

DSA RISK ASSESSMENT REPORT 2024

Confidential
28 August 2024
(Updated 2 October 2024)

CONTENTS

1. FOREWORD	1
2. EXECUTIVE SUMMARY	2
3. INTRODUCTION	4
4. RISK GOVERNANCE	5
5. AIGC, RECOMMENDER SYSTEMS AND RISKS TO USERS' PHYSICAL AND MENTAL WELL-BEING	10
6. ILLEGAL CONTENT: RISKS OF ILLEGAL HATE SPEECH CONTENT	15
7. ILLEGAL CONTENT: RISKS OF TERRORIST CONTENT	22
8. ILLEGAL CONTENT: RISKS OF INTELLECTUAL PROPERTY INFRINGING CONTENT	30
9. ILLEGAL CONTENT: RISKS OF CHILD SEXUAL ABUSE MATERIAL AND CHILD SEXUAL EXPLOITATION AND ABUSE	35
10. ILLEGAL CONTENT: RISKS OF GENDER-BASED VIOLENCE CONTENT	42
11. YOUTH SAFETY: RISKS RELATED TO AGE APPROPRIATE CONTENT AND ONLINE ENGAGEMENT	47
12. YOUTH SAFETY: RISKS RELATED TO AGE ASSURANCE	54
13. CIVIC INTEGRITY: RISKS TO ELECTIONS AND CIVIC INTEGRITY	59
14. CIVIC INTEGRITY: RISKS TO PUBLIC HEALTH FROM MEDICAL MISINFORMATION CONTENT	68
15. CIVIC INTEGRITY: RISKS TO PUBLIC SECURITY FROM HARMFUL MISINFORMATION CONTENT	73
16. FUNDAMENTAL RIGHTS: RISKS TO FUNDAMENTAL RIGHTS	79
ANNEX 1: KEY INFORMATION ABOUT TIKTOK	85
ANNEX 2: HOW TO USE THIS REPORT	87
ANNEX 3: SPECIALIST TEAMS	88
ANNEX 4: STAKEHOLDER ENGAGEMENT	93
ANNEX 5: OVERVIEW OF FACT-CHECKING PROGRAMME	111

1. FOREWORD

TikTok presents here its second annual systemic Risk Assessment Report within the framework of the Digital Services Act ('**DSA**'). TikTok remains steadfast in its mission to inspire creativity and bring joy whilst being a safe and welcoming community. Each day, millions of Europeans come to TikTok to find entertainment, education and fun.

The DSA's focus on transparency and systemic risk assessments aligns with our commitment to accountability and continuous improvement in managing risks. TikTok is dedicated to building a range of effective processes and policies to safeguard our community from the various content and conduct risks that are detailed in this Report. The challenges faced by online platforms are complex and ever evolving. To address these, TikTok strives to develop and support internal teams with expertise across systemic risk areas. Our efforts are further strengthened through close collaboration with regional and country-specific teams across Europe, combining specialised knowledge with local insights to enhance our safety initiatives. This local knowledge has been vital in tackling the significant volume of elections in the EU in 2024 and we are grateful to the European Commission for their engagement and guidance on this important topic.

TikTok has closely monitored existing and emerging risks over the last year, building on the insights gained from our first systemic risk assessment process. We have further developed our risk identification and mitigation strategies to enhance platform safety and security. Three key initiatives have been central to this effort: (1) investing further in stakeholder engagements to strengthen the ways in which we consider external perspectives to risk, which includes specific engagements in relation to the Israel-Hamas war; (2) addressing the known and potential risks associated with Artificial Intelligence, notably AI-generated content; and (3) ensuring we are engaging meaningfully and thoughtfully on the risks associated with our younger users aged between 13 and 17 years old. We expand on each of these areas in more detail below.

TikTok has strengthened its approach to external stakeholder engagement as part of our risk governance. In consideration of the text of Article 34 and Recital 90 DSA, we regularly engage with a wide range of independent experts, researchers, vulnerable groups and civil society organisations who inform our risk identification and mitigation design activities. We also do this through initiatives such as our European Safety Advisory Council, which is made up of diverse experts on systemic risks, and the TikTok Youth Council. The Youth Council comprises teenagers from around the world and provides a forum where we can hear directly from our younger users. We use these perspectives to inform changes we make to create the safest possible experience for our community. The Council's priorities for this year are on strengthening our approach to both teen well-being and inclusion on TikTok. We have included a summary of these important interactions at Annex 4 to this Report.

In last year's Risk Assessment Report, we identified emerging risks associated with AI-generated content and these risks have continued to evolve. TikTok has already proactively deployed processes to identify and mitigate such harmful AI-generated content, resulting in new and improved mitigations documented in this Report, and we are expecting to further strengthen our measures in the coming year.

TikTok continues to prioritise the safety and well-being of its younger users, and has implemented new and updated policies and enforcement measures to support youth safety on the Platform. Key initiatives include the introduction of additional privacy settings, and expanded screen time interventions to promote responsible digital habits. TikTok has also launched a global multi-stakeholder initiative focused on age assurance in the digital environment. This initiative, involving industry partners and leading safety and privacy organisations, aims to create a holistic dialogue on age assurance complexities and establish working groups to address key issues, balancing child protection with safe online participation. TikTok remains committed to creating a safe and supportive environment for its younger users.

Our progress in the year since the implementation of the DSA has been significant, but we recognise that safety is an ongoing commitment. We are dedicated to monitoring our efforts and upscaling our mitigation strategies iteratively and based on the best available research. Looking forward, we welcome feedback on our approach from regulators, researchers and civic society, as such inputs are crucial to refining our risk management strategy.

Adam Presser

Director, TikTok Technology Limited

Global Head of Trust and Safety, TikTok

2. EXECUTIVE SUMMARY

Introduction

This Report summarises the results of TikTok's second annual Systemic Risk Assessment within the framework of the DSA. TikTok submitted its first Systemic Risk Assessment Report to the European Commission in September 2023. This Report uses the terms '**Year 1**' to refer to the 2023 report and assessment, and '**Year 2**' to reflect this updated assessment of systemic risks, and any new or improved mitigations that have been implemented from September 2023 to August 2024.

Methodology

In Year 1, TikTok implemented a risk assessment framework to analyse the systemic risks present on its Platform. That framework scored each inherent and residual risk on a five-level scale, based on the effectiveness of existing mitigation policies, systems, and procedures. The framework guided decisions on when and how to apply further mitigations in a manner that is reasonable, proportionate, and effective.

Building on this foundation, in Year 2, TikTok strengthened this framework by incorporating elements of the ISO31000 risk scoring methodology. The Year 2 methodology examines the design and operating effectiveness of mitigations/controls to produce a comprehensive risk assessment and mitigation plan.

AIGC, Recommender Systems and risks to physical and mental well-being

TikTok has conducted a thorough risk assessment, which identified two priority factors that pose potential risks across many of its assessments, these being content generated by Artificial Intelligence Technology ('AIGC') and the impact of recommender systems. TikTok also sets out its approach to analysing the potential serious negative consequences to users' physical and mental well-being.

Risk assessment results

TikTok has taken a diligent and cautious approach to the analysis of systemic risks that may arise from the design, functioning, use or mis-use of the Platform. TikTok's methodology remains substantially the same as last year. TikTok has classified the risks applicable to the Platform into 11 **Risk Modules** and then organised each Risk Module into 1 of 3 tiers to show how it prioritises its ongoing work to address those risks ('Tiers').

These Tiers reflect TikTok's current assessment of the priority level for each systemic Risk Module, considering existing policies, systems, and procedures for risk mitigation. This Tiering system is integral to TikTok's resource allocation. Special emphasis has been placed on risks affecting vulnerable groups and those risks more likely to evolve over the next 12 months. This overall approach ensures that TikTok is prepared to mitigate risks effectively and proportionately, maintaining a strong focus on community safety.

SUMMARY OF RISK ASSESSMENT RESULTS		
Tier 1 risks	Tier 2 risks	Tier 3 risks
<ul style="list-style-type: none"> • Risks of child sexual abuse material and child sexual exploitation • Risks related to age assurance • Risks related to age appropriate content and online well-being • Risks to elections and civic integrity • Risks of gender-based violence content 	<ul style="list-style-type: none"> • Risks of terrorist content • Risks of illegal hate speech content • Risks to public security from harmful misinformation content • Risks to fundamental rights 	<ul style="list-style-type: none"> • Risks of intellectual property infringing content • Risks to public health from medical misinformation

In Year 2, TikTok observed some changes in the risks to users of Medical Misinformation. This is due to its conclusion that the risk has now pivoted from content related to the Covid-19 pandemic to content related to a group of less widespread risks (See section 14). Medical Misinformation has therefore moved from a Tier 2 risk to a Tier 3 risk. TikTok also observed an increase in the potential for risks to fundamental rights as a result of the ever-evolving challenges of moderating complex hate speech content, driven by global conflicts but impacting content consumed in the EU. Risks to fundamental rights have therefore moved from a Tier 3 risk to a Tier 2 risk. All other risks remain in the same Tier, with the same emphasis of prioritisation as in Year 1. TikTok will continue to mature and improve its approach to the identification and mitigation of systemic risks where these can stem from

the design, functioning or (mis)use of the Platform and its related systems, with particular regard to emerging risks.

3. INTRODUCTION

This DSA Risk Assessment and Mitigation Report (the '**Report**') has been prepared in compliance with Article 42(4)(a),(b) and (e) of the DSA. This Report summarises the results of TikTok's second systemic risk assessment under Article 34 and the mitigation measures that have been put in place under Article 35 DSA. It has been prepared by TikTok Technology Limited ('**TikTok Ireland**') in relation to the operation in the European Union ('**Europe**' or '**EU**') of its online platform named TikTok (referred to as either '**TikTok**' or the '**Platform**' depending on the context within this Report), which has been designated as a Very Large Online Platform ('**VLOP**') under the DSA. The DSA requires VLOPs to conduct systemic risk assessments at least once annually. This Report has been reviewed and approved by the board of directors of TikTok Ireland, following consultation with TikTok's head of the independent compliance function.

TikTok has enhanced its report structure and contents since its Year 1 Report. This Year 2 Report:

- Expands upon the risk identification and risk description from the Year 1 Report, with a dedicated section for an analysis of additional risks identified during Year 2;
- Includes an Annex detailing all relevant external stakeholder engagements and how they have led to risk identification or risk mitigations across each systemic risk in order to make observing compliance with Article 42(4)(e) as straightforward for the reader as possible;
- Addresses **Youth Safety** in two distinct sections to reflect the underlying risk assessment methodology (one for age appropriate content and online engagement; and a second for age assurance);
- Refers to '**Younger Users**' and 'Youth' rather than 'Minor(s)' as it did in Year 1. 'Younger Users' and 'Youth' refer to users aged 13-17. '**Underage Users**' refers to anyone accessing the Platform who is under the age of 13; and
- Includes a distinct '**Risk Environment**' section with an assessment of how intentional manipulation of the service may influence systemic risks in line with Article 34.2.

And also note:

- The section 'Key information about TikTok' can now be found in Annex 1 which sets out information about TikTok's content moderation systems, data related practices, advertising systems, recommender systems; and
- The Year 2 Report sets out the results of each Risk Module, grouped into topic-based risk categories (in the following order: illegal content, youth safety, civic integrity and fundamental rights), rather than grouping them by Tier-based prioritisation as TikTok did in Year 1.

This Report provides the results of TikTok's Year 2 systemic risk assessment through the following sections:

- TikTok's methodology and approach to the governance of systemic risks (see section 4);
- An analysis of the interplay of systemic risks and two horizontal priority areas, being risks related to AIGC and recommender systems (see section 5);
- An analysis of how TikTok has addressed its analysis of risk of 'serious negative consequences to [users'] physical and mental well-being', (see section 5); and
- The summarised results of TikTok's analysis of each systemic risk area, broken down (as it was in Year 1) thematically across illegal content, youth safety and well-being, civic integrity, and fundamental rights. The summary of each of these areas of systemic risk includes an assessment of the severity and probability of each risk. The Report also includes an assessment of the effectiveness (in reducing residual risk), reasonableness and proportionality of TikTok's mitigation measures. The Report incorporates identification of further mitigation actions for the year ahead (see sections 6-16).

This Report is supplemented by a **Stakeholder Engagement Report** in Annex 4, which summarises core stakeholder engagements that TikTok has conducted. These engagements are done to identify emerging risks and develop proportionate mitigation responses based on the best available information and scientific insights, in line with Recital 90 DSA.

4. RISK GOVERNANCE

A. INTERNAL RISK GOVERNANCE

TikTok recognises that effective risk management is critical to providing a safe and welcoming online platform. TikTok operates a risk management framework that emphasises a safety-first approach, integrates human rights standards, and fosters cross-functional collaboration.

TikTok strives to employ effective, reasonable, and proportionate risk mitigation measures tailored to identified systemic risks, whilst avoiding unnecessary restrictions to platform use or fundamental rights. Proactive risk management and detection strategies are also employed, even though general monitoring for illegal content is not required under the DSA.

Balancing user safety with other fundamental rights, TikTok follows the principle of proportionality, assessing risks at the intersection of safety, privacy, protection of youth, and freedom of expression and information. Acknowledging the complexity, TikTok is committed to achieving a fair balance for its users.

TikTok's policy-driven approach to risk management is reflected in its community guidelines, which establish a common code of conduct for all users and set out TikTok's content moderation practices ('**Community Guidelines**'). Continuous improvement is fostered through internal training and

awareness programs, which drive well-informed decision-making and sound governance. Raising user awareness, particularly concerning disinformation campaigns, is also emphasised.

B. SYSTEMIC RISK ASSESSMENT METHODOLOGY

TikTok has built on and enhanced its Year 1 risk assessment methodology to inform its approach in Year 2. In TikTok’s Year 1 Risk Assessment Report, TikTok organised the results of the risk assessments into three tiers. These tiers represented TikTok’s assessment of the priority that a systemic risk category demands, having taken account of existing policies, systems and procedures for mitigating the risk. Over the last year, these tiers have informed TikTok’s view of when further mitigations are needed, and what kind of mitigation would be reasonable and proportionate.

The Tiering system allows TikTok to account for the level of inherent risk in certain systemic risk areas and therefore prepare for unforeseen events and the severity of potential harm to its user community, with a specific focus on vulnerable or marginalised groups. This ensures TikTok is resourced to put in place necessary controls and mitigations for a residual risk which is reasonable, proportionate, and effective.

TikTok has strengthened the detail of its risk scoring methodology (which is based on ISO31000 standards) since Year 1, as it has further leveraged subject matter expertise both internally and through external engagements, and considered new data sets in its risk scoring analysis. TikTok is ready to further develop this methodology should any guidance become available from the European Commission and once it has had the opportunity to consider how other platforms have approached this process.

The methodology comprises six key phases, all of which have been reviewed, and repeated for Year 2.

Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
Systemic risk scoping and definition	Risk identification	Inherent risk analysis and assessment	Identification and assessment of mitigations	Calculation of residual risk and analysis of future improvements	Reporting, governance, and review

Phase 1: Systemic risk scoping and definition

TikTok has developed and maintained a systemic risk taxonomy that outlines the scope of the systemic risks that could stem from the design, functioning, use or mis-use of its Platform. In Year 1, TikTok developed a process for scoping and defining risks based on an analysis of three key elements:

- The four primary systemic risk categories, as described in the DSA;
- A legal and textual interpretation of key DSA terms, other relevant EU law and jurisprudence, and relevant international legal instruments; and
- A functional interpretation made by TikTok’s internal teams.

In Year 1, this process informed the development of a systemic risk taxonomy that included twelve distinct Risk Modules, grouped under four primary categories. In Year 2, following careful review for any developments that would require a change to the taxonomy, TikTok decided that it remained appropriate to use the Year 1 taxonomy without amendment. Therefore, as in Year 1, the four primary categories are **'Illegal Content'**, **'Youth Safety'**, **'Civic Integrity'**, and **'Fundamental Rights'**, which are assessed in 12 distinct modules to ensure a comprehensive analysis of the risks and mitigations (**'Age Appropriate Content'** and **'Online Engagement'** have been summarised together in this Report due to their overlapping risks and corresponding mitigations, and **'Age Assurance'** is reported on separately).

Phase 2: Risk identification

Phase 2 involves identifying the specific ways in which each risk area may manifest on TikTok. TikTok reviewed the risks identified in Year 1 and confirmed that they remain applicable in Year 2 and updated and improved the detail and specificity of risk identification within each module in Year 2. TikTok also identified any emerging or new risks.

This risk identification is informed by 3 key processes:

- Consideration of internal sources (e.g. data insights from content moderation practices);
- Consideration of external sources (e.g. stakeholder engagement, law enforcement reports, and credible independent research); and
- An assessment of potential emerging risks (e.g. proactive risk assessment, internal research, stakeholder engagement).

In identifying how the systemic risk may manifest on the Platform, TikTok assesses how it is influenced by the factors laid out in Article 34(2) DSA and other relevant recitals. Article 34(2) instructs platforms to assess 'the design of recommender systems and any other relevant algorithmic system; content moderation systems; applicable terms and conditions (policies) and their enforcement; systems for selecting and presenting advertisements ('ads'); [and] data related practices.' Where the risks are localised or there are linguistic differences, TikTok takes these into consideration when assessing risk and corresponding mitigations.

In the Year 2 Risk Assessment, TikTok has also set a heightened focus on two horizontal priority areas that are of a cross-cutting nature across several risk areas. These are:

- The impact of an increased prevalence of AIGC; and
- The impact of recommender systems on the content that users consume.

TikTok has examined the 'potential for serious negative consequences to users' physical and mental well-being' (both of adults and minors), which is specified in Art. 34(1)(d), in the process of risk identification in all Risk Modules. A range of such risks were detected in relation to all Risk Modules, apart from Risks of Intellectual Property Infringing Content.

TikTok's selection of the priority considerations has been informed by internal risk detection analysis and an analysis of priorities expressed by stakeholders, including the European Commission.

Phase 3: Inherent risk analysis and assessment

Within the assessment of inherent risk, TikTok has considered severity and probability for each Risk Module, prior to considering the impact of any implemented mitigation measures.

To assess inherent risk, severity and probability are quantified and weighted before being multiplied against each other. To quantify each factor:

- **Severity** is quantified by assessing the scope (i.e. number of users or affected persons in Europe), scale (ranging from individual impacts to consequences to the broader society), the nature of potential impact (physical, psychological or otherwise), and impacted demographics (which may include vulnerable groups). This calculation of severity is then weighted to ensure that severity is accurately reflected and that lower scores in one or two factors do not overshadow high scores elsewhere; and
- **Probability** is quantified by assessing likelihood and historical volume. Likelihood is assessed based on factors such as whether (and how much) the risk has already been observed and whether it is expected to occur in the future. Historic volume is determined based on an assessment of historic occurrence on the Platform. Probability is therefore largely a measure of proxies; given that controls are in place, the probability of events occurring without any controls in place cannot be directly measured.

The calculated scores for severity and probability are multiplied to produce an inherent risk score for each Risk Module on a scale which comprises - as it did in Year 1 - 1 (very low), 2 (low), 3 (moderate), 4 (material) and 5 (high).

Phase 4: Identification and assessment of mitigations

TikTok has identified the mitigation measures that it has implemented relevant to each Risk Module, with reference to the types of mitigations set out in Art. 35(1)(a)-(k) DSA, noting any other additional mitigations that have been implemented in Year 2. Mitigation measures (i.e. controls) are assessed for their design and operating effectiveness. An analysis of the effectiveness of the mitigation is then used to calculate residual risk.

Phase 5: Calculation of residual risk and analysis of future improvements

Residual risk represents the amount of risk that remains for each risk area after the impact of relevant controls/mitigations on reducing inherent risk is considered. It is calculated by evaluating inherent risk against the effectiveness of the mitigations for each Risk Module to produce a residual risk for each Risk Module on a scale which comprises 1 (very low), 2 (low), 3 (moderate), 4 (material) and 5 (high). This analysis generates areas for the enhancement of mitigations, which form the work plan for future improvements.

Phase 6: Reporting, governance, and review

Each Risk Module is then reviewed and approved by subject matter experts and leaders to ensure robust governance and strong accountability, as follows:

1. The Risk Module is first submitted to a specialist Risk Assessment Review Group for analysis and assessment. The Risk Assessment Review Groups' process, including the Compliance

function's input, is designed to ensure that all risks referred to in Article 34 are identified and properly reported on;

2. The results are then reported to the Online Safety Oversight Committee, TikTok's steering group of cross-functional leaders which has been in place throughout its DSA programme; and
3. Finally, the results are submitted to the Board of Directors of TikTok Technology Ireland for review and approval. This includes a decision from the Board on the prioritisation Tier of each risk.

C. RISK ENVIRONMENT

TikTok is one of a number of online platforms whose users face similar, but not identical risks. These risks often transcend the boundaries of individual platforms and involve complex interactions between online experiences and real-world events, facilitating personal expression and the receiving and imparting of information and ideas. This is a central value of the EU, a fundamental right, and a core pillar of European democracy and social life. Each platform is expected to operate on its own analysis of the appropriate balance between safety and fundamental rights for its users. TikTok takes an approach that prioritises caution in areas where risks are more likely to impact real-world harms, or where they may be more likely to impact Younger Users.

TikTok strives to only host content within the bounds of relevant laws and the rules it imposes on itself and its users. TikTok therefore aims, through its use of automated and human moderation measures, to proactively identify and remove content before it is seen by any user. However, operating within such complexity and considering the inherent risks of human expression, it is not possible to eliminate all risks, as mitigating one risk often increases another (e.g. severely restricting hate speech would magnify risks to freedom of expression). Therefore, TikTok aims for a balanced and proportionate approach to reducing risks continuously, in line with Recital 86 DSA.

While TikTok is not specifically aimed at minors or predominantly used by them, TikTok makes it a priority to address how risks may impact these users, please see sections 11 and 12 of this Report.

TikTok has a comprehensive crisis management plan to address unforeseen challenges. TikTok maintains a dedicated incident management team to address urgent issues and to contain and minimise harm. Additionally, TikTok is committed to learning from past incidents, adapting its strategies, and fortifying its Platform against future risks.

D. WHAT DOES TIKTOK DO TO COMBAT INTENTIONAL MANIPULATION OF THE SERVICE?

TikTok aims to foster an environment where genuine interactions and content thrive. However, bad actors may attempt to manipulate the service through deceptive behaviours like account impersonation, spam activity, and fake engagement. This can range from individual behaviour, such as creating a fake account, to sophisticated and coordinated influence operations.

To combat these issues, TikTok prohibits - and works to disrupt - intentional manipulation and inauthentic use of the Platform. This involves detecting and removing deceptive accounts by assessing technical signals [REDACTED]. [REDACTED]. TikTok's verified account badge helps protect users by providing a reliable indication of high-profile accounts.

Expert teams at TikTok focus on detecting, investigating, and disrupting covert influence operations.

[REDACTED] Suspicious accounts and behaviours are automatically removed, and inauthentic engagement signals are ignored in estimating account or video popularity. TikTok reports up to monthly on the number of accounts actioned under its policies against covert influence operations in its public transparency centre.

TikTok recognises that covert influence operations will continue to evolve and that networks may attempt to re-establish a presence on the Platform. As new deceptive behaviours emerge, TikTok is committed to evolving its response, strengthening enforcement capabilities, and publishing its findings for scrutiny and more effective cross-platform risk mitigation.

E. STAKEHOLDER ENGAGEMENTS

TikTok has actively engaged with a broad spectrum of stakeholders to strengthen its risk assessment and mitigation strategies under the DSA. This proactive approach is central to TikTok's commitment to user safety and the protection of fundamental rights on its Platform. Over the past year, TikTok has consulted with a diverse range of experts, including members of its European Safety Advisory Council (**ESAC**), policymakers, civil society organisations, and content creators. Through close collaboration with stakeholders across Member States, TikTok can better understand and respond to local contexts, ensuring that its safety initiatives are not only globally robust but also finely tuned to address specific regional concerns. These partnerships enable TikTok to combine its existing Trust and Safety teams' expertise with further local knowledge, creating more effective and targeted risk mitigations, including campaigns that support a safer online experience for different communities. These engagements have directly informed TikTok's development of mitigation measures addressing systemic risks and the emerging risks associated with AIGC.

Annex 4 highlights TikTok's comprehensive strategy for engaging with stakeholders, which has been integral to refining its safety policies and operations. It details how TikTok's collaborations with external experts have directly influenced the Platform's approach to identifying and mitigating risks. These efforts include targeted consultations on emerging threats, the creation of specialised advisory councils, and ongoing dialogues with civil society. By leveraging this extensive network of partnerships, TikTok has been able to develop and implement measures that are reasonable, proportionate, and effective in addressing the complex challenges associated with hosting user generated content ('UGC').

F. HOW TIKTOK COLLABORATES WITH DSA TRUSTED FLAGGERS

Trusted Flaggers have a key role to play in risk identification and TikTok welcomes the designations of three organisations as Trusted Flaggers under Article 22 DSA in the past few months. In the Year 1 Report, TikTok stated as a mitigation measure that it would 'engage with organisations when they are newly designated as 'trusted flaggers' by EU member states and undertake outreach to onboard such entities to ensure efficient and priority processing of such reports via TikTok's dedicated channels as part of the Trusted Flagger Engagement Strategy'. As a cross-risk year 2 mitigation, TikTok remains committed to this and continues to be ready to engage with such partners as they are designated. In the meantime, TikTok continues to take action on reports received from its Community Partner Channel and trusted subject matter experts.

5. AIGC, RECOMMENDER SYSTEMS AND RISKS TO USERS' PHYSICAL AND MENTAL WELL-BEING

Introduction

Through its risk assessment process, TikTok has identified that AIGC and recommender systems are two critical factors that pose potential risks across several systemic Risk Modules. In response, TikTok has - both pre-dating the DSA entering into force and since that time - adapted and implemented a series of mitigation strategies to address these risks comprehensively across the Platform. These measures are designed to safeguard the user experience, and ensure the Platform's resilience against emerging threats in these areas. Since these factors influence specific Risk Modules in unique ways, they are addressed in relevant sections throughout the Report.

TikTok acknowledges the importance of thoroughly assessing the risk of possible 'serious negative consequences to users' mental and physical well-being'¹. TikTok has evaluated these risks to both Younger Users and other users when identifying risks in all Risk Modules. TikTok has identified a range of such risks in Risk Modules concerning Illegal Content, Youth Safety and Civic Integrity and evaluated how such risks may arise through use of the Platform.

A. THE IMPACT OF INCREASED USE OF AIGC

Definition and cross-module risks

In its Year 1 Report, TikTok identified potential emerging risks associated with AIGC in the online platform ecosystem. TikTok assessed that potential risks could arise through the sharing and dissemination of deceptive - but extremely realistic - images, videos and audio, potentially making it more difficult to distinguish between fact and fiction on the Platform. TikTok also foresaw potential new opportunities for bad actors to create and disseminate illegal content (including image-based abuse). Since the last risk assessment, these risks have materialised online, as there has been significant growth in the number of providers of tools to produce AIGC and a rapid increase in the amount of AIGC created.

As AIGC tools can be used to create content that appears just the same as UGC, the use of such tools creates potential risks to the integrity and reliability of content hosted on TikTok. '**Deep fakes**' - extremely realistic photos, videos or voices of real people - have emerged as a particular area of concern, given their risk of harm to depicted individuals. Additionally, AIGC technologies can be used to facilitate the creation of illegal content (e.g. synthetic Child Sexual Abuse Material, '**CSAM**') and image-based sexual abuse.

AIGC tools are now widely available to create manipulated content at low cost and high speed. AIGC is not inherently harmful, but the increasing accessibility of these tools and the nascence of best practice mitigation strategies within them magnifies the chance of harmful content being created and disseminated.

¹ Article 34(1)(d) DSA.

Cross-risk mitigations

In Year 2, TikTok has continued to implement measures across both its UGC and its advertising features to mitigate potential risks associated with AIGC.

Regarding UGC, TikTok's Community Guidelines prohibit any AIGC that would violate its regular policies in respect of illegal and violative content. In its update to the Community Guidelines earlier this year, TikTok made changes to reflect recent developments in AIGC, labelling and to reinforce existing policies. For example, AIGC uploaded by users that portrays CSAM is prohibited in the same way that any CSAM is prohibited, regardless of how it is created. TikTok has also developed specific policies regarding the creation and dissemination of AIGC, including:

- Requiring users to disclose content that is either completely generated or significantly edited by AI and contains realistic-appearing scenes or people;
- Prohibiting AIGC that depicts realistic-appearing people under the age of 18;
- Prohibiting AIGC that depicts the likeness of adult private figures without their permission;
- Prohibiting misleading AIGC that appears to come from an authoritative source;
- Prohibiting misleading AIGC that depicts a crisis event; and
- Prohibiting misleading AIGC that depicts a public figure who is being degraded or harassed, engaging in criminal or anti-social behaviour, or being politically endorsed or condemned.

With regard to TikTok's advertising features, TikTok's regular UGC policies apply, alongside additional advertising-specific rules. TikTok has developed specific AIGC Advertising Policies, which prohibit ads:

- Containing AIGC or suspected AIGC in relation to Crisis Events or Elections;
- Containing speech fabricated deep fakes;
- Using AIGC or suspected AIGC to display events that did not happen, or mislead users on true/real events;
- Using AIGC or suspected AIGC to make it look like it comes from an authoritative source;
- That contain the likeness of adult private figures without their permission;
- That depict public figures in deceptive ways (e.g. AIGC falsely depicting an endorsement by a public figure is banned);
- Featuring AIGC political figures; and
- Which TikTok knows or suspects features AIGC imagery depicting people aged 18 or under.

TikTok is prepared to further iterate and improve these policies as AIGC trends and risks evolve, and enforcement capabilities evolve.

For the purposes of detecting and enforcing against the use of AIGC to create violative content in either a UGC or ads context, TikTok has implemented tools for users and advertisers to voluntarily label their content as 'AI-generated'. TikTok was the first platform to launch an AIGC labelling tool, which over 37 million creators have used since September 2023 worldwide. From Q3 2023 to Q2 2024, 1,668,376 unique users in the EU have labelled their content with the 'Creator labelled as AI-generated' label, accounting for 5,864,626 unique videos. Additionally, any AIGC that is made with TikTok's AI effects is automatically labelled. If users believe that content seen is in fact undisclosed AIGC, they can report it to TikTok, who may then remove the content if it is otherwise violative of its Community Guidelines regarding AIGC.

TikTok also works with industry groups and civil society to detect undisclosed AIGC. TikTok participates in the Coalition for Content Provenance and Authenticity ('C2PA'), an open technical standard and content provenance solution that can provide information in a piece of content's metadata about its origins and whether AIGC models were used to create or edit it. This technology helps TikTok to detect AIGC at scale when content with C2PA metadata is uploaded to the Platform. Through its use of C2PA, TikTok has been the first video sharing platform to automatically label AIGC that has been made on other platforms. From Q3 2023 to Q2 2024, 4,286,027 videos on TikTok were auto-labelled with the AIGC tag of 'AI-generated' in the EU.

TikTok is also developing improved internal AIGC detection tools to enhance the scale, reach and accuracy of its enforcement against violative AIGC. [REDACTED]

[REDACTED] TikTok's approach has been informed by research conducted by the Trust & Safety Product Policy and Issue Policy teams on academic, industry, and advocacy literature. TikTok has also consulted external bodies and subject matter experts, such as the Partnership on AI.³

B. RECOMMENDER SYSTEMS

Definition and cross-module risks

The main feature on TikTok that uses a recommender system is the For You Feed ('FYF').⁴ This feature is the primary means through which users consume video content on the Platform. The FYF uses a personalised recommendation system to help each user discover content, creators, and topics, while ensuring that content is interesting and relevant to each user. In determining what is recommended, the FYF's recommender system takes account of user interactions, content information, and user information.

