Australian Code of Practice on Disinformation and Misinformation

TikTok

Annual Transparency Report

January 2023 – December 2023

Table of Contents

Summary

Introduction

TikTok is committed to nurturing creativity in a safe, supportive and authentic environment. In a global community, it is natural for people to have different opinions, but we seek to operate on a shared set of facts and reality.

The Integrity and Authenticity (**I&A**) policies within our Community Guidelines prohibit harmful misinformation, impersonation, and coordinated or synthetic, manipulated misleading content. Violative videos are removed from the platform, and these efforts are detailed in TikTok's quarterly Community Guidelines Enforcement Reports.

Key initiatives undertaken by TikTok to support long-term risk mitigation include:

- investment in machine learning models to ensure extensive coverage of nuanced misinformation threats.
- detection and removal of inauthentic visual and audio trends to help combat manipulated, edited and deepfake content.
- maintaining a database of fact-checked claims enabling human moderators to accurately identify misinformation content.
- launching an anti-misinformation program in partnership with accredited fact-checkers to assess inauthentic narratives on third party platforms and restrict them from TikTok.
- developing a first-of-its-kind AI-generated content label launched in September 2023 to help people identify any realistic AIGC, with stringent rules and new technologies to proactively address AIGC related misinformation.

We work with International Fact-Checking Network-accredited fact-checking partners to enforce our rules against harmful misinformation. This work is particularly important during elections and other civic processes, as it enables us to verify claims and take action in line with our Community Guidelines. As a precautionary measure, while content is subject to review by our fact-checking partners, it remains on the platform but is not eligible for recommendation on the For You Feed. This helps to ensure that content produced in the context of an election period can be thoroughly assessed by independent fact-checking partners before misinformation enforcement decisions are made. While this process may sometimes result in delayed content moderation decisions, depending on the context and the complexity of the topic, it is designed to reduce the risk of mismoderating legitimate political discourse. Further information on our work to manage the risks associated with harmful misinformation while supporting freedom of expression is set out in the Appendix.

Notable Highlights

In 2023, TikTok implemented a number of global and Australia-specific initiatives to combat misinformation and ensure a safe user experience. These included:

- Collaborating with the NSW Electoral Commission and the Australian Electoral Commission to support the New South Wales State Election and the Indigenous Voice to Parliament Referendum respectively.
- Providing Public Service Announcements (**PSAs**) for the NSW State Election through hashtags and search terms, reminding users of our Community Guidelines and directing users to the NSW Electoral Commission's website.
- Launching the 2023 Australian Referendum Hub, directing users to a dedicated page with authoritative information on the Referendum.
- Implementing new front-end product safety features for the Voice Referendum, such as a Search Guide and Notice Tags for both short-form videos and TikTok LIVE.
- Maintaining fact-checking partnerships with the Australian Associated Press (**AAP**) to prevent the spread of misleading information.
- Deploying additional Arabic- and Hebrew-speaking moderators in response to the Israel-Hamas war, and launching search interventions which are triggered when users search for terms related to this topic (e.g., “Israel”, “Palestine”).

Throughout the 2023 reporting period, we experienced fluctuations in our <24 hour video removal rate in both Q1 and Q3, as well as our removal before content receives any views rate in Q3 (Fig. 1A). This can in part be attributed to a concomitant increase in third-party fact-checking escalations, particularly regarding conspiracy theories including those related to the Indigenous Voice to Parliament Referendum. Fact-checking played a critical role in mitigating the spread of misinformation on-platform throughout the Referendum campaign period. In the 6-week period leading up to polling day on 14 October 2023, we escalated approximately 1,700 videos to fact-checkers, and enforced approximately 380 of them (i.e. removing the video from the For You feed or from the platform completely). Many of these videos were found to be propagating conspiracy theories, in violation of our Community Guidelines. Given the nuances of the Referendum debate, some of this escalated content required more extensive review to verify factual accuracy, which impacted our overall video removal metrics. These video removal rates subsequently stabilised in Q4, as represented in Fig. 1B below.

Fig. 1A below shows the volume of content removed for violating our harmful misinformation policies in Australia by quarter in 2023, as well as our performance metrics on removal efficiency.

Quarter (2023)	Total videos removed	Proactive removal[1]	Removal before content receives any views	Removal within 24 hours of content posted to platform
January - March	6,737	97.80%	76.90%	12.60%
April - June	10,721	96.30%	77.70%	71.50%
July - September	6,134	92.45%	43.85%	16.22%
October - December	4,919	94.20%	47.25%	41.74%

Fig.1A: Summary of Removal of Harmful Misinformation Violations in 2023 (Australia)

Fig. 1B below shows that our proactive removal rates for Harmful Misinformation in Australia have consistently remained above 90% throughout 2023, as compared to proactive removal rates as low as 66.10% in Q1 2022.

