NORDIS – NORdic observatory for digital media and information DISorders

Platform responsibility, legitimacy building and harmful social media content

Report on two scientific papers

Date: 23-08-2023

Final version

NORDIS

Action No:	2020-EU-IA-0189
Project Acronym:	NORDIS
Project title:	NORdic observatory for digital media and information DISorders
Start date of the project:	01/09/2021
Duration of the project:	26
Project website address:	https://datalab.au.dk/nordis
Authors of the deliverable	Mervi Pantti, University of Helsinki Yang Xu, University of Helsinki
Activity number	Activity 3
Task	Task 3.4

Reviewers: Mathias Holm Tveen (Aarhus University, Denmark), Marina Charquero Ballester (Aarhus University, Denmark), Laurence Dierickx (University of Bergen, Norway)

Funding:
The research was funded by EU CEF grant number 2394203.

Executive Summary

This report stems from research conducted at the University of Helsinki, Media and Communication Studies (Pantti & Pohjonen, 2023; Xu, 2023). The report discusses content moderation from the perspective of powerful platform companies (Big Tech) in the context of increasing demands for accountability. Multinational platform corporations such as Microsoft, Alphabet/Google (YouTube), Meta (Facebook, Instagram, WhatsApp), Twitter (currently X), and Bytedance (TikTok) wield significant political, economic, social, and infrastructural power globally. Recent years have witnessed a public backlash against the platforms for causing social harm because of their profit interest, privacy risks, and biased algorithmic systems (e.g., Zuboff, 2019). There have been demands from policymakers, NGOs, citizens, and former employees for Big Tech companies to demonstrate public accountability concerning disinformation and hate speech published on their platforms.

European Union (EU) has taken several steps to limit the spread of online disinformation. The 2022 Strengthened Code of Practice on Disinformation replaced the 2018 self-regulatory Code and was signed by several platform companies, including Facebook and TikTok. The Digital Services Act (DSA) (2022) requires large online platforms to remove illegal content based on European and national laws and increase their efforts to fight misinformation and disinformation campaigns. In the Nordic context, the Nordic Think Tank for Tech and Democracy has demanded speedy efforts to regulate social media platforms (A Nordic approach to democratic debate in the age of Big Tech (norden.org).

In this context, characterized by a growing political will in Europe and the Nordics to hold platforms accountable, this report focuses on how online platforms have responded to increasing criticism and regulatory efforts. The report analyzes corporate texts (blogs and community guidelines) and investigates how social media platforms perceive their responsibilities and how their discursive legitimation processes have evolved. While corporate public communication has a promotional leaning, it also allows researchers to explore how the platforms articulate their legitimacy and frame their activities.

In particular, the report focuses on how Western platform companies responded to the surge of disinformation following Russia's war in Ukraine and how the Chinese social media

company TikTok's community guidelines have evolved concerning harmful content. We chose these two cases to demonstrate how digital platforms respond to accountability demands. Following the Russian war in Ukraine, Western digital platforms were under exceptional pressure from the EU to counter Russian disinformation. Chinese social media company TikTok has faced increasing security and disinformation concerns in Europe which recently led the company to pledge to counter disinformation more effectively to live up to the EU code of practice on disinformation (Chee 2023).

1.0 Platforms responding to accountability demands

Platforms wield significant power today in shaping political and civic processes. It is through the platforms that people, especially young people, currently experience the social world (van Dijck & Poell, 2013; van Dijck et al., 2018). Such power comes with responsibility but accountability in the tech sector has for long fallen short. While platforms communicate corporate social responsibility, their responses to disinformation and other harmful content have been reactive. While social media platforms constitute a 'public square' and have aspects of public good, as multinational private companies, they are principally driven by commercial interests (Gillespie, 2018; Helberger et al., 2018).

Typically, their policies and practices concerning content moderation have been triggered by 1) shocks and scandals, 2) (anticipated) legislation and regulation, and 3) pressure from stakeholders. Following the COVID-19 pandemic and Russia's invasion of Ukraine, platforms have taken an exceptionally wide array of steps to counter mis/disinformation. A study by Carnegie (2021) on fourteen platforms found a surge in public announcements of interventions to counter mis/disinformation since 2019; however, most of the measures were 'soft measures' (such as content labeling) with questions of effectiveness (Yadav, 2021).

Today, platforms have come under growing global criticism for disrupting democracy as they have not been able to prevent the spread of harmful content. Over the years, several societal actors including scholars, NGOs, journalists, and national governments have urged companies to counter and sanction disinformation (e.g., Larson, 2018), as seen in the criticism of their failure to control disinformation during elections and remove hate speech in countries such as Myanmar and Ethiopia, or more recently in Ukraine (Pohjonen, 2019; Udupa et al., 2012). The usual criticism of social media platforms is that they act on problems reactively and do too little. They lack transparency and information on how they counter disinformation and prioritize some countries or languages over others in their disinformation responses (Wagner et al., 2022). For instance, platforms' responses following the invasion of Ukraine and governmental pressure in Europe led to criticism from civil society activists for the uneven application of human rights across global conflicts. A petition signed by 31 civil liberties organizations argued that other crisis situations where lives are at stake had not received the same amount of support:

> While we recognize the efforts of tech companies to uphold democracy and human rights in Ukraine, we call for long-term investment in human rights, accountability, and a transparent, equal and consistent application of policies to uphold the rights of users worldwide. Once platforms began to take action in Ukraine, they took extraordinary steps that they have been unwilling to take elsewhere. (EFF, 2022)

The relationship between platforms, content moderation, and freedom of speech has been historically antagonistic and contested. Recent crises, the pandemic and war in Ukraine, arguably *represent a shift that brought social media platforms to closer alignment with the Western government's concerns*. For online platforms, these crises have emerged as critical incidents to re-establish their legitimacy. Following platforms' expansive and quick response to ban and remove harmful COVID-19 content, one commentator noted that 'Techlash has been put on pause' (Meserole, 2020).

