent safeguarding agency, to better understand young people's engagement with potentially harmful challenges and hoaxes. While not unique to any one platform, the effects and concerns are felt by all – and we wanted to learn how we might develop even more effective responses as we work to better support teens, parents, and educators. During this project, we surveyed more than 10,000 teens, parents, and teachers from around the world. The report was written by Dr. Zoe Hilton, Director and Founder of Praesidio Safeguarding with the

support of a panel of 12 leading youth safety experts from around the world and is publicly available online. The report findings informed updates to our Community Guidelines and resources for users and guardians detailed here.

- Partnering with Netsafe and Professor Amanda Third to host a webinar to provide an educational session for parents in NZ - this included a strong focus on protecting children online, including education about our family pairing feature, age restriction information and community controls. Amanda is an expert in user-centred, participatory research, and her work investigates children's and young people's technology practices.
- Our engagement with the Australian Associated Press to create and deliver targeted digital media literacy education to a small group of content creators, including from New Zealand. The aim was to empower the participants to confidently source reliable and factual information, understand how to interrogate that information effectively, and encourage them to use those skills to create factual, reliable and educative content. The creators selected have large followings, with the project aiming to have a wide reach and impact in educating audiences in how to identify mis and disinformation.
- Partnering with third party fact checking organisations Agence France Presse and Australian Associated Press to detect and mitigate the spread of misinformation and disinformation in Oceania markets by leveraging the insights of IFCN-certified fact-checking experts.
- TikTok is currently working with leading experts, NGO's and regulators on a TikTok Creator Safety Summit in November 2022, which will train top creators, including from New Zealand, in the online prevention of Bullying, Harassment and Hateful Behaviour. The creators will then make videos highlighting these key messages to their followers on the app.

**Outcome 13.** Support independent evaluation of the systems, policies and processes that have been implemented in relation to the Code.

Measure 45. Commit to selecting an independent third-party organization to review the annual compliance reports submitted by Signatories, and evaluate the level of progress made against the Commitments, Outcomes and Measures, as outlined in section 4, as well as commitments made by Signatories in their Participation Form (see Appendix 2). [Opted in]

*TikTok response:*

- As founding signatories to the Aotearoa New Zealand Code of Practice for Online Safety and Harms, TikTok supports the Code's use of a third party organisation to assess all signatories (including TikTok) annual compliance

report to evaluate compliance and progress made against outcomes and measures relevant to each platform. TikTok will prepare its first annual compliance report for 2023 accordingly.