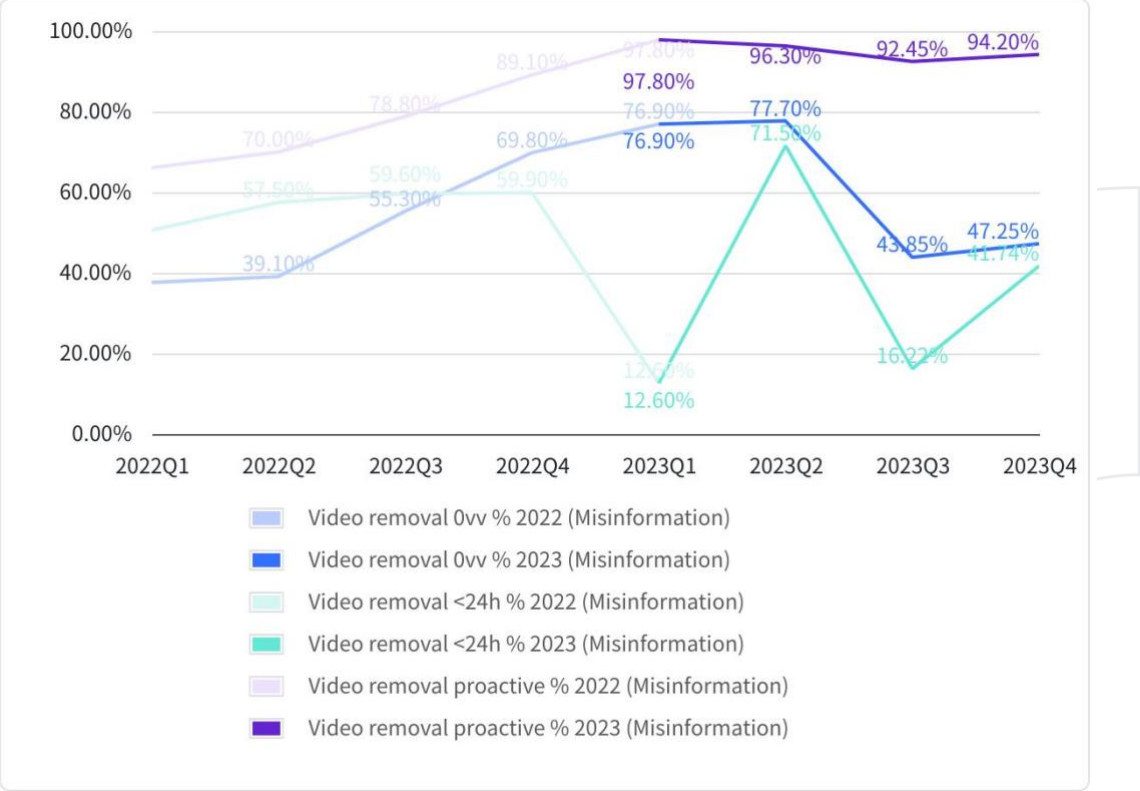

Fig. 1B: Removal of Integrity and Authenticity Violations Breakdown in 2023 vs 2022 (Australia)

[1] "Proactive removal" in this context refers to policy enforcement before it is reported by users.

TikTok's Commitments under the Code

TikTok opts in to all Objectives and Outcomes under the Australian Code of Practice on Disinformation and Misinformation with respect to the TikTok platform.

Objective 1: Provide safeguards against Harms that may arise from Disinformation and Misinformation	
Outcome 1a: Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures Outcome 1b: Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code Outcome 1c: Users can report content or behaviours to Signatories that violate their policies under section 5.10 through publicly available and accessible reporting tools. Outcome 1d: Users will be able to access general information about Signatories' actions in response to reports made under 5.11. Outcome 1e: Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.	Opt in (to all)
Objective 2: Disrupt advertising and monetisation incentives for Disinformation and Misinformation.	
Outcome 2: Advertising and/or monetisation incentives for Disinformation and Misinformation are reduced.	Opt in
Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms	
Outcome 3: The risk that Inauthentic User Behaviours undermine the integrity and security of services and products is reduced.	Opt in
Objective 4: Empower consumers to make better informed choices of digital content.	
Outcome 4: Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.	Opt in

Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.	
Outcome 5: Users are better informed about the source of Political Advertising.	Opt in
Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.	
Outcome 6: Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.	Opt in
Objective 7: Signatories publicise the measures they take to combat Disinformation and Misinformation.	
Outcome 7: The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.	Opt in

The following sections of the report will outline the specific measures, policies and projects undertaken to promote authenticity and counter misinformation on TikTok.

Reporting against 2023 commitments

Objective 1: Provide safeguards against Harms that may arise from Disinformation and Misinformation

Outcome 1a: Reducing harm by adopting scalable measures

In 2023 we implemented new scalable measures (outlined below) and strengthened our policy framework to help safeguard users from potential harms associated with misinformation. We regularly review these measures and ensure that people understand our guidelines, including what kind of content is not allowed on the platform, and when and why we may take action to mitigate any potential risks.

Our policies on misinformation are a subset of our Integrity and Authenticity policies. The 2023 reporting period saw an increase in Integrity and Authenticity policy violations as a proportion of total Community Guidelines violations, partly as a result of numerous significant civic processes, including the Indigenous Voice to Parliament Referendum, during which platforms typically experience higher volumes of misinformation.

Fig. 2A below shows the volume of content violating our Integrity and Authenticity policies in Australia by quarter in 2023, outlining the proportion of Integrity and Authenticity violations against all removed content, as well as our performance metrics on removal efficiency.

Quarter (2023)	Total videos removed	Proactive removal [1]	Removal before content receives any views	Removal within 24 hours of content posted to platform
January - March	8,764	95.90%	72.10%	29.50%
April - June	14,985	96.30%	80.90%	78.40%
July - September	10,672	93.30%	59.50%	48.40%
October - December	12,738	94.20%	70.90%	74.00%

Fig. 2A: Removal of Integrity & Authenticity Violations in 2023 (Australia)

Fig. 2B below outlines an increase in the proportion of Integrity and Authenticity Community Guidelines violations we detected on our platform in Australia in 2023, with Integrity and Authenticity violations accounting for a range of 0.90% - 1.70% of all Community Guidelines violations compared to a range of 0.60% - 0.90% in 2022. Despite the increase in proportion of Integrity and Authenticity violations, our proactive removal rate remains consistently high.

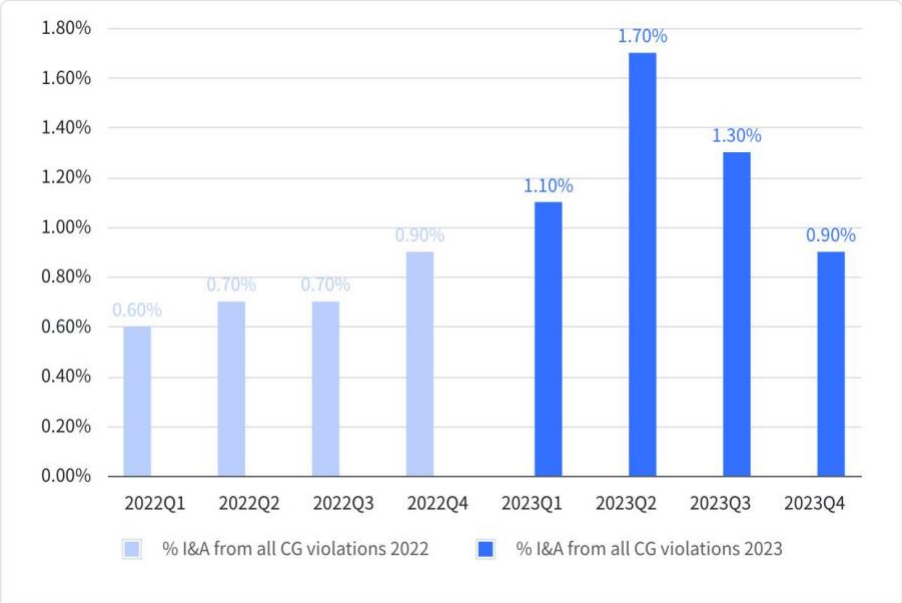

Fig. 2B: Proportion of Integrity and Authenticity violations in 2023 vs 2022 (Australia)

Fig. 2C below shows that our proactive removal rates for Integrity and Authenticity violations in Australia have consistently remained above 90% throughout 2023, as compared to proactive removal rates beginning at 83.60% in Q1 2022.

