European Commission

Digital Services Act: Application of the Risk Management Framework to Russian disinformation campaigns

Internal identification

Contract number: PN/2022/022

EUROPEAN COMMISSION

Directorate-General for Communications Networks, Content and Technology
Directorate F — Platforms Policy and Enforcement
Unit F.2 — Digital Services

Contact: CNECT-DIGITAL-SERVICES@ec.europa.eu

European Commission
B-1049 Brussels

EUROPEAN COMMISSION

Digital Services Act: Application of the Risk Management Framework to Russian disinformation campaigns

Directorate-General for Communications Networks, Content and Technology

2023 EN

EUROPE DIRECT is a service to help you find answers to your questions about the European Union

Freephone number (*):
00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you)

LEGAL NOTICE

PDF	ISBN 978-92-68-04968-6	doi:10.2759/764631	KK-09-23-294-EN-N

Manuscript completed in August 2023

1st edition

The European Commission is not liable for any consequence stemming from the reuse of this publication.

Luxembourg: Publications Office of the European Union, 2023

TABLE OF CONTENTS

1. Executive Summary

During the first year of Russia's illegal war in Ukraine, social media companies enabled the Kremlin to run a large-scale disinformation campaign targeting the European Union and its allies, reaching an aggregate audience of at least 165 million and generating at least 16 billion views. Preliminary analysis suggests that the reach and influence of Kremlin-backed accounts has grown further in the first half of 2023, driven in particular by the dismantling of Twitter's safety standards.

The largest social media platforms made commitments to mitigate the reach and influence of Kremlin-sponsored disinformation. Overall, these efforts were unsuccessful. Over the course of 2022, the audience and reach of Kremlin-aligned social media accounts increased substantially all over Europe. These circumstances raise questions not only about European Union defences against Russia's information warfare but also about the integrity of the European election in June of 2024.

In the meantime, Europe has established new policy and law to address these vulnerabilities. In response to European Commission guidance, most major platforms signed a new Code of Practice on Disinformation in June 2022.[1] Shortly thereafter, the EU passed the Digital Services Act (DSA) – a new landmark regulation of online platforms that enters into force in 2023.

This study evaluates how the DSA's rules can be used to guard against the Kremlin's disinformation campaigns and protect the dignity, safety and free expression of EU citizens. We evaluated Kremlin disinformation campaigns across all major platforms in more than 10 European languages over a period of almost a year. These data sets were then analysed using the compliance framework contained in Articles 34 and 35 of the DSA that require risk assessment and mitigation.

The conclusions are clear. We find that the Kremlin's ongoing disinformation campaign not only forms an integral part of Russia's military agenda, but also causes risks to public security, fundamental rights and electoral processes inside the European Union. Moreover, we observe that disinformation is only one weapon in the Kremlin's information warfare arsenal. The Kremlin's operations on online platforms often build on other inflammatory or deceptive content, and a range of malign behaviours designed to silence opponents and suppress the truth about the war in Ukraine.

These risks were mitigated intermittently by the platforms in particular aspects of Russian disinformation about the war. But their efforts did not effectively impede the growth and influence of Kremlin information warfare generally. Effective mitigation was not yet required by law under the DSA during the period of study in 2022. However, most of the platforms were signatories to the Code of Practice as of June 2022.

Under the Code, online platforms committed to a broad set of measures that could have mitigated some of the Kremlin's malign activities. However, the evidence suggests that online

[1] Guidance on Strengthening the Code of Practice on Disinformation, 26 May 2021, European Commission, https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation.

platforms failed to implement these measures at a systemic level. Moreover, the Code is not designed to mitigate a full scale, state-sponsored information war propagated by thousands of accounts engaged in coordinated tactics. Consequently, in many cases the mitigation measures introduced by online platforms failed to account for the Kremlin's malign intent and full scope of information warfare tactics employed on online platforms. For instance, no platform introduced policies addressing *all* or even most Kremlin-operated accounts. In addition, platforms fundamentally ignored cross-platform coordinated campaigns.

As a result, the Russian Federation continues to operate vast networks of social media accounts propagating deceptive, dehumanising and violent content and engaging in coordinated inauthentic behaviour. Indeed, we find that the reach of Kremlin-sponsored disinformation inside the EU has grown since February 2022. In absolute numbers, pro-Kremlin accounts continue to reach the largest audiences on Meta's platforms. Meanwhile, the audience size for Kremlin-backed accounts more than tripled on Telegram. In addition, we found that no platform consistently applied its terms of service in repeated tests of user notification systems in several Central and Eastern European languages.

The rules provided by the DSA hold great potential to reign in Kremlin disinformation campaigns and other state-sponsored attacks on the democratic integrity and fundamental rights. But they must be applied quickly and effectively in order to help mitigate these coordinated attacks on European democracy.

2. Introduction

On 24 February 2022, Russia attacked all of Ukraine, eight years after Russian troops entered Crimea and Ukraine's Donbas regions. Russia's military strategy has since not only resulted in harrowing violence in Ukraine—it also extended to online spaces, enabling acts of information warfare far beyond Ukraine's borders. Kremlin operatives have deliberately manipulated the features of social media platforms to spread disinformation and influence public opinion.