Historically, digital platforms have tried to balance the conflicting demands of governments and civil society groups to moderate and remove illicit content by positioning themselves as neutral intermediaries who are not legally liable or socially responsible for the published content (Gillespie, 2010; Napoli & Caplan, 2017). One exception was the relative global consensus that emerged in response to the use of social media by jihadi groups such as Al-Qaeda and ISIS (Conway et al., 2017). Within a relatively short time, platforms acted to remove violent content through collaborations, such as the Global Internet Forum for Counter-terrorism (GIFCT). These collaborative efforts between social media companies and governments were later expanded in the Christchurch Call following the Christchurch Mosque attacks in New Zealand in 2019 to control other types of terrorist content on social media (Hoverd et al., 2021).

Even before the pandemic and war, there was growing interest and policy pressure in Europe to confirm intermediary responsibility for platforms. The Digital Services Act (DSA), launched in April 2022, sets out new standards for the accountability of platforms regarding harmful content. In this regulatory framework, platforms are required to mitigate risks, such as disinformation and hate speech. The changing nature of the relationship between digital platforms and governments in times of grave crises is written in the DSA's 'crisis response mechanism' (DSA 2022, Art. 27a), which allows the European Commission to intervene in

content moderation decisions and requires 'Very Large Online Platforms (VLOP)' to limit any urgent threats to public security.

1.1. Legitimacy building following the Russian invasion of Ukraine

Following the Russian war in Ukraine, digital platforms acted on disinformation by blocking Russian state-affiliated media locally or globally (HRW 2022). Platforms adopted similar content moderation changes one after another in response to public pressure – as the platforms tend to closely watch each other's actions and create joint normative visions by unifying their policies of what is 'harmful' and 'how and when they should intervene' (Gillespie et al., 2020; Caplan & Danah, 2018). Scott and Kern (2022) have claimed that these decisions to take a stand against Russia could 'fundamentally change the companies' relationships with governments that are being forced, in real-time, to acknowledge the power that social media wields in a time of war.' From an alternative perspective, however, these decisions show that the EU and European governments had the power to direct how platforms responded to the Russian invasion of Ukraine. The geopolitical crisis added more momentum to efforts in Europe to demand greater accountability from platforms and, in practice, force them to pick a side.

The regulation of the Council of Europe (EU 2022/350) to suspend broadcasts from Russian state-sponsored media outlets RT and Sputnik was implemented on March 1, five days after Russia invaded Ukraine. Although the securitization of disinformation in various EU documents had started following the annexation of Crimea in 2014 (e.g., European Council 2015), such restrictions had not previously been used to regulate social media platforms in times of crisis, at least at such scale and scope against another major geopolitical player. The regulation was justified as a response to the security threat that Russian disinformation poses to the EU: 'The Russian Federation has engaged in continuous and concerted propaganda actions targeted at civil society in the Union and neighbouring countries, gravely distorting and manipulating facts' (EU 2022/350, 7). Disinformation was interpreted as 'part of a hybrid warfare strategy Russia was using against the EU,' requiring extraordinary measures 'to defend all citizens and infrastructure, as well as their democratic systems,' as stated by the 'Special Committee on Foreign Interference in all Democratic Processes in the European Union, including Disinformation' (INGE) (European Parliament 2022). The

committee concluded that the EU should tighten control on platforms to protect European citizens and democracy.

Other parties, including the Ukrainian government and several national governments added to the pressure by making public statements demanding a crackdown on Russian disinformation on platforms. In particular, Mykhailo Fedorov, Ukraine's Minister of Digital Transformation, launched a shame and guilt-eliciting campaign on social media that pressured all main social media and tech companies to cut ties with Russia (The New York Times, 2022). These high publicity requests placed digital platforms at the centre of the information war and geopolitical conflict, in which they were forced to pick sides and put their preferred impartial stance aside.

Facebook (Meta) and the Chinese video platform TikTok took the lead and blocked access to RT and Sputnik across Europe, along with barring Russian state media from running ads. Twitter followed with a similar ban. The actions extended to information technology companies: Microsoft blocked downloads of the RT app worldwide and Google did the same in Ukrainian territory. In addition, Google-owned YouTube blocked RT and Sputnik channels in Europe and barred their ad revenue on YouTube. Apple blocked RT and Sputnik from the Apple App Store outside Russia and suspended all product sales in Russia. As a distinct measure, Meta platforms Facebook and Instagram issued a change in their regular hate speech policy to allow users in Ukraine to call for death against Russian leaders and troops in the context of the Russian invasion of Ukraine. This decision led to global controversy about the alleged double standards adopted by platforms and critics pointing out that 'Facebook's human rights and free speech rules tend to match up with U.S. policy preferences' (Biddle, 2022).

The corporate communication of the major social media companies during the beginning of the war in Ukraine suggests that digital platforms would prefer to see themselves as collaborating – rather than being in an adversarial relationship – with Western governments. We studied all the blog posts published by Google, Meta (Facebook, Instagram), Twitter, and YouTube between the Russian invasion from late February to late October 2022. Our initial aim was also to include TikTok, the only non-Western company, but TikTok did not directly address the invasion in its corporate blog. For the platform companies, the war represented

an opportunity to rehabilitate their reputations and highlight their key values after facing accusations of not doing enough to prevent the spread of harmful content. The platforms, unsurprisingly, did not address their shortcomings regarding the circulation of disinformation and hate speech but stressed their role as 1) humanitarian actors, 2) cybersecurity experts, and 3) guardians of democracy through technology and innovation.