Fig. 2C: Removal of Integrity and Authenticity Violations Breakdown in 2023 vs 2022 (Australia)

New labels for disclosing AI-generated content

In September 2023, we introduced new labels for users to disclose AI-generated content (see Fig. 5) to make clear to viewers when content is significantly altered or modified by AI technology. The labels help creators showcase the innovations behind their content, and can be applied to any content that has been completely generated or significantly edited by AI.

This measure also makes it easier for users to comply with our Community Guidelines' synthetic media policy, which we introduced in early 2023. The policy requires people to label AI-generated content that contains realistic images, audio or video, in order to help viewers contextualise the video and prevent the spread of potentially false, misleading, or deceptive content.

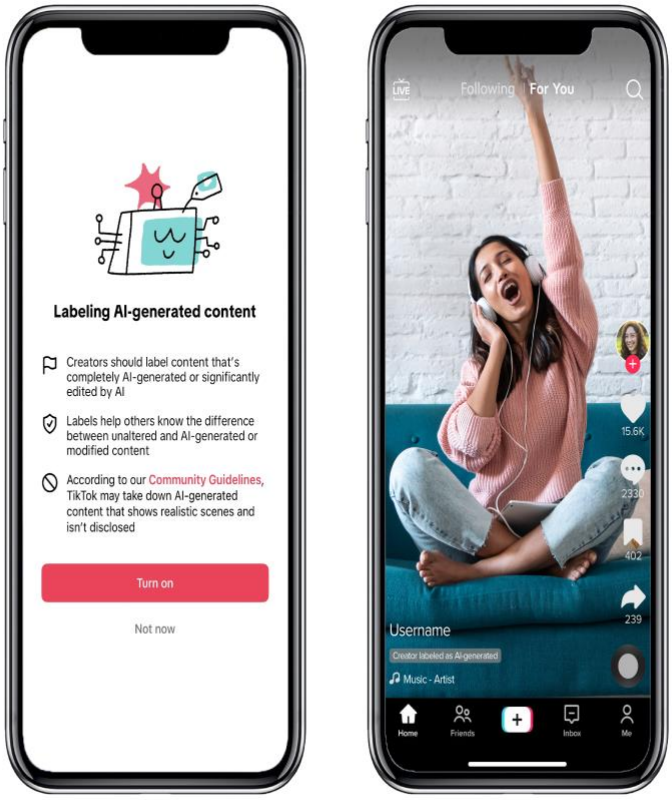

Fig. 3: AI Generated Content Labels

Strengthening enforcement through fact-checking partnerships

Globally, TikTok employs more than 40,000 Trust & Safety professionals who are responsible for ensuring the safety of the TikTok platform. This includes the development, implementation and enforcement of our harmful misinformation policies. During the 2023 reporting period, we collaborated

with 17 fact-checking organisations accredited by the International Fact-Checking Network (IFCN), covering over 50 languages, to ensure the accurate application of our Community Guidelines against misinformation (note: as of the date of reporting, TikTok currently collaborates with 18 such organisations). These partnerships empower our moderation teams to make accurate assessments of potentially misleading claims.

Our 24/7 moderation teams, along with our third-party fact-checking partners, work to review and verify flagged content and accounts. Throughout 2023 we continued to partner with the Australian Associated Press to help us independently review and assess the accuracy of content on our platform in Australia. Where such content is posted and is assessed to be false or deceiving, we remove such content in line with our Community Guidelines, and where fact-checks are inconclusive, we may label and restrict the content from appearing in the For You feed, as detailed in the "For You feed Eligibility Standards" section of our Community Guidelines.

Outcome 1b: Inform users about what content is targeted

TikTok's Community Guidelines are available to users within the app and on our website, and includes detailed descriptions of what constitutes misinformation, what forms of harmful misinformation are not allowed on our platform, and the eligibility criteria for content to appear in users' feeds.

To more clearly inform users about what content constitutes mis/disinformation, in March 2023 we updated our Community Guidelines governing harmful misinformation and disinformation. The March 2023 update significantly expanded the descriptions of content we control within pursuant to our Integrity and Authenticity policies. The revised CGs expand upon our controls on 'Misinformation' more broadly, as well as material related to Civic and Election Integrity, Synthetic and Manipulated Media, and Fake Engagement. The updated CGs also noted that our policies target misleading content as well as that which is inaccurate and false content, and provided information on the type of content that we prevent from being promoted on the For You feed, but do not remove from the platform. This includes:

- General conspiracy theories that are unfounded and claim that certain events or situations are carried out by covert or powerful groups, such as “the government” or a “secret society”.
- Unverified information related to an emergency or unfolding event where the details are still emerging.
- Potential high-harm misinformation while it is undergoing a fact-checking review.
- We also clarified that we allow:

- Statements of personal opinion (as long as it does not include harmful misinformation)
- Discussions about climate change, such as the benefits or disadvantages of particular policies or technologies, or personal views related to specific weather events (as long as it does not undermine scientific consensus)

Specifically in relation to Election Misinformation, we included additional guidance that misinformation related to the following will constitute a violation of our Community Guidelines:

- How, when, and where to vote or register to vote.
- Eligibility requirements of voters to participate in an election, and the qualifications for candidates to run for office.
- Laws, processes, and procedures that govern the organisation and implementation of elections and other civic processes, such as referendums, ballot propositions, and censuses.
- The outcome of an election.

These policies also prevent from promotion on the For Your feed any content containing unverified claims about the outcome of an election that is still unfolding and may be false or misleading. For more information about our Integrity & Authenticity policies, please refer to the Appendix.

Our Community Guidelines are informed through extensive consultations with relevant stakeholders, including NGOs, regulators, academics, subject matter experts, as well as our community. We also ensure that these are regularly reviewed, and where appropriate, updated, and that our community is notified of any major changes.

Outcome 1c: Users can easily report offending content

TikTok is designed so that users can easily report content they consider to be potentially violative of our Community Guidelines. This includes a dedicated category to report misinformation and a selection of sub-categories to choose from. In 2023 we made refinements to these subcategories to include election misinformation, harmful misinformation, as well as deepfake, synthetic media, and manipulated media.