Both inside and outside Russia, the Kremlin's disinformation strategy followed two tactical objectives: suppressing the truth about the war and amplifying lies about an alleged "special operation" to free Ukraine from "Nazism". Inside Russia, the Kremlin moved swiftly to block social media platforms such as Facebook or Twitter and to tighten media censorship in order to cut Russians off from images of the horror their country was inflicting on Ukrainians. At the same time, the Kremlin leveraged its ecosystem of state-controlled media to flood the remaining platforms in Russia with lies and self-serving conspiracies.

Outside Russia, the Kremlin's disinformation strategy followed the same objectives, but it was more subtle. Of course, the Kremlin could not censor the free media of other countries, or block Facebook across the continent to isolate Europeans from the truth. Instead, the Kremlin and its proxies captured growing audiences with highly produced propaganda content, and steered users to unregulated online spaces, where democratic norms have eroded and hate and lies could be spread with impunity.

This is an old playbook: The Kremlin has attempted to manipulate foreign communication systems and public opinion long before the rise of Facebook and Google. The so-called information warfare doctrine goes back to early Soviet times – it builds on "reflexive control."[2] The idea is to shape how adversaries think about an issue, while concealing the activities of manipulation so that the targets remain unaware. Since the 1950s, the Soviet security agency (KGB) hosted a department dedicated to spreading disinformation in other countries, including antisemitic, racist narratives designed to deepen socio-political divides.[3]

However, as this study shows, online platforms have supercharged the Kremlin's ability to wage information war, and thereby caused new risks for public safety, fundamental rights and civic discourse in the European Union. This effort is not limited to shaping opinion about Russian aggression in Ukraine. It is designed to foment political and social instability among its adversaries by stoking ethnic conflict, promoting isolationism, and distracting public attention away from Ukraine and onto domestic affairs. For example, as recently as April 2023, the Washington Post reported on leaked documents detailing the Kremlin's strategy to instigate a new extremist political movement in Germany by tailoring and targeting disinformation campaigns on online platforms.[4]

When the invasion began, Reset organised a rapid response effort of analysts and civil society organisations from across Europe to monitor Kremlin disinformation and the efforts by online platforms to reduce its reach and influence. Meanwhile in Brussels, the European Union was negotiating and finalising the most comprehensive regulatory framework for digital services in the world: the Digital Services Act (DSA). We thus applied the logic underpinning the DSA – that operators of Very Large Online Platforms (VLOPs) must assess and mitigate systemic risks caused by their products – to the Kremlin's disinformation campaign surrounding the war against Ukraine. Our ambitious objective was to create a body of systematic evidence and to explore metrics and methods applied in a case study that could help regulators enforce the new law vis-a-vis online platforms.

Notably, in June 2022 all major platforms except Telegram signed a strengthened Code of Practice on Disinformation based on the European Commission's guidance.[5] In theory, the requirements of this voluntary Code were applied during the second half of 2022 – during our period of study. Companies published the results of compliance efforts in January 2023.[6] This Code includes some commitments analogous to the mitigation requirements codified in Article 35 of the DSA and thus has clear relevance to this analysis. In particular, the Code has measures that (if enforced) would effectively curtail specific high-risk content. However, it was

[2] Chotikul, Diane. 1986. "The Soviet Theory of Reflexive Control in Historical and Psychocultural Perspective: A Preliminary Study." National Security Archive. Accessed 2 June 2023. https://nsarchive.gwu.edu/document/15364-diane-chotikul-soviet-theory-reflexive.

[3] Wilde, Gavin, and Justin Sherman. 2023. "No water's Edge: Russia's information war and Regime Security." Carnegie Endowment for International Peace. Accessed 2 June 2023. https://carnegieendowment.org/2023/01/04/no-water-s-edge-russia-s-information-war-and-regime-security-pub-88644.

[4] Belton, Catherine, Souad Mekhennet, and Shane Harris. 2023. "Kremlin triest to build antiwar coalition in Germany, documents show." Washington Post. Accessed 2 June 2023. https://www.washingtonpost.com/world/2023/04/21/germany-russia-interference-afd-wagenknecht/.

[5] Strengthened Code of Practice on Disinformation, 16 June 2022, European Commission, https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation.

[6] See, Transparency Centre, https://disinfocode.eu/.

not designed to address a systemic information warfare perpetrated by state-backed actors across platforms that includes tactics far beyond the spread of disinformation.

Mindful of the important role the Code might play in shaping some company practice, we chose to look at the broader system of risk assessment and mitigation compliance contained in the DSA. In this way the Code's provisions addressing disinformation are streamlined into the larger regulatory regime. We thus documented the different ways in which pro-Kremlin actors exploited online platforms (including but not limited to disinformation), and created a modular framework for structuring replicable, large-scale data investigations to measure the effectiveness of the companies' mitigation measures. This report is intended to demonstrate how the DSA's compliance regime of risk assessment and mitigation may be applied using a standardised model – both in the abstract and through a specific, data-rich case study. We do not intend to provide a comprehensive enumeration of all potential aspects of risk assessment and mitigation. That is not possible with existing publicly available data sources. Rather, we intend to offer a Baseline Framework of analysis and a set of prototype case analyses upon which the European Commission, the Member State network of Digital Services Coordinators, online platforms and the research community of civil society and academic actors may build over time.