Amidst the accountability pressure targeted at the platforms by Ukraine and the EU, a key theme that the blogs highlighted was the multiple *humanitarian missions* the companies were engaged with to help Ukrainians by collaborating with several governmental and non-governmental agencies. These partnerships effectively communicated that the company/platform was socially responsible. Meta (3 March 2022), for instance, stressed collaboration with leading global humanitarian organizations and detailed its own humanitarian foci, including supporting journalists and human rights activists in Ukraine:

> We're committing $15 million to support humanitarian efforts in Ukraine and neighboring countries. This includes $5 million in direct donations to UN agencies and more than a dozen nonprofits, including International Medical Corps who will be using these funds to deploy mobile medical units to Ukraine and Internews to support at-risk journalists and human rights defenders in the region.

The platforms also underlined their role as *leading cybersecurity actors* during (and preceding) the war in Ukraine. We found that the blogs portrayed platforms as actors whose technological expertise can be used to support Ukraine in the information war and cyber war. More broadly, they defend democracies against cyber threats such as disinformation campaigns, often from authoritarian countries like Russia. The platforms stressed how they were taking measures to support Ukraine by removing misleading and harmful content and providing the technological support needed to protect the Ukrainian government from cyberattacks such as phishing, malware campaigns, espionage, and malign information operations. One such measure was Twitter's new 'crisis misinformation policy' announced on 19 May 2022, supposedly sped up by the public pressure following Russia's invasion. According to the post, their crisis misinformation policy is as follows:

> …a global policy that will guide our efforts to elevate credible, authoritative information, and will help to ensure viral misinformation isn't amplified or

> recommended by us during crises. In times of crisis, misleading information can undermine public trust and cause further harm to already vulnerable communities. [--] While this first iteration is focused on international armed conflict, starting with the war in Ukraine, we plan to update and expand the policy to include additional forms of crisis.

Thus, platforms as cybersecurity actors perceived themselves as active participants in the information war – not as neutral players as might have been the case before – whose interests are aligned with the geopolitical interests of the Western world. In this narrative, the platforms had gained their authority through persistent technological development and sharing their knowledge with others – platform companies, governments, and other societal actors. The emphasis on the long-time efforts to create technologies and policies to safeguard online safety aims to rehabilitate their reputations after facing demands in recent years to take responsibility for spreading harmful content. Whereas the cybersecurity narrative focused on sharing technological expertise to fight against cyber threats, the third narrative posing the platforms as *guardians of democracy* addressed the role platform companies have in defence of Western values more broadly, echoing the earlier debates about Silicon Valley as a 'force for good' globally (Morozov, 2011). The role of the guardian of democracy was articulated as a response to the criticism that the platforms have been a target of in recent years, especially for their inaction and ineptitude in moderating hate speech and disinformation.

Thus, the Ukraine War provided an opportunity to show how technology companies can, once again, be the solution rather than the cause of the problems that democracies face. In particular, the technological innovation represented by these companies can also provide the necessary resources to respond to future crises and defend the free exchange of ideas necessary for democracies to function. Accordingly, Meta explicitly positioned its policies to defend Ukrainian 'rights to speech as an expression of self-defense in reaction to a military invasion of their country,' including actions to block propaganda from state-led media outlets run by Russia, such as RT and Sputnik (Meta, 26 February 2022). This 'ideal' of defending democracy from external threats and state propaganda by authoritarian countries can be

seen in how the social media companies' public relations communication tried to explain their role during the war.

However, in practice, the actions taken by the companies represent more ad hoc decisions taken in response to the changing political situation and growing public and political pressure, as it becomes similar in other crises. Critics have argued that Meta's Facebook and Instagram made more than a dozen policy revisions since the start of the invasion, leading to internal confusion, especially among content moderators working on the front lines of deciding what content is acceptable and what is not (Mac et al., 2022). These narratives nonetheless show how the policies of social media platforms often work reactively with changing political environments and public opinion, trying to find a suitable position in an increasingly strict regulatory environment, especially in the EU.

2.0 The evolving content policies of TikTok

TikTok, a short video social media platform owned by Chinese internet company ByteDance, has generated remarkable global success since its international launch in 2017. The platform has attracted 150 million monthly active users (MAUs) in Europe and has been designated as a 'Very Large Online Platform (VLOP)' by the European Digital Service Act

(TikTok, 2023; European Commission, 2023). The platform has become extremely popular among the younger, especially Generation Z users (people born from 1997 to 2012).

Despite its immense popularity and appeal, the platform has faced multiple criticisms on various issues, including poor content moderation practices in harmful content and misinformation, failing to protect children and youth safety online, and data security concerns due to the platform's Chinese origin. TikTok has been seen as one of the most problematic platforms concerning disinformation: according to *NewsGuard's Misinformation Monitor* (2022), one of five videos delivered in TikTok's search engine contains false information. In recent years, TikTok has also faced several lawsuits across many regions regarding the platform's illegal processing of children's data (POLITICO, 2021; The Guardian, 2023b; TechCrunch, 2023). Moreover, the platform also faces broad distrust from many governments with deep worries about TikTok's connections to the Chinese government and the platform's potential threats to national security. The European Commission, Council of the EU, and the EU Parliament have all banned TikTok from official devices due to security concerns and fear of foreign interference, putting TikTok in a critical position to address issues on platform transparency and accountability (Euronews, 2023). TikTok in a public release stated that the platform seeks to meet its obligation under the EU Digital Service Act and make efforts in content moderation, protecting teens online, and being transparent about the recommendation system (TikTok, 2023).