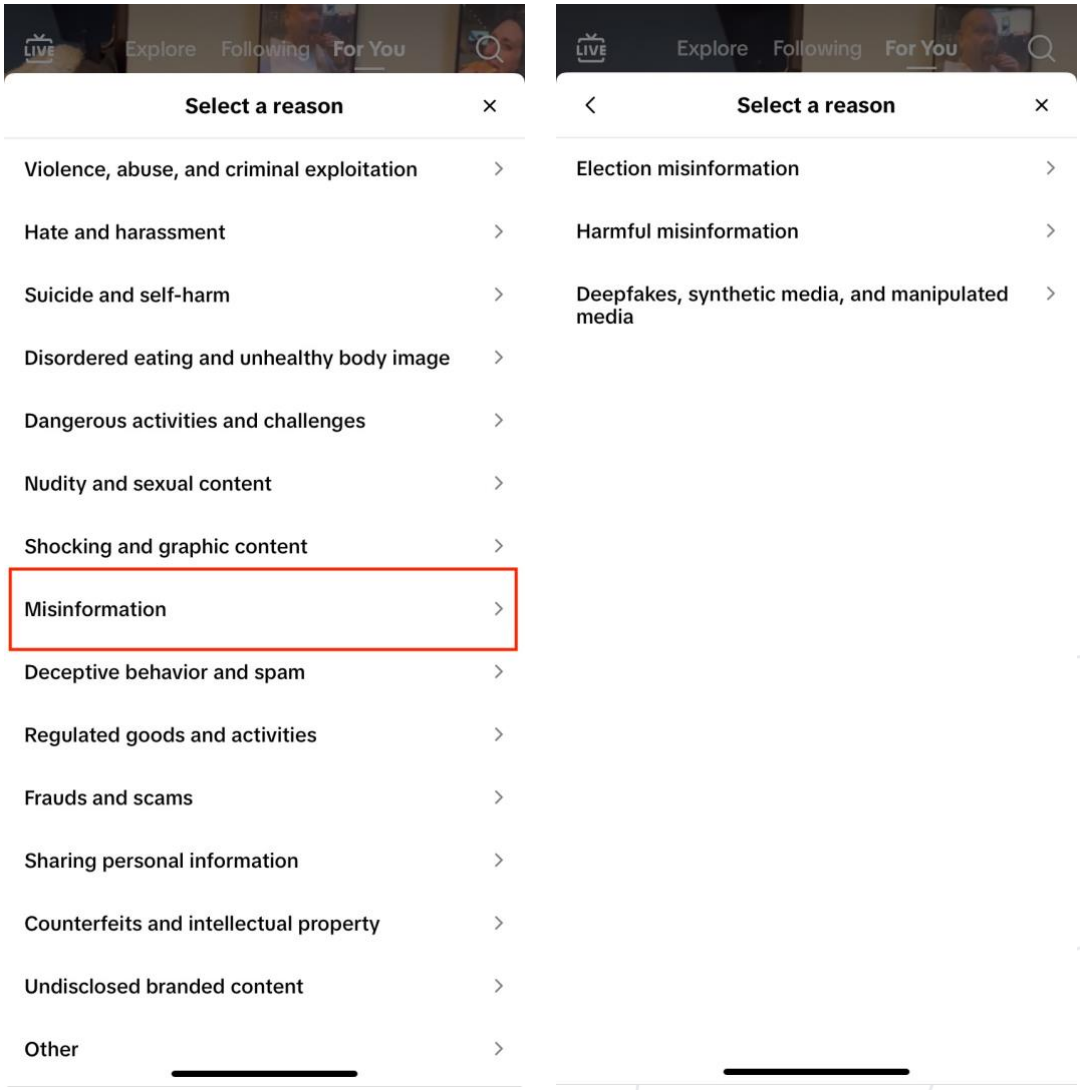

Fig. 4A: User in-app reporting interface

Before a user reports a type of misinformation to us, we make clear what we don't allow under each specific subcategory. This aims to enhance user awareness of our Community Guidelines, reduce ambiguity about permitted content, and strengthen platform safety.

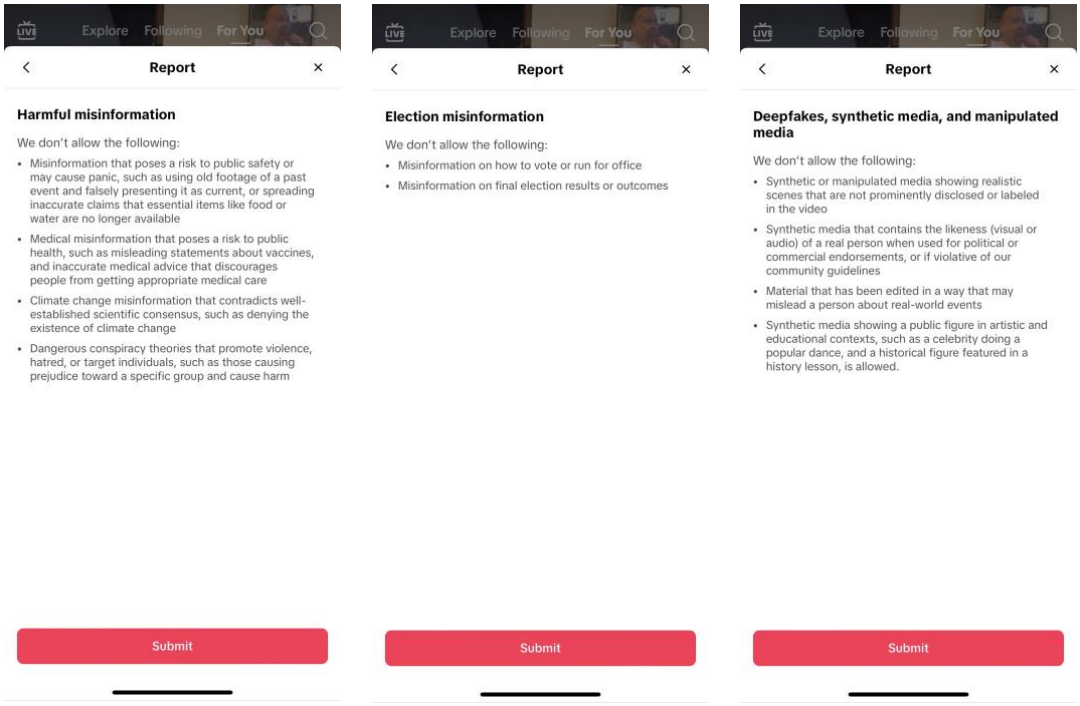

Fig. 4B: User in-app reporting interface

Users can keep track of their reports (including their status) and view their report history under Settings and Privacy > Support > Safety Center. When a user submits a report, we also provide them with the option to hide any further content being shown from the account in the feed.