Concentrated content⁵ is a relevant risk that has been widely documented by civil society.⁶ As recommender systems are designed to offer content based on a user's interests and historical engagement, they might inadvertently present users with a narrow range of content for extended periods of time. Certain types of concentrated content, though not violative of TikTok's Community Guidelines, may cause harm by inadvertently reinforcing a negative personal experience for some

[REDACTED]

³ A non-profit organisation which brings together companies and other stakeholders to establish common codes of practice across the industry.

⁴ TikTok has several other features which are powered by its recommender system, as outlined [here:https://support.tiktok.com/en/using-tiktok/exploring-videos/how-tiktok-recommends-content](https://support.tiktok.com/en/using-tiktok/exploring-videos/how-tiktok-recommends-content).

⁵ Note that this was referred to as 'filter bubbles' in the Year 1 Report.

⁶ <https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content>

viewers. For example, there may be an impact to mental well-being, particularly for Younger Users, associated with concentrated content relating to extreme dieting and body-image-related content.

Given that the type of content that may have a potential negative impact when viewed in high concentrations is not illegal or violative, it is challenging to assess and detect the impact or the appropriate mitigation measures. There are also risks to freedom of expression associated with acting against concentrated content that has a negative impact but which, in isolation, does not violate TikTok's Community Guidelines (e.g. extreme dieting or fitness).

Cross-risk mitigations

TikTok's enforcement strategies mitigate against the risk of amplifying illegal and harmful content through recommender systems. First, TikTok proactively enforces its Community Guidelines against illegal and harmful content, preventing the most egregious content from appearing on users' FYFs. Second, TikTok maintains content eligibility standards for FYF that prioritise safety and are informed by the diversity of community and cultural norms. For example, unverified information awaiting fact-checker review is ineligible for FYF. Similarly, TikTok ensures that content that directly attacks individuals or groups with protected attributes is not eligible to appear in its FYF. Third, TikTok integrates age-appropriate design into FYF; as content created by anyone under 16 years old is ineligible for recommendation and its content classification system rates UGC based on maturity and prevents inappropriate themes from being recommended to Younger Users. Fourth, to mitigate additional risks associated with viral content, TikTok manually reviews all content that reaches a certain level of popularity (by monitoring the volume of video views).

To mitigate risks from concentrated content, TikTok uses dispersion techniques in its FYF. This involves using machine learning models to avoid recommending a series of similar videos on themes that do not violate TikTok's Community Guidelines but are potentially problematic if viewed repeatedly.

TikTok also offers users tools to empower users to influence what they see on their FYF. This includes tools to understand why videos have been recommended, to filter out certain keywords or hashtags to stop seeing certain content, and to 'refresh' their FYF to view a new set of popular content as if they were a new user. At the end of H1 2024, 2,109,240 users had actively filtered hashtags in the EU.

C. POSSIBLE SERIOUS NEGATIVE CONSEQUENCES TO USERS' PHYSICAL AND MENTAL WELL-BEING

TikTok sets out below how it has approached the systemic risk of 'serious negative consequences to the person's physical and mental wellbeing (Art. 34(1)(d) DSA). Given the absence of consensus among health professionals and academics on how social media use may adversely affect user health, TikTok engages experts and monitors associated risks. TikTok observes that these risks arise both from: (1) the risks from the consumption of violative or illegal content; and/or (2) as a result of use of the Platform itself, even though the content being consumed may not in itself be harmful. We set out below further details of how TikTok has analysed these risks.

Risks from the consumption of violative or illegal content

TikTok has addressed such impacts for all its users throughout this assessment, with reference to the areas of systemic risk in the Risk Module to which they relate. Each Risk Module has been

categorised to ensure that the risks to mental and/or physical health are clearly understood and mapped. For example, risks of mental harm, for example due to exposure to radicalising Hate Speech content, are considered in the Risk Module concerning Hate Speech. Risks of physical harm, for example relying on false medical information, are discussed in the Medical Misinformation Risk Module. TikTok also applies extra measures to protect Younger Users from the risks of such content which are considered in Section 11 below.

Risks arising from use of the Platform itself

While TikTok does not believe that use of platforms (in general or of TikTok in particular), is automatically harmful in and of itself, it recognises that some users (in particular Younger Users) could at times and in specific circumstances face serious negative mental and physical health impacts as a result of the way in which they use such platforms.

The nature and causes of such impacts are complex and highly context-dependent, and frequently interact with (or arise from) other areas of systemic risk. While these risks apply to users generally, they are particularly impactful when it comes to Younger Users. TikTok has therefore decided to address such impacts in the Risk Modules on Younger Users. For example, TikTok's analysis of the potential impacts of increased screen time and social media use for Younger Users, which stems from its awareness of the particular challenges youth can face around self-regulation and impulse control, forms part of the Online Engagement Risk Module. TikTok sets out in that Module which mitigations also apply, differentiating between those available to all users and which apply to Younger Users. For example, TikTok offers screentime controls to both adults and Younger Users in order to facilitate the management of time spent on the platform. These controls are turned on by default for Younger Users and also form part of the suite of controls in TikTok's Family Pairing feature. TikTok recognises that the impact of screentime and social media use is a rapidly developing area of research. In particular, TikTok monitors the study of potential links between high levels of screentime and social media use and negative mental health outcomes. In parallel, TikTok is continuing to study potential risk factors in this area and is closely monitoring research developments in order to inform its analysis (including, in the context of Digital Services Act compliance, the European Commission's assessment of this topic).

As its understanding evolves, TikTok's assessment of the risks relating to physical and mental wellbeing will also evolve in order to ensure that appropriate steps are taken to continue to provide a safe environment for all TikTok users.

6. ILLEGAL CONTENT: RISKS OF ILLEGAL HATE SPEECH CONTENT

1. Description of the risk:

a. Description of the risk from Year 1

- TikTok understands the term '**Hate Speech**' in a manner consistent with EU and EU member state laws, in particular, having regard to EU Framework Decision 2008/913/JHA, Article 1, in respect of offences concerning racism and xenophobia (i.e., against a group of persons or a member of such a group defined by reference to sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a

national minority, property, birth, disability, age or sexual orientation);

- The risks associated with Hate Speech may include users attempting to share or disseminate the following content including through video, livestream, comments and in profile information on the Platform:
 - Content claiming individuals or groups with protected attributes are physically, mentally, or morally inferior or referring to them as criminals, animals, inanimate objects, or other non-human entities;
 - Content promoting or justifying violence, exclusion, segregation, or discrimination against them;
 - Content that includes the use of slurs against others (and not the user themselves);
 - Content that targets transgender or non-binary individuals through misgendering or deadnaming; or
 - Content that depicts harm inflicted upon an individual or a group on the basis of a protected attribute.
- TikTok also appreciates that Hate Speech is in direct violation of the principles of liberty, democracy, respect for human rights and fundamental freedoms and, specifically, the right to non-discrimination as outlined in Article 21 of the EU Charter of Fundamental Rights (the 'Charter').

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, TikTok classified Hate Speech as a **Tier 2 priority**.

Summary of risks identified in Year 1

TikTok identified the following risks related to Hate Speech in its Year 1 Risk Assessment:

- **Sharing or dissemination of Hate Speech:** The risk of users attempting to upload, stream or otherwise disseminate illegal hate speech on the Platform;
- **Content moderation systems:** The risk that content moderation systems may not keep pace with rapidly evolving alternative vocabulary can make hate speech difficult to effectively detect and remove from the Platform. However, TikTok notes that this risk must be balanced with the risk that over enforcement may create risks of harm to freedom of expression;
- **Ads:** The risk of advertisers attempting to disseminate Hate Speech on the Platform; and
- **Regional and linguistic complexities:** The risk that Hate Speech can be highly localised in terms of language and region and so, when combined with the risk of a rapidly evolving vocabulary, creates risks to the effectiveness of TikTok's policy coverage and moderation efforts.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of Hate Speech:

- **Evolving Hate Speech content-related risks:** New risks related to trends in Hate Speech online, particularly marginalising speech, or content that indirectly demeans protected groups, such as anti-migrant hate, transphobic, and homophobic content. In the wake of the Israel-Hamas war, there has been a global increase in antisemitism, Islamophobia, and

anti-Arab hate online. As of April 2024, TikTok has removed more than 3.1 million videos and suspended over 140,000 livestreams in Israel and Palestine for violating Community Guidelines, including content related to promoting Hamas, hate speech, violent extremism, and misinformation;

- **Risk of AIGC containing slurs or promoting hateful ideologies:** AIGC poses significant challenges to current Hate Speech detection systems. Certain third-party AI tools may be used to generate content with slurs or hateful ideologies that are subtly altered to evade detection by both human and automated moderation systems. This includes fake videos, audio and memes that spread Hate Speech or extremist content under the guise of humour, often using coded language or symbols. For more detailed information on TikTok’s approach to mitigating risks associated with AIGC, please refer to *Priority Considerations Across all Modules*.

c. Inherent Risk in Year 2

For a detailed analysis on how TikTok assessed the baseline severity and probability for Hate Speech in Year 1, please refer to TikTok’s Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of Hate Speech to be ‘Material’. TikTok attributes this to intensified global conflicts (e.g. Russia and Ukraine and Israel and Hamas) and challenges associated with potential uses of AIGC, which have amplified the potential for Hate Speech content to adversely affect the individuals or groups it targets. TikTok assesses the probability of Hate Speech to be ‘Likely’ in Year 2. Taking this into account TikTok has assessed the inherent risk (that is, the risk without any mitigations in place) of Hate Speech, being shared on the Platform in Year 2 to be ‘Material’.

2. Mitigation Measures:

For detailed information on TikTok’s system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-K	DESCRIPTION
(b) Adaptation of terms and conditions	In addition to the mitigation effectiveness improvements anticipated in the Year 1 Report, TikTok has launched and enforced new Community Guidelines on marginalising speech in order to prevent the dissemination of content that indirectly disparages protected groups. TikTok has designated content that uses stereotypes, insinuation, or indirect statements that may implicitly demean protected groups as 'Feed Ineligible' to strike a balance between mitigating harmful content and allowing free speech. This means the content may remain

	<p>accessible but will not appear on the FYF. Enforcement of this policy is driven by internal teams with subject matter expertise.</p> <p>In its Community Guidelines on Hate Speech and Hateful Behaviours (which includes Hate Speech & Hateful Ideologies, Slurs, Negative Stereotypes and Generalisations, and Disparaging Religion), TikTok has included more examples of protected attributes and groups and has reviewed the examples of prohibited behaviour provided to ensure they are balanced across protected groups.</p>
<p>(c) Adaptation of content moderation processes</p>	<p>In addition to the mitigation effectiveness improvements anticipated in the Year 1 Report, TikTok has delivered a new programme to ensure that all content moderators complete training on anti-bias awareness. This training aims to increase awareness of potential biases in decision-making and equip moderators to recognise and mitigate bias effectively, and was developed with input from anti-bias experts and reviewed by advocacy groups.</p> <p>TikTok also expanded its specialised process for Hate Speech video moderation to enhance enforcement following updates to the Hate Speech policy. This launch includes hate policy specialist moderators who possess relevant language skills and cultural context, who are equipped with moderation technology providing full watch and listen capabilities. This expansion aims to improve overall enforcement quality.</p>  <p>Finally, at the end of 2023 TikTok developed a global policy reminder⁷ related to Antisemitism and Islamophobia. The complexity and evolving nature of Hate Speech may result in new trends, especially in the wake of a conflict situation. This policy reminder ensures that moderators are able to identify and remove Hate Speech content in the wake of the Israel-Hamas war.</p>
<p>(d) Adaptation of algorithmic systems</p>	<p>In addition to the mitigation effectiveness improvements anticipated in the Year 1 Report, TikTok has launched two additional automated moderation models to detect and designate Hate Speech content as ineligible for the FYF, which means that the content will not appear in</p>

⁷ Policy Reminders are documents with refreshed and updated policy information for moderators. These are usually created in response to a crisis event (such as the Israel-Hamas war) or in preparation for an event where TikTok may anticipate specific policy issues (such as Hate Speech before the Euros football). They do not change the policy but rather provide detailed guidance in the instance it is needed.

	<p>the FYF but is discoverable if a user searches for it specifically, to correspond to the new Community Guidelines on Marginalising Speech.</p>
<p>(f) Reinforcing risk detection measures</p>	<p>In its Year 1 Report TikTok stated that it would ‘expand its collaboration with external partners in order to develop enhanced intelligence-gathering capacities in relation to Hate Speech. This should assist in early detection of emerging and new forms of Hate Speech.’</p> <p>[REDACTED]</p> <p>Trust & Safety teams are dedicated to proactive forecasting and detection of risks, including Hate Speech, leveraging regional and subject matter experts and [REDACTED], news alerts, in depth trend reports, and social listening. Every year, TikTok convenes cross-functional teams to identify and mitigate Hate Speech related risks that may arise around events such as Pride, Black History Month, Holocaust Remembrance, and sports events such as the Six Nations.</p>
<p>(g) Initiating or adjusting cooperation with trusted flaggers in accordance with Article 22 and the implementation of the decisions of out-of-court dispute settlement bodies pursuant to Article 21</p>	<p>As stated in Year 1, TikTok is a committed signatory of the EU Code of Conduct on Countering Illegal Hate Speech Online (the 'Hate Speech Code'), and will continue its engagement with the European Commission and key stakeholders as part of this important initiative to combat online Hate Speech.</p> <p>TikTok remains ready to engage with EU 'trusted flaggers' as they continue to be designated (read section 5) and in the meantime, TikTok continues to receive reports of Hate Speech from Community Partners, composed of advocacy groups and experts.</p>
<p>(h) Initiating or adjusting cooperation with other providers of online platforms or of online search engines through the codes of conduct and the crisis protocols referred to in Articles 45 and 48 respectively</p>	<p>As noted in Year 1, TikTok has continued to participate as a signatory to the Hate Speech Code. In collaboration with the European Commission and other participating platforms, TikTok has been actively involved in the revision and improvement of the Hate Speech Code since 2020. TikTok has committed to sign up to the revised Hate Speech Code, which is intended to be adopted under Article 45 DSA later this year.</p> <p>As part of the new revised Hate Speech Code, TikTok has committed to significant new obligations. These include a commitment to review 50% of valid illegal Hate Speech user notices within 24 hours, demonstrating TikTok's dedication to swiftly addressing illegal content on the Platform.</p> <p>For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating the risk of Hate Speech Content, please refer to Annex 4.</p>

(i) Taking awareness-raising measures and adapting their online interface in order to give recipients of the service more information

In its Year 1 Report, TikTok stated that it would 'build on the success of media literacy campaigns, such as the #SwipeOutHate campaigns, and continue such initiatives and consider further media literacy measures to generate awareness of issues relating to illegal content and of the available safety tools.'

TikTok has signed a new [partnership](#) with the Six Nations rugby tournament to continue to run Swipe Out Hate campaigns including for the 2024 editions of the tournament. TikTok ran these campaigns to educate TikTok's users that TikTok does not tolerate hate speech and encourage its users to report hateful content during both the men's and women's Six Nations tournaments.

TikTok has continued to deliver such initiatives and media literacy measures to generate awareness of issues relating to illegal content and of the available safety tools. TikTok has reviewed its communication to users when they violate Hate Speech policies, to help users better understand when and why content represents a Hate Speech violation.

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating the risk of Hate Speech Content, please refer to Annex 4.

3. Residual Risk in Year 2:

Following its assessment of the effectiveness, reasonableness and proportionality of relevant mitigations, TikTok has assessed the residual risk of Hate Speech to be Moderate in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate Hate Speech on the Platform.

This assessment was conducted on the basis of qualitative and quantitative insights, including data on TikTok's enforcement of its Community Guidelines in the EU for the period of Q3 2023 to Q2 2024. Under the Community Guidelines policy of 'Hate Speech & Hateful Behaviour', TikTok removed 1,083,469 total videos in the EU, with 898,809 detected and removed proactively, and 616,327 removed without any views. This high proportion of proactive removal suggests that TikTok's content moderation systems remain effective in detecting and removing Hate Speech on the Platform, helping to mitigate the systemic risk.⁸

This assessment of the effectiveness of TikTok's proactive detection systems is further informed by the following data on user reports, which serve as a useful indication of the volume of Hate Speech that is not proactively detected and removed by TikTok. User reports accounted for a minority of Hate Speech

⁸ The percentage of video removals with zero views decreased throughout 2023 across all policy titles as a result of how TikTok prioritises content for human reviewers. TikTok's approach aims to minimise views of violative content, prioritise expeditious removal of especially egregious content (e.g. CSAM and violent extremism), and ensure consistency and fairness in enforcement. Other indicators such as total content removals reported may be substantively different to Year 1 as TikTok has enhanced its data integration practices by developing CG policy titles that are more specific to systemic risks, localising data to the EU, and regularising an annual reporting period.

detected and removed by TikTok, at 184,660 of the 1,083,469 total videos detected and removed under 'Hate Speech & Behaviour'. TikTok remains committed to continuously working to improve and iterate its proactive detection capabilities, while noting that Hate Speech is a rapidly evolving risk area, with nuance across regions and languages.

TikTok's user reporting tools are an important component of its content moderation process, ensuring that Hate Speech is detected and removed, including new and evolving forms of Hate Speech. The data demonstrates that when TikTok receives user reports of violative Hate Speech, it is actioned efficiently. Under 'Hate Speech & Hateful Behaviour', 144,394 of the 184,660 total videos removed following user reports were removed within two hours of receiving the report. Collectively, TikTok's proactive detection and reporting tools have worked effectively to ensure that 864,390 of the 1,083,469 total videos removed under 'Hate Speech & Hateful Behaviour' were removed within 24 hours of upload.

The effectiveness, reasonableness and proportionality of TikTok's moderation of Hate Speech is indicated by its appeals data, as users can submit appeals against content removals if they believe TikTok has made a mistake. Under the 'Hate Speech & Hateful Behaviour' policy, only 136,526 of the 1,083,469 total videos removed were successfully appealed and reversed.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4 This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok's strategies in these areas.

5. Prioritisation:

When considered in the context of world events, the prevention of the dissemination of Hate Speech content and the mitigation of the evolving risks to users and society remain a key priority for TikTok. TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual risk discussed above. TikTok plans to devote extra resources to combating Hate Speech in the year ahead and as a result it remains a Tier 2 risk. As noted in Year 1, TikTok will continue to keep its prioritisation under review and remains a committed signatory to the revised Hate Speech Code.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(b), Policy review:** TikTok plans to execute a full Policy review to relaunch its Hateful Behaviour Policies to support all moderation teams. [REDACTED]

- [REDACTED]

- **Article 35 (1)(c), Explore better model usage on content detection:** [REDACTED]

[REDACTED] and

- **Article 35 (1)(i), Roll out additional features to empower users against hate:** [REDACTED]

7. ILLEGAL CONTENT: RISKS OF TERRORIST CONTENT

1. Description of the risk:

a. Description of the risk from Year 1

- TikTok understands and interprets the term '**Terrorist Content**' in a manner consistent with EU and EU member state law, and its scope has been defined, in particular, having regard to Regulation (EU) 2021/784 on addressing the dissemination of Terrorist Content Online ('**TCO Regulation**'), and Directive 2017/541 (EU), Articles 3 to 12. Those laws outline the following illegal activities: (i) Terrorist offences; (ii) Offences relating to a terrorist group; (iii) Public provocation to commit a terrorist offence; (iv) Recruitment for terrorism; (v) Providing training for terrorism; (vi) Receiving training for terrorism; (vii) Travelling for the purpose of terrorism; (viii) Organising or otherwise facilitating travelling for the purpose of terrorism; (ix) Terrorist financing; and (x) Other offences related to terrorist activities;
- The risk may arise from users attempting to share or disseminate the following content on or through the Platform including through video, livestream, comments and in profile information:
 - Content that praises, promotes, glorifies, or supports violent acts or extremist organisations or individuals, or content seeking to raise funds or obtain assistance for such entities;
 - Content that encourages participation in, or intends to recruit individuals to, violent extremist organisations; and/or
 - Content with names, symbols, logos, flags, slogans, uniforms, gestures, salutes, illustrations, portraits, songs, music, lyrics, or other objects meant to represent violent extremist organisations or individuals.
- TikTok also appreciates that Terrorist Content is in direct violation of the right to life, liberty and security Art. 2 of the Charter which enshrines the right to life, and that Art. 6 protects the right to liberty and security, and that such rights may be undermined by the dissemination of Terrorist Content on the Platform. TikTok also recognises that efforts to moderate Terrorist Content must be accurate, balanced, and reasonable to ensure that such efforts do not disproportionately impact on other fundamental rights under the Charter, in particular the rights to freedom of expression and information, data protection and non-discrimination, and freedom of thought, conscience and religion, as well as TikTok's freedom to conduct a business.

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, TikTok classified Terrorist Content as a **Tier 2 priority**.

This section summarises additional risks identified in Year 2. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk detection processes, and extensive consultations with external stakeholders.

Summary of risks identified in Year 1

TikTok identified the following inherent risks as part of its Year 1 Risk Assessment:

- **Exposure to illegal, violent, and offensive content:** The risk that users' exposure to harmful content can have negative effects on mental health and potentially lead to desensitisation to violence. Some users that are susceptible to mental illness may also be vulnerable to radicalisation pathways;
- **Content moderation evasion tactics:** The risk that terrorist and violent extremist entities routinely evolve the way they communicate online, including through otherwise benign symbols and vocabulary. There is a risk that TikTok's content moderation systems may not be familiar with vocabulary related to Terrorist Content where it is rapidly evolving; and
- **Intentional manipulation of TikTok via impersonation:** The risk that users can manipulate their account to pose as another individual in a deceptive manner. Impersonation may enable terrorist organisations to create false accounts that appear trustworthy which can result in the recruitment and promotion of radical or extremist beliefs.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of Terrorist Content:

- **Risks related to the evolving nature of terrorist use of UGC platforms:** TikTok continues to monitor and evaluate the risks identified in its Year 1 Risk Assessment. In Year 2, TikTok identified several priority areas reflective of the dynamic and evolving nature of how terrorist and violent extremist entities innovate to exploit UGC-platforms. Notably, the Hamas terror attacks of 7 October 2023 and the resulting Israel-Hamas war were significant events that shaped these priorities. As of April 2024, TikTok has removed more than 3.1 million videos and suspended over 140,000 livestreams in Israel and Palestine for violating Community Guidelines, including content related to promoting Hamas, hate speech, violent extremism, and misinformation. Over the past year, TikTok saw a rise in prevalence of Osama Bin Laden's manifesto, 'Letter to America' (the '**Letter**'). Elements of the manifesto included the dissemination of violent extremist content, discussing the conflict in a manner that initially evaded content moderation. This has highlighted the risk of not all violent extremist manifestos being automatically detected and removed, particularly when parts of the manifesto are used to discuss non-violative topics. The promotion and discussion of the Letter on TikTok was problematic, as it could be mis-used to further extremist agendas, promote violence, or incite fear and hatred without necessary context and condemnation. Discussions on sensitive issues related to Terrorist Content must be handled responsibly, and the uploader must provide counter-speech, condemnation, or awareness of the harms of terrorism;
- **Evolving use of multiple platform features to share violative content:** TikTok has identified a growing trend of users exploiting account biographies and profile information to disseminate

content that violates its Community Guidelines on violent extremism. Violative behaviour among violent extremist users is not limited to a single type of content, but is spread across various platform features. Notably, these users frequently use videos, comments, and user profiles to propagate their messages. This issue has been highlighted through feedback from experts, escalations concerning terrorist and violent extremist activities, and an internal analysis of policy breaches; and

- **Advertising policies:** Prior to the Israel-Hamas war, TikTok experienced a low incidence of content in ads related to terrorism or terrorist organisations. Although this remains uncommon, there has been an increase in ads with humanitarian messages related to the conflict. TikTok has recognised a potential concern regarding ads that could reference terrorism or related activities, potentially seeking to exploit the situation. Moreover, there has been a rise in ads requesting donations, underscoring the need for thorough verification processes. While neither ads with humanitarian messages nor ads requesting donations are inherently risky, TikTok's existing advertising policies were nevertheless reassessed to ensure the platform's policies adequately address these complexities without suppressing calls for humanitarian aid and support.

c. Inherent Risk in Year 2

For a detailed analysis on how TikTok assessed the baseline severity and probability for Terrorist Content in Year 1, please refer to TikTok's Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of Terrorist Content to be 'Material' and the overall probability to be 'Possible' in Year 2. Taking this into account, TikTok assesses the inherent risk (that is, the risk without any mitigations in place) (that is, the risk without any mitigations in place) for Terrorist Content to be 'Moderate'.

2. Mitigation Measures:

For detailed information on TikTok's system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-K	DESCRIPTION
(b) Adaptation of terms and conditions	<p>TikTok updated its Community Guidelines to include the 'sharing of manifestos' of violent extremist entities as an important example of content not allowed on the Platform.</p> <p>TikTok engaged its Safety Advisory Councils across the globe, as well as experts in antisemitism and Islamophobia to ensure</p>

	<p>[REDACTED]</p> <p>TikTok continues to monitor the conflict closely and will continue to adapt moderator guidance and enforcement processes as needed to address evolving issues.</p> <p>TikTok has also updated its moderator guidance across violent extremism, hateful behaviour, illegal activities and regulated goods multiple times since the start of the conflict to ensure TikTok moderators are as well-equipped as possible to identify and action violating content related to the conflict and that enforcement reflects TikTok’s Community Guidelines. TikTok reviewed existing moderator guidance on violent and hateful organisations and individuals to ensure entities violating Community Guidelines are promptly and thoroughly removed from the Platform. TikTok worked closely with external partners to update moderator guidance to ensure alignment with enforcement approaches. TikTok also convened the first issue-specific Safety Advisory Council meeting in December 2023 to focus on responses to the Israel-Hamas war.</p>
<p>(d) Adaptation of algorithmic systems</p>	<p>TikTok has been working on improving its system that automatically removes content that violates TikTok’s violent extremism policies, or directs it for human review, if the content appears violating but is below a certain confidence threshold. This allows TikTok to reduce the reach of violating content. For example, data on TikTok’s enforcement of its Community Guidelines in the EU shows that TikTok removed 247,114 videos without any views under its ‘Violent & Hateful Orgs & Individuals’ policy, and 61,194 videos without any views under its ‘Violent Behaviours & Criminal Activities’ policy between Q3 2023 and Q2 2024.</p>
<p>(e) Adaptation of advertising systems</p>	<p>On 16 October 2023, TikTok updated its advertising policies, permitting only recognised NGOs to advertise fundraising initiatives related to the Israel-Hamas war. These NGOs must verify their identity before they are able to place ads on TikTok. Ads must not reference terrorist organisations or violate any other existing Community Guidelines or advertising policies.</p>

	<p>In February 2024, TikTok updated its advertising policies to allow ads to advocate for victims related to the Israel-Hamas war, provided that these ads do not violate other advertising policies in doing so.</p>
<p>(f) Reinforcing risk detection measures</p>	<p>As mentioned in Year 1 as a key mitigation improvement, TikTok continues to monitor risks associated with Terrorist Content and improve its detection capabilities. In early 2024, TikTok updated its existing risk management processes to better identify trends in emerging risks to the platform.</p> <p>This process now systematically identifies, categorises, and prioritises online trends on a recurring basis. It also establishes oversight and accountability for mitigation efforts aimed at containing these trends and risks. This improved process ensures that both short-term strategies (such as escalation management) and long-term solutions (such as detection iteration) work together effectively.</p> <p>Since the revamp, TikTok has processed 57 high-risk trends, ensuring the removal of associated harmful content, making it undiscoverable through search, providing updated guidance to moderation teams, and updating proactive detection mechanisms. When a new trend under Terrorist Content policies is identified, relevant stakeholders are brought together to discuss mitigation strategies, understand the root cause of the content appearing, and implement short-term plans to control the trend and minimise user exposure.</p>
<p>(g) Initiating or adjusting cooperation with trusted flaggers in accordance with Article 22 and the implementation of the decisions of out-of-court dispute settlement bodies pursuant to Article 21</p>	<p>TikTok remains ready to engage with EU 'trusted flaggers' as they continue to be designated (read section 5) and in the meantime, TikTok continues to receive reports from Community Partners and subject matter experts.</p>
<p>Case study: Israel-Hamas war</p> <p>TikTok activated its crisis protocols immediately in response to the Hamas attack on 7 October 2023, swiftly establishing 24/7 control rooms and mobilising over 40,000 safety professionals worldwide. It</p>	

was clear from the start that content about the conflict would be visible to, and affect, users in Europe. Additional moderation resources in Arabic and Hebrew were deployed swiftly to prevent the spread of violent, hateful, or misleading content. TikTok's Crisis Management playbook, consisting of six phases of detection, triaging and readiness, activation, response, de-escalation and recovery, and closure guided the response. By 6:30 am (Irish time) on Saturday 7 October 2023, the detection phase was activated, and within minutes, the event was escalated for assessment.

Application of TikTok's Crisis Protocol to the Israel-Hamas war

By 7:25 am on 7 October, TikTok declared the situation a crisis and moved to the activation phase. By 7:30 am, a cross-functional response team was deployed, initiating strategies such as 24/7 moderation coverage in Arabic and Hebrew. This included creating dedicated groups to manage the crisis and additional reviews of uploads for violative content. A Crisis Committee, including senior leadership from the EMEA region, was established by 10:00 am, and senior global leadership was informed due to the conflict's severity and global impact.

TikTok's response was driven by its command centre, consisting of 200 members from 14 global teams, primarily operating out of Dublin, with members also located in Israel, the Middle East, and the USA. The command centre ensures Platform safety, coordinates crisis management across Trust & Safety teams, provides regular updates to senior leadership, and identifies and implements both short-term and long-term risk mitigation strategies. The team remains vigilant during any declared crisis, addressing fluctuations in content moderation and emerging risks to maintain the Platform's integrity during the ongoing Israel-Hamas war.

On 9 October 2023, based on the evolution of the crisis, TikTok decided to implement restrictions on content by suspending user-generated short-form video and LIVE content originating from Israel and Palestine from being eligible for recommendation on the For You Page for users in Europe. This was a preventative measure to mitigate the risk of violative LIVE and other video content (e.g. Hate Speech, Terrorist Content) being widely disseminated on the Platform. TikTok also introduced new restrictions on LIVE eligibility for users based in Israel and Palestine, to mitigate the risk of bad actors using the Platform to post violative content.

The de-escalation and recovery phase of any crisis involves daily reviews of information sources and emerging trends to monitor the crisis's evolution and TikTok's response. As the volume of violative content stabilises, the frequency of meetings and updates to senior leadership is reduced, but the cross-functional team continues monitoring for new risks. In the closure phase, a root cause analysis is conducted once the crisis impact on the Platform diminishes and metrics return to pre-crisis levels.

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness and proportionality of TikTok's mitigations relevant to the systemic risk of Terrorist Content, TikTok has assessed residual risk to be 'Moderate' in Year 2. TikTok commits to expand and invest in mitigation measures to identify and mitigate Terrorist Content on the Platform.

This assessment was conducted on the basis of qualitative and quantitative insights, including data on TikTok's enforcement of its Community Guidelines in the EU for the period of Q3 2023 to Q2 2024. Under the Community Guidelines policy of 'Violent & Hateful Orgs', TikTok removed 369,712 total videos in the EU, with 346,449 detected and removed proactively, and 247,114 removed without any views. Under the 'Violent Behaviours & Criminal Activities' policy, TikTok removed 163,006 total videos in the EU, with 138,569 detected and removed proactively, and 61,194 removed without any views. This high proportion of proactive removal suggests that TikTok's content moderation systems remain effective in detecting and removing Terrorist Content on the Platform, helping to mitigate the systemic risk.