Content policies, often in the name of community guidelines or community standards, are sets of principles, rules, and standards for platforms to classify online content and behaviors that are allowed, disfavored, and prohibited. These rules often provide legible justifications for platforms' content moderation work and reveal platforms' key ideas, attitudes, and values towards the governance of online space. Platforms keep their community policies up to date as both the platform and users change. *Changes in these policies reveal the platform's stance on specific issues, the controversies the platform faces, and the external forces the platform needs to respond*.

The documents collected for the study utilized Wayback Machine powered by Internet Archive (https://archive.org/web/), a nonprofit platform that provides access to archived digital materials. Twelve versions of TikTok's community guidelines were found throughout

the data collection of this study, with the first version published in August 2018 and the latest version published in March 2023. Among all twelve versions, five versions of TikTok's community guidelines (August 2018, January 2020, December 2020, February 2022, and March 2023) were given the most attention due to the significant volume and structural change. The study focused on two policy problems/areas: misinformation and minor safety. In addition to community guidelines, other documents (public releases, transparency reports, and user guides) were analyzed associatively to provide a complete picture of TikTok's policy improvements.

2.1 On information disorder

Over several updates of TikTok's community guidelines, the problem of information disorder has experienced several reframings where different terms and definitions were applied to the target content and behaviors. NORDIS uses the term 'information disorder' to include all forms of misleading information with or without harmful intentions (Wardle & Derakhshan, 2017). The first change occurred in January 2020 when TikTok adopted the terms 'misleading information' and 'misinformation' in its content policy under the problem area of Integrity and authenticity. The platform removes misinformation that 'could cause harm to an individual's health or wider public safety' and 'content distributed by disinformation campaigns.' In the later August and December updates, TikTok added 'misinformation related to emergencies that induce panic,' 'digital forgeries (synthetic media or manipulated media),' 'coordinated inauthentic activities,' and 'conspiratorial content' to the forbidden list, further expanding the type of content and behavior related to misinformation.

One crucial and decisive factor that accelerated TikTok's policy changes from 2020 to 2022 was the advent of the COVID-19 pandemic and the US 2020 election, where TikTok became the epicenter of health misinformation and misinformation about the US election. The unprecedented, rapid proliferation of Covid-19 misinformation globally posed severe pressures for social media platforms (World Health Organization, 2020). Besides the general pressure of the COVID-19 pandemic as an international public health emergency, the government also significantly pressures online platforms regarding policy changes. TikTok joined the EU's Code of Practice on Disinformation in June 2020, following the commitments to take voluntary steps to counter the spread of disinformation (TechCrunch, 2020).

Meanwhile, the European Commission set out an emergent joint communication and monitoring program in June 2020 to tackle COVID-19 misinformation, requesting online platforms to submit monthly reports about their actions (European Commission, 2020). The platform highlighted a series of efforts, including reviewing its misinformation policies and moderation practices, directing users toward health information from authoritative sources, removing misinformation, and restricting advertising to trusted sources (TikTok, 2020).

In the US election 2020, TikTok was greatly pressurized by the US government under the growing criticism of the platform's Chinese origin and its potential threat to US national security. Months after the Trump Administration's Executive Order which proposed sanctions on TikTok in August 2020, the platform released a series of measures targeting the 2020 US election, including launching an in-app guide for users to access authoritative information and enhancing misinformation policies tackling election interference (TikTok, 2020). Under the section 'election integrity' In TikTok's safety center (https://www.tiktok.com/safety/en/election-integrity/), the platform provided detailed information about the moderation approaches to election misinformation with clear rationales. Although TikTok has never explicitly stated the pressures that prompted the platform's policy change, we believe that on misinformation policy, the platform was pressured by crisis events (COVID-19 pandemic and health misinformation) and political leadership (Trump administration's order and election misinformation).

In February 2022, TikTok reframed the problem as 'Harmful misinformation,' emphasizing misinformation that causes significant harm (serious physical injury and psychological trauma, large-scale property damage, and undermining of public trust) while exempting certain types of misinformation such as 'simply inaccurate information, myths, commercial or reputational harm.' Meanwhile, the February 2022 policy introduced a separate policy for the 'For You Feed (FYF)' — a page based on the platform's personalized recommendation system. Despite not receiving algorithmic promotion on the platform, ineligible content for the FYF is still searchable and viewable. In the February 2022 guidelines, conspiratorial content that is against widely-accepted beliefs and casts blame on a group or entity, or potentially misleading and harmful content about current, still-developing events are not eligible for platform recommendation.

In the latest policy update in March 2023, TikTok further specified its moderation approach as the 'three pillars' structure in which content is categorized as Not Allowed (forbidden), For You Feed Ineligible (Not receive platform recommendation), and Allowed. Like February 2022's version, misinformation is regulated based on the severity of real-life effects/impacts posed to the public. While some types of misinformation (such as climate change misinformation and dangerous conspiracy theories) are forbidden, other types of misinformation were given a less restrictive measure, making the rules ambivalent. General conspiracy theories, for instance, which TikTok defined as 'unfounded' claims about 'certain events or situations carried out by covert or powerful groups,' are FYF ineligible (not recommended by the platform but still searchable and viewable). Statements of personal opinion (as long as it does not include harmful misinformation) are labeled as 'allowed' even though they can alter the viewer's beliefs.