Reporting misinformation is not limited to short-form videos. We enable users to also report misinformation across other features of the platform, including comments on videos, direct messages they receive from other users, accounts, sounds, hashtags and auto-suggestions generated when they search for something on TikTok. Users can also report LIVE videos and comments on livestreams if they encounter content that violates our Community Guidelines.

We also have reporting channels for non-users to report potentially harmful material. Our reporting form hosted on our website enables direct reports for our immediate review and action. Instructions for our publicly accessible reporting tools are available on our website.

Outcome 1d: Information about reported content available

Our Transparency Centre serves as a central hub to understand how TikTok moderates content, develops products, and protects user data. It provides users and the broader public with access to data and periodic reports, including:

- Community Guidelines Enforcement Reports: Quarterly insights into our efforts to enforce guidelines and terms of service.

- Information Requests Reports: Biannual data on user information requests from governments and law enforcement, along with our responses.
- Government Removal Requests Reports: Biannual data on requests from government agencies to restrict content and our actions in response.
- Intellectual Property Removal Requests Reports: Biannual data on requests to remove copyrighted and trademarked content, along with our responses.

These reports are published in multiple languages, are available for download in machine-readable formats from our Transparency Centre, and can be visualised in interactive charts and graphs.

Our latest Community Guidelines Enforcement Report for the period October 2023 - December 2023, published in March 2024, summarises our capabilities to proactively detect and remove violative material from our platform in Australia.

Quarter (2023)	Total videos removed	Proactive removal rate	Removal before content receives any views	Removal rate within 24 hours
January - March	792,784	95.60%	65.00%	81.10%
April - June	832,120	96.50%	77.70%	88.70%
July - September	741,846	95.40%	70.10%	85.50%
October - December	1,222,046	95.70%	74.70%	88.30%

Fig. 5: Video Removals in 2023 (Australia)

Outcome 1e: Information about recommender engines

The content people see on TikTok is generated by our community and recommendations are based on the content people have previously engaged with. Using signals such as view counts, likes, and shares, the recommendation algorithm creates a prediction score to rank videos to potentially recommend.

Our support page provides detailed information to users about how content is recommended across TikTok and how users can influence what they see on the platform. We have also provided additional information in our Transparency Centre.

Aside from the signals a user provides by how they interact with content on TikTok, there are additional tools we have built to help our community better control what kind of content is recommended to them. These include:

- **Not interested**: A user can long-press on the video in your For You feed and select 'Not interested' from the pop-up menu. This will let us know they are not interested in this type of content and we will limit how much of that content we recommend.
- **Video keyword filters**: A user can add keywords – both words or hashtags – they'd like to filter from their For You feed.
- **For You feed refresh**: To help discover new content, a user can refresh the For You feed, which provides an entirely new side of TikTok for them to explore.

Objective 2: Disrupt advertising and monetisation incentives for disinformation.

We place considerable emphasis on proactive content moderation and the vast majority of the violative content we remove is taken down before it is reported to us or receives any views. We are also committed to continuing to keep pace with evolving issues that affect our users.

Our work since the last report continues to reflect our strong commitment to combatting disinformation on our platform and to providing transparency to our wider community about the measures we take. We are in the process of developing and launching more granular misinformation policies and policies to govern AI-generated content in the coming months.

Transparency and Scrutiny of Advertising

Ads must comply with and are reviewed against our ad policies before being allowed on our platform. These policies specifically prohibit misleading, inauthentic and deceptive behaviours.

We continue to engage with external stakeholders in order to increase the effectiveness of our scrutiny of ad placements. As a Global Alliance for Responsible Media (GARM) member, we also remain committed to upholding the GARM Framework and, as part of that, removing harmful misinformation from monetisation.

Like all users of our platform, participants in content monetisation programs must adhere to our Community Guidelines, including our Integrity and Authenticity policies. Those policies make clear that we do not allow activities that may undermine the integrity of our platform or the authenticity of our users. They also make clear that we remove content or accounts, including those of creators, which contain misleading information that causes significant harm or deceptive behaviours. In certain scenarios, we may remove a creator's access to a creator monetisation feature.

Our policies and approach

Our Integrity and Authenticity policies within our Community Guidelines are the first line of defence in combating harmful misinformation and deceptive behaviours on our platform.

Paid ads are also subject to our ad policies and are reviewed against these policies before being allowed on our platform. Our ad policies specifically prohibit inaccurate, misleading, or false content that may cause significant harm to individuals or society, regardless of intent. They also prohibit other misleading, inauthentic and deceptive behaviours. Ads deemed in violation of these policies will not

be permitted on our platform, and accounts deemed in severe or repeated violation may be suspended or banned.

We also have other, existing ad policies that focus on certain topics where the risk of disinformation may be higher. By way of example, our Covid-19 ad policy prohibits ads that seek to take advantage of Covid-19 to push sales, for example by manipulating consumers' fear or anxiety, or spreading harmful misinformation to push sales. As well as ensuring ads relating to Covid-19 do not spread harmful misinformation, we also promote authoritative sources of information.

We are continually reflecting on whether there are further focused areas for which we should develop new policies. Our ad policies require advertisers to meet a number of requirements regarding the landing page. For example, the landing page must be functioning and must contain complete and accurate information, including about the advertiser. Ads may not be approved if the product or service advertised on the landing page does not match that included in the ad.

We make various brand safety tools available to advertisers to assist in helping to ensure that their ads are not placed adjacent to content they do not consider to fit with their brand values. While any content that is violative of our Community Guidelines, including our Integrity & Authenticity policies, is removed, the brand safety tools are designed to help advertisers to further protect their brand. As a GARM member, we believe in its mission and have adopted GARM's Brand Safety Floor and Suitability Framework (the **GARM Framework**).

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.

TikTok remains committed to preventing, detecting and deterring inauthentic user behaviours on our platform. These efforts include removing inauthentic accounts, tackling fake account engagement and disrupting Covert Influence Operations **(CIO).** In 2023, we disrupted a total of 46 networks globally.

We have enhanced our ability to detect CIO through a dual-pronged strategy that focuses on enhancing detection efforts by integrating insights gained from extensive global investigations, as well as developing strategic partnerships with third-party intelligence providers to complement existing in-house capabilities. We also consult with members of our Safety Advisory Council to gain insight into, and obtain advice on, our efforts to detect covert influence operations as we continually work to improve our efforts in this regard.