The report begins with an overview of the framework's analytical model. It has a two part structure – Risk Assessment and Risk Mitigation – that follows the logic of the DSA. The Regulation establishes the categorical objectives of public protection through mitigating risks to fundamental rights, public safety, electoral process and more. Within this, we have attempted to establish standards of measurement that are rooted in human rights law and combine a set of qualitative contextual analysis and quantitative metrics to assess the real-world risks posed by online platforms and the effectiveness of the mitigation measures taken by companies to comply with the law. Where necessary, we indicate the absence of available data that would be required to conduct a full analysis and highlight the ability of the regulator to compel this data. In recognition of the critical importance of replicability across different Member States, languages, and platforms, we have attempted to distil the model into the simplest form possible without losing analytical insight.

The report proceeds to apply this abstract model to the Kremlin's disinformation campaign across online platforms in all of its facets. Through systematic qualitative analysis, we found that the behaviours of Kremlin-backed accounts on online platforms, and the content they were disseminating, caused risks that are within the definitional scope of the DSA, including risks to public safety, fundamental rights and civic discourse, as well as an increased risk of illegal content disseminating on digital services. Through quantitative analysis, we subsequently assessed whether these risks caused by Kremlin disinformation qualified as systemic, and whether they would thus warrant mitigation by online platforms. Lastly, we assessed how effective different types of measures introduced by online platforms have been thus far in mitigating the observed risks, ranging from changes in Terms & Conditions to content moderation and algorithmic recommendation systems.

We conclude the case study with an interpretation of this data that identifies trends that persisted across the information ecosystem throughout 2022. The top lines of our findings are these:

- More than one year into the war, the Kremlin's operations on online platforms continue to cause severe risks to public safety, fundamental rights and civic discourse. Kremlin-backed accounts continue to propagate disinformation about the war and content designed to dehumanise or incite hatred against Ukrainians, women, or LGBTIQ communities. These disinformation campaigns also attempt to interfere with the ability of citizens in the EU to speak freely and receive verifiable information about the war.
- The scale of these risks is significant: Overt Kremlin-backed accounts that are not blocked inside the EU have a total subscriber number of at least 165 million across major platforms. In less than a year, their content was viewed at least 16 billion times.
- In addition to social media accounts under the Kremlin's direct control, a growing ecosystem of Kremlin-aligned accounts propagates the same type of content across the EU. The reach of these pro-Kremlin networks has more than doubled since the war began.
- In absolute numbers, pro-Kremlin accounts continue to reach the largest audiences on Meta's platforms. However, their audiences only grew marginally on Facebook and Instagram compared to other platforms. The subscriber numbers of pro-Kremlin channels more than tripled on Telegram since the start of the war, more than doubled on TikTok and rose by almost 90 percent on YouTube.
- In response to these growing risks, platforms introduced very few changes at the level of Terms and Conditions. The exceptions were restrictive policies targeting Russian state media accounts, and some narrow content policies regarding the denial of war crimes, or the publication of personal information about prisoners of war. Pre-existing policies covering incitement to violence and hate speech were applied inconsistently across platforms, languages, and time-periods.
- The narrow scope of relevant mitigation policies enabled pro-Kremlin accounts to circumvent them successfully. The Kremlin information operations continue to deceive vast audiences, to spread hate, and to incite violence. They employed a range of behaviours aimed at artificially inflating the reach and perceived popularity of Kremlin narratives, as well as to silence pro-Ukrainian voices.
- The Kremlin's disinformation strategy was tailored to the entire ecosystem of online platforms. However, the policies of each platform only accounted for their own product, ignoring cross-platform manipulation. As a result, the Kremlin was able to exploit diverging levels of content moderation by funneling audiences to the least regulated environments and by coordinating information operations across platforms with impunity.
- The evidence we were able to collect suggests that platforms' content moderation efforts in response to the war were ineffective. Platforms rarely reviewed and removed more than 50 percent of the clearly violative content we flagged in repeated tests in several Central and Eastern European languages.
- We developed a novel metric to test the impact of algorithmic recommendation systems on the distribution of Kremlin disinformation: the Non-Follower Engagement (NFE) metric. Overall, we found that measures designed to restrict the algorithmic reach of, for instance, Russian state media accounts were fairly effective in reducing engagement with manually curated sets of accounts, but did not reduce the risks of Kremlin disinformation at systemic level.

The purpose of this study is not to evaluate existing EU policy instruments, but to develop methodologies for risk assessments and metrics for risk mitigation under the DSA framework (Articles 34 and 35), and then to apply them to the Kremlin's activities on online platforms. However, it is worthwhile noting that the mitigation measures stipulated in the Strengthened Code of Practice on Disinformation (see, e.g. Commitment 14, 16, 18, 23 and many others) could have addressed several of the risks identified in this case study. For instance, the Tactics, Techniques, and Procedures listed under Commitment 14 cover many of the behaviours employed by Kremlin-backed accounts on online platforms. However, the evidence we collected suggests that online platforms intermittently applied pre-existing policies, rather than implementing their commitments under the Code at systemic level. These deficient practices are documented in the platforms' published reports in January 2023 and further weakened by the absence of Telegram and the withdrawal of Twitter from the Code. At the same time, the Code is not built to mitigate the activities of a state-backed actor operating a multi-faceted campaign encompassing numerous tactics beyond the spread of disinformation that worsen systemic risks across multiple categories.