From the changes in February 2022 onward, TikTok's misinformation policy has experienced a shift towards a more impact-focused approach in which the platform focused on the real-life 'harm' that misinformation caused to individuals and the public. Nevertheless, the concept of 'harm' is not neutral and is subject to contesting values and individual interpretations. Misinformation policy guided by the concept of 'harm' can be vulnerable to disinformation in the form of personal claims, questions, or entertainment content (Baker et al., 2020). Moreover, we believe that TikTok's three-pillar moderation approach developed since February 2022 still contains loopholes for harmful misinformation. NewGuard Misinformation Monitor's investigation indicated that 20% of videos contain toxic misinformation through TikTok's search engine (NewsGuard, 2022). TikTok's misinformation policy has a structural weakness that allows certain types of misinformation to be searchable and viewable across the platform. Finally, TikTok's misinformation policy has gradually followed the footsteps of other large social media platforms by taking an ambivalent stance between 'ensuring freedom of speech' and 'countering misinformation,' as challenges persist within the platform industry where more context is needed when assessing misinformation.

2.2 On children and youth safety

TikTok's content policies regarding youth safety and well-being have gained greater importance over the years. From 2018 to 2022, the volume of policy text concerning children

and youth has grown from 153 words (August 2018) to 1014 words (February 2022). Compared to the August 2018 version, where the policy only touched upon the problem of 'child safety infringement,' problems were greatly diversified in the January 2020 version, where the platform framed the problem as 'minor safety' with five distinct categories: Nudity and sexual exploitation involving minors, underage delinquent behaviors, child abuse, grooming behavior, and sexualization of minors. Later in the December 2020 update, TikTok further expanded its reach by reframing the issues into new categories such as physical and psychological harm to minors and crimes against children. Finally, in the February 2022 version, TikTok moved the problem of 'minor safety' to the forefront of the document after the introduction, marking its significance to the platform's content regulation.

On a platform level, TikTok has adopted several measures to protect minors' online safety over the years. One of them is offering limited app experience to underage users. In December 2020, TikTok updated its policies by setting different age limits. On top of the minimum age requirement, which is set at 13 years, TikTok limited direct messaging, hosting live stream, and featuring on the For You Page to users above 16. Virtual gifting features (purchasing, sending, or receiving virtual gifts) are limited to users over 18 years old (TikTok, 2019). In the US, TikTok introduced 'TikTok for Younger Users' to accommodate users under 13 with a restrictive app experience (viewing curated videos only) (TikTok, 2019). Later in January 2021, TikTok announced a series of changes, tightening up settings for underage users, including the default privacy settings, comment features, Duet and Stitch settings, video download settings, and account suggestion settings (TikTok, 2021). In October 2022's policy update, the platform raised the age limit of the live stream feature (TikTok LIVE) from 16 years to 18 years and older as the platform planned to cater to adult-only streams (TikTok, 2022).

Nevertheless, the platform does not have any age verification tools for newcomers and has been facing repercussions for its malpractices in keeping the age restriction strict. On April 4th, 2023, TikTok was fined in the UK for illegally processing data from children under 13 using the platform without parental consent (The Guardian, 2023a). A similar penalty was given by The European Data Protection Board (EDPB) in August 2023 for failing to ensure age verification and illegally processing data of children under 13 (The Guardian, 2023). In

TikTok's US Congressional hearing on March 23rd, 2023, the CEO Shou Zi revealed that besides the 'age gating' method (user provides their age themselves when asked), the platform scans public videos to determine the age (TechCrunch, 2023; Tech Policy Press, 2023). However, TikTok's CEO did not further explain the technical details behind the alleged 'video scanning' method. What we see from TikTok's trajectory reflects the industrial struggle to conduct effective age verifications while protecting user privacy. TikTok's approach of a 'neutral, industry-standard age gate' requires more strengthening and transparency.

On parental control and youth digital well-being, TikTok introduced 'Family Pairing' in April 2020, enabling parents to link their children's accounts and set controls such as managing screen time and restricting direct messages (TikTok, 2022). In March 2023, Family Pairing received new features on screen time management and notification settings (TikTok, 2023). Later in June, the platform added the content filtering tool to its Parent Pairing feature, allowing parents to customize content for their children (TikTok, 2023). Despite all these improvements, parental control is not a default setting on TikTok. According to euCONSENT's report published in 2021, technical proficiencies often bar parents from using parental control tools effectively. Automated or default parental control settings as a starting point for delegating controls are beneficial for parents with less technical skills (Smirnova et al., 2021). In alignment with the Nordic Think Tank's recommendation (2A), we believe that TikTok shall implement default parental controls to safeguard children and youth online safety.

3.0 Discussion and conclusion

In this report, we combined two scientific research on the interplay between platform values, platform discourses, and the external pressures that hold platforms accountable. In the first study, we saw platform value discourses serve as an (instrumental) necessity of legitimization, both in the sense of platforms' efforts to maintain their integrity, build public trust, and construct a positive public image. As online platforms are under growing pressure to respond to the accountability demands from governments and civil society actors, reporting in public blogs and working on their community guidelines functions as an act of

discursive legitimation. Moreover, value discourses are highly performative as they are closely associated with the platform power. Drawing from Reed's (2013) conception of power and its different dimensions, power has a performative dimension in which situated actions (well-timed acts) and interaction (in tune with the situation) exert control over actors and their future actions. Values, aims, and missions expressed in corporate texts have consequences. They shape platform companies' future actions and give governmental and civic actors grounds to hold them responsible based on their own commitments. Increasing pressure has led to a 'responsibility turn' in platform companies' public communication (Flew, 2018; Mager & Katzenbach 2021). Covid-19 and the war in Ukraine have been critical incidents for the platforms to re-establish their societal role, even if it means that platforms need to tweak their narratives in a different direction than what they have long committed to. The geopolitical crisis and Pandemic added more momentum to efforts in Europe - and Nordics - to demand greater accountability from platforms.