Where our teams have a high degree of confidence that an account is engaged in inauthentic coordination of content creation or amplification, uses deceptive practices to deceive/manipulate platform algorithms or coordinated mass reporting of non-violative opposing content/accounts and is engaged in or is connected to networks we took down in the past as part of a CIO, it is removed from our Platform in accordance with our CIO policy.

When we investigate and remove these operations, we focus on behaviour and assessing linkages between accounts and techniques to determine if actors are engaging in a coordinated effort to mislead TikTok's systems or our community. We know that CIOs will continue to evolve in response to our detection and networks may attempt to re-establish a presence on our platform. We continue to iteratively research and evaluate complex deceptive behaviours on our platform and develop appropriate product and policy solutions as appropriate in the long term. We publish the details of all of the CIO networks we identify and remove within our transparency reports, here.

Our Integrity & Authenticity policies, which address fake engagement, do not allow the trade of services that attempt to artificially increase engagement or deceive TikTok's recommendation system. We do not allow our users to facilitate the trade of services that artificially increase engagement, such as selling followers or likes, or to provide instructions on how to artificially increase engagement on TikTok.

If we become aware of accounts or content with inauthentically inflated metrics, we will remove the associated fake followers or likes. Content that tricks or manipulates others as a way to increase engagement metrics, such as "like-for-like" promises and false incentives for engaging with content

is ineligible for our For You feed. To know more about our approach on disrupting CIO, please refer to the Appendix.

Objective 4: Empower consumers to make better informed choices of digital content.

We take a multi-pronged approach to enabling users to make better informed choices of content on TikTok including implementing in-app features to provide timely, accurate and authoritative information to users regarding major civic events such as elections and referenda, and content relating to public health issues.

Combatting the spread of misinformation for the Indigenous Voice to Parliament Referendum

In light of the national significance and importance of the Voice to Parliament Referendum in October 2023, TikTok dedicated additional Trust and Safety resources to enact a broad spectrum of mitigative measures to detect and prevent the spread of harmful referendum-related misinformation during the Referendum campaign.

Our teams relied on both internal and external intelligence sources to address general misinformation trends, including ones logged in the Australian Electoral Commission's Disinformation Register. This ensured our teams were able to effectively identify and act upon known misinformation trends.

In addition, we created a Referendum Hub in collaboration with SBS's National Indigenous Television channel (**NITV**) and implemented front-end product features such as a Search Guide[2] and a Notice Tag.[3] These features ensured our users who were seeking or viewing content associated with the Referendum were able to access authoritative and factual sources of information from community partners and the Australian Electoral Commission.

---

[2] An information panel that appears when users search referendum-related terms which briefly explains what the Referendum is about. When the information panel is clicked, it will lead users to the Referendum Hub.

[3] A small text-based banner that appears on the bottom of videos associated with the Referendum. When the text based banner is clicked, it will lead users to the Referendum Hub.

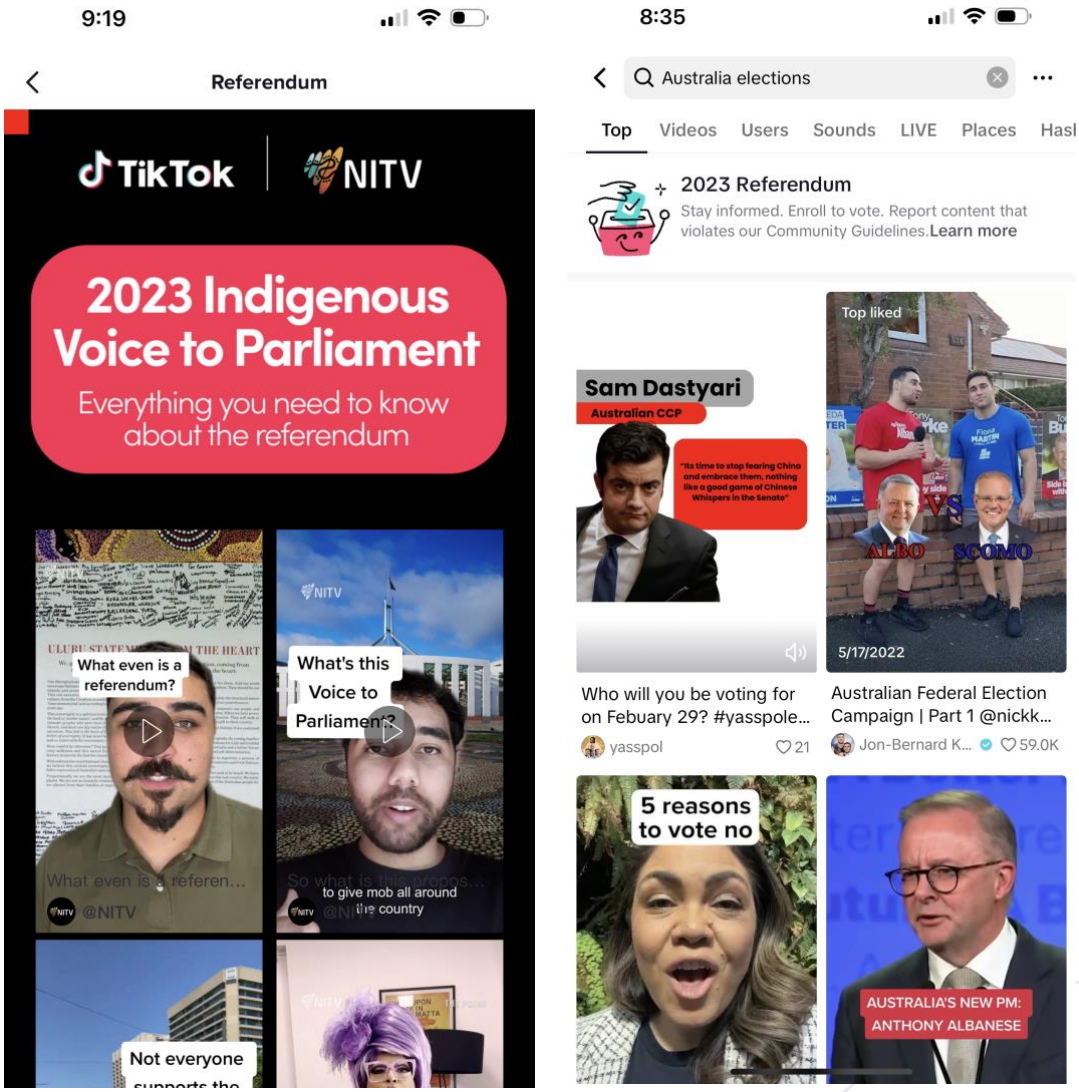

Fig. 6: Referendum Hub (Left), Search Guide (Right)

These measures saw strong engagement from our community. During this time:

- our Search Guide received approximately 350,000 impressions.
- our Referendum Hub was viewed approximately 76,000 times.
- Notice Tags were applied to approximately 191,000 TikTok video and photo posts, with approximately 49.5 million impressions.
- Notice Tags applied to TikTok LIVE-related content received approximately 400,000 impressions.