This assessment of the effectiveness of TikTok's proactive detection systems is further informed by the following data on user reports, which serve as a useful indication of the volume of Terrorist Content that is not proactively detected and removed by TikTok. User reports accounted for a minority of Terrorist Content detected and removed by TikTok, at 23,263 of the 369,712 total videos detected and removed under 'Violent & Hateful Orgs & Individuals', and 24,437 of the 163,006 total videos removed under 'Violent Behaviours & Criminal Activities'. TikTok remains committed to continuously working to improve and iterate its proactive detection capabilities, while noting that Terrorist Content is a rapidly evolving risk area, with nuance across regions and languages.

TikTok's user reporting tools are an important component of its content moderation process, ensuring that Terrorist Content is detected and removed, including new and evolving forms of Terrorist Content. The data demonstrates that when TikTok receives user reports of violative Terrorist Content, it is actioned efficiently. Under 'Violent & Hateful Orgs & Individuals', 18,384 of the 23,263 total videos removed following user reports were removed within two hours of receiving the report. Similarly, under 'Violent Behaviours & Criminal Activities', 20,211 of the 24,437 total videos reported were removed within two hours. Collectively, TikTok's proactive detection and reporting tools have worked effectively to ensure that 300,163 of the 369,712 total videos removed under 'Violent & Hateful Orgs & Individuals' were removed within 24 hours of upload. Similarly, 123,204 of the 163,006 total videos removed under 'Violent Behaviours & Criminal Activities' were removed within 24 hours.

The effectiveness, reasonableness and proportionality of TikTok's moderation of Terrorist Content is indicated by its appeals data, as users can submit appeals against content removals if they believe TikTok has made a mistake. Under the 'Violent & Hateful Orgs & Individuals' policy, only 45,117 of the 369,712 total videos removed were successfully appealed and reversed. Similarly, under 'Violent Behaviours & Criminal Activities', only 17,450 of the 163,006 total videos removed were successfully appealed and reversed.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok's strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual

risk discussed above. TikTok plans to devote extra resources to combatting the risk of Terrorist Content in the year ahead and as a result it remains a Tier 2 risk. As noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(b), Enhanced policies:** TikTok plans to review and enhance four of its policies to better address terrorist and violent extremist content. [REDACTED]
- [REDACTED]
- **Article 35(1)(c), Content moderation:** [REDACTED]
- **Article 35(1)(e), Adapting advertising systems:** TikTok will continue to ensure the readiness of its human and automated moderation systems to prevent emerging and evolving Terrorist Content from being monetized on the Platform.

8. ILLEGAL CONTENT: RISKS OF INTELLECTUAL PROPERTY INFRINGING CONTENT

1. Description of the risk:

a. Description of the risk from Year 1

- TikTok understands the term '**IP-Infringing Content**' to mean content that is created and disseminated in breach of copyright or other intellectual property ('**IP**') rights;
- TikTok's approach to actioning IP-Infringing Content is structured in particular in accordance with relevant EU laws on copyright, including Directive (EU) 2019/790 (the '**Copyright Directive**');
- IP-Infringing Content includes the following content, whether uploaded as a video, livestream, or in profile information on the Platform:
 - Content that reproduces, and disseminates on the Platform, the original work of another person or entity without that person's or entity's permission and which does not fall in one of the copyright exceptions. This content may include music works and audio files, artistic works (e.g. photographs, paintings, drawings, and other original visual renderings), and audio-visual recordings;
 - Content that contains the unauthorised use of a trademark or service mark in connection with goods or services in a way that is likely to cause confusion, deception or mistake about the source, origin, sponsorship or affiliation of the associated goods and/or services; or
 - Content that advertises or promotes counterfeit products.
- In addition, TikTok also notes that Art. 17 of the Charter enshrines the right to protection of intellectual property.

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, TikTok classified IP-Infringing Content as a **Tier 3 priority** in its priority tiering system.

This section summarises additional risks identified in Year 2. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk detection processes, and extensive consultations with external stakeholders.

Summary of risks identified in Year 1

TikTok identified the following inherent risks as part of its Year 1 Risk Assessment:

- **Sharing and dissemination of IP-Infringing Content:** The risk that users disseminate IP-Infringing Content;
- **Terms of Service and Community Guidelines:** The risk that these rules may be misinterpreted by users (intentionally or unintentionally), leading to a risk of users uploading and disseminating IP-Infringing Content on the Platform;
- **Moderation of IP-Infringing Content:** The risk that TikTok may inaccurately remove content that the rightsholder has informed TikTok that it in fact wishes to remain online; and
- **Ads:** The risk that advertisers share ads content which disseminates IP-Infringing Content.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of IP-Infringing Content.

- **AIGC:** As third party AIGC technology becomes more accessible and sophisticated, TikTok has observed an increase in claims that AIGC posted by users is IP-Infringing Content. There are challenges with moderating this type of content, for example, it can be difficult for rights holders to demonstrate that their copyright has been infringed by AIGC content, and therefore difficult for moderators to make an assessment on infringement; and
- **Counterfeit goods:** The risk that counterfeit goods are marketed to users on the Platform, in violation of the Community Guidelines. Bad actors may provide instructions to users on TikTok directing them to third-party platforms where counterfeit goods can be purchased. Counterfeit goods may pose potential health and safety risks to consumers.

c. Inherent Risk in Year 2

For a detailed analysis on how TikTok assessed the baseline severity and probability for IP-Infringing Content in Year 1, please refer to TikTok’s Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of IP-Infringing Content has been assessed to be ‘Moderate’. TikTok assesses the probability of IP-Infringing Content in Year 2 to be ‘Possible’. Taking into account the severity and probability in Year 2, TikTok has assessed the inherent risk (that is, the risk without any mitigations in place) of IP-Infringing Content being shared on the Platform in Year 2 to be ‘Moderate’.

2. Mitigation measures:

For detailed information on TikTok’s system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-K	DESCRIPTION
(a) Adaptation of feature or platform design	TikTok has made changes to the way IP-Infringing Content can be manually discovered on the Platform by blocking specific search terms. When users search for certain terms, no results will be shown, and they will receive a message notifying them that unauthorised streaming of digital content and promotion of counterfeit goods violates TikTok’s Community Guidelines. 

	<p>[REDACTED]</p> <p>TikTok has optimised its external webforms and internal systems to streamline the process through which a rightsholder or reporter can request that future copies of reported and removed infringing videos are prevented from reappearing on the Platform. TikTok’s copyright webform has also been improved to reduce the incidence of rights holders or reporters providing insufficient evidence.</p>
<p>(b) Adaptation of terms and conditions</p>	<p>TikTok has launched a dedicated Music Terms of Service to help users understand how music can be legally used on the Platform.</p>
<p>(c) Adaptation of content moderation processes</p>	<p>TikTok has adapted its content moderation workflows for IP notice and takedown requests to mitigate potential risks associated with AIGC. This has included:</p> <ul style="list-style-type: none"> ● Establishing an internal guideline for processing IP reports against AIGC music; ● Revising standard operating procedures to instruct moderators on how to handle a piece of AIGC reported to TikTok for copyright or trademark infringement; ● Guidance has been provided to moderators to help them better identify when reports of IP-Infringement have been made against AIGC; and ● Moderators have been advised on scenarios involving claims against AIGC that require escalation to legal teams for further risk assessment. This includes grey area 'edge cases' that may be difficult for a moderator to assess without specialised legal knowledge. <p>TikTok will remain vigilant as intellectual property law evolves in response to AIGC and will implement updates to its policies where appropriate.</p>
<p>(e) Adaptation of advertising systems</p>	<p>TikTok has improved its automated counterfeit ad detection system [REDACTED]</p>
<p>(f) Reinforcing risk detection measures</p>	<p>As mentioned in the Year 1 Report as a key planned mitigations effectiveness improvement, TikTok has continued to further develop its internal tools to detect and prevent copyright infringing activities. These improvements include:</p> <ul style="list-style-type: none"> ● Based on the trends and characteristics of counterfeit goods sales ads, TikTok has provided regular training to moderators,

	<p>alongside regular improvements to the Platforms automated detection tools;</p> <ul style="list-style-type: none"> • TikTok has optimised the complaint path for IP rights holders, making it clearer and more accessible, and iterated the complaint interface on its web page to make it easier to understand. TikTok has also fixed some known bug issues to make the complaint-filing process smoother; and • TikTok has optimised its internal ticketing system for receiving and responding to IP-infringement reports, improving its efficiency by revising policies and workflow, and simplifying documentation. <p>Additionally, TikTok has provided more rightsholders with access to TikTok's copyright tools, which ensure the unavailability of protected works on the Platform where rightsholders have provided TikTok with the relevant and necessary information. Blocked content cannot be accessed by users in the region(s) in which it is blocked.</p>
<p>(g) Initiating or adjusting cooperation with trusted flaggers in accordance with Article 22 and the implementation of the decisions of out-of-court dispute settlement bodies pursuant to Article 21</p>	<p>TikTok remains ready to engage with EU 'trusted flaggers' as they continue to be designated (read section 5). TikTok treats reports from the Copyright Information and Anti-Piracy Centre ('CIAPC') with priority since February 2024. This was in advance of the Finnish non-profit appointed as a Trusted Flagger by the Finnish Digital Services Coordinator on 7 March 2024. To date CIAPC has not submitted any reports under Article 16 DSA.</p> <p>For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating the risk of Intellectual Property Infringing Content, please refer to Annex 4.</p>

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness and proportionality of TikTok's mitigations relevant to the systemic risk of IP-Infringing Content, TikTok has assessed residual risk to be 'Low' in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate IP-infringing Content on the Platform.

This assessment was conducted on the basis of qualitative and quantitative insights, including global data from TikTok's Intellectual Property Removal Requests Report¹⁰ for the period of 1 July to 31 December 2023. TikTok provides reporting tools to rights holders and their representatives through which they can notify TikTok of potential copyright or trademark infringements and, where appropriate, have the content taken down. In the reporting period detailed above, TikTok received 272,555 total

¹⁰ <https://www.tiktok.com/transparency/en/intellectual-property-removal-requests-2023-2/>.



copyright removal requests globally, with 56.1% assessed as successful. In the same period, TikTok received 29,188 trademark removal requests globally, with 57.4% assessed as successful.

This data indicates the volume of copyright or trademark violating content that is not detected and removed by TikTok's proactive systems. At 152,830 total successful copyright removal requests and 16,758 total successful trademark removal requests for the period 1 July to 31 December 2023, this represents an extremely small share of overall content uploaded and shared on the Platform. TikTok is committed to proactively preventing IP-infringing content from being uploaded to the Platform, and efficiently and diligently processing removal requests from appropriate parties.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok's strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual risk discussed above. TikTok plans to devote extra resources to combating IP-Infringing Content in the year ahead and as a result it remains a Tier 3 risk. As noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(i), Awareness raising measures:** TikTok is committed to improving its approach to mitigate IP-Infringing Content by creating additional resources to educate users who receive infringement notices and enhancing collaboration with industry groups to improve the reporting process for rights-holders;
- **Article 35(1)(c), Reviews of Guidelines:** In addition to current regular policy reviews, TikTok will implement a regular review for Community Guidelines updates and moderator guidance to ensure effectiveness and relevance;
- **Article 35(2)(e), Adapting advertising systems:** TikTok is planning to implement new stringent measures to identify and remove users who consistently violate the Platform's ad policies in respect of IP, including implementing enforcement actions against repeat offenders and fraudulent behaviours. TikTok also commits to regularly updating its internal ad policies and processes to reflect changes in industry standards and legal requirements regarding IP. Last, TikTok is planning to develop and maintain a process for expeditiously receiving and addressing rights holders' complaints regarding IP-infringing content in ads; and
- **Article 35 (1) (f), Processing illegal content orders:** Following development of a robust Intellectual Property system for high risk events scenario-planning and readiness, TikTok will continue to strengthen communication channels with rightsholders and cross-functional working groups within Trust and Safety.

9. ILLEGAL CONTENT: RISKS OF CHILD SEXUAL ABUSE MATERIAL AND CHILD SEXUAL EXPLOITATION AND ABUSE

1. Description of the risk:

a. Description of the risk from Year 1

- TikTok understands the terms child sexual abuse material ('**CSAM**'¹¹) and child sexual exploitation and abuse ('**CSEA**') in a manner consistent with EU and EU member state laws. The scope of CSAM and CSEA risks have been defined in particular having regard to Directive 2011/93/EU, Articles 3 to 7 respectively in relation to: offences concerning sexual abuse; offences concerning sexual exploitation; offences concerning child pornography; solicitation of children for sexual purposes; and incitement, aiding and abetting, and attempts to commit such offences.
- TikTok considers that the dissemination of CSAM content may involve users of the Platform attempting to, whether in relation to video, photo or livestream content or in relation to the abuse of account settings:
 - share, re-share or offer to trade or sell, or direct users of the Platform to obtain or distribute CSAM content;
 - generate and/or share self-generated CSAM;
 - disseminate content that depicts, solicits, glorifies, or encourages child abuse imagery including nudity, sexualised minors, or sexual activity with minors;
 - disseminate content that depicts, promotes, normalises, or glorifies pedophilia or the sexual assault of a minor; or
 - post information on their account profile, including their username, handle, profile picture and profile bio, that contains CSAM.
- TikTok considers that the activities associated with CSEA behaviour may involve adult users of the Platform attempting to:
 - build an emotional relationship with a minor in order to gain the minor's trust for the purposes of future or ongoing sexual contact, sexual abuse, trafficking, or other exploitation;
 - solicit real-world contact between a minor and an adult or between minors with a significant age difference;
 - solicit minors to connect with an adult on another online platform, website, or other digital space for sexual purposes;
 - solicitation of nude imagery or sexual contact, through blackmail or other means of coercion; or
 - post information on their account profile, including their username, handle, profile picture and profile bio that involves CSEA behaviour.
- TikTok notes that Article 34 of the United Nations Convention on the Rights of the Child enshrines the right of the child to protection from all forms of sexual exploitation and sexual abuse, and that such risks should be assessed with the best of interests of the child as the primary consideration, in accordance with Article 24 of the Charter. TikTok notes that in 2021,

¹¹ Although various legal texts still refer to the term 'child pornography', in line with the [Guidelines for the Protection of Children from Sexual Exploitation and Sexual Abuse \(known as the 'Luxembourg Guidelines'\)](#), TikTok considers CSAM to be the more appropriate term.

the United Nations Committee on the Rights of the Child underlined that these rights must be equally protected in the digital environment.¹²

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, TikTok classified CSAM and CSEA as a **Tier 1 priority** in its priority tiering system.

This section summarises additional risks identified in Year 2. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk detection processes, and extensive consultations with external stakeholders.

Summary of risks identified in Year 1

TikTok identified the following risks related to CSAM and CSEA in its Year 1 Risk Assessment:

- **Sharing and dissemination of CSAM:** The risk of users attempting to share and disseminate CSAM on the Platform; and
- **Carrying out or participating in CSEA:** The risk of users attempting to use the Platform to carry-out or participate in CSEA.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of CSAM and CSEA.

- **Sextortion and grooming:** Sextortion is a growing threat to Younger Users of online platforms across the globe, with data demonstrating a rise in abuse targeting teenage boys and young adult males¹³. Grooming also continues to be a risk to Younger Users of all genders. Sextortion and grooming can lead to online and offline risks to the safety and well-being of users and encourage the proliferation of CSAM production and distribution networks. The incidence of sextortion and grooming on TikTok is low, but these risks may be challenging to detect and moderate given the following factors:
 - **Subtle tactics:** Groomers build trust over time, making their conversation (in comments related to UGC) hard to distinguish from regular interactions. Automated systems often struggle with the nuanced and contextual nature of these conversations;
 - **Anonymity:** Predators frequently use fake profiles and multiple accounts to evade detection, and often employ mass communication tactics to reach more victims;
 - **Underreporting:** Victims may not report grooming or sextortion due to fear, shame, or lack of awareness of their ability to make such a report; and/or
 - **Abuse patterns:** TikTok has identified [REDACTED]
- **Difficulty in detecting covert and suspicious behavioural signals:** Online predators may conceal their behaviour so as to bypass moderation efforts by interacting with and searching

¹² [UN General Comment No. 25 \(2021\)](#) on Children's Rights in Relation to the Digital Environment.

¹³ Source: [Missingkids.org](#)

for content created by Younger Users without explicitly violating Community Guidelines. Online platforms could be exploited by adults with predatory intent, by connecting with like-minded adults via comments, likes, and follower lists, potentially forming 'offender communities'. TikTok's current systems have historically been targeted at content-level review, but their systems are now being enhanced and complemented to identify holistic account-level signals;

- **Increased volume of reporting:** As TikTok's detection systems can now identify more content that needs to be reviewed for potential reporting to the National Center for Missing and Exploited Children ('**NCMEC**'). In 2023, TikTok sent 590,376 reports to NCMEC, whereas in 2022 TikTok sent 288,125. This increase in flagged content requires a more focused approach, prioritising the most serious cases over a simple 'first in, first out' method, to ensure the swift handling of the most urgent matters. The increase in the volume of sensitive content also requires attention to ensure continued well-being of moderators through strong support systems. Central to TikTok's approach is accurate reporting and effective cooperation with law-enforcement, including requiring all NCMEC-related content to be reviewed by human moderators. This adds significant pressure on human staff and limits the use of automated processes. Although TikTok's enhanced detection has resulted in an increased volume of reporting, it remains low compared to other platforms;¹⁴
- **Ads containing child abuse cases:** Advertising systems may be mis-used to share ads featuring AIGC depictions of minors that were involved in real-life abuse or abduction cases. While these ads are removed under existing Youth Safety policies, this content is still treated as an emerging risk in Year 2.

c. Inherent Risk in Year 2

For a detailed analysis on how TikTok assessed the baseline severity and probability for CSAM and CSEA in Year 1, please refer to TikTok's Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of CSAM and CSEA to be 'Material' and the overall probability to be 'Possible' in Year 2. Taking this into account, TikTok assesses the inherent risk (that is, the risk without any mitigations in place) for CSAM and CSEA to be 'Moderate'.

Mitigation Measures:

Please refer to the Year 1 Risk Assessment Report for detailed information on TikTok's system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-K	DESCRIPTION

¹⁴ <https://www.missingkids.org/content/dam/missingkids/pdfs/2023-reports-by-esp.pdf>

<p>(a) Adaptation of feature or platform design</p>	<p>In addition to blocking searches associated with CSEA and CSAM, TikTok deploys in-app warnings and deterrence resources designed by CSEA prevention organisations¹⁵ to deter users from violative behaviour related to CSEA and CSAM.</p>
<p>(b) Adaptation of terms and conditions</p>	<p>TikTok's Community Guidelines have been updated to classify hyper-realistic AIGC sexual content which depicts minors as CSAM, ensuring its prompt removal and reporting of such content to NCMEC.</p>
<p>(c) Adaptation of content moderation processes</p>	<p>In addition to the mitigation effectiveness improvements anticipated in the Year 1 Report, TikTok has updated its AI-enabled detection system to identify and flag potential CSAM and CSEA content for review.</p> <p>[REDACTED]</p> <p>TikTok also uses several third-party CSAM detection systems and has expanded its scope of auto enforcement to reduce moderation exposure to explicit CSEA content.</p> <p>[REDACTED]</p>
<p>(e) Adaptation of advertising systems</p>	<p>TikTok has made a number of updates to the policies and enforcement actions that govern its ads, including:</p> <ul style="list-style-type: none"> ● Prohibition on Child Abuse References: Since January 2024, TikTok prohibits references to historical child abuse cases in ads and landing pages¹⁶ on Google Play or Apple App Store, including those made using AIGC in order to protect the dignity of victims. Exceptions are made for anti-child abuse awareness campaigns and certain movies/TV shows, documentaries/trailers, provided they comply with TikTok's Minor Safety Policies; and ● Misleading AIGC Policy: Launched in June 2024, this policy prohibits AIGC ads that depict minors, political figures, or authoritative sources. An exception is made for 'baby generator' ads, where AIGC images of minors are allowed, but the use of real photos or identities of minors is strictly prohibited. All content, including AIGC images, must adhere to TikTok's existing Minor Safety and Adult & Sexual Content policies; and ● AIGC app ads: Ads promoting AIGC creation tools must link to Google Play or Apple App Store pages, with the verification provided by those third parties. This mitigates the risk of bad actors promoting unregulated AIGC tools.

¹⁵ <https://www.stopitnow.org.uk/>

¹⁶ A landing page is a standalone page created specifically for an advertising campaign, designed to direct visitors towards a clear action such as downloading an app.

<p>(f) Reinforcing risk detection measures</p>	<p>TikTok’s Centralised Child Safety Team is composed of specialists from child safety backgrounds, including former prosecutors, government, intelligence, and law enforcement officials. This team and its internal partners handle the moderation and escalation of potential CSAM content. [REDACTED]</p> <p>Trend management tools are used and aim to provide timely and efficient solutions for escalation handling with comprehensive video recall conditions and precise strategy quality and lifecycle management.</p> <p>As mentioned in the Year 1 Report as a key mitigation improvement, TikTok has recently updated its specialised 'Computer Vision' model, an AI-powered detection system designed to identify visual material potentially containing CSAM and CSEA. This system directs flagged content to TikTok's Child Safety Team and for review and to report to NCMEC. [REDACTED]</p>
<p>(g) Initiating or adjusting cooperation with trusted flaggers in accordance with Article 22 and the implementation of the decisions of out-of-court dispute settlement bodies pursuant to Article 21</p>	<p>As mentioned in the Year 1 Report, TikTok remains ready to engage with EU 'trusted flaggers' as they continue to be designated (read section 5). TikTok continues to receive reports from Community Partners and subject matter experts, including a system for NGOs and other partner entities to report CSAM and CSEA for expedited decision making. This system is managed by TikTok’s Incident Management team and within the Law Enforcement channel. In the first half of 2024, 541 escalations took place from law enforcement authorities to the Child Safety Team for the EMEA (Europe, Middle East and Asia) region.</p>
<p>(i) Awareness raising measures</p>	<p>TikTok is committed to the ‘Voluntary Principles to Counter Child Sexual Exploitation and Abuse’ from the We Protect Global Alliance, of which it is a signatory in order to help build collaboration across platforms. TikTok’s active participation in the WeProtect Global Alliance fosters collaboration across sectors – including NGOs, safety firms, and governments – facilitating the sharing of knowledge and best practices, and strengthening the collective response to CSEA.</p> <p>For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating the risk of CSAM and CSEA, please refer to Annex 4.</p>

3. Residual Risk in Year 2:

[REDACTED]

Following its assessment of the effectiveness, reasonableness and proportionality of relevant mitigations, TikTok has assessed residual risk of CSAM and CSEA being shared on the Platform to be 'Moderate' in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate CSAM and CSEA on the Platform.

This assessment was conducted on the basis of qualitative and quantitative insights, including data on TikTok's enforcement of its Community Guidelines in the EU for the period of Q3 2023 to Q2 2024. Under the Community Guidelines policy of 'Safety & Civility - Youth Exploitation & Abuse', TikTok removed 1,541,762 total videos in the EU, with 1,491,028 detected and removed proactively, and 1,118,478 removed without any views. Under 'Youth Safety & Well-Being - Youth Exploitation & Abuse', TikTok removed 1,289,147 total videos in the EU, with 1,238,736 detected and removed proactively, and 866,929 removed without any views. This high proportion of proactive removal suggests that TikTok's content moderation systems remain effective in detecting and removing CSAM and CSEA on the Platform, helping to mitigate the systemic risk.

This assessment of the effectiveness of TikTok's proactive detection systems is further informed by the following data on user reports, which serve as a useful indication of the volume of CSAM and CSEA that is not proactively detected by TikTok's proactive systems. User reports accounted for a minority of CSAM and CSEA detected and removed by TikTok, at 50,734 of the 1,541,762 total videos detected and removed under 'Safety & Civility - Youth Exploitation & Abuse' and 50,411 of the 1,289,147 total videos detected and removed under 'Youth Safety & Well-Being - Youth Exploitation & Abuse'. TikTok remains committed to continuously working to improve and iterate its proactive detection capabilities.

TikTok's user reporting tools are an important component of its content moderation process, ensuring that CSAM and CSEA is detected and removed, including new and evolving forms of CSAM and CSEA. The data demonstrates that when TikTok receives user reports of CSAM and CSEA content, it is actioned efficiently. Under 'Safety & Civility - Youth Exploitation & Abuse', 43,616 of the 50,734 total videos removed following user reports were removed within two hours of receiving the report. Under 'Youth Safety & Well-Being - Youth Exploitation & Abuse', 43,290 of the 50,411 total videos removed following user reports were removed within two hours of receiving the report. TikTok's proactive detection and reporting tools have worked effectively to ensure that 1,355,823 videos were removed under TikTok's 'Safety and Civility - Youth Exploitation & Abuse' policy and 1,103,357 videos were removed under 'Youth Safety & Wellbeing - Youth Exploitation & Abuse' within 24 hours of upload.

The effectiveness of TikTok's moderation of CSAM and CSEA is indicated by its appeals data, as users can submit appeals against content removals if they believe TikTok has made a mistake. Under the 'Safety & Civility - Youth Exploitation & Abuse' policy, only 64,804 of the 1,541,762 total videos removed were successfully appealed and reversed. Under 'Youth Safety & Well-Being - Youth Exploitation & Abuse', only 27,552 of the 1,289,147 total videos removed were successfully appealed and reversed.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok's strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual risk discussed above. TikTok plans to devote extra resources to combatting CSAM and CSEA in the year ahead and as a result, CSAM and CSEA remains a Tier 1 risk. As noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(c),** [REDACTED]
[REDACTED] TikTok is enhancing its detection of CSAM by developing new models to better identify AIGC CSAM, [REDACTED]
- **Article 35(1)(c), Safety strategies targeting suspicious predatory behaviour:** [REDACTED]
- **Article 35(1)(f), Increase coverage of Third Party APIs and Keyword Lists:** [REDACTED]
- **Article 35(1)(c), Improve operational efficiencies to manage increased volume in reporting:** [REDACTED]
- **Article 35(1)(e), Adapting advertising systems:** TikTok commits to establishing a robust risk governance process for CSAM in relation to its ads products. TikTok will also continue to monitor the robustness of its ad policies with respect to CSAM to ensure that they cover emerging CSAM risks identified in internal risk assessments, including new forms of online abuse that could fall within the scope of CSAM. Last, TikTok will continue to provide training to relevant teams, helping them to identify violative content under any updated and enhanced ad policies regarding CSAM.

10. ILLEGAL CONTENT: RISKS OF GENDER-BASED VIOLENCE CONTENT

1. Description of the risk:

a. Description of the risk from Year 1

- TikTok understands that neither the DSA nor EU law contains a single or comprehensive

definition of gender-based violence. TikTok therefore understands gender-based violence to be the perpetration, support or incitement of 'any type of harm [...] against a person or group of people because of their factual or perceived sex, gender [...] and/or gender identity' ('GBV');¹⁸

- Directive (EU) 2024/1385 on combating violence against women and domestic violence defines 'violence against women' as 'all acts of gender-based violence directed against a woman or a girl because she is a woman or a girl or that affect women or girls disproportionately, that result in or are likely to result in physical, sexual, psychological or economic harm or suffering, including threats of such acts, coercion or arbitrary deprivation of liberty, whether occurring in public or in private life'. This definition is consistent with, yet narrower than, TikTok's interpretation of GBV set out above, as violence against women is only one form of GBV. TikTok has therefore assessed the nature of GBV risks on its Platform using its definition above;
- In the specific context of the Platform, the risk relating to GBV may involve users attempting to share or disseminate content depicting or involving the following types of behaviour on or through video, livestream, comments and in profile information on the Platform:
 - non-consensual sexual acts that are real or fictional, including rape, molestation, and non-consensual touching;
 - promoting violence, exclusion, segregation, discrimination, and other harms on the basis of a protected attribute (i.e. such as gender, gender identity or sex);
 - threatening or expressing a desire to cause physical injury to a person or a group (where gender is a relevant factor); and
 - degrading someone or expressing disgust on the basis of their personal characteristics or circumstances, such as their physical appearance, intellect, personality traits, and hygiene (where gender is a relevant factor) (together, '**GBV Content**').
- In addition, TikTok notes that Article 21 of the Charter prohibits discrimination based on sex and sexual orientation, and that Article 23 enshrines the right to equality between men and women. TikTok further notes that GBV, in particular violence against women and domestic violence, violates fundamental rights such as the right to human dignity, the right to life and integrity of the person, the prohibition of inhuman or degrading treatment or punishment, the right to respect for private and family life, personal data protection, and the rights of the child, as enshrined in the Charter.²⁰

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, TikTok classified GBV Content as a **Tier 1 priority** in its priority tiering system.

This section summarises additional risks identified in Year 2. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk detection processes, and extensive consultations with external stakeholders.

¹⁸ The Council of Europe's [definition](#) is based on the [Explanatory Report](#) to the Convention on preventing and combating violence against women and domestic violence.

¹⁹ Note that TikTok considers violative content related to hate speech against people due to their sexual orientation in the separate Hate Speech section of this Report.

²⁰ [Proposal for a Directive on combating violence against women and domestic violence. Recital 8.](#)

Summary of risks identified in Year 1

TikTok identified the following inherent risks as part of its Year 1 Risk Assessment:

- **Sharing or dissemination of GBV Content:** Users may attempt to use the Platform to share GBV Content;
- **Content moderation systems:** Given that GBV can be nuanced and highly localised across regional, cultural and linguistic differences in the EU, there is a risk that TikTok's automated detection systems may not keep pace with evolving vocabulary and GBV behaviours, such as coded and localised gendered slurs; and
- **Intentional manipulation of TikTok via impersonation:** Bad actors may use fake accounts to harass and silence victims, particularly in the context of Sexual Exploitation and Abuse. They may impersonate individuals, including victims, to post harmful content and discredit victims by spreading rumours or questioning their experiences. Impersonators may also solicit or share explicit content under the guise of someone the victim trusts, worsening the harm of GBV.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of GBV Content.

- **AI-generated image based sexual abuse:** Image Based Sexual Abuse ('IBSA') is the creation, manufacture, or distribution of nude, partially nude, or sexually explicit content without the consent of the person in the content, for the purpose of sexualising their body, or portraying them in a sexual manner. TikTok considers the harm of IBSA to be the same whether it includes real or manipulated media given the psychological impact to the survivor. The risk involves possessing, distributing, or providing instructions on how to create or access AIGC or altered intimate images that were created or distributed without consent;
- **Sextortion:** TikTok identified sextortion as an evolving risk due to NCMEC reports and escalations on the Platform in the sextortion space in the past year, in particular relating to the 'Yahoo Boys' sextortion group;²¹ and
- **Sexual harassment of female public figures:** TikTok identified an increased risk of videos and comments which bully or harass female candidates, journalists, and other public figures during elections in response to NGO reports, and external engagements. The risk is particularly relevant in Year 2 as almost half of the world's population is voting in national elections in 2024²².

c. Inherent Risk in Year 2

For a detailed analysis on how TikTok assessed the baseline severity and probability for GBV in Year 1, please refer to TikTok's Year 1 Systemic Risk Assessment Report.

²¹ Collection of cybercriminals who are the main drivers of financial sextortion on social media platforms. <https://www.wired.com/story/yahoo-boys-scammers-facebook-telegram-tiktok-youtube/>
<https://longreads.com/2023/07/11/inside-the-world-of-nigerian-yahoo-boys-atavist-excerpt/>

²² The sentence '*half of the world's population voting in national elections in 2024*' has been frequently used by the media to define the magnitude of the 2024 super cycle elections (Source: [Reuters](#), [The New Yorker](#)).