Besides crisis events, platforms also react to political pressure and legal frameworks. As we have seen in the second study, guiding principles and values for TikTok's moderation have received a complete reshuffle in the January 2020 and March 2023 update, where the platform has completely reconstructed its community guidelines following a new approach aligned with international legal frameworks. TikTok stated that these principles are informed by the UN Guiding Principles on Business and Human Rights and the Santa Clara Principles, and the platform seeks to align with international legal frameworks, such as the International Bill of Human Rights and the Convention on the Rights of Children. From the changes in TikTok's community guidelines since 2018, we saw the platform's reactivity to political leadership (since the Trump administration's Executive Order of banning the platform nationwide) and legal framework (enforcement of DSA), as these two forces are effective to hold the platform accountable.

Moreover, we also want to remind of Gorwa's (2019) claim that 'platform companies are (still) companies,' in the sense that platforms are profit-making businesses that abide by the logic of the platform market, which is fueled by neoliberalist ideology, narratives, and ideals. Platforms are not legally treated as media, so they are not obligated to follow the same editorial principles or be transparent about them; rather, the quality of the content depends

on a combination of political pressure and the interest of keeping users on the platforms for pure economic gains. Too much bad publicity may cause platforms to change the algorithm in order to more aggressively detect false information (Allcott, Gentzkow, & Yu, 2019). We need to keep in mind that the platform's activity/reactivity is necessary for its business growth, and these blogs, community guidelines, and the ethos within do not necessarily reflect their 'true' actions. Despite that, studying corporate texts allows a perspective to understand how platform companies themselves want to frame their actions and how they see their role in the future. Through these two scientific studies, we conclude that platforms' attempts to frame actions in alignment with their self-claimed 'ideals' is partly a self-protective tactic from the demands for accountability. It remains the responsibility of stakeholders, such as Nordic and European regulators and civic organizations, to closely monitor and evaluate digital platforms' professed actions against harmful content.

References

Allcott, H., Gentzkow, M., & Yu, C. (2019). 'Trends in the diffusion of misinformation on social media', *Research & Politic*s, 6 (2). https://doi.org/10.1177/2053168019848554.

Banchik, A. V. (2021). 'Disappearing acts: Content moderation and emergent practices to preserve at-risk human rights–related content', *New Media & Society*, 23 (6), 1527–44.

Baker, S. A., Wade, M., & Walsh, M. J. (2020). 'The challenges of responding to misinformation during a pandemic: Content moderation and the limitations of the concept of harm', *Media International Australia*, 177 (1), 103-107. https://doi.org/10.1177/1329878x20951301.

Biddle, S, (2022). 'Facebook's Ukraine-Russia rules prompt cries of double standard', *The Intercept*, 14 April, https://theintercept.com/2022/04/13/facebook-ukraine-russia-moderation-double-standard/.

Bornakke, T. (2023, April 17).' A Nordic approach to democratic debate in the age of Big Tech: Recommendations from the Nordic Think Tank for Tech and Democracy', Nordic Council of Ministers. https://www.norden.org/en/publication/nordic-approach-democratic-debate-age-big-tech

Caplan, R., & Boyd, D. (2018). 'Isomorphism through algorithms: Institutional dependencies in the case of Facebook', *Big Data & Society*, 5 (1), 205395171875725. https://doi.org/10.1177/2053951718757253

Chee, F.Y. (2023). 'TikTok promises to ramp up fight against disinformation in EU'. Reuters, 9 February, https://www.reuters.com/technology/tiktok-promises-ramp-up-fight-against-disinformation-eu-2023-02-09/

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2017). 'Disrupting Daesh: Measuring takedown of online terrorist material and its impacts', *Studies in Conflict & Terrorism*, 42 (1-2), 141–60.

DeCook, J. R., Cotter, K., Kanthawala, S., & Foyle, K. (2022). 'Safe from "harm": The governance of violence by platforms', *Policy & Internet*, 14, 63–78.

De Gregorio, G. (2019). 'From constitutional freedoms to the power of the platforms: Protecting fundamental rights in the algorithmic society', *European Journal of Legal Studies*, 11 (2), 65–103.

DSA (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065&qid=1666966938325.

EDMO (2022). 'How Covid-19 conspiracy theorists pivoted to pro-Russian hoaxes', European Digital Media Observatory, 30 March, https://edmo.eu/2022/03/30/how-covid-19-conspiracy-theorists-pivoted-to-pro-russian-hoaxes/.

EU (2022/350). Council Regulation (EU) 2022/350 of 1 March 2022, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R0350.

Euronews. (2023, March 3).' Domino effect as more EU institutions move to ban TikTok on work devices', Euronews. https://www.euronews.com/my-europe/2023/03/03/domino-effect-as-more-eu-institutions-move-to-ban-tiktok-on-work-devices.

European Commision. (2020). COVID-19 disinformation monitoring programme. European Commission. https://digital-strategy.ec.europa.eu/en/policies/covid-19-disinformation-monitoring.

European Commission. (2023, April 25). Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413.

European Commission (2022). 'Digital Services Act: Commission Welcomes Political Agreement on Rules Ensuring Safe and Accountable Online Environment', Press release 23 April, https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545, accessed 21 September 2022.

European Council (2015). European Council Conclusions, 19-20, https://www.consilium.europa.eu/media/21888/european-council-conclusions-19-20-march-2015-en.pdf

European Parliament (2022). European Parliament resolution of 9 March 2022 on foreign interference in all democratic processes in the European Union, including disinformation (2020/2268(INI)), https://www.europarl.europa.eu/doceo/document/TA-9-2022-0064_EN.html.