Fact-checking played a critical role in mitigating the spread of misinformation on-platform throughout the referendum campaign period. In the 6-week period leading up to polling day on 14 October 2023, we escalated approximately 1,700 videos to fact checkers, and enforced approximately 380 of them (i.e. removing the video from the For You feed or from the platform completely). Many of these videos were found to be propagating conspiracy theories, in violation of our Community Guidelines.

Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.

Transparency and Scrutiny of Advertising

Like all users of our platform, participants in content monetisation programs must adhere to our Community Guidelines, including our Integrity and Authenticity policies. These policies include our prohibition on paid political advertising.

Prohibiting Paid Political Ads

TikTok does not allow anyone to place political ads, nor do we allow politicians and political party accounts to place ads. We also prevent Government, Politician, and Political Party Accounts (**GPPPAs**) from accessing our monetisation features and campaign fundraising.

Sharing political beliefs and engaging in political conversation is allowed as organic content, but our policies prohibit users from paying to advertise or promote this content. We allow some cause-based advertising and public service advertising from government agencies, non-profits and other entities, provided they are not politically partisan and make exceptions for governments in certain circumstances, e.g., to promote public health.

Transparency Risk Controls

Where accounts are designated as GPPPAs, those accounts are banned from placing ads on TikTok (with the exception of certain government entities in certain circumstances, as outlined above) and from monetisation features. We publish the details of our GPPPA policy on our website, where we set out who we consider to be a GPPPA and the restrictions on those types of accounts.

We apply an internal label to accounts belonging to a government, politician, or political party. Once an account has been labelled in this manner, a number of policies will be applied that help prevent misuse of certain features, e.g., access to advertising features and solicitation for campaign fundraising are not allowed.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.

We recognise the important role of researchers and subject matter experts in helping to identify misinformation trends and practices.

Advancing our commitment to combat climate misinformation

We remain committed to increasing climate literacy among our global community. As outlined above, climate change misinformation is directly referenced within our Community Guidelines. We permit discussions about climate change, such as the benefits or disadvantages of particular policies or technologies, or personal views related to specific weather events (as long as it does not undermine scientific consensus), but do not allow climate change misinformation that undermines well-established scientific consensus, such as denying the existence of climate change or the factors that contribute to it.

As part of our ongoing commitment, in November 2023, to coincide with the COP28 UN Climate Change Conference, we launched our #ClimateAction campaign with new initiatives and programming. This included the introduction of a $1 million initiative to tackle climate misinformation in support of Verified for Climate, a joint program of the United Nations and Purpose. The initiative helps bring together a network of climate messengers with credibility and experience (known as 'Verified Champions)', including scientists and trusted experts from Brazil, the United Arab Emirates, and Spain who support select TikTok creators in developing educational content to tackle climate misinformation and disinformation.

Ongoing support with Australian Associated Press Fact Check

During the 2023 referendum period, Trust & Safety staff also participated in a closed event hosted by our fact-checking partners, Australian Associated Press, and stakeholders, where representatives of different platforms discussed their respective approaches to content moderation for issues like misinformation and listened to commentary and questions concerning the fact-checking process.

Community Partner Channel

Our Global Community Partner Channel provides selected organisations an additional route for reporting content that they believe breaks our Community Guidelines so that it can be reviewed by

our teams. To date, more than 400 organisations who specialise in a range of safety issues use our Community Partner Channel. In Australia, 25 partners have been introduced to the program, including organisations focusing on combating antisemitism, Islamophobia, hate speech and racism.

Objective 7: Signatories publicise the measures they take to combat Disinformation.

TikTok remains committed to transparency and to ensuring the clarity of our practices for users, law enforcement agencies, governments and the general public. We continue active engagements with government and regulatory bodies, providing visits to our Transparency and Accountability Centres (**TACs**) to see up close how we moderate and recommend content, secure our platform, and protect people's privacy. Our newly opened TAC in Singapore complements our existing TACs in Dublin and Los Angeles, and has hosted hundreds of guests both physically and virtually.

We will continue our efforts to publish quarterly transparency reports and provide newsroom updates. In 2023, we broadened the scope of information included in our Community Guidelines Enforcement Reports, which include reporting the number of removed suspected under-age accounts and information on the identification and removal of covert influence operations globally.

Helping our community access authoritative information on the Israel-Hamas war

In the aftermath of the start of the Israel-Hamas war, we immediately mobilised significant resources and personnel to help maintain the safety of our community and integrity of our platform. Our focus has been on supporting free expression, upholding our commitment to human rights, and maintaining the safety of our community and integrity of our platform during the war. We have provided regular updates on our efforts as part of our commitment to transparency. In the six months since October 7, 2023, we have removed more than 3.1 million videos and suspended more than 140,000 livestreams in Israel and Palestine for violating our Community Guidelines, including content promoting hate speech, violent extremism and misinformation.4 Approximately 155,000 videos propagating dangerous misinformation have been removed globally.

We are continually working hard to ensure that TikTok is a source of reliable and safe information and recognise the heightened risk and impact of misleading information during a time of crisis. As part of our crisis management process, we launched a command centre that brings together key members of our 40,000-strong global team of safety professionals, representing a range of expertise and regional perspectives, so that we remain agile in how we take action to respond to this fast-evolving crisis.

[4] Data covers October 7, 2023 to March 31, 2024.

Since the onset of the war, there has been a rise in misinformation and conspiracy theories relating to the war. We have also seen spikes in deceptive account behaviours and continue to take swift action against fake engagement and accounts, for example, by removing 35 million fake accounts in the month after the start of the war - a 67% increase in the previous month. From October 7 through to the end of 2023, we have removed more than 169 million fake accounts globally, as well as removing approximately 1.2 million bot comments on content tagged with hashtags related to the war.

To help raise awareness and to protect our users, we have also launched search interventions which are triggered when users search for non-violating terms related to the war (e.g., Israel, Palestine). These search interventions remind users to pause and check their sources and also direct them to wellbeing resources.