In Year 2, TikTok assesses the overall severity of GBV to be 'Material'. TikTok assesses the probability of GBV in Year 2 to be 'Possible'. TikTok has therefore assessed the inherent risk (that is, the risk without any mitigations in place) of GBV being shared on the Platform in Year 2 to be 'Moderate'.

2. Mitigation Measures:

Please refer to the Year 1 Risk Assessment Report for detailed information on TikTok's system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-K	DESCRIPTION
(b) Adaptation of terms and conditions	In April 2024, TikTok launched an update to the Adult Physical and Sexual Abuse section in its Community Guidelines. The guidelines enhance accessibility and transparency, clarifying further that TikTok does not allow showing, promoting, or engaging in adult sexual or physical abuse or exploitation. This includes non-consensual sexual acts, IBSA, sextortion, physical abuse, and sexual harassment. The Community Guidelines clearly articulate that the Platform is committed to providing a space that embraces gender equality, supports healthy relationships and respects intimate privacy.
(c) Adaptation of content moderation processes	<p>To continue to mitigate the emerging risk of IBSA, TikTok uses a hash-matching system created in partnership with StopNCII, to proactively identify and remove potential non-consensual intimate imagery content on the Platform.</p> <p>TikTok continues to update keywords to support proactive and automated detection of GBV Content through consultations with external experts and government agencies to ensure keyword lists are up to date.</p> <p>For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating the risk of Gender Based Violence Content please refer to Annex 4.</p>
(i) Awareness measures	<p>A consultation [REDACTED] [REDACTED] resulted in a review of the sexual abuse focused page of the Safety Centre. TikTok recognises that there is language regarding GBV that is not centred on survivors' experiences which can exacerbate</p>

trauma and hinder recovery. As a result, TikTok updated its language to be more survivor-centred.

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating the risk of Gender Based Violence Content please refer to Annex 4.

Case Study: Sextortion

Sextortion risks have evolved in the past year in the way it manifests online. Sextortion is a type of blackmail in which the perpetrator uses threats and intimidation to extort money from their victim. In most cases, the threat is related to the distribution or publication of intimate content of the victim.

In partnership with ██████, TikTok's User Research and Product teams created a survey in November 2023, to get user feedback on their experiences with sextortion across multiple social media platforms. Before this study, TikTok's Risk Analysis team also ran internal checks to analyse the existence of sextortion on TikTok.

This study concluded that TikTok is generally not the Platform of choice for exploitation predators, who are drawn to other social media platforms that allow predators to message users that do not follow them.

Through this partnership with ██████, TikTok gained a more robust understanding of the nature of this risk on its Platform and was able to confirm that existing mitigations for this risk are performing well, including moderation of videos and text to recognise sextortion behaviour, as these tools are based on banks of similar content to capture previously known violations. Of the social media platforms listed in the survey, TikTok has the lowest rate of sextortion.

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness and proportionality of TikTok's mitigations relevant to the systemic risk of GBV, TikTok has assessed residual risk to be 'Moderate' in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate GBV on the Platform.

This assessment was conducted on the basis of qualitative and quantitative insights, including data on TikTok's enforcement of its Community Guidelines in the EU for the period of Q3 2023 to Q2 2024. Under the Community Guidelines policy of 'Sexual Exploitation & Gender-Based Violence', TikTok removed 617,156 total videos in the EU, with 426,161 detected and removed proactively, and 136,813 removed without any views. This high proportion of proactive removal suggests that TikTok's content moderation systems remain effective in detecting and removing GBV Content on the Platform, helping to mitigate the systemic risk.

This assessment of the effectiveness of TikTok’s proactive detection systems is further informed by the following data on user reports, which serve as a useful indication of the volume of GBV Content that is not detected and removed by TikTok’s proactive systems. User reports accounted for a minority of GBV Content detected and removed by TikTok, at 190,995 of the 617,156 total videos detected and removed under ‘Sexual Exploitation & Gender-Based Violence’. TikTok remains committed to continuously working to improve and iterate its proactive detection capabilities, while noting that GBV Content is a rapidly evolving risk area, with nuance across regions and languages.

TikTok’s user reporting tools are an important component of its content moderation process, ensuring that GBV Content is detected and removed, including new and evolving forms of GBV. The data demonstrates that when TikTok receives user reports of violative GBV Content, it is actioned efficiently. Under ‘Sexual Exploitation & Gender-Based Violence’, 166,689 of the 190,995 total videos removed following user reports were removed within two hours of receiving the report. Collectively, TikTok’s proactive detection and reporting tools have worked effectively to ensure that 509,501 of the 617,156 total videos removed under ‘Sexual Exploitation & Gender-Based Violence’ were removed within 24 hours of upload.

The effectiveness, reasonableness, and proportionality of TikTok’s moderation of GBV Content is indicated by its appeals data, as users can submit appeals against content removals if they believe TikTok has made a mistake. Under the ‘Sexual Exploitation & Gender-Based Violence’ policy, only 35,804 of the 617,156 total videos removed were successfully appealed and reversed.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok’s strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual risk discussed above. TikTok plans to devote extra resources to combating GBV Content in the year ahead and as a result it remains a Tier 1 priority. As noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(b), GBV policies:** [REDACTED]

- **Article 35(1)(b), AIGC IBSA:** [REDACTED]

- **Article 35(1)(i), User education:**

11. YOUTH SAFETY: RISKS RELATED TO AGE APPROPRIATE CONTENT AND ONLINE ENGAGEMENT

1. Description of the risk:

a. Description of the risk from Year 1

- TikTok is a platform accessible to minors and therefore gives careful and appropriate consideration to the risk encompassed by Article 28(1) DSA, which requires it to put in place appropriate and proportionate measures to ensure a high level of privacy, safety and security for minors on its service. TikTok's minimum age requirement is 13 years old. TikTok uses the term '**Underage Users**' to refer to minors under the age of 13 who may attempt to access the Platform. '**Younger Users**' and '**Youth**' refer to TikTok users with a declared age of between 13 to 17 years old inclusive. However, TikTok is not specifically aimed at or predominantly used by minors aged under 18 years old in Europe; and
- TikTok acknowledges that a balance must be found between the design of the measures to address risks to Younger Users and the Fundamental Rights of legitimate users of any age. Those rights include the right to freedom of expression, which includes the right of users to express themselves freely on the Platform, receive and impart information and ideas to aid their individual learning and development and to exercise autonomy. Accordingly, TikTok seeks to avoid unnecessary and disproportionate impacts on such legitimate users in the design of its Platform and the specific Youth Safety measures summarised in this Report;
- TikTok recognises the following Youth Safety risks:
 - '**Content Risk**', also referred to as risks related to '**Age Appropriate Content**': Younger Users might access or view content on the Platform that is not age-appropriate. Related risks may involve negative effects on physical or mental well-being; and
 - '**Conduct Risk**' and '**Contact Risk**' (collectively, risks related to '**Online Engagement**'): In creating and posting content on the Platform, or by engaging with content posted by others, Younger Users may engage in inappropriate behaviour or potentially encounter inappropriate behaviour from other users, such as inappropriate comments, bullying or behaviour amounting to child sexual exploitation.²³ TikTok notes that this risk only arises in relation to users who are active, rather than passive (i.e., those that only watch content and take no action in relation to it).

²³ CSAM and CSEA risks are dealt with in the above section of this Report on risks of child sexual abuse material and child sexual exploitation.

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, risks related to Age Appropriate Content and Online Engagement were assessed together with risks related to Age Assurance as 'Online Protection of Minors and Associated Risks'. Based on this assessment, TikTok classified risks related to Youth Safety as a **Tier 1 priority** in its priority tiering system.

Summary of risks identified in Year 1

TikTok identified the following inherent risks as part of its Year 1 Risk Assessment:

- **Creating and publishing content ('Conduct Risk')**: Younger Users may post content without understanding the consequences, including violating Community Guidelines, sharing illegal content, or private information, leading to loss of privacy, embarrassment, or distress;
- **Proactively engaging with content ('Content Risk')**: Younger Users may engage with other creators' content, including commenting or using Stitch or Duet tools, potentially leading to risky or harmful behaviour;
- **Others engaging with Younger Users and their content ('Contact Risk')**: Published content can attract interactions from other users, such as sharing, commenting, or using Stitch or Duet tools, which may result in inappropriate behaviour towards youth, causing embarrassment or distress; and
- **Potential impacts to well-being**: Younger Users may experience negative effects to their well-being, such as spending excessive time on the Platform or feeling pressure to post content and receive likes, impacting their self-esteem and overall well-being.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of risks related to Age Appropriate Content and Online Engagement.

- **Screen time**: The risk to mental and physical well-being of excessive use of the Platform was highlighted in its Year 1 Risk Assessment. In Year 2, TikTok has identified some specific risk factors that it considers to be relevant:
 - **Online platforms' ease of use**: The ease of using an online platform may make it difficult for Younger Users to intentionally take breaks as they may struggle with self-regulation and impulse control due to their developmental stage; and
 - **Incentives to use an online platform and engage with more content or users**: For Younger Users, who may be focused on peer acceptance and feedback as part of their identity development, the desire for positive social rewards (e.g. likes on a video) may further drive their online engagement and screen time;²⁴
- **Risks related to AIGC**: AIGC is an emerging area with certain specific concerns related to Youth Safety:
 - **Additional content harms**: Younger Users may create and publish AIGC featuring other individuals that is inappropriate or in violation of our Community Guidelines, which

²⁴ TikTok relied on expert and academic research to assess this risk, including research produced by the Digital Wellness Lab, American Association of Pediatrics, National Sleep Foundation, Pew Research Center, American Psychological Association, and Common Sense Media.

could lead to distress, embarrassment, loss of privacy or societal harm if this content is viewed by other users. They may also be misled by AIGC, making Younger Users potentially more susceptible to harm from misinformation, more susceptible to misunderstanding satire or humour, and/or more susceptible to adopting unrealistic standards for personal physical appearance and video quality;

- **Increased engagement with Younger Users' content:** AIGC may increase engagement with Younger Users' content by making it more appealing or interesting. This increased attention might overwhelm Younger Users, who may not be equipped to handle it safely; and
- **Well-being impact and social comparison:** Early research indicates there are risks to Younger Users' well-being associated with AIGC, including content created using off-platform AIGC tools, in particular around self-esteem, body image, and mental health. A lack of agreed knowledge and industry best practices to mitigate these effects is a risk factor.
- For more detailed information on TikTok's approach to mitigating risks associated with AIGC, please refer to *Priority Considerations Across all Modules*.

c. Inherent Risk in Year 2

For a detailed analysis on how TikTok assessed the baseline severity and probability for Online Engagement and Age Appropriate Content in Year 1, please refer to the 'Online Protection of Minors' section in TikTok's Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of risks related to Online Engagement and Age Appropriate Content to be 'Material'. TikTok assesses the probability of risks related to Online Engagement and Age Appropriate Content to be 'Likely' in Year 2. Taking into account the severity and probability in Year 2, TikTok has assessed the inherent risk of risks (that is, the risk without any mitigations in place) related to Online Engagement and Age Appropriate Content in Year 2 to be 'Material'.

2. Mitigation Measures:

For detailed information on TikTok's system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-I	DESCRIPTION
(a) Adaptation of feature or platform	As mentioned in the Year 1 Report, a key mitigation TikTok implemented was screen time management features - including 60 minute a day

<p>design</p>	<p>default screen time limits for Younger Users and availability of that feature for all users. In the last year, TikTok further integrated and expanded its screen time management features. These features are also now available on the web app version of the Platform (i.e., TikTok.com).</p> <p>TikTok has also implemented delayed screen time interventions in order to increase the interventions' effectiveness. Screen time interventions are often triggered while users are watching a video, which forces the disruptive decision of finishing what they are watching or adhering to their screen time limits. Some users will want to finish watching the video they're currently viewing and will ignore the pop-up. Consequently, TikTok has launched a delay, which prevents the app from showing screen time management interventions until just after a user navigates away from the video, or has watched the video for 60 seconds. If they are not currently watching a video (via Home tab <i>or</i> Search), then there's no change to the trigger logic. This is intended to operate so that users are more likely to pay attention to the screen time intervention.</p> <p>TikTok has also launched the STEM Feed, a specialised feed dedicated to showcasing content related to Science, Technology, Engineering, and Math. The STEM Feed is enabled by default for all Younger Users, while others can activate it through content preferences. It's designed as a space where users can discover reputable, educational content outside of the FYF.</p> <p>As mentioned in Year 1 as a key mitigation improvement, TikTok also expanded its Content Classification system (which rates and then displays UGC based on maturity of the user, based on declared age) to include LIVE content. This is used to prevent livestream content with overly mature themes from reaching Younger Users.</p> <p>Since the Year 1 Risk Assessment Report, a new registration pop-up has been introduced for new 16-17 year olds (and a pop-up was presented to existing 16-17 year olds with public accounts) to inform them about the implications of public accounts, with the private account option pre-selected and requiring the 16-17 year old user to make a selection. If no selection is made, the 16-17 year old user's account is defaulted to private.</p>
<p>(b) Adaptation of terms and conditions</p>	<p>In May 2024, TikTok updated its Community Guidelines to address emerging and ongoing risks:</p> <ul style="list-style-type: none"> • Youth Safety and Well-being: TikTok updated its Youth Safety and Well-being policies around appropriate and inappropriate content, including detailed enforcement outcomes ('not allowed', 'restricted', 'FYF ineligible') and updated key terms related to

	<p>Youth Safety such as CSAM, grooming, body exposure, and disordered eating;</p> <ul style="list-style-type: none"> • AIGC: TikTok updated its AIGC guidelines and policies to remove any AI-generated visual or audio representations of younger users if the platform becomes aware of such instances. • Harmful Body Idealisation: In June 2024 TikTok implemented a Harmful Body Idealisation policy which makes ineligible for the FYF content that promotes idealised body aesthetics associated with harmful weight management behaviours, such as disordered eating and anabolic steroid use. TikTok may undertake proactive ad hoc searches for this type of content from time to time to enforce this policy.
<p>(c) Adaptation of content moderation processes</p>	<p>In addition to the Year 1 content moderation mitigations still in place, TikTok has implemented an early detection model. In 2023, TikTok updated the early detection model that it uses to identify and stop potentially dangerous trends and challenges. [REDACTED]</p> <p>[REDACTED] Moderators can then refer to this database in their content moderation decisions for guidance.</p>
<p>(d) Adaptation of algorithmic systems</p>	<p>As mentioned in Year 1 as a key mitigation improvement, TikTok has improved its existing Content Classification process. In addition to (a) above, it has also done so by refining its approach to better prioritise accuracy of the automatic labelling of mature content , minimise views of violative content, and remove egregious content quickly. TikTok has developed and integrated a Content Level [REDACTED] system for all content channels accessed by logged-out users in the TikTok app. Content Level [REDACTED] refers to content that is appropriate for general audiences to view. This aims to ensure that Younger Users using the Platform without logging in are only exposed to Age Appropriate Content. TikTok’s Restricted Mode feature, available to all users and also as part of Family Pairing, provides users with the ability to restrict their viewing to Content Level [REDACTED] content. TikTok continues to invest in strengthening its content classification system to provide an age-appropriate experience for Younger Users.</p> <p>TikTok has also changed how certain categories of content appear on the FYF by updating its models to prevent the display of concentrated content for topics such as dietary discussions, weight loss, sadness, grief, loneliness, and hopelessness.</p>
<p>(e) Reinforcing risk detection measures</p>	<p>As mentioned in Year 1 as a key mitigation improvement, TikTok continues to monitor existing metrics, model efficacies, and related safety features by sampling videos viewed by teenagers and reviewing them manually to understand if teenagers are viewing content that is appropriate for them to view based on Content Classification Policies.</p>

	<p>This sampling and reviewing process occurs regularly and is key to detect emerging risks impacting Younger Users.</p>
<p>(i) Awareness measures</p>	<p>TikTok has launched a global Youth Council to reinforce its commitment to enhancing safety measures for teens on the Platform and to hear from teens directly. This initiative responds to recent global research indicating that teens and their parents want more collaborative opportunities with online platforms. Created in partnership with Praesidio Safeguarding, the Youth Council consists of 15 teens from diverse backgrounds and countries, including the United States, United Kingdom, Brazil, Indonesia, Ireland, Kenya, Mexico, and Morocco. The group, which first convened in December 2023 and met again in February 2024 with TikTok Chief Executive Officer Shou Chew, has prioritised teen well-being and inclusion for the year. The Youth Council's insights help TikTok to enhance resources for Younger Users and to refine its risk detection measures and enhance the overall user experience for young people.</p> <p>TikTok is updating its Youth Guide (also known as/now renamed Youth Portal) to include new safety features, based on feedback from the Youth Council on reporting and blocking, and is introducing new GIFs to help Younger Users locate these features. Members of the Youth Council provided input and asked TikTok to share more information about reporting and blocking content and users to better understand what happens after a report is made.</p> <p>For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating risks related to Age Appropriate Content and Online Engagement please refer to Annex 4.</p>

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness and proportionality of TikTok's mitigations relevant to the systemic risk related to Age Appropriate Content and Online Engagement, TikTok has assessed residual risk to be 'Moderate' in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate risks related to Online Engagement and Age Appropriate Content on the Platform.

This assessment was conducted on the basis of qualitative and quantitative insights, including data on TikTok's enforcement of its Community Guidelines in the EU for the period of Q3 2023 to Q2 2024. Under the Community Guidelines policy of 'Youth Safety and Well-Being', TikTok restricted 8,155,040 total videos in the EU, with 7,967,430 detected and restricted proactively, and 5,871,273 restricted without any views. This high proportion of proactive removal suggests that TikTok's content moderation systems remain effective in detecting and restricting content endangering Younger Users on the Platform, helping to mitigate the systemic risk.

- **Article 35(1)(c), Content Levels:** TikTok also plans to enhance the content levels across the Platform for its Younger Users

12. YOUTH SAFETY: RISKS RELATED TO AGE ASSURANCE

1. Description of the risk:

a. Description of the risk in Year 1

TikTok understands 'Age Assurance' in the context of Article 34(1) DSA and the previously addressed risks to Younger Users related to Age Appropriate Content and Online Engagement. While Younger Users are entitled to freedom of expression, access to information and other fundamental rights associated with use of the Platform, TikTok considers it reasonable and proportionate to prohibit access for Underage Users and permit Platform access for Younger Users while restricting their access to specific features of Platform experiences. 'Age Assurance' refers to the tools and technologies employed by TikTok to prohibit or restrict Youth access to the Platform, and thereby, help protect Underage Users from the risks associated with Online Engagement and Age Appropriate content. TikTok's approach is consistent with relevant international legal frameworks relating to the protection of child safety (e.g. UN Convention on Rights of the Child), which recognise the unique vulnerabilities of children and teenagers.

This module addresses the systemic risk that (1) Underage Users may mis-state their age at registration and that they may gain access to the Platform as a result; and (2) that Younger Users may mis-state their age at registration and not receive an age appropriate experience on the Platform, together the **Mis-Stated Age Risk**. Additionally, this risk includes the possibility of inaccurate age determinations at other stages of the Age Assurance process, such as in underage moderation and appeals.

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, risks related to Age Assurance were assessed together with risks related to Age Appropriate Content and Online Engagement as 'Online Protection of Minors and associated risks'. Based on this assessment, TikTok classified risks related to Youth Safety as a **Tier 1 priority** in its priority tiering system.

This section summarises additional risks identified in Year 2. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk detection processes, and extensive consultations with external stakeholders.

Summary of risks identified in Year 1

TikTok identified the following inherent risks as part of its Year 1 Risk Assessment:

- **Mis-stated age:** Underage Users may attempt to gain access to the Platform, and Younger Users may attempt to avoid an age-appropriate experience by mis-stating their age at the point of registration;
- **Parents or guardians misunderstanding Age Assurance:** Parents or guardians may not appreciate or understand the minimum age requirement, and age-appropriate restrictions, and support or help their child to bypass age assurance measures (e.g. through appeals); and
- **Unreasonable and disproportionate impacts on legitimate users:** Age Assurance tools and processes may not adequately and proportionately protect other fundamental rights, such as adult users' rights to freedom of expression (e.g. erroneous removal of adult users as suspected underage accounts).

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of risks related to Age Assurance.

- **Mis-Statement Age Risk in moderation processes:** When TikTok detects potential mis-stated age accounts, human moderators are tasked with identifying the user's age, however individual cases can occasionally be challenging for moderators to discern. There is therefore a risk of harm to the rights, particularly freedom of expression, of account owners if their age is incorrectly assessed as being underage, and their account is removed from (or restricted on) the Platform (though users have the right to appeal); and
- **Use of AIGC to circumvent the age appeals process:** Users can appeal TikTok's decision to remove them from the Platform on the basis that it has concluded that they are an Underage User. This process involves using a choice of methods to provide proof of age. There has not been any reported instance of AIGC being used to intentionally create fake appeal data in order for a user to appear older than they are. [REDACTED]

[REDACTED] Without effective enforcement, Underage Users may gain access to the Platform, and Younger Users may avoid age restrictions.

c. Inherent Risk in Year 2

For a detailed analysis of how TikTok assessed the baseline severity and probability for risks related to Age Assurance in Year 1, please refer to TikTok's Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of risks related to Age Assurance to be 'Material'. TikTok assesses the probability of risks related to Age Assurance in Year 2 to be 'Likely'. Taking into account the changes in severity and probability in Year 2, TikTok has assessed the inherent risk (that is, the risk without any mitigations in place) related to Age Assurance in Year 2 to be 'Material'.

2. Mitigation Measures:

For detailed information on TikTok's system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-I	DESCRIPTION
(a) Adaptation of feature or platform design	<p>TikTok implemented additional steps to protect its age appeals process against the use of AIGC by Underage or Younger Users to misrepresent their age. If users want to confirm their age through the 'Selfie with ID' option, they must upload a live photo of themselves holding their ID directly from the camera function. This helps to mitigate the risk that users upload manipulated or AIGC images from their photo album. Similarly, if Younger Users want to confirm their age through TikTok's 'photo with a parent/guardian or trusted adult' option, the various requirements would be difficult for AIGC to replicate (i.e. selfie with trusted adult holding paper detailing proof of age, date of birth, and unique 6-digit TikTok code, and the photo must also be taken from the camera function). TikTok's 'credit card authorisation' appeal option for users aged over 18 would be close to impossible for AIGC to replicate.</p> <p>To complement these three existing age appeals options, TikTok implemented an age estimation technology for users aged 18 and over, as mentioned in the Year 1 Report. This new appeal option has been developed by Yoti, a specialised partner that was selected based on its superior technological capabilities, regulatory approval and industry support. Users can provide a selfie taken through a secure image capture function that is sent to Yoti, which uses its technology to conduct an age estimation before returning the result to TikTok and deleting the selfie from their systems. The secure image capture function will prevent users from uploading AIGC or manipulated images that may interfere with age estimation by misrepresenting their age. This tool expands appeal options to users who do not want, or are unable to, use a credit card or ID for age confirmation.</p>
(c) Adaptation of content moderation processes	<p>As mentioned in the Year 1 Report as a planned mitigation, in January 2024, TikTok updated its 'Suspected Underage User' moderation guidance to reflect two distinct policies; 'Suspected U13 user' and [REDACTED]. This change allows moderators to identify [REDACTED] separately. Content tagged with one of these two new policy titles will be reviewed by human moderators, with</p>

the content removed and account banned if the account is found to be that of an Underage User.

Prior to this change, only one policy was applied (i.e. 'Suspected Underage User'), which did not differentiate between [REDACTED] and so, providing separate policy titles will allow reviewers to consider the varying accuracy of respective age signals in deciding on enforcement. This policy adjustment supports more accurate enforcement against underage accounts. Additionally, using separate internal tags for 'Suspected U13 Age' and [REDACTED] means that TikTok can monitor and improve U13 detection in a more granular way.

In advance of bifurcating its 'Suspected Underage User' policy title, TikTok launched new moderation guidance that provided additional guidance for moderators to use in identifying [REDACTED].

While these factors have long featured in TikTok's underage moderation, the new guidance reflects advances in TikTok's understanding of the risk, and incorporates trends and tactics observed in Underage Users. This ultimately helps moderators to attach the appropriate policy title under TikTok's new bifurcated policies, supporting appropriate action on underage accounts.

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness and proportionality of TikTok's mitigations relevant to the systemic risk related to Age Assurance, TikTok has assessed residual risk to be 'Moderate' in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate risks to Age Assurance on the Platform.

This assessment was conducted on the basis of qualitative and quantitative insights, including data on TikTok's enforcement of its Community Guidelines in the EU for the period of Q3 2023 to Q2 2024.²⁵ TikTok's mitigations aim to detect and remove Underage Users, and detect and restrict Younger Users. TikTok detected and removed 3,232,672 Underage Users from the Platform in the EU from Q3 2023 to Q4 2023 .

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok's strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual risk discussed above. TikTok plans to devote extra resources to mitigating risks related to Age Assurance in the year ahead and as a result it remains a Tier 1 risk. As noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(c), Continually improve and monitor quality of moderation of suspected Underage Users and Younger Users:** TikTok will improve visual signals-based guidelines and other instructions provided to moderators when reviewing suspected Underage User and Younger User accounts that reflect Trust & Safety's assessment of risks and recent trends seen in enforcement;
- **Article 35(1)(c), Bias mitigation:** TikTok will strengthen mechanisms to better identify, and reduce, bias in age assurance processes;
- **Article 35(1)(c), Monitoring of age detection models:** TikTok will continue to monitor and assess its age detection models for Underage Users and Younger Users in order to further enhance age assurance processes;
- **Article 35(1)(f),** [REDACTED]
- **Article 35(1)(j),** [REDACTED]

²⁵ TikTok can provide data on U13 accounts detected and removed for Q3 and Q4 2023, but not the remainder of the Year 2 reporting period. Going forward, TikTok intends to ensure that it can provide this data for the full reporting period, as per the CGER data provided in other risk sections of this report.

13. CIVIC INTEGRITY: RISKS TO ELECTIONS AND CIVIC INTEGRITY

1. Description of the risk:

a. Description of the risk in Year 1

- b. TikTok understands the risk to be the actual and foreseeable negative effects on election processes and on Civic Integrity arising from the dissemination of verifiably false or misleading content related to an election or other civic events in the EU (such as referenda or a census) (together, '**Election Misinformation**');
- c. Election Misinformation risk may arise from attempts to share or disseminate the following content on or through the Platform, whether as short video, comment, livestream or within their profile information:
 - d. Misleading information about how to vote or register to vote or the qualifications to vote or run in an election;
 - e. Misleading information about the date of an election or other civic process (e.g. stating that an election is on a later date than it is scheduled for);
 - f. Misleading information about how to participate in a census or eligibility requirements for participating in a census;
 - g. Content that advances false claims related to the technical eligibility requirements for current political candidates and sitting elected government officials to serve in office;
 - h. False claims of election fraud (such as voting machines being tampered with to favour a candidate or political party);
 - i. Content that falsely claims that an election has been or will be rigged, so the results cannot be trusted;
 - j. Misinformation or/and conspiracy theories about candidates, candidate impersonation and related issues that may impact civic integrity; or
 - k. Synthetic and manipulated media (e.g. modified using model technology) featuring public figures (such as a government official, politician, business leader, or celebrity) may impact Civic Integrity if mis-used for political endorsements or other purposes (TikTok refers to this as '**Synthetic and Manipulated Media**').
- l. TikTok also acknowledges that Articles 39 and 40 of the Charter enshrine the rights for every EU citizen to vote and to stand as a candidate at elections to the European Parliament and at municipal elections, respectively, and that such Fundamental Rights may be undermined by Election Misinformation.

Summary of risks identified in Year 1

TikTok identified the following risks as part of its Year 1 Risk Assessment:

- a. **Election misinformation:** The risk of attempts to share or disseminate election misinformation that could adversely influence an election, undermine trust in democratic processes, and/or cast doubt on the legitimacy of election outcomes. The specific types of election misinformation identified by TikTok in Year 1 were:
 - b. Misinformation about how to vote;

- c. Misinformation about qualifications to run in an election;
- d. Misinformation about the date of an election or other civic processes;
- e. False claims about technical eligibility requirements for current political candidates and sitting elected officials;
- f. Misleading claims of election fraud; and/or
- g. Misinformation and/or conspiracy theories about candidates or candidate impersonation.
- h. **Synthetic and manipulated media:** The risks of mis-use of synthetic and manipulated media to depict public figures in misleading ways, such as for political endorsements;
- i. **Fake engagement:** The risk of covert influence operations, fake engagement and spam activity that attempts to disrupt election processes, amplify certain viewpoints in elections or sow social disharmony;
- j. **Ads:** The risk that ads contain election misinformation.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of risks to Elections and Civic Integrity.

- a. **Geopolitical factors:** The emergence and continuation of two prominent wars (i.e. Ukraine-Russia and Israel-Hamas) and the coincidence of a high number of global elections in 2024 has increased the complexity of identifying and moderating Election Misinformation. These events have necessarily led to a higher volume of political content on the Platform, more complex and nuanced political opinions being shared and new opportunities for bad actors to promote polarising narratives. There is a risk of policy gaps if trends and evolving patterns of discourse outpace policy updates. Without a significant reinforcement of internal resources and processes, added geopolitical complexity creates further risks that Election Misinformation is shared and amplified on the Platform;
- b. **Additional potentially harmful types of Election Misinformation:** In the context of heightened geopolitical complexity, TikTok has identified additional types of Election Misinformation that could cause harm on the Platform. This includes misleading information about (or the outright denial of) legitimate election outcomes, false claims misrepresenting authoritative sources of civic information (e.g. the text of a parliamentary bill), unverified claims that may harm users' trust in election/civic processes, procedures or outcomes or that of election workers, and manipulated media that falsely appears to come from reputable news organisations. The potential negative impacts on election and civic processes remain the same: adversely influencing public debate, discouraging participation, and damaging trust in democratic processes. The risks were identified through a combination of expert consultations, media monitoring, internal expertise, platform preparations for the 2024 elections, and engagement with relevant communities and events;
- c. **Commercial disinformation:** TikTok has also identified commercial disinformation as a new risk in Year 2. Commercial disinformation refers to influence operations with a financial intent or motivation. This includes disinformation-as-a-service sold by third-party vendors, and misleading marketing campaigns presented as UGC. Commercial disinformation can pollute the information environment and reduce user trust in the Platform. This risk was identified through TikTok's collaboration with third party vendors to monitor the emerging capabilities of threat actors, alongside industry reports and other publicly available investigative reporting;
- d. **Risks related to AIGC:** The risk of Election Misinformation is further aggravated by emerging and evolving (mis)uses of AIGC. Risks include users sharing synthetic and manipulated media

that falsely depicts a public figure in a situation that did not occur, impersonating reputable news sources, and creating content (e.g. AIGC video) that claims to provide 'evidence' that substantiates Harmful Misinformation narratives. Additionally, AIGC potentially makes it easier for bad actors (e.g. coordinated influence operations) to increase the quality and volume of dis- and misinformation production. Internal escalations, consultations with external experts, and TikTok's preparations in advance of the 2024 election cycle, among other detection mechanisms, have also identified the potential risk that AIGC technologies are exploited by covert influence operatives in order to automate malign influence operations. The risks were identified through a combination of expert consultations, media monitoring, internal expertise, platform preparations for the 2024 elections, and engagement with relevant communities and events;

- e. **Electoral misconduct:** TikTok has identified electoral misconduct as a new Election Misinformation risk. This risk is that users may share content that promotes or provides instruction on illegal participation and electoral interference, including voter intimidation, and calls for disrupting legitimate election outcomes through violence. Such misconduct can lead to serious harm, including off-platform violence and the subversion of democratic processes. This risk was identified through internal escalations, preparations for the 2024 election cycle, and consultations with experts and civil society groups;
- f. **Risks from evolving state-affiliated media activities:** TikTok has identified risks associated with the evolution of state-affiliated media strategies used to interfere with the information environment. This includes efforts to bypass and/or evade TikTok's system for labelling 'State-Affiliated Media' accounts by deploying localised franchises and rebranded entities, using influencers to amplify state-affiliated narratives, and/or using individual media employees to propagate controlled messaging;
- g. **Hacked materials:** Hack-and-leak operations may negatively impact or undermine democratic elections by disturbing civic processes, escalating violence, further polarising society, and/or reducing trust in the information ecosystem. In the context of high-priority elections world-wide, TikTok has identified an increased risk of the Platform being used to disseminate hacked materials, with an adverse impact on civic integrity;
- h. **Political ads:** TikTok does not allow political advertising in Europe on the Platform. With almost half of the world's population voting in national elections in 2024, bringing a higher volume of potential political content to the Platform, there is a heightened risk that TikTok encounters difficulties in detecting and enforcing against political content in ads; and
- i. **Advertising:** TikTok offers 'Promote', a product feature that allows users to purchase advertising inventory for their UGC, ensuring a certain level of traffic for their content. As with other types of advertisements, 'Promote' is subject to TikTok's updated advertising policies, meaning that TikTok's prohibition on political advertising applies to the 'Promote' feature. Although users are informed that using Promote will mean the content must comply with our Community Guidelines and advertising policies, some users still might not consider it advertising when promoting their own UGC.