EFF (2022). 'Civil Liberties Groups Urge Social Media Platforms to Better Protect Free Flow of Information in Crisis Zones', Electronic Frontier Foundation, 12 April, https://www.eff.org/press/releases/human-rights-groups-urge-social-media-platforms-better-protect-free-flow-information.

Franceschi-Bicchierai, L. (2021). 'Facebook says 'Death to Khamenei' posts are OK for the next two weeks', *Vice*, 23 July, https://www.vice.com/en/article/v7exzm/facebook-says-death-to-khamenei-posts-are-ok-for-the-next-two-weeks.

Frenkel, S., Issac, M., & Mac, R. (2022, February 1). 'How Facebook is morphing into Meta', *The New York Times*. https://www.nytimes.com/2022/01/31/technology/facebook-meta-change.html

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A., & Myers West, S. (2020). 'Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates', *Internet Policy Review*, 9 (4), https: https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy.

Gillespie, T. (2010). 'The politics of '"platforms"', *New Media & Society*, 12 (3), 347–64.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven, CT: Yale University Press.

Glegg, N. (2020). 'Combating COVID-19 misinformation across our apps', Meta, 25 March, https://about.fb.com/news/2020/03/combating-covid-19-misinformation/.

Gorwa, R. (2019). 'What is platform governance?', *Information, Communication & Society*, 22(6), 854-871. https://doi.org/10.1080/1369118x.2019.1573914.

Hallinan, H., Scharlach, R., & Shifman, L. (2022). 'Beyond neutrality: Conceptualizing platform values', *Communication Theory*, 32 (2), 201–22.

Helberger, N., Pierson, J., & Poell, T. (2018). ‘Governing online platforms: From contested to cooperative responsibility’, *The Information Society*, 34 (1), 1–14.

Hoverd, W., Salter, L., & Veale, K. (2021). ‘The Christchurch Call: Insecurity, democracy and digital media – Can it really counter online hate and extremism?’, *SN Social Sciences*, 1 (2), https://doi.org/10.1007/s43545-020-00008-2.

HRW (2022) Human Rights Watch. Russia, Ukraine, and Social Media and Messaging Apps. Questions and Answers on platform Accountability and Human Rights Responsibilities, https://www.hrw.org/news/2022/03/16/russia-ukraine-and-social-media-and-messaging-apps.

Larson, H. J. (2018). ‘The biggest pandemic risk? Viral misinformation’, *Nature*, 562(7727), 309. https://doi.org/10.1038/d41586-018-07034-4

Pohjonen, M. (2019). ‘A comparative approach to social media extreme speech: Online hate speech as media commentary’, *International Journal of Communication*, 13, 3088–103.

Mager, A., & Katzenbach, C. (2021). ‘Future imaginaries in the making and governing of digital technology: Multiple, contested, commodified’, *New Media & Society*, 23 (2), 223–36.

Meta. (2022, February 26). Meta's ongoing efforts regarding Russia's invasion of Ukraine [Press release]. https://about.fb.com/news/2022/02/metas-ongoing-efforts-regarding-russias-invasion-of-ukraine/

Meserole, Chris (2020). ‘Covid-19 and the future of ‘Techlash’, https://www.brookings.edu/articles/covid-and-the-future-of-techlash/

Morozov, E. (2011). ‘America's internet freedom agenda’, *New Perspectives Quarterly*, 28 (2), 61-63. https://doi.org/10.1111/j.1540-5842.2011.01248.x

Napoli, P., & Caplan, R. (2017). ‘Why media companies insist they’re not media companies, why they’re wrong, and why it matters’, *First Monday*, 22 (5), 1–16.

NewsGuard. (2022, September 11). ‘Beware the 'New Google:' TikTok's search engine pumps toxic misinformation to its young users’, Misinformation Monitor: September 2022. https://www.newsguardtech.com/misinformation-monitor/september-2022/

Pantti, M. & Pohjonen, M. (2023) ‘Social media platforms responding to the invasion of Ukraine’. In M. Mortensen and M. Pantti (eds.), *Media and War in Ukraine*. Peter Lang.

POLITICO. (2021, January 22).’Italy orders TikTok to stop using children's data’, *POLITICO*. https://www.politico.eu/article/italy-orders-tiktok-to-stop-using-childrens-data/

Scott, M., & Kern, R. (2022, March 2). POLITICO, https://www.politico.com/news/2022/03/02/social-media-goes-to-war-00012993

Smirnova, S., Livingstone, S., & Stoilova, M. (2021, September). Understanding of user needs and problems: A rapid evidence review of age assurance and parental controls. EuCONSENT. https://euconsent.eu/download/understanding-of-user-needs-and-problems-a-rapid-evidence-review-of-age-assurance-and-parental-controls/

Stockmann, D. (2022). ‘Tech companies and the public interest: The role of the state in governing social media platforms’, *Information, Communication & Society*, https://doi.org/10.1080/1369118X.2022.2032796

TechCrunch. (2020, June 22). ‘TikTok joins the EU's Code of Practice on disinformation’, *TechCrunch*. https://techcrunch.com/2020/06/22/tiktok-joins-the-eus-code-of-practice-on-disinformation/

TechCrunch. (2023a, March 23). ‘TikTok CEO says company scans public videos to determine users' ages’, *TechCrunch*. https://techcrunch.com/2023/03/23/tiktok-ceo-says-company-scans-public-videos-to-determine-users-ages/#:~:text=Beyond%20investigating%20flagged%20accounts%2C%20the,tend%20to%20interact%20with%20content.