Initial quantitative analysis suggests that compliance with the Code of Practice was insufficient also in the months after they submitted their first progress reports in January. In fact, the reach of pro-Kremlin accounts has increased between January and May of 2023, with average engagement rising by 22 percent across online platforms. However, this increased reach was largely driven by Twitter, where engagement grew by 36 percent after CEO Elon Musk decided to lift mitigation measures on Kremlin-backed accounts, arguing that "all news is to some degree propaganda." Shortly after the end of our monitoring period, Twitter withdrew from the Code of Practice. By contrast, and in apparent compliance with the Code, average engagement with pro-Kremlin accounts dropped by 20 percent on Facebook, and remained largely unchanged on the remaining platforms. As regards the posting activity of pro-Kremlin accounts, we observed a notable increase of 34 percent on TikTok, whereas activity on Telegram dropped by 22 percent after both activity and reach had been rising exponentially throughout the previous monitoring period. This sudden trend reversal may in part be a result of Telegram's belated implementation of EU sanctions against Russian state media.

In general, it seems advisable to implement the Code of Practice as a Code of Conduct under Article 35 – backed up by strong regulatory enforcement. The DSA offers unique new tools for addressing efforts by authoritarian third country governments to interfere in democratic processes inside the European Union. As this study highlights, there is a high risk that the Russian Federation in particular will continue its efforts at interfering in electoral processes in the European Union, including 2024's elections to the European Parliament. Therefore, a Code of Conduct under the DSA should be further complemented with measures tailored specifically to mitigate state-backed disinformation and information operations. This requires comprehensive investigation of all online identities, behaviours and content types employed by the Kremlin and affiliated actors, as well as criteria for assessing their risks. This study draws on the field of international human rights law to develop criteria for risk assessments, and applies them to Russia's activities on online platforms.

In sum, we found that the overall effect of platform policies on the Kremlin disinformation campaign did not significantly reduce the risks that are laid out in the DSA. In specific cases, such as the channels of official Russian government institutions and state media, the reach

and influence of disinformation were highly constrained both by suspension and algorithmic demotion. However, adjacent accounts with similar content picked up the audience and exploited amplification by platform commercialization features without impunity. Those services that attempted more comprehensive mitigation were undermined by others that did not. The absence of systemic level policy responses on any platform, the failure to address cross-platform exploitation, and the inconsistent application of those mitigation measures that were applied, have resulted in significant impact for Kremlin disinformation campaigns.

3. Baseline Framework

The primary purpose of the Baseline Framework for Digital Services Risk Management (hereafter, “Baseline Framework”) documented here is to offer a simple but viable model for researchers to apply the principles of Articles 34 and 35 of the DSA – Risk Assessment and Risk Mitigation – to any of the systemic risks referenced in the regulation. These risks are specified in Article 34 of the Act, but the method of assessing the level of risk to the public is not. The first task then is to develop replicable risk assessment methods that can document the impact of online platforms on the ability of European citizens to assert their fundamental rights, live safely and vote freely. Similarly, in Article 35, the Act directs providers of Very Large Online Platforms and Very Large Online Search Engines to apply “reasonable, proportionate and effective mitigation measures” for each of the risks identified in Article 34. Again, there is guidance as to the ways in which these policies might be applied, but the research community is yet to develop standardised methods for assessing the relative success or failure of these mitigation measures.

The Baseline Framework presented here is a structure for collaboration across the research community to collect and organise qualitative and quantitative evidence of risk on online platforms. We present a methodology that is both modular and extensible. It is suitable for narrowly tailored investigations of particular case studies. But it can also be used to address large and diverse bodies of evidence across multiple platforms and risk vectors to provide more comprehensive assessments. The central components offer:

1) a qualitative system for evaluating the severity of a particular risk factor in context;
2) a quantitative method for determining the scale and intensity of the risk factor;
3) a qualitative method for identifying whether or not a platform has designated a mitigation measure for a particular risk; and
4) a set of quantitative techniques for assessing the effectiveness of mitigation measures deployed to address these risks.

We intend this Baseline Framework as a working model for iterative improvement, collaborative adaptation, and increased utility over time as more researchers apply its logic and methods to new investigations and contribute knowledge to a growing community of practice.

a. Baseline Framework Part 1: Risk Assessment

In Article 34, the DSA prescribes the risks that platforms must assess in clear terms. It is worth quoting in full to illustrate the scope of the requirement.