We remain committed to engagement with experts across the industry and civil society, such as our Safety Advisory Councils, and cooperation with law enforcement agencies globally in line with our Law Enforcement Guidelines, to further safeguard and secure our platform during these difficult times.

Concluding remarks

TikTok is committed to upholding its obligations under the Australian Code of Practice for Disinformation and Misinformation. This report has highlighted our continued efforts - both locally and globally - to combat misinformation through a range of new and strengthened measures and policy frameworks. We remain dedicated to safeguarding our users from the risks associated with harmful misinformation.

We recognise that misinformation is an evolving issue, and with new tools and advancements in technology, including with respect to generative AI, we are committed to proactively enhancing our methods of detecting and mitigating risks associated with such content, as well as assessing options for partnerships to determine the origin of content. We also continue to partner closely with experts to ensure we are able to stay ahead of emerging trends, and effectively mitigate risks associated with AI-related misinformation.

As TikTok's user base continues to grow, we strive to enable creativity in a safe environment, as well as to support genuine discussions on global affairs, politics and health. We deeply value the trust of our community and are committed to providing a transparent, authentic platform experience globally.

Appendix

Approach to Disinformation and Misinformation

Our misinformation policies apply to content regardless of the poster's intent, as the content's harm is the same either way. Hence, they cover both "disinformation" (which is intentionally shared to mislead) and harmful misinformation that may not have been shared with the goal of deceiving people.

Like others in our industry, we do not prohibit people from sharing personal experiences, simply inaccurate myths, or misinformation that could cause reputational or commercial harm, in order to balance creative expression with preventing harm.

Policy on Misinformation

In a global community, it is natural for people to have different opinions, but we seek to operate on a shared set of facts and reality. **We do not allow inaccurate, misleading, or false content that may cause significant harm to individuals or society, regardless of intent.** Significant harm includes physical, psychological, or societal harm, and property damage. It does not extend to commercial and reputational harm, nor does it cover simply inaccurate information and myths. We rely on independent fact-checking partners and our database of previously fact-checked claims to help assess the accuracy of content.

Content is ineligible for the FYF if it contains general conspiracy theories or unverified information related to emergencies. To be cautious, content that warrants fact-checking is also temporarily ineligible for the FYF while it is undergoing review.

To help you manage your TikTok experience, we add warning labels to content related to unfolding or emergency events which have been assessed by our fact-checkers but cannot be verified as accurate, and we prompt people to reconsider sharing such content.

Misinformation includes inaccurate, misleading, or false content.

Significant harm includes severe forms of:

- Physical injury and illness, including death
- Psychological trauma
- Large-scale property damage
- Societal harm, including undermining fundamental social processes or institutions, such as democratic elections, and processes that maintain public health and public safety

Conspiracy theories are beliefs about unexplained events or involve rejecting generally accepted explanations for events and suggesting they were carried out by covert or powerful groups.

NOT allowed

- Misinformation that poses a risk to public safety or may induce panic about a crisis event or emergency, including using historical footage of a previous attack as if it were current, or incorrectly claiming a basic necessity (such as food or water) is no longer available in a particular location
- Medical misinformation, such as misleading statements about vaccines, inaccurate medical advice that discourages people from getting appropriate medical care for a life-threatening disease, and other misinformation that poses a risk to public health
- Climate change misinformation that undermines well-established scientific consensus, such as denying the existence of climate change or the factors that contribute to it
- Dangerous conspiracy theories that are violent or hateful, such as making a violent call to action, having links to previous violence, denying well-documented violent events, and causing prejudice towards a group with a protected attribute
- Specific conspiracy theories that name and attack individual people
- Material that has been edited, spliced, or combined (such as video and audio) in a way that may mislead a person about real-world events

FYF ineligible

- General conspiracy theories that are unfounded and claim that certain events or situations are carried out by covert or powerful groups, such as “the government” or a “secret society”
- Unverified information related to an emergency or unfolding event where the details are still emerging
- Potential high-harm misinformation while it is undergoing a fact-checking review

Allowed

- Statements of personal opinion (as long as it does not include harmful misinformation)
- Discussions about climate change, such as the benefits or disadvantages of particular policies or technologies, or personal views related to specific weather events (as long as it does not undermine scientific consensus)

Since the start of the war on 7 October 2023, we have been working diligently to remove content that violates our policies. We have set out below some of the main threats both observed and considered in relation to the war and the actions we have taken to address these.

(I) Spread of misinformation

We believe that trust forms the foundation of our community, and we strive to keep TikTok a safe and authentic space where genuine interactions and content can thrive. TikTok takes a multi-faceted approach to tackling the spread of harmful misinformation, regardless of intent. This includes our: Integrity & Authenticity policies (I&A policies) in our Community Guidelines (CGs); as well as our external partnerships with fact-checkers, media literacy bodies, and researchers. We support our moderation teams with detailed misinformation policy guidance, enhanced training, and access to tools like our global database of previously fact-checked claims from our IFCN-accredited fact-checking partners, who help assess the accuracy of content.

Since 7 October 2023, there has been a rise in misinformation and conspiracy theories relating to the war. We have also seen spikes in deceptive account behaviours and continue to take swift action against fake engagement and accounts, for example, by removing 35 million fake accounts in the month after the start of the war - a 67% increase on the previous month.

(II) Covert Influence Operations (CIO)

TikTok's I&A policies do not allow deceptive behaviour that may cause harm to our community or society at large. This includes coordinated attempts to influence or sway public opinion while also misleading individuals, our community, or our systems about an account's identity, approximate location, relationships, popularity, or purpose. We have specifically-trained teams on high alert to investigate CIO and we provide quarterly updates on the CIO networks we detect and remove from our platform in our Community Guidelines Enforcement Reports (CGER).

We have assigned dedicated resourcing within our specialist teams to proactively monitor for CIO in connection with the war. While we have not identified any CIO specifically targeting the war during this reporting period, we reported on a CIO relevant to the region in our Q1 (Jan-March 2023) CGER where we identified and removed a network operated from Israel that targeted Israeli audiences. We are currently investigating a number of operations and will publish details of any CIO networks we identify and remove. While we currently report the removals of covert influence networks in the quarterly CGER, in the coming months, we will also introduce dedicated CIO reports to further increase transparency, accountability, and cross-industry sharing.

We know that CIOs will continue to evolve in response to our detection and networks may attempt to re-establish a presence on our platform, which is why we continually seek to strengthen our policies

and enforcement actions in order to protect our community against new types of harmful misinformation and inauthentic behaviours.