1. Inherent Risk in Year 2

For a detailed analysis of how TikTok assessed the baseline severity and probability for risks related to Age Assurance in Year 1, please refer to TikTok's Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of risks related to Age Assurance to be 'Material'. TikTok assesses the probability of risks related to Age Assurance in Year 2 to be 'Likely'. Taking into

account the changes in severity and probability in Year 2, TikTok has assessed the inherent risk (that is, the risk without any mitigations in place) related to Age Assurance in Year 2 to be 'Material'.

2. Mitigation Measures:

For detailed information on TikTok's system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-I	DESCRIPTION

(b) Adaptation of terms and conditions

In Year 2, TikTok expanded policy coverage and updated its Community Guidelines in a number of areas, including:

- Launch of the 'Unverified Election Claims' Community Guidelines policy which prevents unverified claims about an election from appearing in the FYF (e.g. a premature claim that all ballots have been counted or tallied). This may occur when a new claim has been identified and sent to be fact checked- the label preventing the content from appearing in the FYF is applied until the result of the fact check, when either the label or content is removed;
- Launch of the 'Election Misconduct' policy, which prohibits content that encourages or instructs on illegal actions related to elections, including interfering with the electoral process or advocating for the disruption of a legitimate election outcome through unlawful means, such as a coup;
- Launch of the 'Distribution of Hacked Materials' policy, which prohibits the sharing of hacked materials that may lead to disturbances in civic processes, escalation of violence, polarisation, or reduced trust in the information ecosystem;
- Launch of the 'Misrepresented Civic Sources' policy, prohibits content that severely misleads users about important civic information, which could harm the information ecosystem, undermine public participation in civic processes, and reduce awareness of political issues; and
- Updated TikTok's Community Guidelines and advertising policies regarding misleading AIGC and synthetic media to more specifically identify and prohibit specific types and uses of AIGC in UGC and ads (including AIGC that could reduce trust in civic processes and institutions or inflame tensions during crises), ensuring robust policy coverage. These specific Community Guidelines changes are detailed in the cross-risk mitigations outlined for AIGC in the 'Priority Considerations' section.

(c) Adaptation of content moderation processes

TikTok has built a new task prioritisation feature for its misinformation moderation processes. This change means TikTok has begun to move from a 'first-in-first-out' task moderation system to a system that prioritises matters for moderator review based on a machine learning model risk score. The new system will allow TikTok's teams to quickly action the highest risk and most likely violative misinformation first, as opposed to the order in which first appeared in the process to be actioned. Misinformation that falls under TikTok's database of previously fact-checked claims will be immediately actioned, while new instances of potential new Election Misinformation will be sent to a relevant fact-checker partner for review.

As mentioned in the Year 1 Report, TikTok has expanded its Fact-Checking Programme to ten partners. These organisations provide fact-checking coverage in 22 official EU languages: [Agence France-Presse](#) (AFP), [dpa Deutsche Presse-Agentur](#), [Demagog](#), [Facta](#), [Faktograf](#), [Lead Stories](#), [Logically Facts](#), [Newtral](#), [Poligrafo](#) and [Science Feedback](#). For detailed information on TikTok's Fact-Checking Programme, see Annex 5.

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating risks to Elections and Civic Integrity, please refer to Annex 4.

TikTok has reviewed and updated its fact checker ratings. This update includes providing more veracity rating options to fact checkers when they are assessing content, including new and updated rating categories [REDACTED]

[REDACTED] Having a better spectrum of ratings is more reflective of the range of content escalated, and provides options to tag grey-area content. Previously, fact checkers didn't have specific categories to rate content, particularly around evolving events and crises where things may have been rated 'unconfirmed' or 'disputed'. (e.g. 'inconclusive') which resulted in less nuanced assessments and moderation actions. This update has been accompanied by improved definitional language for fact-checkers regarding the scope of each rating and the topical sub-classification options.

Given added Election Misinformation risks and complexity associated with the Israel-Hamas and Russia-Ukraine wars, TikTok has launched new guidance to help moderators apply TikTok's policies to more complex and evolving types of violative misinformation content. Previously, moderators did not have sufficient guidance around how to interpret fact checker ratings. TikTok has also continued to routinely update moderator guidelines to account for local languages and political nuances in anticipation of upcoming elections.

<p>(e) Adaptation of advertising systems</p>	<p>TikTok has launched four new, more granular misinformation policies for TikTok’s advertising products in Europe on medical misinformation, dangerous misinformation, fake news, and dangerous conspiracy theories. [REDACTED]</p> <p>In May 2024, TikTok also updated its policies regarding state-affiliated media accounts to limit their reach on global events and affairs. Now, when TikTok identifies such accounts trying to reach audiences outside their home country, their content will be ineligible for recommendation, meaning the content will not appear in the FYF. Additionally, if these accounts advertise, they will only be allowed to do so within their own country. These changes will be applied to all markets where TikTok’s state-controlled media labels are available.</p> <p>In Year 2, TikTok has launched dedicated political advertising content moderation processes in all major monetised EU election markets. Moderators receive market-specific election training, allowing for more nuanced and enhanced enforcement of TikTok’s prohibition against political advertising in EU member states.</p> <p>[REDACTED]</p> <p>Additionally in Year 2, TikTok has developed new automated models [REDACTED] to enhance proactive moderation by the Platform’s advertising moderation system.</p>
<p>(f) Reinforcing risk detection measures</p>	<p>In Year 2, TikTok has deployed improved machine learning models to facilitate the effective detection of potential Election Misinformation. The primary detection model now takes into account relevant sub-policies for more accurate identification of violative content (e.g. Election Misinformation, and Medical Misinformation). This means that the model now takes account of different possible types of misinformation and their respective varying characteristics, rather than purely basing its assessment on a binary outcome.</p>

<p>(g) Initiating or adjusting cooperation with trusted flaggers in accordance with Article 22 and the implementation of the decisions of out-of-court dispute settlement bodies pursuant to Article 21</p>	<p>TikTok remains ready to engage with EU 'trusted flaggers' as they continue to be designated (read section 5) and at the same time, TikTok also continues to receive reports from Community Partners and subject matter experts.</p>
<p>(h) Initiating or adjusting cooperation with other providers of online platforms or of online search engines through the codes of conduct</p>	<p>TikTok has created a rapid response system ('RRS') to streamline the exchange of information among civic society organisations, fact-checkers, and online platforms as part of TikTok's commitments as signatories of the Code of Practice on Disinformation ('CoPD'). The rapid response system was a time-bound dedicated framework for cooperation among signatories during the 2024 European Parliament elections, allowing non-platform signatories to flag time sensitive content, accounts or trends that may present threats to the integrity of the electoral process. There have been 5 RRS reports received between the beginning of May and the EU election week in early June.</p>
<p>(i) Awareness-raising measures</p>	<p>As part of its ongoing Election Integrity Programme described in Year 1, starting in 2023 and throughout 2024, TikTok implemented in-app Election Centres for each EU member state providing access to authoritative information, in the lead-up to, and during, the 2024 EU Parliamentary Elections. These Centres were developed in collaboration with external stakeholders in a bespoke way for each member state and in the local language. One part of this included collaborating with fact checkers to run media literacy campaigns in EU member states. These campaigns have included embedding a media literacy video focused on critical thinking skills in the EU Election Centre page of each member state. These videos are unique for each member state, with localised detail and context and support users' ability to identify misinformation and manipulated media.</p> <p>For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating risks to Elections and Civic Integrity, please refer to Annex 4.</p>

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness, and proportionality of TikTok's mitigations relevant to the systemic risk of Election Misinformation, TikTok has assessed residual risk to be 'Moderate' in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate Election Misinformation on the Platform.

TikTok has considered relevant data including data on its enforcement of its Community Guidelines in the EU for the period of Q3 2023 to Q2 2024, data under its [Covert Influence Operations reporting](#), and data collected in relation to the 2024 EU Parliamentary election to support its assessment of the effectiveness of its mitigations against the systemic risk of Election Misinformation.

Under the Community Guidelines policy of 'Civic & Election Integrity', TikTok removed 75,335 total videos in the EU, with 73,930 detected and removed proactively, and 71,841 removed without any views. This high proportion of proactive removal suggests that TikTok's content moderation systems

remain effective in detecting and removing Election Misinformation on the Platform, helping to mitigate the systemic risk.

This assessment of the effectiveness of TikTok's proactive detection systems is further informed by the following data on user reports, which serve as a useful indication of the volume of Election Misinformation that is not proactively detected and removed by TikTok's systems. User reports accounted for a minority of the Election Misinformation detected and removed by TikTok, at 1405 of the 75,335 total videos detected and removed under 'Civic & Elections Integrity'. TikTok remains committed to continuously working to improve and iterate its proactive detection capabilities, while noting that Election Misinformation is a rapidly evolving risk area, with nuance across regions and languages.

TikTok's user reporting tools are an important component of its content moderation process, ensuring that Election Misinformation is detected and removed. The data demonstrates that when TikTok receives user reports of violative Election Misinformation content, it is actioned efficiently. Under 'Civic & Election Integrity', 1,383 of the 1405 total videos removed following user reports were removed within

two hours of receiving the report. Collectively, TikTok's proactive detection and reporting tools have worked effectively to ensure that 74,268 of the 75,335 total videos removed under this policy were removed within 24 hours of upload.

The effectiveness of TikTok's moderation of Election Misinformation is indicated by its appeals data, as users can submit appeals against content removals if they believe TikTok has made a mistake. Under the 'Civic & Election Integrity' policy, only 1,885 of the 75,335 total videos removed were successfully appealed and reversed.

TikTok also reports on its efforts to detect and remove covert interference operations ('CIO'). TikTok's global data shows that the Platform detected and removed 51 separate new CIO networks from 1 July 2023 to 30 June 2024. These efforts were reinforced by the removal of 23,923 accounts from 1 January to 30 June 2024 associated with previously disrupted networks attempting to re-establish their presence.

TikTok has also collected data regarding user engagement with the Election Centre pages set-up in the context of the EU Parliamentary Elections of June 2024. The Election Centres were linked to users through labels on relevant videos (e.g. videos that engaged with election topics), LIVE streams, and search, with the pages containing authoritative information tailored to each EU member state. From 6 May to 9 June 2024, Election Centre labels received over 4.5 billion views (4,753,349,948) on video and LIVE, with 8,808,681 clicks through to see more information in the Election Centre. Similarly, Elections Centre labels on relevant searches, received over 63 million views (63,613,989) and 46,261 clicks.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok's strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual risk discussed above. TikTok plans to devote extra resources to combatting risks of Election Misinformation in the year ahead and as a result it remains a Tier 1 risk. As noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(c)(i)(f), Elections Integrity Programme:** TikTok will continue to deliver its Elections Integrity Programme for forthcoming elections in EU member states, elevating authoritative sources, civic information and media literacy through in-app product features (e.g. the Election Centre). [REDACTED]
- **Article 35(1)(c), Synthetic and manipulated media:** TikTok plans to take steps to strengthen enforcement in relation to Synthetic and Manipulated Media, [REDACTED]
- **Article 35(1)(e), Adaptation of advertising systems:** TikTok plans to implement a new policy to restrict suspected political UGC where it is suspected of being unlabelled branded content. Additionally, TikTok plans to update its GPPPA Election Campaign Fundraising Restriction policy in order to improve operability and enforcement. TikTok will also continue to increase the readiness of human and automated moderation systems for emerging election risks, and continue monitoring for urgent election-related incidents, to ensure that election and political-related content cannot be monetised on the Platform; and

- **Article 35(1)(c), Election misinformation:**

14. CIVIC INTEGRITY: RISKS TO PUBLIC HEALTH FROM MEDICAL MISINFORMATION CONTENT

1. Description of the risk:

a. Description of the risk in Year 1

- TikTok understands the risk to be the actual and foreseeable negative effects arising from the dissemination of verifiably (by a recognised medical authority such as the World Health Organisation) false or misleading Medical Misinformation, such as misleading statements about vaccines, inaccurate medical advice that may cause imminent negative health effects such as discouraging people from getting appropriate medical care for a life-threatening disease, and other misinformation that poses a risk to public health (together '**Medical Misinformation**');
- This risk may arise from attempts to share or disseminate the following content on or through the Platform, whether as short video, comment, livestream or within their profile information:
 - Content undermining the existence or severity of COVID-19 (e.g. that the COVID-19 pandemic is a hoax/scam/exaggerated);
 - Medical Misinformation regarding transmission and prevention of COVID-19 (e.g., that COVID-19 tests cause adverse effects/illness, or that face masks are harmful or will cause illness);
 - Medical Misinformation regarding vaccines, including COVID-19 vaccines (e.g., that COVID-19 vaccines change people's DNA, RNA, or genetic makeup; or that COVID-19 vaccines will be used for mind control/to track people); and
 - Medical Misinformation related to serious medical conditions/life-threatening diseases, including but not limited to COVID-19, HIV/AIDS, Ebola, strokes, cancer, heart diseases, tuberculosis, diabetes, and zika, and other similar viruses or conditions as they may arise; or
 - Other Medical Misinformation regarding holistic/homoeopathic remedies (e.g., that drinking or inhaling cleaning/corrosive substances as a preventative or treatment for any disease or that drinking or eating a herbal remedy can treat cancer or other life threatening illnesses).
- In addition, TikTok notes that Article 35 of the Charter enshrines for everyone the right of access to preventive health care and a high level of human health protection as part of EU policies and activities, which may be undermined by Medical Misinformation.

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, TikTok classified Medical Misinformation as a **Tier 2 priority** in its priority tiering system.

This section summarises the risks identified in TikTok's Year 1 Risk Assessment Report and outlines additional risks identified in Year 2. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk detection processes, and extensive consultations with external stakeholders.

Summary of risks identified in Year 1

TikTok identified the following risks as part of its Year 1 Risk Assessment:

- **The risk that content undermines the existence or severity of COVID-19** (e.g. that the COVID-19 pandemic is a hoax/scam/exaggerated);
- **The risk of Medical Misinformation** regarding transmission and prevention of COVID-19 (e.g., that COVID-19 tests cause adverse effects/illness, or that face masks are harmful or will cause illness);
- **The risk of Medical Misinformation regarding vaccines**, including COVID-19 vaccines (e.g., that COVID-19 vaccines change people's DNA, RNA, or genetic makeup; or that COVID-19 vaccines will be used for mind control/to track people);
- **The risk of Medical Misinformation related to serious medical conditions/life-threatening diseases**, including but not limited to COVID-19, HIV/AIDS, Ebola, strokes, cancer, heart diseases, tuberculosis, diabetes, and zika, and other similar viruses or conditions as they may arise; and
- **The risk of other Medical Misinformation regarding holistic/homoeopathic remedies** (e.g., that drinking or inhaling cleaning/corrosive substances as a preventative or treatment for any disease or that drinking or eating a herbal remedy can treat cancer or other life threatening illnesses).

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of Medical Misinformation.

- **Misinformation leading to potentially life-threatening harm:** TikTok has seen an increase in Medical Misinformation claims referring to conditions such as cancer, heart disease, HIV/AIDS, abortion, birth control, pregnancy, serious disabilities, birth defects, strokes and certain chronic illnesses such as Alzheimers or diabetes, that could potentially result in serious harm. While this was identified in Year 1, health discussions and interest in topics have shifted away from COVID-19 and include more broadly:
 - Misinformation concerning the conditions of a disease;
 - Misinformation concerning treatment plans (including dissuading individuals from seeking appropriate treatment and medical care); and
 - Misinformation concerning the promotion of unproven remedies.

This type of Medical Misinformation can arise on the Platform, particularly when these topics intersect with other viral trends or developments.

- **Misinformation regarding non-life threatening treatments and alternative prevention/treatment options:** While noting that not all Medical Misinformation will lead to a risk of significant or immediate negative health consequences for individuals and/or society at

large, certain claims about health issues and/or treatments can still present a moderate risk and a proportionate approach is still required. Specifically, TikTok has identified the potential for an increase in content which may encourage the use of alternative prevention or treatment methods which are medically unsupported or unproven, or content which misrepresents authoritative sources (i.e. selectively referencing certain scientific data to support a conclusion that is counter to the findings of the study).

c. Inherent Risk in Year 2

For a detailed analysis of how TikTok assessed the baseline severity and probability for Medical Misinformation in Year 1, please refer to TikTok’s Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of Medical Misinformation to be ‘Moderate’ in Year 2. TikTok assesses the probability of Medical Misinformation in Year 2 to be ‘Unlikely’. Taking into account severity and probability in Year 2, TikTok has assessed the inherent risk (that is, the risk without any mitigations in place) of Medical Misinformation being shared on the Platform to be ‘Low’.

2. Mitigation Measures:

For detailed information on TikTok’s system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-K	DESCRIPTION
(b) Adaptation of terms and conditions	In June 2024, TikTok updated its Community Guidelines to prohibit moderate harm health misinformation, which means false or misleading content regarding the treatment or prevention of injuries, conditions, or illnesses that are <u>not</u> immediate or life-threatening. Alongside this, TikTok updated its moderator guidelines to reflect this change for moderators to align decisions with this new policy.
(c) Adaptation of content moderation processes	Since Year 1, TikTok has updated its fact checker rating classifications to ensure moderation teams are able to identify Misinformation with more accuracy (read section 13 mitigations). Additionally, TikTok has developed updated moderator guidance to align with updates to the moderate harm health misinformation policy.
(f) Reinforcing risk detection measures	TikTok has committed to continued risk monitoring for Medical Misinformation as mentioned in Year 1 as a key mitigation improvement.

	TikTok's primary detection model used for detecting general misinformation, including Medical Misinformation, was updated to factor in the numerous different policy titles from its training data to increase precision.
(g) Initiating or adjusting cooperation with trusted flaggers in accordance with Article 22 and the implementation of the decisions of out-of-court dispute settlement bodies pursuant to Article 21	TikTok remains ready to engage with EU 'trusted flaggers' as they continue to be designated (read section 5) and in the meantime, TikTok continues to receive reports from Community Partners and subject matter experts.

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness and proportionality of TikTok's mitigations relevant to the systemic risk of Medical Misinformation, TikTok has assessed residual risk to be 'Low' in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate Medical Misinformation on the Platform.

To support TikTok's assessment of the effectiveness of its mitigations against the systemic risk of Medical Misinformation, TikTok has considered data on its enforcement of its Community Guidelines in the EU for the period of Q3 2023 to Q2 2024, and additional data related to user interactions with video notice tags in order.

Under the Community Guidelines policy of 'Misinformation', TikTok removed 517,353 total videos in the EU, with 502,285 detected and removed proactively, and 411,072 removed without any views.²⁶ This high proportion of proactive removal suggests that TikTok's content moderation systems remain effective in detecting and removing Medical Misinformation on the Platform, helping to mitigate the systemic risk.

This assessment of the effectiveness of TikTok's proactive detection systems is further informed by the following data on user reports, which serve as a useful indication of the volume of Medical Misinformation that is not proactively detected by TikTok's proactive systems. User reports accounted for a minority of Medical Misinformation detected and removed by TikTok, at 15,068 of the 517,353 total videos detected and removed under the 'Misinformation' policy. TikTok remains committed to continuously working to improve and iterate its proactive detection capabilities, while noting that Medical Misinformation is a rapidly evolving risk area, with nuance across regions and languages.

²⁶ TikTok's 'Misinformation' policy prohibits various types of misinformation, including health, climate change, conspiracy and public safety related misinformation. This data has been used for the purposes of both the Medical Misinformation and Harmful Misinformation risk sections. Election Misinformation is actioned separately under TikTok's 'Civic and election integrity' policy.



TikTok's user reporting tools are an important component of its content moderation process, ensuring that Medical Misinformation is detected and removed. The data demonstrates that when TikTok receives user reports of violative content, it is actioned efficiently. Under the 'Misinformation' policy, 9,183 of the 15,068 videos removed following user reports were removed within two hours of receiving the report. Collectively, TikTok's proactive detection and reporting tools have worked effectively to ensure that 406,980 of the 517,353 total videos removed under the 'Misinformation' policy were removed within 24 hours of upload.

The effectiveness, reasonableness and proportionality of TikTok's moderation of Medical Misinformation is indicated by its appeals data, as users can submit appeals against content removals if they believe TikTok has made a mistake. Under the 'Misinformation' policy, only 52,208 of the 517,353 total videos removed were successfully appealed and reversed.

Beyond content removals, TikTok also uses video notice tags to connect users with authoritative sources of information on issues such as Covid-19 and vaccines. Quantitative data reflects strong reach for these notices. Covid-19 notices had 1,894,574,236²⁷ views, notices related specifically to the Covid-19 vaccine received 985,920,146²⁸ views, and monkeypox tags accrued 10,156,710²⁹ total views.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok's strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead to mitigate the risk. TikTok continues to be vigilant to existing risks identified in last year's assessment, including risks identified around medical misinformation that poses significant harm to individuals including but not limited to Covid-19 misinformation and vaccine misinformation and has identified additional emerging areas of risk, albeit the overall likelihood of harm has assessed to be comparatively lower to the TikTok community as the risk environment is less concentrated following the end of the pandemic. TikTok has closely considered its risk environment and the inherent and residual risk discussed above and has amended this risk to a Tier 3 risk in this Year 2 prioritisation. However, as noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(c), Content moderation update:** [REDACTED]
- **Article 35(1)(b), Policy development:** TikTok plans to develop a new policy for low harm and

²⁷ Data updated 2 October 2024

²⁸ Data updated 2 October 2024

²⁹ Data updated 2 October 2024

high prevalence medical misinformation.

15. CIVIC INTEGRITY: RISKS TO PUBLIC SECURITY FROM HARMFUL MISINFORMATION CONTENT

1. Description of the risk:

a. Description of the risk in Year 1

- TikTok understands the risk to be the actual or foreseeable risks to public safety or security arising out of Harmful Misinformation/content which may relate to: armed conflicts and emerging conflicts; acts of terrorism; natural and manmade disasters (such as floods, earthquakes, hurricanes, fires, landslides, environmental or industrial accidents); and other emergency situations that may induce panic, including in relation to current/unfolding events, such as civil unrest (such as protests or riots);
- This risk may arise from attempts to share or disseminate the following content on or through the Platform, whether as short video, comment, livestream or within their profile content that includes:
 - Misinformation making verifiably false and harmful claims regarding natural and manmade disasters (such as floods, earthquakes, hurricanes, fires, landslides, environmental or industrial accidents);
 - Misinformation making verifiably false and harmful claims regarding unfolding shooting events and mass murders;
 - Misinformation making verifiably false and harmful claims regarding public demonstrations or protests;
 - Repurposing old video content, making verifiably false and harmful claims that the event or video is new/current and likely to trigger societal panic (e.g., misleadingly repurposing footage of a bombing or armed attack out of context);
 - Making verifiably false and harmful claims that basic necessities (e.g., food, water.) or services (e.g., banks, cash machines) are no longer available in a particular location, causing hoarding;
 - Stating 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; or
 - Incitement to violence and criminal acts, such as property damage.
- In addition, TikTok notes that Article 6 of the Charter enshrines for everyone the right to liberty and security of person, and Article 12 of the Charter enshrines the right to freedom of peaceful assembly and to freedom of association at all (in particular in political, trade union and civic matters), and that protection of these rights in particular could be restricted or jeopardised in the context of Public Security Risks. TikTok also recognises that in the context of Public Security Risks, individuals retain other fundamental rights under the Charter, in particular the rights to freedom of expression and information, and TikTok's freedom to conduct a business, which must be balanced in a proportionate manner in the context of addressing Public Security Risks.

b. Risk Identification: changes to the risk profile since Year 1:

In its Year 1 Risk Assessment, TikTok classified Public Security Risks as a **Tier 2 priority** in its priority tiering system.

This section summarises additional risks identified in Year 2. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk detection processes, and extensive consultations with external stakeholders.

Summary of risks identified in Year 1

- **Sharing and dissemination of Harmful Misinformation:** The risk that TikTok may be used to attempt to share and disseminate Harmful Misinformation, posing threats to public security in the EU (e.g. inciting riots and civic unrest). This may manifest across TikTok's content formats, including short video or photo, LIVE, comment, or profile features; and
- **Synthetic and Manipulated Media:** The risk that this type of content may be shared on the Platform with additional acute risks of deception and harm; for example, by depicting false instances of current/ongoing violence causing off-platform panic. Content using the likeness of real people may be shared to mislead users about real-world events.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of Harmful Misinformation.

- **AIGC:** TikTok identifies new manifestations of harmful and misleading AIGC as part of its ongoing risk identification of synthetic and manipulated media. This risk is monitored closely, in the context of high-profile events and emergencies including elections, armed conflicts, and crises. Potential new manifestations of misleading AIGC include instances where content has been created using an AIGC tool and which features:
 - False depictions of public figures being endorsed or attacked by specific groups or individuals;
 - False depictions of public figures admitting to criminal activity or otherwise making damaging admissions;
 - Catastrophic or crisis events taking place that did not actually occur;
 - Depictions of youth in ways that might re-victimise or fetishise them or violate privacy, for example by depicting victims of crimes or tragedies, or sexualised portrayals of young people;
 - Content that appears to come from authoritative or official sources such as news outlets/entities; public institutions and peer reviewed research when it does not;
 - Portrayals of public figures being supported or endorsed by individuals or groups that did not do so;
 - Competing claims about whether content is authentic or unverified AIGC, particularly that falsely depicts private conversations between public figures; and
 - Impersonation, privacy and non-consensual depiction of both private and public individuals.

- **Rapidly evolving events:** TikTok closely considered the content risks which arise from unforeseen and unexpected rapidly evolving events. Content about these events becomes outdated quickly and can potentially be unsafe for users who search for live content. TikTok considers an event to be rapidly evolving if it meets all of the following criteria:
 - Details about the event can change so quickly that, in less than 24 hours from being posted, shared content about situational status and recommended guidance could be considered outdated and potentially unsafe for users to follow;
 - It is distinct from misinformation in that the now out-of-date content was not originally misinformation when posted; and
 - It is an event of multinational impact or concern.
- **Climate change misinformation:** As part of its ongoing efforts to identify and mitigate risks, TikTok defines climate change misinformation as content that denies the existence of climate change or the human activities contributing to it, disputes the scientific consensus on climate change, or spreads false claims about climate change solutions. These narratives undermine public understanding and trust in climate science and hinder efforts to address the global climate crisis.

c. Inherent Risk in Year 2

For a detailed analysis of how TikTok assessed the baseline severity and probability for Harmful Misinformation in Year 1, please refer to TikTok's Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of Harmful Misinformation to be 'Moderate' in Year 2. TikTok assesses the probability of Harmful Misinformation in Year 2 to be 'Possible'. Taking into account severity and probability in Year 2, TikTok has assessed the inherent risk (that is, the risk without any mitigations in place) of Harmful Misinformation being shared on the Platform to be 'Moderate'.

2. Mitigation Measures:

For detailed information on TikTok's system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2

MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-K	DESCRIPTION
(b) Adaptation of terms and conditions	<p>As part of TikTok’s ongoing risk mitigation for the risk of ‘repurposed media’ content, in Year 2 TikTok updated its ‘Repurposed Media’ policy now includes the removal of unedited media content that is presented out of context and may mislead a person about a developing topic of public importance, such as showing a crowd at a music concert and suggesting it is a political protest.</p> <p>TikTok added a policy on ‘Misrepresented Authoritative Sources’. This policy prohibits content that promotes misleading correlations or conclusions related to authoritative information that is recognised and trusted, such as reports from research institutions. An example would be content that selectively references certain scientific data to support a conclusion that is counter to the findings of the study.</p> <p>TikTok has updated its AIGC and Edited Media Policy to include prohibitions on: depictions of public figures engaging in or admitting to negative behaviours; claims about public figures’ personal characteristics of political or public importance; false endorsements or condemnations; information falsely appearing to come from authoritative sources; and misleading depictions of public importance, including fictitious crisis events.</p>
(c) Adaptation of content moderation processes	<p>TikTok updated its task prioritisation feature and fact checker rating classifications to ensure moderation teams are able to identify Misinformation with more accuracy (read section 13 mitigations). Additionally, TikTok has developed updated moderator guidance to align with these updates.</p>
(f) Reinforcing risk detection measures	<p>As mentioned in Year 1 as a key mitigation improvement, TikTok updated its primary risk detection model, used for detection of general misinformation as well as Medical Misinformation, to better detect sub-categories of misinformation (see Section 13 Mitigations)</p> <p>TikTok also developed an improved system for receiving and utilising third-party intelligence from Fact Checking partners, streamlining the review of actionable intelligence relating to on-platform risk.</p>

(i) Awareness-raising measures and adoption of online interface in order to give recipients of the service more information

TikTok created and launched search guides to minimise the discoverability of Harmful Misinformation regarding rapidly evolving events. These search guides appear to users on the Platform in response to searches relating to rapidly evolving events and assist users in finding authoritative and reliable information. This functionality, launched for the Israel-Hamas war and the Crocus Hall attacks (a terrorist attack carried out against civilians at Moscow’s Crocus City Hall on 22 March 2024), can be quickly deployed for crisis events that may be particularly susceptible to misinformation risks and to critical information becoming quickly outdated.

TikTok collaborated with fact-checking partners and local media literacy bodies to produce in-app media and digital literacy campaigns to help inform and educate users about identifying and reporting Harmful Misinformation. In August 2023, TikTok rolled out an additional 13 new media literacy campaigns in Europe on topics identified as priority areas for increased digital literacy for TikTok users in Europe.

During major events, including COP28 and Earth Month, TikTok adopts an online interface in the form of a pop-up at the top of the search bar, to direct users to an in-app page with some of the most relevant content about climate action.

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to identifying and mitigating Public Security Risks, please refer to Annex 4.

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness and proportionality of TikTok’s mitigations relevant to the systemic risk of Harmful Misinformation, TikTok has assessed residual risk to be ‘Moderate’ in Year 2.