> …This risk assessment shall be specific to their services and proportionate to the systemic risks, taking into consideration their severity and probability, and shall include the following systemic risks:
>
> (a) the dissemination of illegal content through their services;
>
> (b) any actual or foreseeable negative effects for the exercise of fundamental rights, in particular the fundamental rights to human dignity enshrined in Article 1 of the Charter, to respect for private and family life enshrined in Article 7 of the Charter, to the protection of personal data enshrined in Article 8 of the Charter, to freedom of expression and information, including the freedom and pluralism of the media, enshrined in Article 11 of the Charter, to non-discrimination enshrined in Article 21 of the Charter, to respect for the rights of the child enshrined in Article 24 of the Charter and to a high-level of consumer protection enshrined in Article 38 of the Charter;
>
> (c) any actual or foreseeable negative effects on civic discourse and electoral processes, and public security;
>
> (d) any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being.
>
> 2. When conducting risk assessments, providers of very large online platforms and of very large online search engines shall take into account, in particular, whether and how the following factors influence any of the systemic risks referred to in paragraph 1:
>
> (a) the design of their recommender systems and any other relevant algorithmic system;
>
> (b) their content moderation systems;
>
> (c) the applicable terms and conditions and their enforcement;
>
> (d) systems for selecting and presenting advertisements;
>
> (e) data related practices of the provider.
>
> The assessments shall also analyse whether and how the risks pursuant to paragraph 1 are influenced by intentional manipulation of their service, including by inauthentic use or automated exploitation of the service, as well as the amplification and potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions.
>
> The assessment shall take into account specific regional or linguistic aspects, including when specific to a Member State.

The Regulation gives further guidance in the Recitals. Recital 79 describes how platforms might measure a threshold of risk above which it must be mitigated. The factors should include

the severity and probability of potential impact, the reach of the risk (e.g. size of audience affected), whether the negative impact can be reversed and the level of difficulty of applying a remedy. In Recital 80, the Act provides guidance on the types of illegal content that must be addressed under Article 34(1)(a).

The central requirement of Article 34 is to observe content and behaviour on these services that implicates the different categories of risk and submit an assessment of those where the level of severity rises to the level of "systemic risk." For the analyst, this instruction requires answering this question: At what point does content or conduct on a service cross the threshold, resulting in an "actual or foreseeable negative effect" that is severe enough to be considered systemic in relation to the risk factors specified in the regulation?

Determining a systemic level of severity requires the application of a proportionality test to measure the threshold of risk as a function of both qualitative and quantitative indicators. As most VLOPs host user generated content without preemptive review, there will always be content that constitutes potential risk to one or more categories in Article 34. However, because of the scale and the wide variation in the audience size and intensity of exposure to any given piece of content, it does not follow that a single instance of potentially harmful content constitutes systemic risk. *Severity is a function of the relationship between the qualitative assessment of the risk posed by the content in context and a quantitative measure of the reach and/or intensity of exposure of audiences to that content.*

It follows then that a risk may reach a systemic level in different ways. The higher the level of risk inherent in the content in context, the smaller the audience required to reach a systemic level. And by contrast, the lower the level of risk inherent in the content in context, the larger the audience required to reach a systemic level.

The features and Terms and Conditions of the product provided by the platform (i.e. Article 34(2)(a-e)) may further influence these measures of systemic risk. These factors are closely related to the mitigation measures VLOPs and VLOSEs may apply, subject to Article 35 requirements, to address the risks assessed pursuant to Article 34. Content moderation standards, changes to recommender algorithms, or modifications to data-driven audience curation could therefore lower the risk below the systemic threshold year over year. However, these features must be shown empirically to reduce risk in order to alter subsequent risk assessments. Other factors, such as the intentional manipulation of the platform service (e.g. through the purchase of fake engagements, the coordination of fake accounts, impersonation, or the automated production of content) may exacerbate the quantitative scale of risk regardless of their prohibition by Terms and Conditions. These product features and contextual factors should be considered in the conduct of the qualitative and quantitative assessment process.

i. Assessment Process

Measuring risk levels against the severity threshold for systemic risk requires a two step process for any given instance of content or behaviour observed on the platform.

Step 1 – Qualitative Evaluation: Following the logic described above, our operational model for risk assessment begins with a qualitative assessment of content or behaviour observed in context on the VLOP based on the risk categories in Article 34.

In order to assess the qualitative severity of risk posed by a particular type of content, we apply a proportionality test based upon a modified version of the Rabat Plan of Action published by the UN Office of the High Commissioner for Human Rights in 2012. The original Plan provides guidance rooted in human rights law on the appropriate balance between restrictions on incitement to violence/hate and the principles of freedom of expression.[7] The modified version we have adopted here takes the spirit and form of its proportionality test and customises it for the purpose of the Article 34 risk assessment – while the Rabat Plan was originally designed to assess the proportionality of legal restrictions on speech, we find that it equally serves to evaluate less severe interventions, such as reductions of the algorithmic amplification of specific types of content.

This method of protecting the public from harm while simultaneously protecting freedom of expression calls for evaluation of the probability of real world harms. It is a model that incorporates both qualitative and quantitative assessment that can be used across all of the risk categories defined in Article 34 to determine severity. The more of these factors that indicate the potential for harm, the more likely the content may be judged to carry systemic risk. We suggest a 5-point formulation of analytical reference points that tracks a similar concept in the Rabat Plan of Action:

1. Context of the Statement
2. Speaker's Position or Status or Intent
3. Content and Form of the Statement
4. Reach, Size, Characteristics of the Audience
5. Likelihood or Imminence of Harm

By way of example, when we apply this proportionality test to Russian state media accounts on online platforms, we can conclude that war propaganda published by these outlets exceeds the systemic risk threshold of severity. The *context* of the speech – an ongoing war of aggression – makes clearer the probability that it will incite violence and provides insight into how the speech will be understood by the audience. The *position and status* of the speaker – official channels of state propaganda – provides important context about their *intent*. The evaluation of the *content and form* of the speech demonstrates highly produced audio, video, images and text that threatens or incites violence, engages in dehumanising provocation, and spreads intentional disinformation. These state media accounts *reach* large audiences that are targeted at particular languages and geographies likely to be impacted by the war. The *likelihood* of harm is a function of the cumulative evidence assessed across the other elements of the test – wartime propaganda containing incitement to violence broadcast through professional channels that reach large audiences.