TechCrunch. (2023b, April 4). ‘TikTok hit with $15.7M UK fine for misusing children's data’, *TechCrunch*. https://techcrunch.com/2023/04/04/tiktok-uk-gdpr-kids-data-fine/

Tech Policy Press. (2023, March 24). ‘Transcript: TikTok CEO testifies to Congress’, *Tech Policy Press*. https://techpolicy.press/transcript-tiktok-ceo-testifies-to-congress/

TikTok. (2019a, December 3). Updating our gifting policies to protect our community [Press release]. https://newsroom.tiktok.com/en-us/updating-our-gifting-policies-to-protect-our-community

TikTok. (2019b, December 14). TikTok for Younger Users [Press release]. https://newsroom.tiktok.com/en-us/tiktok-for-younger-users

TikTok. (2020a, April 15). TikTok introduces Family Pairing [Press release]. https://newsroom.tiktok.com/en-us/tiktok-introduces-family-pairing

TikTok. (2020b, July 15). Response from TikTok to the European Commission Communication on Covid-19 Disinformation. https://ec.europa.eu/newsroom/dae/redirection/document/69157

TikTok. (2020c, August 5). Combating misinformation and election interference on TikTok [Press release]. https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok

TikTok. (2021, January 31). Strengthening privacy and safety for youth on TikTok [Press release]. https://newsroom.tiktok.com/en-us/strengthening-privacy-and-safety-for-youth

TikTok. (2022, October 17). Enhancing the LIVE community experience with new features, updates, and policies [Press release]. https://newsroom.tiktok.com/en-us/enhancing-the-live-community-experience

TikTok. (2023a, February 17). Investing for our 150m strong community in Europe [Press release]. https://newsroom.tiktok.com/en-eu/investing-for-our-150-m-strong-community-in-europe

TikTok. (2023b, March 1). New features for teens and families on TikTok [Press release]. https://newsroom.tiktok.com/en-eu/new-features-for-teens-and-families-on-tiktok

TikTok. (2023c, June 27). Updating Family Pairing and establishing TikTok's Youth Council [Press release]. https://newsroom.tiktok.com/en-eu/tiktok-family-paring-and-youth-council

TikTok. (2023d, August 4). An update on fulfilling our commitments under the Digital Services Act [Press release]. https://newsroom.tiktok.com/en-eu/fulfilling-commitments-dsa-update

TikTok. (2018, August 28). Community Guidelines. Wayback Machine. https://web.archive.org/web/20180831022137/http://support.tiktok.com/?ht_kb=community-policy

TikTok. (2020, January 8). Community Guidelines. Wayback Machine. https://web.archive.org/web/20200713031545/https://www.tiktok.com/community-guidelines?lang=en

TikTok. (2020, August). Community Guidelines. Wayback Machine. https://web.archive.org/web/20201203004513/https://www.tiktok.com/community-guidelines?lang=en

TikTok. (2020, December). Community Guidelines. Wayback Machine. https://web.archive.org/web/20220208013210/https://www.tiktok.com/community-guidelines?lang=en

TikTok. (2022, February). Community Guidelines. Wayback Machine. https://web.archive.org/web/20220308060657/https://www.tiktok.com/community-guidelines?lang=en

TikTok. (2022, April). Community Guidelines. Wayback Machine. https://www.tiktok.com/community-guidelines/en/mental-behavioral-health/?cgversion=2023

The Guardian. (2023a, April 4). 'TikTok fined £12.7m for illegally processing children's data'. *The Guardian*, https://www.theguardian.com/technology/2023/apr/04/tiktok-fined-uk-data-protection-law-breaches

The Guardian. (2023b, August 14). 'TikTok to be fined for breaching children's privacy in the EU'. *The Guardian*, https://www.theguardian.com/technology/2023/aug/04/tiktok-to-be-fined-for-breaching-childrens-privacy-in-eu

The New York Times. (2022, March 12). 'Shaming Apple and Texting Musk, a Ukraine minister uses novel war tactics'. *The New York Times*, https://www.nytimes.com/2022/03/12/technology/ukraine-minister-war-digital.html

Twitter (X). (2022, May 19). Introducing our crisis misinformation policy [Press release]. https://blog.twitter.com/en_us/topics/company/2022/introducing-our-crisis-misinformation-policy

Udupa, S., Gagliardone, I., & Hervik, P. (2012). *Digital hate: The global conjuncture of extreme speech*. Indiana University Press.

van Dijck, J., Nieborg, D., & Poell, T. (2019). 'Reframing platform power', *Internet Policy Review*, 8 (2), https://doi.org/10.14763/2019.2.1414.

van Dijck, J., & Poell, T. (2013). 'Understanding social media logic', *Media and Communication*, 1 (1), 2-14. https://ssrn.com/abstract=2309065

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

Wagner, K., Deutch, J., & Zuidijk, D. (2022, March 4). 'Putin propaganda machine undercut by social media blackout', Bloomberg. https://www.bloomberg.com/news/articles/2022-03-04/putin-propaganda-machine-undercut-by-social-media-blackout#xj4y7vzkg

Wardle, C., & Derakhshan, H. (2017, September 27). Information disorder: Toward an interdisciplinary framework for research and policy making (DGI(2017)09). The Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

World Health Organization. (2020, September 23). Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation: Joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC. World Health Organization. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

Xu, Y. (2023). '*Power, values and accountability of digital platforms: An in-depth look into TikTok's official discourses'* [Working paper in preparation]. Faculty of Social Sciences, University of Helsinki.

Yadav, K. (2021, January 25). 'Platform interventions: How social media counters influence Operations', Carnegie Endowment. https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698

Zuboff, S. (2019). 'Surveillance capitalism and the challenge of collective action', *New Labor Forum*, 28 (1), 10-29. https://doi.org/10.1177/1095796018819461