To support TikTok’s assessment of the effectiveness of its mitigations against the systemic risk of Harmful Misinformation, TikTok has considered relevant data on its enforcement of its Community Guidelines in the EU from Q3 2023 to Q2 2024, and additional data relating to video notice tags. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate Harmful Misinformation on the Platform.

Under the Community Guidelines policy of ‘Misinformation’, TikTok removed 517,353 total videos in the EU, with 502,285 detected and removed proactively, and 411,072 removed without any views.³⁰ This high proportion of proactive removal suggests that TikTok’s content moderation systems remain effective in detecting and removing Harmful Misinformation on the Platform, helping to mitigate the systemic risk.

³⁰ TikTok’s ‘Misinformation’ policy prohibits various types of misinformation, including public safety, conspiracy, health, and climate change related misinformation. This data has been used for the purposes of both the Harmful Misinformation and previous Medical Misinformation risk sections. Election Misinformation is actioned separately under TikTok’s ‘Civic and election integrity’ policy.

This assessment of the effectiveness of TikTok’s proactive detection systems is further informed by the following data on user reports, which serve as a useful indication of the volume of Harmful Misinformation that is not proactively detected by TikTok’s proactive systems. User reports accounted for a minority of Harmful Misinformation detected and removed by TikTok, at 15,068 of the 517,353 total videos detected and removed under the ‘Misinformation’ policy. TikTok remains committed to continuously working to improve and iterate its proactive detection capabilities, while noting that Harmful Misinformation is a rapidly evolving risk area, with nuance across regions and languages.

TikTok’s user reporting tools are an important component of its content moderation process, ensuring that Harmful Misinformation is detected and removed. The data demonstrates that when TikTok receives user reports of violative content, it is actioned efficiently. Under the ‘Misinformation’ policy, 9,183 of the 15,068 videos removed following user reports were removed within two hours of receiving the report. Collectively, TikTok’s proactive detection and reporting tools have worked effectively to ensure that 406,980 of the 517,353 total videos removed under the ‘Misinformation’ policy were removed within 24 hours of upload.

The effectiveness, reasonableness and proportionality of TikTok’s moderation of Harmful Misinformation is indicated by its appeals data, as users can submit appeals against content removals if they believe TikTok has made a mistake. Under the ‘Misinformation’ policy, only 52,208 of the 517,353 total videos removed were successfully appealed and reversed

Beyond content removals, TikTok also uses video notice tags to connect users with authoritative sources of information. Quantitative data reflects strong reach; Holocaust video notice tags received 360,164,504³¹ total views.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok’s strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual risk discussed above. TikTok plans to devote extra resources to combating Public Security Risks in the year ahead and as a result it remains a Tier 2 priority. As noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(c), Content moderation tooling:** [REDACTED]

³¹ Data updated 2 October 2024

- **Article 35 (1)(c),** [REDACTED]
- **Article 35(1)(b), Enforcement on misinformation:** [REDACTED]
- **Article 35(1)(i), Awareness raising:** TikTok is planning on launching media literacy campaigns in nine more countries, meaning that by the end of the year, it will have run localised campaigns in all 27 member states;
- **Article 35(1)(c), Content moderation updates:** [REDACTED]
- **Article 35(1)(b), Policy development:** [REDACTED]

16. FUNDAMENTAL RIGHTS: RISKS TO FUNDAMENTAL RIGHTS

1. Description of the risk

a. Description of the risk in Year 1

- TikTok understands the risks to the exercise of 'Fundamental Rights' on its Platform to comprise the rights set out below, as enshrined in the Charter;
- Following consideration of any actual or foreseeable negative effects for the exercise of fundamental rights as protected under the Charter, TikTok has determined the following fundamental rights as most relevant to its Platform: (1) the right to human dignity; (2) the right to non-discrimination; (3) the right to freedom of expression; (4) the right to private and family life; (5) the right to the protection of personal data; and (6) the right to consumer protection; and
- Within the context of this Report, it is not possible to provide a detailed analysis of each Fundamental Right described above (and as noted above, various other fundamental rights are considered in this Report in the context of other risks).

b. Risk Identification: changes to the risk profile since Year 1

In its Year 1 Risk Assessment, TikTok classified risks to Fundamental Rights as a **Tier 3** priority in its priority tiering system.

This section summarises additional risks identified in Year 2. In the spirit of continuous improvement, these areas of focus result from the risk assessment conducted in Year 1, internal risk detection processes, and extensive consultations with external stakeholders. These additional risks are described below:

Summary of risks identified in Year 1

- **Content risks:** The risk that content uploaded by TikTok's users may undermine Fundamental Rights by being discriminatory or dehumanising and/or seeking to scam or defraud other users or other violations of rights;
- **Conduct risks:** The risk of users misunderstanding or ignoring behaviour that is prohibited on the Platform set out in its Terms of Service, Community Guidelines, Privacy Policy and Health, Safety and Transparency Centres; and
- **Moderation risks:** The risk that, in the interest of safety, TikTok's content moderation systems and human moderators may over-moderate by mistakenly acting on speech that doesn't violate Community Guidelines or under-moderate by failing to recognise evolving discriminatory language or other violating content.

Additional risks identified in Year 2

In Year 2, TikTok identified the following additional risk areas, which it considers to be reflective of the dynamic and evolving nature of risks to Fundamental Rights:

- **AI models:** There is an inherent risk that features and advancements in content moderation models may be developed in a biased or discriminatory manner. This risk arises from the potential development of these features and policies using insufficiently inclusive data sets, safeguards, and safety controls;
- **Armed conflicts:** This is a new emerging risk that is related to the rise in conflicts in various parts of the world resulting in risks to users and human rights defenders. The Israel-Hamas war highlighted room for enhancements in how our policies can best uphold human dignity with regard to content depicting victims of conflict or hostages; and
- **Bias:** Manual policy enforcement may be negatively impacted by political bias, in particular in the context of increased political developments and events, such as the numerous elections in 2024. This bias may manifest itself at the enforcement stage, thus undermining freedom of expression and the right to nondiscrimination, especially when there are increased levels of political and contentious content.

c. Inherent Risk in Year 2

For a detailed analysis of how TikTok assessed the baseline severity and probability for risks to Fundamental Rights in Year 1, please refer to TikTok's Year 1 Systemic Risk Assessment Report.

In Year 2, TikTok assesses the overall severity of risks to Fundamental Rights risks to Fundamental to be 'High' in Year 2. TikTok assesses the probability of risks to Fundamental Rights Year 2 to be 'Possible'. Taking into account severity and probability in Year 2, TikTok has assessed the inherent risk (that is, the risk without any mitigations in place) to Fundamental Rights to be 'Material'.

2. Mitigation Measures:

For detailed information on TikTok's system of mitigations/controls (under DSA Art. 35 (1)(a)-(k)) that applied when its designation as a VLOP came into force, please refer to the Year 1 Risk Assessment Report. That Year 1 Report identified further mitigation improvements, progress on which (and any other such mitigations) is listed below. In the spirit of continuous improvement, these areas of focus

result from the risk assessment conducted in Year 1, internal risk monitoring and detection processes, and extensive consultations with external stakeholders.

Implementation of Additional Mitigation Effectiveness Measures in Year 2	
MEASURES IN ACCORDANCE WITH DSA ART. 35(1) A-K	DESCRIPTION
(a) Adaptation of platform features or design	<p>As mentioned in the Year 1 Report, TikTok operates an Inclusion Advisory Committee. TikTok’s Inclusion Advisory Committee is an internal group made up of employees across the globe with experiences of vulnerable and marginalised communities, including those focused on disability, women’s empowerment, racial and ethnic identity, LGBTQ+ inclusion, and more. TikTok has worked with the Inclusion Advisory Committee to ensure marginalised and vulnerable groups’ perspectives are considered in feature design, leading to projects like understanding LGBTQ+ experiences with filters and effects, and developing an Accessibility Roadmap for more inclusive product features.</p>
(b) Adaptation of terms and conditions	<p>TikTok also made changes to the Community Guidelines by addressing multiple unfolding and ongoing armed conflicts throughout 2023-24 including adding to their violent actors designation lists, clarifying public interest expectations such as for counter speech by providing examples to moderators, updating guidance on dangerous conspiracy theories and bullying.</p> <p>Also, in relation to the potential risk that inadequate mental and physical health support leads to leakage of violative content on TikTok and/or over moderation of content, the following actions were taken:</p> <ul style="list-style-type: none"> ● A well-being assessment of metric demands to surface potential mitigations (e.g., restructuring to allow for well-being time). This assessment was conducted and resulted in the creation of a user guide outlining the importance of wellness in relation to psychological recovery intended to provide all managers, team leaders and content moderators a clear understanding of the available wellness principles; and ● Sharing research and best practice with partners and peers and engaging in cross-industry collaboration on solutions. TikTok is a member of TSPA³² and currently engages in their cross-industry group on moderator well-being.
(c) Adaptation of content moderation processes	<p>In November 2023, following a series of discussions with the Jewish Creators Roundtable, TikTok formed a taskforce to enhance its efforts in combating antisemitism, islamophobia, homophobia, and other forms of</p>

³² <https://www.tspa.org/about-tspa/supporters/>

	<p>harm toward marginalised communities. The taskforce aimed to improve the Platform's training methods, focusing on updating implicit bias training with input from external partners. They also implemented new reminders about hate policies and introduced e-learning modules to reinforce these updates. These efforts reduced moderator errors and improved accuracy, with over 75% of moderators finding the training highly effective.</p> <p>TikTok strives to embed human rights standards in policy development and enforcement. For example, in Year 2, TikTok has issued moderator guidance to differentiate Hate Speech from advocacy, thereby protecting freedom of expression. TikTok's Platform Fairness team also conducted the following:</p> <ul style="list-style-type: none"> • Changed the term 'prostitution' to 'sexual services' in Community Guidelines to reflect Human Dignity recommendations and minimise penalisation of sex workers; • Mapped risks related to human dignity and discrimination, identifying markets prone to Anti-LGBTQ+ sentiment, Anti-immigrant sentiment, punishment of content creators, and censorship of dissent, and proposed recommendations on proportionate content moderation responses; and • Addressed potential biases in policy development and moderation through impact assessments to ensure fairness in Community Guidelines enforcement.
<p>(d) Adaptation of algorithmic systems</p>	<p>In May 2024, the Platform Fairness team revised its Fairness Guidelines by providing specific guidance for different model types, like classification, and included a new section on AIGC. Instead of giving broad advice for all models, the new guidelines offer detailed instructions, which are now part of the model review process. If a model is found to be high risk against this review process, these guidelines are used for fairness testing. To revise these guidelines, TikTok looked at images based on gender expression, skin tone, and body size, and checked how these attributes were distributed across different categories. For each test, TikTok generated 50 images covering various areas where fairness issues often come up, such as jobs, wealth, education, crime, beauty, and more. This helps ensure TikTok's models treat people fairly and align internal processes with Community Guidelines.</p>
<p>(f) Reinforcing risk detection measures</p>	<p>TikTok enhanced its implicit bias training for moderators, particularly around antisemitism and anti-Islamophobia and integrated international human rights standards like the International Covenant on Civil and Political Rights and Rabat Plan in policy reviews. The objective was to protect freedom of expression and prevent discrimination.</p>

Other - right to data protection

TikTok facilitates the right of access through the 'Download Your Data' feature, allowing users to request a copy of their personal data, with additional requests handled via a privacy webform. The right of portability is supported by enabling users to transfer their TikTok data to third parties through an API, in compliance with the EU General Data Protection Regulation ('GDPR') and the DSA. Users can exercise the right of erasure by deleting their data or requesting deletion via in-app controls and reporting features. For rectification, users can correct their information through in-app controls or request changes from TikTok. The right of objection allows users to contest data processing for public interest or legitimate interests via a webform. Finally, the right of restriction permits users to request data processing limitations under specific conditions, such as accuracy disputes or legal claims.

3. Residual Risk in Year 2:

Following an assessment of the effectiveness, reasonableness and proportionality of TikTok's mitigations relevant to the systemic risk to Fundamental Rights, TikTok has assessed residual risk to be 'Moderate' in Year 2. TikTok commits to continuing to expand and invest in mitigation measures to identify and mitigate risks to Fundamental Rights on the Platform.

This assessment was conducted on the basis of qualitative and quantitative insights, including data on TikTok's enforcement of its Community Guidelines in the EU from Q3 2023 to Q2 2024. Systemic risks to Fundamental Rights are relevant across all risk sections, and data in those respective sections was consulted in assessing the effectiveness of TikTok's mitigations. For example, the removal of content under TikTok's 'Harassment & Bullying' policy, as documented in the Hate Speech section, is relevant to TikTok's mitigation of risks to the fundamental right to Human Dignity.

Additionally, TikTok has considered data from other relevant CGER policy categories for this risk section. First, under the Community Guidelines policy of 'Suicide & Self Harm', TikTok removed 622,924 total videos in the EU, with 600,489 detected and removed proactively, and 521,259 removed without any views. Second, under the 'Human Exploitation' policy, TikTok removed 64,797 total videos in the EU, with 58,227 detected and removed proactively, and 48,376 removed without any views. Third, under the 'Sexual Activity & Services' policy, TikTok removed 4,213,632 total videos in the EU, with 3,496,944 detected and removed proactively, and 1,461,933 removed without any views. Fourth, under the 'Shocking & Graphic Content' policy, TikTok removed 1,717,094 total videos in the EU, with 1,661,537 detected and removed proactively, and 1,398,038 removed without any views. This high proportion of proactive removal suggests that TikTok's content moderation systems remain effective in detecting and removing violative content negatively impacting Fundamental Rights on the Platform, helping to mitigate the systemic risk.

This assessment of the effectiveness of TikTok's proactive detection systems is further informed by the following data on user reports, which serve as a useful indication of the volume of content negatively impacting Fundamental Rights that is not proactively detected and removed by TikTok's proactive systems. User reports accounted for a minority of relevant violative videos detected and removed by TikTok; at 22,435 of the 49,292 total videos detected and removed under 'Suicide & Self Harm', 6,570 of the 64,797 total videos removed under 'Human Exploitation', 716,688 of the 4,213,632 total videos

removed under 'Sexual Activity & Services', and 55,557 of the 1,717,094 total videos removed under 'Shocking & Graphic Content'. TikTok remains committed to continuously working to improve and iterate its proactive detection capabilities, while noting that Fundamental Rights is a rapidly evolving risk area, with nuance across regions and languages.

Fundamental rights, particularly Freedom of Expression, may also be negatively impacted by content moderation that is not reasonable and proportionate. TikTok's appeals mechanism allows users to challenge content removal decisions if they believe TikTok has made a mistake. The reasonableness and proportionality of TikTok's content moderation is demonstrated in the very low share of content removal restrictions that result in successful appeals. Under 'Suicide & Self Harm, 57,439 of the 622,924 total videos removed were successfully appealed. Under 'Human Exploitation', 5,493 of the 64,797 total videos removed were successfully appealed. Under 'Sexual Activity & Services', 95,305 of the 4,213,632 total videos removed were successfully appealed. And under 'Shocking & Graphic Content', 147,776 of the 1,717,094 total videos removed were successfully appealed.

4. Key stakeholder engagement:

For detailed information on how TikTok has engaged with external stakeholders to inform its approach to risk identification and risk mitigation, please refer to Annex 4. This Annex provides an in-depth overview of the collaborative efforts and consultations that have shaped TikTok's strategies in these areas.

5. Prioritisation:

TikTok has reported above on its progress in the last year, and below it states key additional actions for the year ahead. TikTok has closely considered its risk environment and the inherent and residual risk discussed above. TikTok plans to devote extra resources to combatting risks to Fundamental Rights in the year ahead and as a result it has been amended to a Tier 2 risk in Year 2. As noted in Year 1, TikTok will continue to keep its prioritisation under review.

6. Planned further mitigation effectiveness improvements:

- **Article 35(1)(c), Anti-bias in content moderation:** TikTok plans to revamp its training materials for moderators to continually ensure moderation decisions are fair and unbiased. The training update will incorporate new deep dive training on: gender and sexuality, age assurance, race and ethnicity, and divisive political and religious content.
- **Article 35(1)(e), Advertising systems:** TikTok plans to incorporate additional reviews on ads selected and presented to ensure ads are fair and unbiased.
- **Article 35(1)(c),** [REDACTED]
- **Article 35(1)(f) Monitoring of Human Rights Risks:** [REDACTED]

ANNEX 1: KEY INFORMATION ABOUT TIKTOK

What is TikTok?

TikTok's mission is to inspire creativity and bring joy. TikTok is available in many countries globally and its global headquarters are in Los Angeles and Singapore, and its offices include New York, London, Dublin, Paris, Berlin, Dubai, Jakarta, Seoul, and Tokyo. TikTok had on average 142 million monthly active users in the EU between August 2023 and January 2024. TikTok is offered primarily as a mobile app, web app, mobile browser and web browser.

What are the main features of TikTok?

TikTok's video content gives users quick and engaging content experiences whenever they want it. Content is served based on interests and user engagement so entertainment is always personal and connects people from all around the world through shared humour, interests and passions. A non-personalised feed is also available to users in the EU. TikTok's features include video, photo, livestream and comments. Users can share, skip, swipe, like, comment on, or replay videos. Users can also Duet (side by side) with the video of another creator or Stitch another creator's content into their video. Users can send virtual gifts to creators whose content they like.

How does the For You page feed work?

The For You Feed ('FYF') is a unique TikTok feature that uses a personalised recommendation system to allow each community member the ability to discover a breadth of content, creators, and topics. In determining what gets recommended, the system takes into account factors such as likes, shares, comments, searches, diversity of content, and popular videos.

TikTok maintains content Eligibility Standards for the FYF that prioritise safety and are informed by the diversity of TikTok's community and cultural norms. TikTok makes certain content ineligible for the FYF as it may not be appropriate for a broad audience and may also make some of this content harder to search for. In order to prevent harmful content from appearing in the FYF, TikTok operates a range of processes and provides an additional layer of controls for minors. The FYF is designed to ensure that users are given opportunities to discover new interests and are not presented with types of content that can create harm when encountered in heavy concentrations.

What does advertising look like on TikTok?

Businesses show ads on TikTok to reach the people they care about in a creative and meaningful way. This helps keep TikTok free for users. TikTok publishes a Guide to Ads and Your Data and is committed to being transparent with its users about how it collects, uses and shares data for ads. In Europe, minors do not see ads based on profiling but will see generic ads.

TikTok's advertising policies determine the type of products and services that can be advertised on TikTok. Users will see different kinds of ads when they use TikTok. Users can interact with the ad in much the same way as content posted by other users. For example, users can share, skip, swipe, like,

or replay an ad. Users can also comment on an ad if the advertiser enables that feature for a particular ad. If users consider an ad to be violative of TikTok's advertising policies, they can report it to TikTok.

How does TikTok identify and take action on violative content?

TikTok operates proactive and systematic measures to identify, remove or restrict access to content and accounts that violate its Community Guidelines, Terms of Service or advertising policies. TikTok's content moderation is fundamental to its overarching risk management strategy as it underpins its ability to respond effectively to existing and emerging risks. TikTok's risk management strategy places considerable emphasis on proactive content moderation where it endeavours to detect and remove violative content before it is reported by users or third parties.

TikTok operates its content moderation processes using automated and manual (human) means in accordance with the following four key principles, which provide that TikTok will:

1. Remove violative content from the Platform that breaks its rules (whilst noting that TikTok does not allow several types of mature content themes, notably nudity and sexual activity which includes, but is not limited to, pornography);
2. Age-restrict mature content (that does not violate its Community Guidelines but which contains mature themes) so that it is only viewed by adults (18 years and older);
3. Maintain FYF eligibility standards to help ensure that any content promoted by its recommendation system is appropriate for a broad audience; and
4. Empower its community with information, tools, and resources.

TikTok has implemented automated and manual content moderation systems and processes as well as a range of other safety features that are developed, maintained and applied by a range of teams. All video, photo and text-based content uploaded to the Platform are subject to a real time, technology-based automated review. While a video is undergoing this review, it is visible only to the uploading user/creator. Users can report user-generated and ad content which they consider to be violative of the Community Guidelines or TikTok's advertising policies (as applicable).

TikTok's Trust and Safety teams action reports of violative content made by users, non-users and by third parties who form part of TikTok's Community Partner Channel (similar to trusted flaggers under DSA). These teams perform a manual review against TikTok's Moderation Policy Framework, which provides moderators with necessary detail on how to apply TikTok's Community Guidelines. Additionally, in compliance with Article 16 DSA, users can report user generated and ad content that they consider to be illegal under European or Member State law.

TikTok takes action in relation to content that it considers violative, which may include a review by its Trust and Safety team to determine whether the content should be removed or made ineligible for the FYF according to the Community Guidelines. Decisions can be appealed. TikTok also operates a 'strikes' policy for accounts that repeatedly post violative content, which could result in various account level actions up to, and including, a permanent account ban. TikTok adopts a range of processes to review the accuracy of decisions when moderating potentially violative content, in order to ensure that content moderation measures are accurate, effective and proportionate (and in particular to ensure they do not disproportionately impact on users' rights to freedom of expression and information).

TikTok's policies and processes for detecting and removing violative ads are similar to those set out above.

An overview of TikTok's data related practices

TikTok offers a suite of privacy settings and controls for users of the Platform. In addition, it has adopted a variety of technical, contractual and organisational measures. TikTok endeavours to ensure the integrity, confidentiality and security of user personal data and to prevent unauthorised access or disclosure of personal data. For example, TikTok implements a multi-tier system of technical access controls (such as system entry controls and technical access controls). TikTok also has in place an incident management program, and forensic capabilities to ensure recovery in case of an incident and the timely restoration of data while ensuring redundancy of the TikTok infrastructure through a backup methodology. Operation and network security are also ensured through the adoption of controls aligned to industry standards, including vulnerability scanning and network monitoring. TikTok has contractual arrangements, with additional supplementary measures in place with group entities and its external service providers to ensure GDPR compliance. TikTok has implemented a number of internal organisational and policy measures to further control access and use of personal data.

TikTok's Privacy Policy applicable to European users of the Platform (and which more broadly covers the European Economic Area, the United Kingdom and Switzerland) provides information and transparency regarding TikTok's processing activities, security measures, and data retention. It also explains to users their rights regarding their personal data, including information on their rights of access, objection, restriction, deletion, rectification and portability of their personal data. This Privacy Policy was updated in November 2023 to address DSA requirements related to sharing data with researchers. Other changes included additional detail for users on the scope and consequences of having a public account and additional information regarding the measures in use to enforce our terms, guidelines and policies.

ANNEX 2: HOW TO USE THIS REPORT

The geographic scope of this Report:

- This Report has been prepared in compliance with Articles 34, 35 and 42 DSA and specifically in accordance with Article 34(2), which requires TikTok to consider the regional and linguistic aspects of any systemic risk in Europe;
- Therefore, the contents of this Report should not be relied upon as representative of TikTok's position outside Europe.

The legal purpose of this Report:

- This Report has been prepared for the limited and specific purposes of Articles 34, 35 and 42 DSA. As such, this Report is not intended to be a definitive statement of TikTok's position on the matters covered, as they may relate to other laws and regulations in Europe;
- This Report should not therefore be relied upon for any other regulatory or litigation purpose,

whether inside or outside Europe.

The contents of this Report:

- This Report covers TikTok's risk assessments which were completed by 28 August 2024 and which rely on information collated prior to that date;
- This Report summarises the results of TikTok's detailed risk assessments. It is not intended, and nor should it be treated as, a comprehensive or exhaustive overview of the detailed analysis undertaken in those underlying risk assessments.

Further resources:

- For further information on how TikTok complies with the DSA, please see its European Online Safety Hub (available at: <https://www.tiktok.com/euonlinesafety/en/>);
- For further information about TikTok's voluntary transparency reporting which shows it enforces its Community Guidelines, please see TikTok's online Transparency Centre (available at: <https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2024-1/>).

ANNEX 3: SPECIALIST TEAMS

TikTok has a diverse network of internal teams, all of whom have a part to play in its strategy to risk management. The teams below may fulfil their functions in supporting TikTok with its response to one or more Risk Module or systemic risk, as follows:

Trust & Safety Risk & Response team: Trust & Safety's Risk & Response team involves the following key functions, which are summarised as follows:

- Incident Management ('IM'): The IM team is a multidisciplinary team of experts who work together to identify, detect, and mitigate risk in response to escalation scenarios. The IM team provides 24/7 incident management and risk handling coverage and support, which primarily involves removal of violating content from the Platform.
- Law Enforcement Response team: TikTok has a dedicated Law Enforcement Response team (known as 'LERT') with responsibility for reviewing law enforcement and governmental requests in line with applicable law and our policies.
- Law Enforcement Outreach: To enhance education efforts and establish direct communications channels with law enforcement agencies, TikTok has a dedicated Law Enforcement Outreach team. This team frequently engages with national law enforcement authorities ('LEAs') across the EU and with international agencies (such as EUROPOL and INTERPOL, and with trusted law enforcement agencies outside the EU) to educate officials on data request submission processes that comply with applicable law and TikTok's policies. An important aspect of outreach engagement is to understand new and emergent trends that help to inform our risk detection and safeguarding strategies.
- Emergency Response Team: TikTok's Emergency Response Team (known as 'ERT') provides 24/7 coverage to ensure any escalations involving imminent harm or the risk of death or serious

physical injury to a person, can be assessed and sent to law enforcement to prevent that harm, as permitted by applicable law. These requests relate to extreme risk, which may involve imminent terrorist-related risks. The team is comprised of specialists who have considerable investigatory and incident-handling experience within the technology industry, a number having previously worked within law enforcement agencies, including the UK's National Crime Agency.

- Risk Analysis team: The Risk Analysis team is staffed by experienced professionals with backgrounds in cyber intelligence and risk detection. This team's role is to monitor risks both in real time and over a period of time to detect emerging trends and patterns, and their activities include monitoring open-source resources and reporting to cross-functional colleagues on potential emerging risks.
- Child Safety team: TikTok's Child Safety Team (known as '**CST**') operates 24/7 to detect, remove and report suspected child sexual exploitation and abuse material to the National Centre for Missing and Exploited Children (NCMEC) and law enforcement authorities. Reports regarding children assessed to remain at imminent and ongoing risk are escalated via priority channels to NCMEC and authorities. The team is comprised of specialists who have considerable investigatory and child safeguarding experience in the technology industry, having previously worked as prosecutors, within law enforcement agencies and child protective and social services.
- Global IP Ops: The purpose of Global IP Ops is to quickly and effectively address any takedown requests regarding alleged copyright and trademark infringements on UGC and to remove any reported content assessed to be infringing with utmost accuracy, in accordance with our terms and conditions / policies and applicable legal requirements. Our aim is to provide a safe environment for our users, uphold the protection of intellectual property rights and safeguard creativity.

Trust & Safety Product Policy: TikTok's Trust & Safety Product Policy team involves four key functions (Issue, Regional, Feature Policy, and Outreach & Partnerships ('**OPM**')), which are summarised as follows:

- Issue Policy Team: The Trust & Safety Product Policy team has a dedicated Issue Policy team that is responsible for TikTok's safety policies. The Issue Policy team includes various specialist teams including:
 - Youth Safety and Well-being Issue Policy Team: The dedicated Youth Safety and Well-being ('**YSW**') policy team are experts in adolescent development, education, and children's rights who consider how Younger Users may be uniquely affected by content, interactions, and platform design features in ways that are developmentally different from experiences for adults. In connection with TikTok's policies and platform features, the YSW policy team's overarching goal is to best protect young people's unique developmental life stages while accounting for diverse global experiences. The team works closely with the other Product Policy Teams, T&S Product, Legal, and other XFN Teams on youth-specific strategies to address harms captured across all T&S policies and feature designs.
 - Content Classification: The Content Classification Policy Team are experts in content classification and maturity ratings and have deep experience and knowledge of existing rating and classification systems, including in Europe, that exist for television, film and games; emerging research into online media and digital entertainment; and developmental psychology. Many of these experts have previous experience rating content at companies like YouTube and Netflix.

- Integrity & Authenticity: The Integrity & Authenticity Team has subject matter expertise on Integrity & Authenticity issues and specific expertise on Election Misinformation issues. Various team members have also undertaken specialist study and research, and have deep sectoral experience in policy development and implementation in this complex and fast-evolving space. The Integrity & Authenticity Issue Policy team leads in the formulation of our policies to combat Election Misinformation.
- Violence & Aggression: Within the Issue Policy team that covers Violence & Aggression, we have sub-teams focused on each of:
 - Exploitation and Abuse: Within this team we have subject matter experts on Exploitation and Abuse who have undertaken specialist study and research, and who have deep sectoral experience in policy development in this field. Together with the Harassment and Hateful Behaviour Issue Policy team, the Exploitation and Abuse Issue Policy team leads in the formulation of our policies that combat harmful content pertaining to GBV Content and plays a crucial role in framing our approach to combating these issues.
 - Violent Behaviours & Dangerous Actors: The Violent Behaviours & Dangerous Actors team has subject matter experts who have undertaken specialist study and research, and who have deep sectoral experience in policy development in this field. The Violent Behaviours & Dangerous Actors team leads in the formulation of our policies that combat harmful content pertaining to Terrorist Content and plays a crucial role in framing our approach to combating these issues.
 - Harassment & Hateful Behaviour Issue Policy: Similarly, the Harassment & Hateful Behaviour Issue Policy team aims to ensure our policies minimise verbal animus and anger directed toward others, in particular marginalised communities. The Harassment & Hateful Behaviour team works closely with the Violent Behaviours & Dangerous Actors team, who together play a crucial role in framing our approach to combating Hate Speech.
- Regional/Local Policy teams: Within our wider Trust & Safety Product Policy team, we also have Regional Policy teams in each region who represent a specific country or region. The Regional Policy teams provide invaluable insight, cultural context, local knowledge and understanding of the interplay between global and local policies, and drive local discussions and engagement on content policy issues. The Regional Policy teams work closely with the Issue Policy teams including, for example, to share local/regional information related to violent extremism-related incidents or threats, hate-speech related incidents or in their local markets, provide policy guidance to moderation teams and ensure that violating content after an incident does not spike in their relevant markets.
- Feature Policy teams: The Feature Policy teams are responsible for developing moderation policies that are specific to the individual products or features that form part of the Platform (such as Comments), as well overseeing account-level policy moderation issues. Feature Policy teams work closely with both the Issue Policy teams and Regional/Local Policy teams.

- Trust & Safety Outreach and Partnerships Management team ('OPM'): The OPM team is responsible for managing our relationships with existing partners and outreach initiatives and for developing new outreach partnerships. This work is extremely important to our 'Think Global - Lean Local' strategy, as it enables our Trust & Safety teams to access external expertise and build partnerships at a regional and global level, which facilitates those teams in developing policies and practices that are objective, principled and operable. OPM leads several key partnerships and strategic collaboration with external experts, specialists and academics in the field of countering violent extremism, hate speech, minor safety, etc.