This mode of qualitative analysis cannot achieve an absolute measurement of risk or harm. However, over time and with replication, it should produce a consistent body of cases that may be measured relatively against one another to determine if the same conclusion of severity that reaches systemic levels of risk is justifiable based on past precedent. It may also yield

[7] United Nations, Office of the United Nations High Commissioner for Human Rights. 2013. "The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence." A/HRC/22/17/Add.4. Accessed on 2 June 2023. https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action.

heuristics or benchmarks for assessing qualitative levels of harm and intensity of risk – drawing upon the adjacent field of human rights law and practice. This is an area that would benefit from additional deliberation among stakeholders.

Step 2 – Quantitative Evaluation: Once a particular type of content has been assessed qualitatively to have the potential for systemic risk, the process can move to quantitative measures. The second part of the evaluation seeks to measure – as precisely as possible with available data – the size of audience, the prevalence of exposure and engagement among audience segments, and the influence of algorithmic promotion by the platform or through the exploitation of recommender systems.

This part of the model features standardised methods that may be applied consistently with comparable results both within and across platforms. The primary challenge is selecting and isolating the appropriate sample of data to evaluate. This can be keyed on *actors* posting content that carry the potential of systemic risk, *behaviours* (by actors or platforms) that result in increased distribution or promotion of content carrying the potential for systemic risk, and/or examples of *content* carrying potential of systemic risk.

We recommend keeping this exercise simple and applying the well-tested research heuristic of ABC for structuring evidence for quantitative assessment – starting with a qualitative analysis of Actors, Behaviour, Content.[8] This approach maps onto the modified Rabat Framework, but it requires adding Distribution as a cross-cutting, quantitative variable to determine the scale of impact and the propagation characteristics of specific risks. Ideally, it also requires adding Effect to gauge real or potential negative impact. In assessing impact, we limit our analysis to on-platform effects as measured by our metrics. A comprehensive assessment of off-platform effects is beyond the scope of this paper. These variables are always interactive. But for purposes of comparative evaluation of risk across platforms or relative evaluation of risk over time on the same platform, it is necessary to categorise and quantify data such that each aspect can be assessed. This permits, for example, the evaluation of Russian state media accounts (Actors), the application of methods to circumvent VLOP Terms and Conditions (Behaviours), the posts reflecting incitement to violence or threat to security (Content), and the amplification of that content through algorithmic recommendation (Distribution), as well as negative impact on or off platform (Effect). In the Appendix of this study, we provide a taxonomy of how these categories can be broken down into subcategories of variables for analysis.

This method has an important virtue for future research – modularity. It can be broken down into smaller pieces of research that can be conducted in consistent ways by different kinds of researchers. And it can be done with validity at small or large scale – with a recognition, of course, that the smaller the scale, the higher the level of qualitative severity that will be necessary to make a finding of systemic risk. This modular approach will allow an entire community of academic and civil society researchers to participate effectively in support of the platforms' self-assessments and the regulatory oversight conducted across the EU.

[8] Francois, Camille. 2019. "Actors, Behaviors, Content: A Disinformation ABC." Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, Graphika and Berkman Klein Center for Internet & Society at Harvard University, 20 September. https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf, accessed 2 June 2023.

The simplest way to begin the quantitative analysis is to isolate a particular group of accounts (Actors) to document how they try to amplify or suppress content (Behaviour), what kinds of text, image, audio or videos they post (Content) and measure their reach and engagement (Distribution). This account selection can be done in different ways: 1) representative sample; 2) curated set of accounts with similar features (e.g. state media, accounts with a minimum level of reach, accounts that self-identify as journalists or government, etc.); 3) accounts that engage with one another consistently to form a network; or 4) accounts posting similar content or that share specific, designated features or keywords associated with a particular Article 34 risk category. For any given set of accounts, the analysis will also select parameters such as platform(s) for analysis, language, geo-coding (i.e. from where the content is accessed), and time period. In order to perform analyses that yield meaningful results, samples must be selected based on reasoned criteria and variables isolated for measurement and comparison.

The next step is data collection. The availability of data will vary widely prior to the enforcement of Article 40 data access provisions of the Act. Platforms themselves, of course, may access any and all data in their possession. They are only limited in so far as they seek to conduct comparison with other platforms outside of common ownership. Regulators may require whatever data they deem necessary in order to apply the provisions of the regulation – including comparing similar data across more than one platform. Researchers in academia and civil society are the most limited. In some cases, the data is publicly available via an API. In other cases, the data must be collected in ways that limit scale of analysis. In all cases, care must be taken to ensure compliance with relevant data protection and data security requirements. In the future, new regulation is likely to significantly expand data access. In the meantime, targeted analyses from independent researchers may serve as catalysts for regulators to demand larger data sets from VLOPs to evaluate findings more holistically.

ii. Assessment Metrics

For any given modular analysis, different metrics may be best depending upon the variable that is isolated and the research question. These are the metrics that we recommend considering for quantitative risk assessment. This list will grow over time with further research that builds upon this Baseline Framework.