Trust & Safety Product:

- Trust & Safety Product: TikTok's Trust & Safety Product team has deep expertise and practical operational experience in designing and implementing safety programmes. The team has deep sectoral experience in policy implementation, threat identification, user safety and crisis handling, and works closely with our Product Policy teams. This level of experience and expertise, and close collaboration with the Product Policy teams is crucial in the formulation of effective strategies to detect and combat systemic risks. This team leads on the detection of and enforcement against harmful content and the implementation of TikTok's safety-by-design strategy. Within the T&S Product Team is a dedicated function for product implementation of Age Assurance measures. The Age Assurance team's efforts are focused on ensuring TikTok has greater confidence in user age for higher risk products and experiences, enabling us to build age-appropriate content and feature experiences.
- Covert Influence Operations team: The Covert Influence Operations Team is a key team working to disrupt external attempts to engage in covert influence operations by manipulating the TikTok Platform and/or harmfully misleading our community.
- LIVE Safety: TikTok's LIVE Safety team also comprises product managers, algorithm engineers, engineers and data scientists working in close collaboration, along with the LIVE Policy team, to design safety strategies, as well as launch features and models. We also monitor for new livestream risk trends and/or potential viral harmful content via media monitoring and risk monitoring mechanisms.
- Regional Safety & Integrity team: Within our Trust & Safety team, we have Regional Safety & Integrity teams. Led from Dublin, the role of the EMEA Regional team is to monitor and manage risks at a regional level, including across the EU, and to design, coordinate and project manage cross-functional risk mitigation programmes, involving multiple other teams (e.g. such as programmes focused on proactively managing, scenario-planning and mitigating risks around large events or major sports tournaments, events/dates/anniversaries). The EMEA Regional Safety & Integrity team plays an important role in risk detection and identification and longer-term mitigation at a regional and local level through its proactive scenario-planning, programme management and wider cross functional collaboration.
- Trust & Safety Operations (content moderation teams): An important part of content moderation involves human review by our trained content moderation teams. No matter the time of day, if content is reported or flagged for review, our teams are on standby to help take appropriate action. To ensure 24/7 coverage, we operate a 'follow-the-sun' model. Through human review, we also improve our machine learning systems as moderators provide feedback to the technology, helping to capture emerging content trends and improve our future detection capabilities.

- Account moderation teams: While the technical measures described above in this section are a core aspect of our approach on the Platform, an important part of moderation involves human review by our trained content moderation teams. Suspected underage accounts, including those reported via the in-app reporting channel, are reviewed by the moderation teams handling account-level moderation.
- Product Team: Sitting within the Product organisation, the Privacy & Responsibility Team concentrates on: minor protection and age-appropriate access; data transparency, management and access; and user choices over personalisation, interaction, and discoverability.
- Global Business Solutions - CoreOps: The CoreOps team conducts further investigations into advertisements, as required, to investigate hidden risks. The CoreOps team may action the advertiser account depending on the severity of the violation found.
- Global Privacy Operations: This multilingual team, which is led from Dublin, provides support to our users for any privacy-related inquiries including requests to exercise the GDPR rights of access, deletion, restriction etc.
- Public Policy and Government Relations: TikTok's regional and local Public Policy and Government Relations teams play an important role in bringing local and regional context to our decision-making, including feeding input from regulators, government departments and other institutions and external stakeholders to Trust & Safety, Product and other parts of TikTok. Our work in this area is also supported by TikTok's Public Policy Subject Matter Experts team, which has deep expertise at the intersection of public policy, content regulation and harmful content issues, and who provide support on the public policy aspects of TikTok's work in this space.

ANNEX 4: STAKEHOLDER ENGAGEMENT

1. INTRODUCTION	94
2. STAKEHOLDER ENGAGEMENT: Purpose and Role	95
3. ILLEGAL CONTENT: Risks of Illegal Hate Speech Content	96
4. ILLEGAL CONTENT: Risks of Terrorist Content	98
5. ILLEGAL CONTENT: Risks of Intellectual Property Infringing Content	99
6. ILLEGAL CONTENT: Risks of Child Sexual Abuse Material and Child Sexual Exploitation and Abuse	100
7. ILLEGAL CONTENT: Risks of Gender Based Violence Content	101
8. YOUTH SAFETY: Risks Related to Age Appropriate Content and Online Engagement	103
9. YOUTH SAFETY: Risks related to Age Assurance	105
10. CIVIC INTEGRITY: Risks to Elections and Civil Integrity	107
11. CIVIC INTEGRITY: Risks to Public Health from Medical Misinformation	108
12. CIVIC INTEGRITY: Risks to Public Security from Harmful Misinformation Content	109
13. FUNDAMENTAL RIGHTS: Risks to Fundamental Rights	110
14. Generative AI	111

1. INTRODUCTION

TikTok recognises that the complex nature of potential risks arising from platforms hosting and sharing UGC requires constructive engagement with a wide range of outside experts and stakeholders. Seeking advice and input across a wide array of stakeholders has been a cornerstone of the development of TikTok's safety efforts prior to the DSA and continues to be a central pillar of its risk governance efforts. These engagements are integral to TikTok's process of developing and refining its safety policies and operations and continuously improving the Platform's capacity to reduce and mitigate the risk of harm. TikTok's stakeholder engagement strategy reflects a commitment to dynamically adapting its services based on continuous feedback and collaboration.

Over the past year, TikTok has been consistently proactive in engaging with stakeholders across various regions and contexts to ensure that it has a comprehensive understanding of the diverse challenges and opportunities that arise from its nature, scale and impact. In alignment with Articles 34, 35, and Recital 90 of the DSA, TikTok has - and will continue to - prioritise external stakeholder engagement in its risk assessment and mitigation strategies.

This Report details TikTok's strategy to engage with stakeholders to ensure that measures to mitigate systemic risks on the Platform are reasonable, proportionate, and effective, whilst also reinforcing users' right to freedom of expression. This strategy involves maintaining ongoing dialogues with users, policy makers, industry experts and civil society organisations. TikTok's own ESAC is also key as that body provides critical expert input into both identifying risks specific to TikTok and in advising upon specific mitigation implementation.

This stakeholder engagement report covers the following topics:

- The relevance of stakeholder engagements in effectively identifying and mitigating risks on TikTok;
- An analysis of the most relevant stakeholder engagements conducted in Year 2 of the DSA, organised by TikTok's understanding of the systemic risks outlined by the DSA; and
- Insights from consultations with ESAC members regarding TikTok's approach to DSA systemic risk assessments.

2. STAKEHOLDER ENGAGEMENT: PURPOSE AND ROLE

Developing Trust and Safety policies and processes to support the many communities that are celebrated on TikTok is not something it does alone, as TikTok purposely seeks to engage with as many different perspectives as possible at all steps of the development process.

This dialogue is important for a number of reasons. While TikTok has an experienced and knowledgeable Trust & Safety team, external perspectives are essential to help it more deeply understand and proactively address topics of concern. TikTok is committed to listening to and including independent voices and to ensure that it makes decisions rooted in fact and aligned with international best practices. These partnerships come from a range of entities, such as NGOs, industry experts, academia, researchers, creators and the voice of lived experience. TikTok manages these engagements in a variety of ways, including ESAC and a separate Youth Council (which launched on 25 March 2024), one off consultations, long term partnerships, roundtable discussions, focus groups and community engagements. Diversifying the ways and means in which TikTok

gathers these external inputs allows it to support its Trust and Safety teams in a range of different ways, appropriate to policy areas of focus,

Third party engagement is also important in building trust across stakeholders and importantly with TikTok's user community. With expert input, TikTok can transparently demonstrate its efforts to understand the topic in question and to develop a solution that is balanced and factual. This process provides reassurance in the content decisions that TikTok makes and provides additional safeguards that prevent harm before it happens, improving overall user experience. TikTok also provides post-implementation feedback to its external partners where possible so that those partners can understand and observe the impact of their engagement.

3. ILLEGAL CONTENT: RISKS OF ILLEGAL HATE SPEECH CONTENT

TikTok is committed to ensuring the safety and well-being of its users by undertaking stakeholder engagements to understand and mitigate Hate Speech.

a. Marginalising speech and behaviour

Between December 2023 and April 2024, TikTok conducted nine consultations on marginalising speech and behaviour with 35 experts from 14 countries. As part of this work, TikTok engaged with at least one member of every TikTok content advisory council across the globe including **ESAC**. TikTok also spoke with community organisations and content creators affected by Hate Speech on the Platform, as well as behavioural scientists, and freedom of expression experts. Based on the concerns raised by the external experts around disproportionate restriction of speech, TikTok re-defined and limited the scope of its Marginalising Speech and Behaviour policy. The policy drafters clarified what type of content the policy should address, and provided more clear and detailed instructions on how to categorise content under this policy. The risk of over enforcement may create risks of harm to freedom of expression, which this work sought to limit. The team defined and limited the use case scenarios for the policy to only periods with a high risk of harm and increased Hate Speech and is developing extra guardrails for the implementation of the policy to assure consistent and reliable decision-making. The policy launched in May 2024. TikTok subsequently updated its Community Guidelines, which now states that 'Some content that uses stereotypes, insinuation, or indirect statements that may implicitly demean protected groups' is ineligible for the FYF.

b. Hate Speech mitigation

TikTok conducted a dedicated consultation in June 2024 to seek advice and feedback on its risk assessment and management approach with a specific focus on TikTok's current approach to identifying and mitigating Hate Speech. TikTok consulted with [REDACTED] [REDACTED] validated TikTok's effective use of keyword filter tools that empower users to block hateful comments on content and livestreams proactively.

TikTok has continued to participate as a signatory to the EU Code of Conduct on Countering Illegal Hate Speech Online. In collaboration with the European Commission and other participating platforms, it has been actively involved in the revision and improvement of the Hate Speech Code

since 2020. TikTok has committed to sign up to the revised Hate Speech Code, which is intended to be adopted under Article 45 DSA later this year.

TikTok undergoes yearly 'Hate Speech test' evaluations by participating EU based NGOs, who conduct Hate Speech monitoring over a four-to-six week period to assess the Platform's response time and reliability in addressing reported content. While the Code is being finalised for DSA adoption and formal monitoring is paused, TikTok has participated in the SafeNet programme run by INACH, an independent platform monitoring initiative. TikTok regularly engages with INACH and the participating organisations on this initiative to receive their feedback, until formal monitoring commences again.

As part of the new revised Hate Speech Code, TikTok has committed to significant new obligations. These include a commitment to review 50% of valid illegal Hate Speech user notices within 24 hours, demonstrating TikTok's dedication to swiftly addressing illegal content on the Platform.

TikTok also held roundtables with Muslim, Jewish, and LGBTQ creators to understand their experiences and address toxicity on the Platform. TikTok's Law Enforcement Outreach team engaged with the European Counter Terrorism Centre and the EU Internet Referral Unit to address potential Terrorist Content, [REDACTED]

c. Antisemitism and Islamophobia emerging from the Israel-Hamas war

In order to validate its approach to enforcing Hate Speech policies within the context of the Israel-Hamas war, TikTok engaged with 14 experts from Jewish, Arab, and Muslim communities, as well as freedom of expression and human rights specialists. This included consulting [REDACTED]. TikTok used this expert feedback to inform its enforcement of its Hate Speech policies related to the conflict ensuring it is as balanced as possible, particularly with regards to religious expression.

d. Protections of Public Figures

TikTok engaged with over 40 experts, including members of ESAC, to review its definition of public figures and their level of protection under TikTok's Bullying & Harassment Policies. Like other platforms, TikTok strives to strike a balance between allowing freedom of expression - and legitimate criticism of public figures - without neglecting its responsibilities to protect these public figures from abuse. TikTok defines public figures as: 'Adults (18 and older) with a significant public role, such as a government official, politician, business leader, or celebrity.' TikTok does not consider people under the age of 18 as public figures. This approach was accepted and endorsed by the experts consulted. The majority of experts supported distinguishing between public and private figures, but the extent and method were debated. Proponents of a higher bar for criticising public figures argued that such individuals, by their role, invite scrutiny, and that limiting criticism could harm democratic debate. Supporters of greater protections were concerned about the chilling effect on women and vulnerable communities, the mental health of public figures, and the potential for toxic discourse on online platforms. Even among those who disagreed on the level of protection for public figures, there was consensus that the definition of a public figure could be refined, with a majority agreeing on increased protections in certain areas, such as body shaming. There was broad agreement that minors should not be considered public figures. This review provided valuable insights and highlighted diverse perspectives on the issue. This work has validated our broad approach to protections of public

figures under our Harassment and Bullying policies, and is continuing to inform the planned evolution of this policy with potential tweaks to the policy possible later in 2024.

e. Reporting Hate Speech

TikTok has over 30 EU-based organisations to monitor and report suspected Hate Speech as part of its Community Partner Channel. TikTok is aware that Hate Speech is a cross-platform issue, as actors and content can spread between platforms, leading to a cumulative incident which one platform would not be able to identify on its own.

f. Campaigns to raise awareness

TikTok signed a new partnership with the Six Nations rugby tournament to continue to run Swipe Out Hate campaigns including for the 2024 editions of the tournament. TikTok ran these campaigns to educate TikTok's users on its zero tolerance approach to hate and encourage its users to report hateful content during both the men's and women's Six Nations tournaments.

4. ILLEGAL CONTENT: RISKS OF TERRORIST CONTENT

The moderation of terrorist and violent extremist activity and content online is an especially nuanced and dynamic area which requires deep understanding and linguistic knowledge of ideologies, political speech, and violent extremist use of technologies. As such, TikTok engages with a variety of subject-matter experts to develop proactive and reactive policies to address the risk of Terrorist Content on its Platform.

a. High-risk events and crisis readiness in the EU

TikTok has actively engaged in several strategic initiatives in order to enhance its response to EU terrorist attacks and crises. In February 2024, TikTok participated in a Terrorist Content Online Table Top exercise with the EU Internet Referral Unit (EU IRU), Europol, member states' law enforcement representatives, and observers from the UK and New Zealand. This exercise tested readiness for large-scale terrorist attacks, focusing on content moderation and proactive escalations.

In September 2023, during Referral Action Day, TikTok supported Europol's European Counter Terrorism Centre and the EU IRU, alongside 11 EU Member States, in targeting and assessing 2145 pieces of suspected terrorist and violent extremist content.

TikTok also contributed to the SIRIUS EU Electronic Evidence Situation Report, collaborating with Europol and providing input on improving cooperation between EU Member States and social media platforms. Meetings with the SIRIUS unit, including a session at TikTok's Dublin office in April 2024, reinforced this collaboration. TikTok provided strategic guidance to law enforcement agencies across Europe ahead of the Olympic Games and Euro Football championships, which subsequently facilitated data requests and emergency response strategies.

TikTok also engaged with counterterrorism and violent extremism agencies from various EU Member States, including France, Germany, Spain, Belgium, Netherlands, Italy, and Ireland. These engagements aimed to educate law enforcement on how TikTok's processes comply with relevant

DSA provisions and to establish efficient communication channels. Turning to the risk of a crisis event happening in the context of an election of which there is a significant volume in 2024 - TikTok is a member of the Code of Practice on Disinformation (CoPD) Working Group on Crisis and also co-chairs the Elections subgroup to collaborate with the European Commission and signatories on trends and mitigation strategies. Also, TikTok is an active member and participant of the EU Internet Forum. Through this engagement, TikTok engages with industry and government for dialogue and the exchange of best practices.

b. Threats related to evolving Terrorist Content

TikTok has worked extensively with Tech Against Terrorism to continue understanding terrorist and violent extremist behaviour and tactics online, adversarial shifts and mitigation strategies. Since September 2023, TikTok has conducted 7 workshops with Tech Against Terrorism to address:

- Educational, Documentary, Scientific, and Artistic Policies to assess challenges around such as counterspeech content, human rights implications and content moderation evasion techniques;
- 'Borderline content', or content that is not evidently related to Terrorist Content, activity, or groups, or not obviously violative of Community Guidelines but may be shared by a terrorist entity;
- Violent speech against designated terrorist entities; and
- The strengthening of human rights safeguards when identifying and removing Terrorist Content online.

TikTok engaged with [REDACTED] from February 2024 to address terrorist and violent extremist threats in Germany and Austria. This initiative identified local signals of hateful ideologies and actors in Germany in order to enhance moderation accuracy by providing localised guidance on signals that can be detected in order to identify violations at scale. This collaboration also identified specific audio files and songs as potentially violative of TikTok's policies.

c. ESAC advice on emerging threats

ESAC, together with TikTok's Safety Advisory Council members across the Middle East and the Asia Pacific regions, have been continuously consulted in relation to ongoing crises such as the Israel-Hamas war so that TikTok can ensure that its crisis response is as well informed as possible. For example, in December 2023 TikTok convened the first issue-specific Safety Advisory Council meeting (with members from across all regional councils attending) on terrorist entities and activity online.

In June 2024, TikTok conducted a dedicated consultation with [REDACTED] [REDACTED] provided critical feedback on assessments of borderline content and conduct-related risks related to known behavioural patterns of terrorist use of UGC platforms.

5. ILLEGAL CONTENT: RISKS OF INTELLECTUAL PROPERTY INFRINGING CONTENT

TikTok collaborates closely with external experts and stakeholders to manage the risks associated with

IP-Infringing Content to ensure compliance with EU and national laws. By engaging with these specialists, including content creators and IP-rights holders themselves, TikTok strives to effectively combat the unauthorised use of protected material, such as music, art, audio-visual recordings, trademarks, and patents.

a. Industry collaborations and regulatory engagements for IP-Infringing Content

TikTok's Trust & Safety team and Government Relations Public Policy Safety teams actively engage with various third-party industry bodies and external experts to assess and mitigate the risks associated with IP-infringing content. These collaborations provide crucial information and insights that inform TikTok's policies and ensure that its Community Guidelines are both objective and balanced. Key engagements include those with the Audiovisual Anti-Piracy Alliance (AAPA), an industry group consisting of rights owners and broadcasters, which focuses on tackling audiovisual piracy in Europe through lobbying, supporting law enforcement, and building partnerships; [REDACTED]

TikTok cooperated with the Copyright Information and Anti-Piracy Centre (CIAPC), a Finnish non-profit association, which was appointed as a Trusted Flagger under DSA Article 22 by the Finnish Digital Services Coordinator, Traficom on 7 March 2024. Since February 2024, TikTok has processed CIAPC's IP reports with priority, although CIAPC has not submitted any reports under Article 16 DSA to date.

TikTok engages with regulatory authorities within the EU such as France's Arcom and Italy's AGCOM in relation to our IP protection measures (including compliance with the EU Copyright Directive).

6. ILLEGAL CONTENT: RISKS OF CHILD SEXUAL ABUSE MATERIAL AND CHILD SEXUAL EXPLOITATION AND ABUSE

TikTok actively engages with stakeholders focused on controlling CSAM dissemination and preventing CSEA related harms. TikTok maintains close relationships with a number of expert child safety groups to identify potential risks before they reach TikTok as they are detected across other platforms and services.

a. Key industry partnerships

TikTok is a member of the WeProtect Global Alliance in order to help build collaboration across platforms. TikTok's active participation in the Alliance fosters collaboration across sectors – including NGOs, safety firms, and governments – facilitating the sharing of knowledge and best practices, and strengthening the collective response to CSEA. TikTok is also a member of the Internet Watch Foundation (IWF) and the Safer Internet Centers. TikTok regularly receives and acts on keyword lists and hash lists from organisations including the National Center for Missing and Exploited Children (NCMEC) and IWF to ensure that internal detection tooling is running on the basis of most up-to-date emerging risk signals.

b. Information sharing

TikTok holds a board seat on the Tech Coalition, an industry group dedicated to fighting CSEA online. TikTok participates regularly in – and leads several working groups – within the Coalition. TikTok has

also led work for the Tech Coalition and other tech companies in relation to hackathons, conferences, and other events. The output of such activities serves to increase TikTok's information ability to recognise and remove CSAM more quickly, for example regarding grooming behaviours in regional contexts. With the Tech Coalition, TikTok has contributed to the creation of a risk identification rubric in order to identify potential areas of increased risk of CSAM or CSEA, for example whether a product enables real-time communication or whether it enables text or multimedia based content sharing and interactions.

c. Identifying emerging risks

TikTok participates in consultations and meetings [REDACTED]

[REDACTED] to better understand and identify sex offender behaviour and juvenile sex offender psychology to proactively address and prevent instances of CSEA on TikTok.

TikTok has also undertaken consultations with InHOPE on AIGC CSAM and grooming patterns associated with AIGC CSAM. This helped TikTok create appropriate language for its CSAM policy. TikTok also participated in INHOPE's conference, sending staff to present and learn best practices from those managing CSAM hotlines.

TikTok also engages with ConnectSafely for insights and/or recommendations on relevant youth safety topics to help TikTok better understand emerging risks, trends and/or relevant world events that might impact users on its Platform.

d. Third-party detection tools and keyword lists

TikTok's Child Safety Team meets monthly with NCMEC to coordinate on best methods for CSAM and child endangerment/abuse response. TikTok receives direct feedback on the efficacy of its reporting, as well as information when reports have led to specific children being identified and helped. TikTok made 288,125 reports to NCMEC in 2022 and in 2023, that number grew to 590,376. TikTok also participates in NCMEC's Take It Down initiative to share known hashed CSAM content to help prevent CSAM from being disseminated on the app. Some of the CSAM Detection Systems that TikTok deploys and implements are integrating PhotoDNA and the Google Content Safety Toolkit, along with ingesting hash lists from the IWF. These enable TikTok to remove previously-identified CSAM at point of upload, before it has been viewed.

7. ILLEGAL CONTENT: RISKS OF GENDER BASED VIOLENCE CONTENT

TikTok is committed to the prevention of Gender Based Violence ('GBV') Content occurring on its Platform. GBV Content disproportionately affects women and marginalised communities, and addressing such hateful content is crucial for promoting gender equality and ensuring the safety and well-being of all individuals. In the past year, TikTok has engaged with expert stakeholders from leading nonprofits that are working to prevent GBV and to support survivors. These engagements have supported the identification and categorisation of granular issues under the broader umbrella of GBV Content, supported the roll out of new tools (for example a hash matching system), new resources in the Safety Center, and assisted with raising awareness, especially for Younger Users. Finally, a cross-platform partnership has led to a larger case study on the risk of sextortion.

a. Sextortion

Sextortion is a type of blackmail in which the perpetrator has (or purports to have) sexually explicit content about a user and then uses threats (to distribute that content) and intimidation to extort money from that user. Sextortion is a known risk that TikTok prohibits, and one that requires continuous monitoring due to its evolving nature as bad actors seek new strategies. The number of global sextortion cases across the internet and several platforms is increasing rapidly. Reports to the NCMEC more than doubled in 2023 across public and electronic service providers, rising to 26,718 compared to 10,731 the year before. As a result, TikTok partnered with Toluna to carry out a user survey in multiple countries to better understand the scope of the problem on its Platform. A case study on this topic is included in the GBV section of the Report above.

b. Non-Consensual intimate imagery and image based sexual abuse

TikTok continues its partnership with StopNCII for proactive detection when an image is uploaded, and so that TikTok can receive and enforce on known examples of non-consensual intimate imagery. TikTok then employs a hash matching system to proactively identify and remove potential non-consensual intimate images on the Platform.

TikTok attended two conferences, the Virtual Summit on Deepfake Abuse and Silicon Saviors or Digital Threats? Exploring Tech’s Impact on Domestic and Sexual Violence, which informed its identification of the risk of image based sexual abuse. Image Based Sexual Abuse ('IBSA') is the creation, manufacture, or distribution of nude, partially nude, or sexually explicit content without the consent of the person in the content, for the purpose of sexualizing their body, or portraying them in a sexual manner. TikTok decided to approach the harm of AIGC images the same regardless of whether it includes real or manipulated media, as the impact on the survivor is the same. TikTok updated its policies to make this clear.

c. Gendered Harassment and Youth

TikTok co-hosted a roundtable discussion on gendered harassment and the impact of misogynistic content on youth together with Dr. Mary Anne Franks from the George Washington University Law School and Cyber Civil Rights Initiative. Representatives from a range of nonprofits attended, including experts from [REDACTED]

[REDACTED]. These representatives work on specific issues within GBV and each brought varied and specialised perspectives to the discussion. Their insights highlighted nuances and granular aspects of this complex issue, which inform TikTok’s broader risk identification strategy. Educating youth about respect, consent, and healthy relationships can have a lasting impact, reducing the prevalence of GBV long term. TikTok further discussed youth focused GBV prevention with Futures without Violence. These discussions highlighted nuances relevant to TikTok’s policies at the intersection of youth safety and GBV content.

d. Language promoting GBV

TikTok held policy consultations on the risks associated with the promotion of GBV content with each of the following experts: [REDACTED]

[REDACTED] These meetings identified and categorised the harm levels of different forms of language that promote GBV Content and the psychological impact on users viewing such content,

which informed TikTok's ongoing policy-drafting process.

e. Safety Center and Survivor Resources

TikTok held two meetings with [REDACTED] to consult with them on the existing resources available to survivors in the Safety Center. A consultation with [REDACTED] revealed that TikTok's Safety Centre pages on sexual abuse could be improved by using more survivor-centred language in order to avoid exacerbating survivors' trauma and hindering their recovery. TikTok partners with local organisations, helplines and social services in individual EU Member States to tailor its in-app support resources for survivors. These include rape crises centers across multiple EU member states.

f. LGBTQIA+ Community

TikTok works with [REDACTED] to gain insight into the lived experiences of the LGBT+ community so that it can mitigate the risk of GBV in a socially and culturally sensitive manner. LGBT+ individuals often experience GBV differently, and transgender and non-binary people are at a higher risk of violence. These partners can also report abusive content to TikTok and help to identify new trends. TikTok's Community Partner Network entities that assist with this are [REDACTED]

8. YOUTH SAFETY: RISKS RELATED TO AGE APPROPRIATE CONTENT AND ONLINE ENGAGEMENT

TikTok is committed to ensuring the safety and well-being of its Younger Users by actively engaging with expert stakeholders to understand and mitigate risks related to Age Appropriate Content and risks related to Online Engagement. Collaborating with experts has enabled TikTok to identify and mitigate emerging risks, such as detecting mature content and preventing young audiences from accessing inappropriate material. Stakeholder input has been vital in addressing exposure to harmful content, discoverability by bad actors, and mental health impacts from features like comments and subscriptions.

a. Age Appropriate Content

TikTok closely observes the work of global content ratings bodies to inform its policies and processes to prevent Younger Users' exposure to content that can be harmful or inappropriate. These bodies are public organisations that classify films, videos, and other online content and motion pictures into age-based ratings. There are several global ratings bodies whose guidance TikTok actively monitors, including but not limited to the BBFC, the Netherlands Institute for the Classification of Audiovisual Media, the Motion Picture Association, the TV Parental Guidelines in the US and the Korea Media Rating Board. By following ratings bodies globally, TikTok identifies when changes to national guidelines are made so that any applicable changes may be reflected in TikTok's Content Classification system.

b. Online Engagement

TikTok is part of multistakeholder forums which bring together representatives from countries,

intergovernmental organisations, civil society organisations and the technology industry to identify, implement and share innovative solutions to better protect children online. The consultations and emerging topics and trends from this inform TikTok's identification of risks to minor safety.

c. Campaigns for Younger Users

TikTok recognises the importance of reaching Younger Users in ways that resonate with their unique cultural and social circumstances. To achieve this, TikTok partners with local organisations in individual EU countries to tailor its approach.

In Italy, TikTok works with the Safer Internet Centre, which promotes a safer and better use of the internet among children, parents, and teachers. TikTok has worked with e-Enfance in France to combat cyberbullying by integrating the national hotline for digital violence victims directly onto its interface. Centro Internet Segura in Portugal operates the Linha Internet Segura, a telephone and online support service for young people online. It is part of TikTok's Community Partner Channel where onboarded organisations can submit reports for prioritised action.

d. Resources for parents on youth safety

TikTok has partnered with the UK-based NGO Internet Matters to develop policies and in-app features across the EU-UK market that seek to improve digital well-being. Through this partnership, TikTok creates resources to educate families about the importance of digital well-being and addresses common challenges such as peer pressure online. In partnership with the Family Online Safety Institute, TikTok created The Good Digital Parenting Initiative, offering resources to help parents better understand the available tools and controls that can shape the desired digital environment for their family. The continued collaboration with the Digital Wellness Lab at Boston Children's Hospital further supports these safety improvements. The Lab has consulted on various platform designs and digital tools. Most importantly, the Lab provided insights that contributed to TikTok's design of its screen time management tools, which imposes a default 60-minute limit for Younger Users and several different options for Family Pairing accounts. TikTok is also an early signatory to the Lab's Inspired Internet Pledge, through which it has committed to focusing on well-being for Young Users, and sharing best practices on creating a healthy online environment.

TikTok also conducted consultations and engagements with 13 organisations worldwide to guide and share developments around family pairing product updates.

e. AIGC tools

TikTok partners with the [REDACTED], which is an initiative that brings together youth globally to share their insights and advice to make the internet safer. The Council is a unique space that provides the chance for youth voices to be heard on issues that directly affect them and allows TikTok to engage directly with affected young user groups on various experiences with the app to inform product and policy development.

TikTok's Youth Safety and Well-being Policy team worked in collaboration with the Outreach and Partnerships team to conduct two consultations with youths via [REDACTED] in order to identify potential risks related to AIGC tools. This engagement has informed TikTok's understanding of the risks to self-esteem, body image, and mental health.

TikTok also consulted with various mental health and safety advisors on the use of beauty filters on the

Platform. [REDACTED]

f. TikTok’s Youth Council and teen well-being

On March 25, 2024, TikTok launched a global Youth Council to reinforce its commitment to enhancing safety measures for teens on the Platform. This initiative responds to recent global research indicating that teens and their parents want more collaborative opportunities with online platforms. Created in partnership with Praesidio Safeguarding, the Youth Council consists of 15 teens from diverse backgrounds and countries, including the United States, United Kingdom, Brazil, Indonesia, Ireland, Kenya, Mexico, and Morocco. The group, which first convened in December 2023 and met again in February 2024 with TikTok Chief Executive Officer Shou Chew, has prioritised teen well-being and inclusion for the year. They are contributing to the redesign of TikTok’s Youth Portal and have sought clarity on what happens after an account or video is reported , aiming to ensure a safer and more inclusive environment on the app. The Youth Council’s insights help TikTok refine its risk detection measures and enhance the overall user experience for young people.

TikTok is updating its Youth Guide (now renamed Youth Portal) to include new safety features, based on feedback from the Youth Council on reporting and blocking, and is introducing new GIFs to help Younger Users locate these features. Members of the Youth Council provided input and urged TikTok to share more information about reporting and blocking to better understand what happens after a report is made.

g. Bullying

TikTok engaged with 17 experts through Safety Advisory Council meetings and 1:1 engagements on their opinions on TikTok’s bullying policies and how TikTok can better protect Younger Users from bullying harms. The consultations highlighted a number of key areas where minors are more or uniquely exposed to bullying compared to adults, particularly in the areas of revictimisation through resharing of bullying content and non-physical parental discipline.

h. Multi-stakeholder working group on Younger Users’ use of social media

In 2023, TikTok joined the United Nations Violence Against Children’s Office Protection through Online Participation working group (PoP). This global initiative aims to understand how children and youth use digital platforms to be safe. TikTok is an active member of PoP’s working group (through participation in virtual meetings), and has also supported and attended in-person events that have taken place in Brussels, Paris, and Madrid.

9. RISKS RELATED TO AGE ASSURANCE

While Younger Users have the right to freedom of expression, access to information, and other fundamental rights linked to the use of the Platform, TikTok considers it proportionate to impose some

restrictions on access for Younger Users. Expert consultations have supported TikTok's approach to identifying and mitigating risks related to Age Assurance.

a. Multi-stakeholder initiatives for Age Assurance best practices

TikTok has worked with Microsoft to establish a global multi-stakeholder initiative on age assurance. The aims of this initiative are to facilitate a holistic multi-stakeholder dialogue that explores the complexities and nuances around assessing age in a digital environment and to increase alignment on key principles relating to age assurance. To do this members of the group are drawn from other industry partners, leading safety and privacy organisations such as the children's organisation WeProtect Global Alliance and the privacy think tank, the Centre for Information Policy Leadership. The first milestone of this work was a convening (funded by TikTok) of relevant stakeholders at the end of March 2024 in London. The dialogue underscored the importance of balancing rights and risks, involving experts from various sectors to identify challenges and opportunities. Critical discussions centred on developing evidence-based, child-informed, and risk-based approaches to age assurance, ensuring robust protection mechanisms while enabling safe online participation for minors. The group is also establishing four Working Groups on key themes relating to age assurance, including law and regulation, horizon scanning, global and regional perspectives and risk assessments.

TikTok has engaged with national authorities and regulators, including the AEPD in Spain and the Ministry of Education in France, who are proposing various age assurance solutions. TikTok also consulted with Italy's Garante and AGCOM, and France's CNIL, and continues discussions with [REDACTED] on their emerging solutions to age assurance technology. Additionally, TikTok is part of the Digital Trust & Safety Partnership, which has produced best practice documents for the sector.

b. TikTok's Safety Advisory Council and content classification

In 2023, during an in-person Safety Advisory Council summit with global Youth Safety and Well-being experts, the Child Safety team conducted a session titled 'Autonomy for Younger Users & Content Restrictions.' This session, focused on content classification, provided crucial insights that supported TikTok's existing content classification efforts by incorporating global, evidence-based perspectives. Key insights gathered included the importance of developing localised interventions and policies that account for different acceptability standards for various age groups (13-15 and 16-17) in different countries. Additionally, the session emphasised the need to support Younger Users in making informed decisions and learning while acknowledging that such users develop at different paces and stages.

10. CIVIC INTEGRITY: RISKS TO ELECTIONS AND CIVIC INTEGRITY

TikTok recognises that 2024 has been and will continue to be a critical year for EU and global democratic processes. Citizens may use TikTok to explore and engage with political ideas and interact with their representatives. However, this activity gives rise to risks that election misinformation and/or

misconduct is shared and amplified on the Platform, which could lead to real world effects harming democratic processes. TikTok conducts extensive and continuous engagements with external stakeholders, including fact-checkers, misinformation experts, law enforcement, and potentially impacted groups in order to identify and define Election Misinformation risks and to design and implement mitigations.

a. Identifying emerging risks

TikTok conducts regular consultations with its ESAC, comprising leading external experts on technology and content-related issues. Through its Fact-Checking Programme, TikTok implements various measures to identify emerging misinformation-related risks, including ad hoc policy consultations and trend analysis with fact-checking partners. Eleven fact-checking partners across the EU help TikTok to understand and identify localised risks to election integrity. Additionally, TikTok's Community Partner Channel (CPC) also supports risk identification, by onboarding NGOs who act as trusted flaggers to report suspected harmful content, including election misinformation. TikTok has created a rapid response system to streamline the exchange of information between civil society organisations, fact-checkers, and online platforms, as part of TikTok's commitments, as signatories of the CoPD. The rapid response system is a time-bound dedicated framework for cooperation among signatories during the 2024 European Parliament elections which allows non-platform signatories to flag time sensitive content, accounts or trends that may present threats to the integrity of the electoral process. TikTok has been a proactive participant in the Permanent Taskforce and subgroups including the ones related to Crisis Response, Elections and AIGC set up under the CoPD.

b. Addressing election misconduct, hacked materials, and misrepresented sources

TikTok updated its Community Guidelines in May 2024 following individual consultations with over 40 global experts in democracy, tech policy and freedom of speech. The update added new policies and enforcement regarding unverified elections claims, election misconduct, distribution of hacked materials, and misrepresented civic sources.

c. Fact-checking at scale

TikTok's Fact-Checking Programme is engaged to determine whether potential misinformation content that is not already contained in its database of fact-checked claims is in fact violative. This Programme involves 11 IFCN-verified fact-checkers in the EU, who assess unverified content based on their independent, expert, localised and linguistic knowledge. These fact-checkers include Agence France-Presse (AFP), dpa Deutsche Presse-Agentur (DPA), Demagog.pl, Facta, Faktograf, Logically Facts, Newtral, Maldita, Poligrafo, The Journal, Nieuwscheckers.

During 2023 and 2024, TikTok sought feedback and advice from its Fact-Checking Program members before updating its internal ratings labelling system in July 2024 for content that has undergone a fact checking process. This expanded the rating options available to fact-checkers so that they can label content as either 'false', 'misleading', 'unsubstantiated', 'inconclusive', 'true', or 'out of scope', which informs proportionate enforcement.

For detailed information on TikTok's Fact-checking Programme, and the media literacy campaigns launched in collaboration with these partners, please refer to Annex 5.

11. CIVIC INTEGRITY: RISKS TO PUBLIC HEALTH FROM MEDICAL MISINFORMATION

TikTok held two consultations with academics and one with the World Health Organization (WHO) in December 2023 due to the potential global impact of Medical Misinformation. These consultations were valuable in helping frame TikTok's policy and enforcement updates for both critical and moderate harm health risks that launched in June 2024.

12. CIVIC INTEGRITY: RISKS TO PUBLIC SECURITY FROM HARMFUL MISINFORMATION CONTENT

TikTok recognises that misinformation can be highly localised in individual EU Member States with unique cultural, social, and political nuances. TikTok therefore leverages the insights and expertise of external stakeholders to develop effective and tailored strategies to combat misinformation that poses Public Security Risks on the Platform, ensuring that interventions are culturally sensitive and contextually relevant.

For detailed information on TikTok's Fact-checking Programme, and the media literacy campaigns launched in collaboration with these partners, please refer to Annex 5.

a. Rapidly evolving events

During events that may threaten public security, a vast amount of live or near live information can be shared rapidly. It is crucial to emphasise to users the importance of verifying information with official sources and the availability of tools to report violative content. In response to the unfolding Israel-Hamas war, TikTok collaborated with various fact-checking organisations including Agence France-Presse (AFP), Reuters and Fatabayano as authoritative sources for conflict related information. The in-app search intervention that resulted from these partnerships was used by millions of users, and TikTok scaled the search intervention globally in March 2024 in response to the terrorist attack of the Crocus Hall in Russia.

b. Climate change misinformation

TikTok launched a \$1m initiative to tackle climate misinformation across Spain, Brazil and the UAE during COP28. This Verified for Climate programme is a joint programme between the United Nations and Purpose, and brings together a team of verified champions, including scientists and trusted experts from Brazil, the United Arab Emirates, and Spain who supported select TikTok creators in developing educational content to tackle climate misinformation and disinformation. One such creator, Verified, developed a series of initiatives to assist with the dissemination of accurate climate change educational content as a result. TikTok then co-created campaigns to uplift the voices of creators within Verified's network in order to increase public engagement at key moments, for example during UN Climate Week and COP28.

Verified for Climate partnered with 35 experts to produce 225 educational videos on TikTok, achieving 632+ million impressions and nearly 15.8+ million engagements globally. This partnership resulted in users gaining increased awareness and skills in identifying and reporting climate misinformation.

c. Conferences to improve understanding of risk mitigation strategies

TikTok attended multiple conferences and events to improve its risk mitigation strategies relating to misinformation, including the Global Fact-Checking Conference hosted by the International Fact-Checking Network in (Oslo, June 2022); The Trust & Safety Research Conference (Stanford University, September 2022); and the Association of Internet Researchers Conference (Technological University Dublin, November 2022).

13. FUNDAMENTAL RIGHTS: RISKS TO FUNDAMENTAL RIGHTS

TikTok is dedicated to engaging with stakeholders to uphold Fundamental Rights and address risks to such rights on and or due to its Platform. TikTok is committed to the UN Guiding Principles on Business and Human Rights and its call to conduct human rights due diligence across the lifecycle of its business. In this regard, TikTok has continued to engage in external partnerships with the objective of improving its ability to assess and mitigate risks to fundamental rights. TikTok has also maintained memberships in industry wide forums such as the Business for Social Responsibility's Human Rights Working Group.

a. Balancing fundamental rights in risk assessment and mitigation

TikTok regularly consults with Safety Advisory Council members in Europe, Asia, and Latin America to gain varied perspectives from civic society on best practices for the protection of fundamental rights. A key aspect of protecting these rights is navigating the delicate balance between competing rights and interests. TikTok conducted a dedicated consultation in June 2024 to seek advice and feedback on its risk assessment and management approach in light of the DSA's emphasis on fundamental rights and in particular, freedom of expression. TikTok consulted with [REDACTED] [REDACTED] holds expertise in balancing content moderation with the protection of fundamental rights and the ongoing interest in TikTok's moderation and privacy measures. [REDACTED] shared his view that the process of addressing and navigating the balance of fundamental rights is more important than the final position taken: for example, it can be acceptable to create a risk to freedom of speech due to mitigations against Hate Speech, so long as TikTok's process has meaningfully compared the two risks and assessed that the risk from Hate Speech is more severe than the risk to freedom of expression. This perspective underscores the importance of a deliberative process in policy-making regarding fundamental rights protection, which TikTok is committed to. [REDACTED] provided critical insights, highlighting how the DSA's definition of systemic risks may be interpreted at various levels - social, industry-wide, or specific to individual platforms. [REDACTED] recommended transparency about the complexities of risk assessments, advising against an over-reliance and advocating for qualitative evaluations of risks to the Platform to ensure that a nuanced understanding is gained. [REDACTED] also stressed the importance of engaging stakeholders from different contexts, languages, and cultures to tailor how various manifestations of fundamental rights risks are mitigated on the Platform. [REDACTED] insights validated TikTok's

methodological approach to systemic risk assessments under the DSA, highlighting TikTok's proactive approach to engaging with external stakeholders to identify and mitigate systemic risks.

b. Collaborating to anticipate future risks to fundamental rights

By participating in conferences and collaborating with peers and civil society stakeholders regarding fundamental rights, TikTok can share experiences and learn with peers to anticipate future risks.

In June 2023, TikTok experts attended the RightsCon global conference which convened business leaders, policy makers, general counsels, government representatives, technologists, academics, journalists, and human rights advocates from around the world to tackle pressing issues at the intersection of human rights and technology. TikTok also participated in RightsCon's 'Young leaders summit', in which students and young professionals engage in workshops to empower them to join policy and advocacy discussions around human rights in the digital age. The summit engaged these young leaders in advocacy, policymaking, and movement building for emerging issues, including data protection and privacy, network discrimination and connectivity, digital security, diversity and digital inclusion, and artificial intelligence and algorithmic accountability.

In June 2024, TikTok actively participated in the EU Rights & Risks Forum, an event hosted by Digital Trust and Safety Partnership and Global Network Initiative. The event brought together representatives from platforms, civil society, and academic experts from Europe and beyond. The forum focused on systemic risk assessments under the DSA and their implications for fundamental rights. Over two days of panels and workshops, TikTok, along with other participants, engaged in discussions aimed at informing and refining approaches to risk assessment and stakeholder engagement. The insights gained from this forum have contributed to TikTok's ongoing efforts in these areas.

TikTok's engagements with the WeProtect Global Alliance, Trust & Safety Professional Association, Tech Against Terrorism, Business for Social Responsibility, the Tech Coalition, and launch partners for the Partnership on AI's Framework Responsible Practices for Synthetic Media, mentioned in more detail elsewhere in this Annex, also all inform its approach to the mitigation of risks to fundamental rights.

14. AIGC

TikTok has collaborated with expert external stakeholders to develop TikTok's comprehensive Community Guidelines and policies covering the risks of AIGC.

c. Fact-checking at scale

TikTok's fact-checking partners are authorised to proactively identify to TikTok where they detected AIGC that may constitute Harmful Misinformation. TikTok is then able to take action to remove or label it (depending on the status/outcome of any necessary verification). Those partners will also provide TikTok with intelligence on prominent misinformation that is circulating on other social media platforms or websites that may benefit from verification.

d. Election-related mis- and disinformation

TikTok has joined industry partners as a party to the Tech Accord to Combat Deceptive Use of AI in 2024 Elections (the 'Accord'). Via the Accord, member technology companies are working together to safeguard communities against misleading and deceptive AI in this election year – including, with respect to AIGC, to safeguard against an increased risk of disinformation campaigns, efforts to discredit politicians and the proliferation of fake political endorsements by celebrities. Through this work, TikTok and industry partners aim to collaborate on tools to detect and address fake political online distribution of such AIGC, with attention to the importance of tracking the origin of deceptive election-related content and the need to raise public awareness about the problem.

e. Transparency and responsible AI

TikTok was the first social media platform signatory to the PAI's Responsible Practices for Synthetic Media (the 'Practices'), a first-of-its-kind code of industry best practices for AI transparency and responsible innovation, balancing creative expression with the risks of emerging AI technology. In accordance with its commitments as a launch partner for the practices, TikTok worked on a case study outlining how the Practices informed TikTok's policy making on synthetic media.

f. Labelling AIGC

TikTok is working to develop content provenance practices through implementation of the Coalition for Content Provenance and Authenticity (C2PA) standard in the second or third quarter of 2024. The C2PA standard is an open technical standard and content provenance solution that can provide information in the metadata for a piece of content about its origin and whether AIGC models were used to create or edit it. In 2024, TikTok collaborated with C2PA to integrate its standard into the Platform, enabling automatic AIGC labelling for off-platform content containing C2PA metadata. In 2023, TikTok held two consultations with media literacy experts and a behavioural scientist to develop language for an AIGC label, which was launched in September 2023. Additionally, TikTok conducted three external consultations on the use of satire in AIGC which informed updates to TikTok's synthetic and manipulated media policies, which were implemented in May 2024.

g. Expert consultations

To address and better understand the risks, trends, and opportunities that AIGC poses to youth, TikTok conducted consultations with experts like [REDACTED]

[REDACTED]

ANNEX 5: OVERVIEW OF FACT-CHECKING PROGRAMME

Overview: Below is a high-level overview of TikTok’s Fact-checking Programme (the “**Fact-checking Programme**”):

- **Context:** To deliver on its commitment to tackle Harmful Misinformation, including Election Misinformation, Medical Misinformation, and other Harmful Misinformation, TikTok’s Integrity & Authenticity team has developed a Fact-checking Programme. This programme is led by members of the Trust & Safety team who have deep expertise and experience in Integrity & Authenticity issues, and entails a highly cross-functional approach with input from subject matter experts across various Trust & Safety teams (including our Integrity & Authenticity Policy, Product and Outreach Partnership Management (“**OPM**”) teams), and multiple other teams;
- **Core objective and mission:** The core objective and mission of the Fact-checking Programme is to leverage the expertise of external fact-checking organisations to verify potential misinformation claims, and to enable our Trust & Safety teams to determine the appropriate enforcement action. This programme is grounded in our belief that technology companies should not alone make **decisions** about the veracity of borderline claims. For this reason, we consider that partnering with third-party fact-checking organisations is a critical component of our strategy to effectively tackle Harmful Misinformation;
- **Global framework for regional/local application:** The Fact-checking Programme provides a global overarching framework for fact-checking that can be applied in a consistent manner across a range of markets and languages. Since its launch, it has been expanded substantially to the point where we now work closely with 19 fact-checking partners globally, covering more than 100 countries and 50+ languages, including English, French, German, Spanish, Italian, Dutch, Austrian, Polish and Portuguese, along with a range of other languages. As part of the Fact-checking Programme, we can also secure temporary fact-checking resources when needed (and have done so in the past, for example, around elections).

1. OUR FACT CHECKING PARTNERS

a. Our fact checking partners

The fact-checking partners that we work with play a key role in contributing to the effectiveness and success of the Fact-checking Programme. TikTok therefore carefully considers the selection of which fact-checking partners we work with, based on those who have proven themselves to be reliable and

capable. Within the EU, we partner with 10 fact-checking organisations who provide fact-checking coverage in 22 official EU languages: [Agence France-Presse](#) (AFP), [dpa Deutsche Presse-Agentur](#), [Demagog](#), [Facta](#), [Faktograf](#), [Lead Stories](#), [Logically Facts](#), [Newtral](#), [Poligrafo](#) and [Science Feedback](#). As part of the Fact-checking Programme, we can also secure temporary fact-checking resources when needed (and have done so in the past, for example, during EU Parliamentary elections expanding coverage for Maltese language). Our partners have specialists, including trained journalists, who review and verify reported content. Our moderators then use that independent feedback to take action and where appropriate, remove or make ineligible for recommendation false or misleading content or label unverified content. We have 3 primary criteria when selecting partners. More info [here](#):

- **IFCN accredited:** The [International Fact-Checking Network](#) (“IFCN”) is a global leader in fact-checking excellence that supports more than 100 fact-checking signatories around the world focused on best practices. Through the Fact-checking Programme, we only work with partners who are signatories to IFCN [Code of Principles](#). Signatory status is based on an application process to the IFCN, and applicants are assessed by independent assessors for compliance with 31 criteria. Their assessment is reviewed by the IFCN Advisory Board to ensure fairness and consistency;
- **Relevant experience:** We generally choose partners who have experience working with technology companies, and especially those with deep expertise in fact-checking video content. This means that our fact-checking partners are a good fit for the types and formats of content that may require fact-checking as part of our Fact-checking Programme; and
- **Relevant language skills and support:** We prioritise partners with staff who have deep expertise in the markets they are covering, along with relevant language and topic knowledge.

b. The fact-checking process

As noted above, the core objective of the Fact-checking Programme is to leverage the expertise of external fact-checking organisations to determine enforcement actions on the most harmful and difficult to verify claims. The fact-checking process operates as follows:

- **Detection:** Where content is proactively detected or reported as being in violation of our Harmful Misinformation policies, it is routed and centralised to our dedicated specialised moderation team for integrity-related issues (“**Integrity Operations**”) for review;
- **Integrity Operations Review:** Within our Trust & Safety Operations function, we have established a dedicated Integrity Operations team who moderate suspected misinformation content, which is made up of experienced moderators who then undergo specialised training on Integrity & Authenticity policy issues. Our Fact-Checking Programme is integrated into our specialised misinformation moderation workflow to inform our Integrity Operations team's decisions on whether content violates our misinformation policies;
- **Escalation:** If the Integrity Operations moderators are unsure as to the veracity of the video and believe it may contain misinformation, they have the ability to escalate to a fact-checking partner with expertise in the market who will assess the video, conduct reporting, and verify whether the information is false. Pending the fact-checker review, such videos will not be eligible for recommendation in the Platform's For You Feed. Our fact-checking partners access content which has been flagged for review through a dashboard made available for their exclusive use. The dashboard shows our fact-checkers certain quantitative information about the services they provide, including the number of videos queued for assessment at any one time, as well as the time the review has taken. Fact-checkers can also use the dashboard to see the rating they applied to videos they have previously assessed;

- **Enforcement:** Our fact-checking partners do not take direct actions on the content that they review. After a fact-checker assesses a video, it returns the matter to the Integrity Operations team who then uses the fact-checker insights to determine whether the video violates our Harmful Misinformation policies. In terms of sanctions that may be imposed, violating videos can be removed globally, Not Recommended (which means they will not be eligible for recommendation in the Platform's For You Feed and we take measures to ensure that this content is harder to find in our search functionalities on the Platform), or labelled as unverified content. If content is confirmed to be accurate, and the content is in line with TikTok's Community Guidelines, it will remain on the platform;
- **Repository of fact-checked misinformation:** To ensure it is both efficient and scalable, as part of the Fact-checking Programme, our Integrity & Authenticity team maintains a repository of fact-checked information - such as claims, statements and conspiracy theories, and that have been fact-checked by our external fact-checking partners. This is a key resource for our Integrity Operations team to help them accurately, quickly and efficiently assess the veracity of suspected misinformation as part of their review process. This database is updated on an ongoing basis to include information that has been assessed and determined by our fact-checking partners to be verifiably false or misleading.

2. OTHER SUPPORT AND SERVICES

a. Other support and services

In addition to fact-checking services, we often work with our fact-checking partners where they are able to provide ancillary support to our Fact-checking Programme and misinformation mitigation strategy more widely, including:

- **Policy consulting/trends analysis:** This may involve our fact-checking partners providing us with reports identifying general misinformation trends observed on our Platform and across the industry generally, including new/changing industry or market trends, events or topics that generated particular misinformation or disinformation;
- **Media/digital literacy campaigns:** We can collaborate with fact-checking partners to produce media/digital literacy videos/campaigns to help inform and educate users about identifying and reporting misinformation. These videos can also educate users about specific topics such as civic engagement. For example, in partnership with our fact-checking partners, we ran integrity campaigns in advance of several European major elections in 2023 and 2024:
 - **Media Literacy: stand-alone campaigns in the EU:** Since August 2023, we rolled out an additional 13 new media literacy campaigns in Europe in collaboration with local media literacy bodies and our trusted fact-checking partners on topics identified as priority areas for increased digital literacy for our users in Europe; for example, the Russia-Ukraine war or general critical thinking skills including identifying *misinformation* and *manipulated media and deep fake*. Of these media literacy campaigns:

- Six (6) campaigns launched in [Austria](#), Bulgaria, Czech Republic, Croatia, [Germany](#) and Slovenia were specific to the war in Ukraine, including identifying misinformation and manipulated media reports;
- Seven (7) campaigns launched in [Finland](#), [Ireland](#), [Italy](#), [Spain](#), [Sweden](#), Denmark and [Netherlands](#) were focused on general media literacy and critical thinking skills to fight misinformation and manipulated media online, H5 Example below.

b. Other support and services

In addition to stand-alone media literacy campaigns, starting from 2023 and throughout 2024, we introduced media literacy resources in our Elections Centres including tips about how to develop critical thinking skills, identifying *misinformation*, *manipulated media* and *deep fake*. The media literacy section in our elections centres includes videos from our trusted fact-checking partners (see table below):

- **2023 Slovak Parliamentary Election:** From 4 September 2023, we launched an in-app [Election Centre](#) in collaboration with electoral commissions to provide users up-to-date information about the 2023 Slovak parliamentary elections. The centre contained a section about fighting-fake news, including videos;
- **2023 Polish Parliamentary Election:** From 18 September 2023, we launched an in-app [Election Centre](#) in collaboration with electoral commissions to provide users up-to-date information about the 2023 Polish parliamentary elections. The centre contained a section about fighting-fake news, including videos created in partnership with media local media literacy organisation and fact-checkers [Demagog.pl](#) and [FakeNews.pl](#);
- **2023 Dutch General Elections:** From 25 October 2023, we launched an in-app [Election Centre](#) to provide users up-to-date information about the 2023 Dutch House of Representatives elections. The centre contained a section about fighting-fake news, which linked to media literacy association [isdatechtzo.nl](#) and included videos created in partnership with fact-checking organisations [dpa Deutsche Presse-Agentur](#) and [Nieuwscheckers](#);
- **2024 Finnish Presidential Election:** From 1 January 2024, we launched an in-app [Election Centre](#) to provide users up-to-date information about the 2024 Finnish presidential elections. The centre contained a section about fighting-fake news, which linked to media literacy association [Mediataitokoulu](#) and included videos created in partnership with fact-checking organisation [Logically Facts](#);
- **2024 Irish Referendum:** From 2 February 2024, we launched an in-app [Referendum Centre](#) to provide users up-to-date information about the 2024 Irish Referendum and legislative elections. The centre contained a section about fighting-fake news, which linked to media literacy association [medialiteracyireland.ie](#) and included videos created in partnership with fact-checking organisations [The Journal](#) and [Logically Facts](#);
- **2024 Slovak Presidential Election:** From 3 March 2024, we launched an in-app [Election Centre](#) to provide users up-to-date information about the 2024 Slovak presidential elections. The centre contained a section about fighting-fake news, including videos created in partnership with media local media literacy organization [DigiQ](#);
- **2024 EU Parliamentary Election:** From 28 March 2024, we launched an individual [in-app Election Centre for each EU member state](#) available in 27 EU Languages. Working with electoral commissions, European Parliament, civic society organisations including

Fact-checkers, these Centres connected people with reliable voting information, including when, where, how to vote and, ultimately, the election results themselves. Between March and June 2024, a few countries in the EU, including Croatia, Lithuania, Bulgaria and Belgium, were called to national wide elections. In this specific instances we worked with national electoral commissions to include relevant voting information in our in-app EU Election Centre. All 27 Elections Centres contained a media literacy section that includes videos from our trusted fact-checking and partners local media literacy bodies: [Agence France-Presse](#) (AFP), [dpa Deutsche Presse-Agentur](#) (DPA), [Demagog.pl](#), [Demagog.cz](#), [Facta](#), [Faktograf](#), [Logically Facts](#), [Newtral](#), [Poligrafo](#), [Delfi.lt](#), [The Journal](#), [Nieuwscheckers](#), [Funky Citizens](#), [DigiQ](#), [Ostro](#); and

- **2024 French Parliamentary Election:** From 17 June 2024, we launched an in-app [Election Centre](#) to provide users up-to-date information about the 2024 French Legislative elections. The centre contained a section about fighting-fake news, including videos created in partnership with fact-checking organisations [Agence France-Presse](#) (AFP).

c. Knowledge development

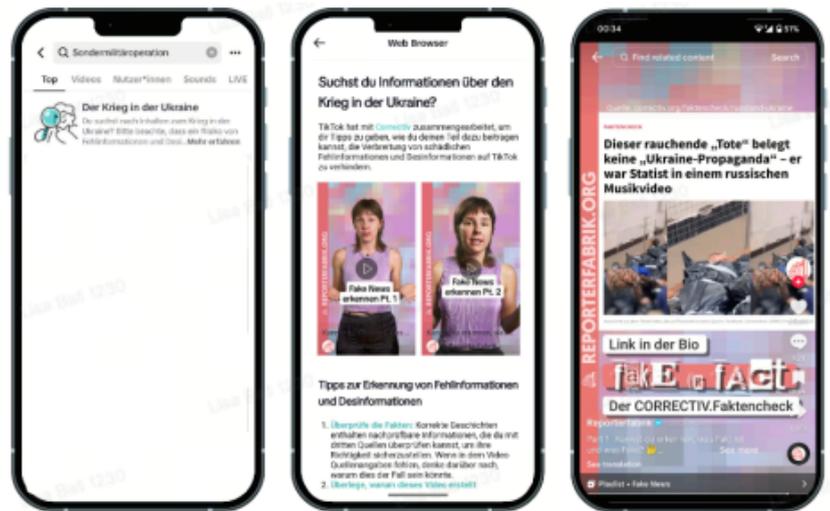
In order to effectively scale the expertise and feedback provided by our fact-checkers globally, we have implemented the measures listed below:

- **Fact-checking repository:** As noted above, we have developed a database of claims which have previously been fact-checked by our global fact checking partners. This contributes to knowledge development and achieves efficiencies because feedback from fact-checkers is more easily scalable and, therefore, avoids our moderators repeatedly sending the same claims for review and verification, and enables them to make swifter decisions on suspected misinformation content;
- **Proactive detection by our fact-checking partners:** Each of our fact-checking partners are authorised to proactively identify content that may constitute Harmful Misinformation on our Platform and suggest prominent misinformation that is circulating on other social media platforms or websites that may benefit from verification; and
- **Fact-checking guidelines:** We create guidelines and trending topic reminders for our moderators on the basis of previous fact-checking assessments. This ensures our moderation teams leverage the insights from our fact-checking partners and helps our moderators make swift and accurate decisions on flagged content regardless of the language in which the original claim was made.

3. TIKTOK MEDIA LITERACY CAMPAIGNS IN EUROPE

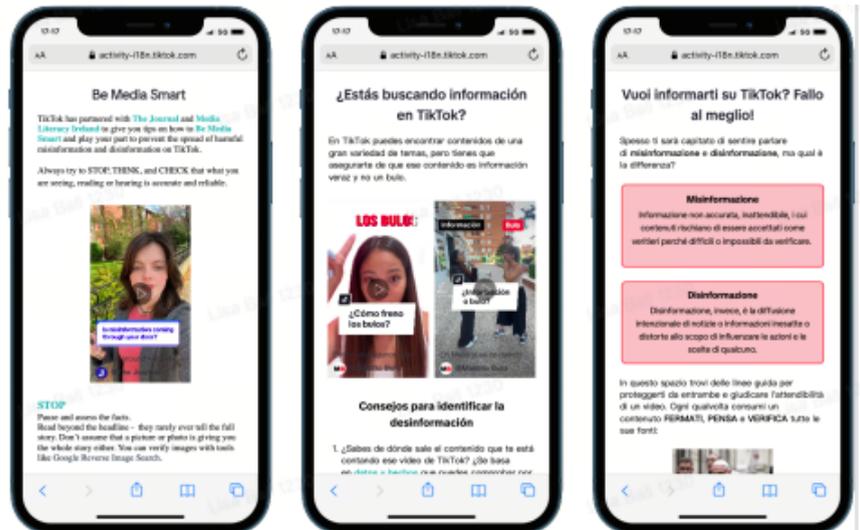
Election Campaign	Relevant TikTok Interface
-------------------	---------------------------

Six (6) campaigns launched in [Austria](#), Bulgaria, Czech Republic, Croatia, [Germany](#) and Slovenia were specific to the war in Ukraine, including identifying misinformation and manipulated media reports.



Germany and Austria

Seven (7) campaigns launched in [Finland](#), [Ireland](#), [Italy](#), [Spain](#), [Sweden](#), Denmark and [Netherlands](#) were focused on general media literacy and critical thinking skills to fight misinformation and manipulated media online, H5 Example below.



Ireland

Spain

Italy

2023 Slovak Parliamentary Election: From 4 September 2023, we launched an in-app [Election Centre](#) in collaboration with electoral commissions to provide users up-to-date information about the 2023 Slovak parliamentary elections. The centre contained a section about fighting-fake news, including videos



2023 Dutch General Elections: From 25 October 2023, we launched an in-app [Election Centre](#) to provide users up-to-date information about the 2023 Dutch House of Representatives elections. The centre contained a section about fighting-fake news, which linked to media literacy association isdatechtzo.nl and included videos created in partnership with fact-checking organisations [dpa Deutsche Presse-Agentur](#) and [Nieuwscheckers](#).



2024 Finnish Presidential Election: From 1 January 2024, we launched an in-app [Election Centre](#) to provide users up-to-date information about the 2024 Finnish presidential elections. The centre contained a section about fighting-fake news, which linked to media literacy association [Mediataitokoulu](#) and included videos created in partnership with fact-checking organisation [Logically Facts](#).



2023 Irish Referendum: From 2 February 2024, we launched an in-app [Referendum Centre](#) to provide users up-to-date information about the 2024 Irish Referendum and legislative elections. The centre contained a section about fighting-fake news, which linked to media literacy association [medialiteracyireland.ie](#) and included videos created in partnership with fact-checking organisations [The Journal](#) and [Logically Facts](#).



2023 Slovak Presidential Election: From 3 March 2024, we launched an in-app [Election Centre](#) to provide users up-to-date information about the 2024 Slovak presidential elections. The centre contained a section about fighting-fake news, including videos created in partnership with media local media literacy organisation [DigiQ](#).



2024 EU Parliamentary Election: From 28 March 2024, we launched an individual [in-app Election Centre for each EU member state](#) available in 27 EU Languages. Working with electoral commissions, European Parliament, civic society organisations including Fact-checkers, these Centres connected people with reliable voting information, including when, where, how to vote and, ultimately, the election results themselves. Between March and June 2024, a few countries in the EU, including Croatia, Lithuania, Bulgaria and Belgium, were called to national wide elections. In these specific instances we worked with national electoral commissions to include relevant voting information in our in-app EU Election Centre. All 27 Elections Centres contained a media literacy section that includes videos from our trusted fact-checking and partners local media literacy bodies: [Agence France-Presse](#) (AFP), [dpa Deutsche Presse-Agentur](#) (DPA), [Demagog.pl](#), [Demagog.cz](#), [Facta](#), [Faktograf](#), [Logically Facts](#),



[Newtral](#), [Poligrafo](#), [Delfi.lt](#), [The Journal](#), [Nieuwscheckers](#), [Funky Citizens](#), [DigiQ](#), [Ostro](#).