Any list of metrics used to assess platform performance is vulnerable to being gamed. There may even be perverse incentives for platforms to weaken performance in one area, such as proactively detecting fake accounts, in order to perform well in other areas, such as the ability to remove fake accounts. However, there is scope for extending the list of metrics, for instance by using experimental approaches to test the platforms' ability to proactively prevent fake account creation. Additionally, the DSA audit mechanism may be designed in ways that disincentivise gaming of the system.

Audience size — the number of followers, fans or subscribers attached to a specific VLOP account or channel. This is a publicly accessible metric for all VLOPs. This metric may be used to indicate the relative significance of an account on a VLOP (*total audience size*) as well as its popularity (*growth rate*) over time. As a risk metric, it may be a misleading

indicator, as it can be gamed, for instance through the purchase of fake followers (a practice which can also be tested). It is a largely static metric as most subscribers are long-term subscribers unlikely to unsubscribe even when an account goes dormant. It also counts only the users who have explicitly opted to receive the content from a particular account or channel.

Exposure — the estimated number of times a particular piece of content has been seen through VLOP distribution. Depending on the platform, exposure can be measured by the total number of post views (for platforms such as Telegram, Twitter, VKontakte), the total number of video views (for all video content on YouTube and TikTok; for video posts only on Facebook and VKontakte). This metric can also be analysed together with Audience and Engagement levels to inform an assessment of artificial amplification.

Engagement — measured as the absolute number of likes, comments, or shares received for a piece of content (post, video, etc) or the average number of likes, comments or shares received by all posts published by an analysed account. Engagement is a primary risk metric as it counts the number of intentional interactions. It goes beyond mere exposure to log the number of users who chose to interact with the content.

Prevalence – the number of exposures to a particular piece of content as a proportion of the total amount of content. The prevalence metric can be taken at the platform level or within specific audience segments.

Amplification factor — measures an account's average engagement for a particular type of content compared to the account's baseline engagement. This metric demonstrates the virality level of a particular type of post, content or account.

Volume — measures the posting activity of a specific VLOP account or channel. Estimated by the absolute number of posts or by posting frequency over time (percentage increase of posting activity). This can be an indicator of increasing risk if the posting activity rate is growing over time or if the posting activity rate indicates the involvement of automation.

Re-channeling — measures the volume of referral traffic towards less regulated platforms such as Vkontakte, Odnoklassniki or Telegram to drive audiences outside of platforms with higher standards of risk mitigation and oversight. Measured as percentage of the total links shared by a social media account.

Toxicity — measured as the proportion of toxic posts or comments in a given sample. As a metric, toxicity represents the level of unmoderated harmful interactions on a platform. We operationalise toxic content as material with a toxicity probability score above 0.8, as estimated by the Perspective API, which defines toxic content as material likely to give offence.

This Risk Assessment framework with its combination of qualitative and quantitative methods maps directly onto the regulatory instruction in Article 34 and may contribute to enhanced collaboration of regulators, VLOPs, and civil society researchers alike. Over time, we expect

benchmarks for the "severity threshold" will develop such that the qualitative assessment using the Rabat Action Plan model coupled with quantitative metrics may produce standardizable findings that can shape the behaviour of VLOPs and the expectations of their users.

In the meantime, there will be a process of iteration – a constant process of replication and learning that will feed back into the entire regulatory system and contribute to outcomes that serve the public. The process of iteration should not deter timely and persistent attempts to address systemic risks immediately. This is especially urgent given the upcoming EU elections in June of 2024. We recommend the community of stakeholders begin immediately to map the landscape of potential threat vectors. This can be done on a modular basis that samples particular data sets, isolates relevant variables, and applies appropriate assessment metrics. In the aggregate across the research community, this will contribute to a prototyping process that rapidly informs standards for a community of practice.

b. Baseline Framework Part 2: Mitigation Analysis

Under Article 35 of the DSA, content or behaviour on the VLOP or VLOSE that amounts to a systemic risk must be addressed with mitigation measures. These must be *"reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to Article 34, with particular consideration to the impacts of such measures on fundamental rights."* Article 35 names eleven separate measures or processes that may be relevant to this task. The next step in our framework, therefore, is to measure the effectiveness of mitigation measures on the systemic risks that have been identified in the previous section.

We model our Mitigation Analysis methods principally on three of the measures (Article 35(1)(b-d)) under which almost all of the others may be categorised. Notably, these are also referenced as factors that play back into risk assessment (Article 34(2)(a-c): **Terms and Conditions** (also called policies or standards), **Content Moderation Processes** (implementation of the policies to suspend accounts and to remove or label content), and **Algorithmic Recommender Systems** (implementation of policies and standards related to ranking, searchability, demotion and promotion).

i. Assessment Process

The method is straightforward. First, we evaluate the extent to which a platform *has* a policy, either stated explicitly in its Terms and Conditions or public statements that address a specific systemic risk identified in the risk assessment. Our evaluation at this layer considers both individual platforms as well as the ecosystem of platforms involved in the risk assessment. This is relevant insofar as platforms without policies to address a particular risk may be used to circumvent platforms that do. If a platform does not have a policy that is sufficient to cover a particular system risk, we judge that it cannot be compliant with the Article 35 standard of "reasonable, proportionate, and effective mitigation measures."

Within this first step, we ask two additional types of questions for evaluation that inform the effectiveness of the policy under the Article 35 standard.

- *Preparedness: Do adequate policies exist prior to the emergence of risks? Is there evidence that (in)sufficient resources were deployed for risk mitigation for a reasonable expectation of likely risk?*
- *Transparency: Are policies and processes published publicly (or submitted to regulators in line with Articles 34 and 42)? Is information available about how and whether mitigation measures are applied, to what accounts, and with what effect?*

In many cases, the answers to these questions may not be public and must be inferred from comparative analysis of risk mitigation over time. In any event, transparency is a binary measure that may always be factored into the mitigation analysis.

It follows, then, that if a platform does have a published policy, the second step in the mitigation analysis is to evaluate whether it is reasonable, proportionate and effective as a mitigation measure. Here, we are looking at a few key questions: Are policies *responsive* to the emergence of risks? How long does it take for policies to be enforced? Was the level of resource deployed in response commensurate with the emergent risk factor? The answers to these questions can be inferred from quantitative analysis of the performance of the mitigation measure.

Each mitigation measure for any aspect of the systemic risk categories in Article 34 can be quantitatively evaluated. For platforms to have fulfilled their obligations under Article 35 and mitigated systemic risks, their efforts must be sufficiently reasonable, proportionate, and effective. To determine where that threshold lies in each particular case, we refer to the guidance given by Recital 79. In particular, we consider the severity, probability, and prevalence of the systemic risk in question (gleaned from the Risk Assessment analysis), and whether the negative impact of the risk in question can be reversed or how difficult it is to apply a remedy.

For the same reasons as described in the Risk Assessment analysis, the measure of effectiveness and proportionality that applies to mitigation is tied to the inverse relationship between severity and exposure. This means that where there is a very high probability for a systemic risk to lead to negative consequences, the number of individuals potentially impacted need not be high to require strong mitigation measures for the risk to be reasonably, proportionately, and effectively mitigated. Similarly, where a risk is neither likely nor severe and easy to remedy, fewer or less intense content moderation measures are required to fulfil the “reasonable, proportionate, and effective” threshold of Article 35. This scale of severity is already implicitly represented in most platform Terms and Conditions. The highest risk content – such as threats of violence, explicit racism, child predation, scams, etc. – is prohibited at any level of distribution and removed when it is detected. Content with more moderate levels of risk is demoted in curation algorithms, labelled, or removed from search functionality.

It is here important to clarify that exposure (total number of views of a particular piece of content) and prevalence (number of pieces of one type of content as a proportion of the total amount of content) are consistently used as measures of both risk and mitigation by platforms and independent researchers alike. However, they can be misleading. High-risk content is not evenly distributed on a platform. It is concentrated in particular audience segments. Therefore exposure and prevalence must be considered not principally in the aggregate, but more

importantly in terms of the actual effects on the audiences with the highest levels of exposure and for whom the risk is most severe. To address this issue, it is advisable to conduct mitigation analyses on specific audiences chosen due to their likelihood of exposure to the high risk content.

Finally, it is important to note that mitigation metrics can go too far. Platforms are required under both Article 34 and 35 to assess and mitigate risk "with particular consideration to the impacts of such measures on fundamental rights." As such, it is important to evaluate the "false positive" rate of mitigation measures as well as the effectiveness of its regimes of rapid redress for instances of error in either content moderation or algorithmic recommendation policies.

Using a set of **mitigation metrics** listed below, we can structure replicable investigations to measure the effectiveness of platforms' risk mitigation systems. These metrics work in parallel to the risk metrics. In our case study below, we demonstrate how these metrics work together. The risks posed by Kremlin disinformation are inversely related to the effectiveness of platforms' countermeasures. Any measurement of one is simultaneously a measurement of the other.

ii. Mitigation Metrics

Speed and consistency of removal — In their policies, platforms spell out what is prohibited on their platform and may result in removal of posts or a suspension/ban on violating accounts. This basket of metrics covers speed and consistency of removal in different circumstances:

- The platforms *proactively identify and remove* vast quantities of violative material; here the relevant metrics are the time between posting and removal, and the exposure and engagement with the content before removal.
- Content moderation seeks to enforce platform policies, without unduly restricting freedom of speech. *False positive and false negative* rates will help determine that the balance has been appropriately struck.
- Users and regulators will expect *linguistic equity*—that platforms provide the same level of service (e.g. content moderation) across languages and countries.

Deamplification — Platform mitigation measures that seek to reduce risk without removing actors or content from the service entirely use features such as algorithmic down-ranking, the removal of recommendation, searchability, and/or monetization from particular accounts. This metric measures the effectiveness of these reduced visibility measures in the *percentage reduction of Exposure, Engagement, Toxicity and Algorithmic reach.*

Non-Follower Engagement — Measures how many users interacted with a specific piece of content as a result of VLOP recommendations or algorithmic sorting, rather than subscriptions. This metric also approximates the relative extent to which non-subscribers to a particular account interact with content from that account compared to its own subscribers ("algorithmic reach"). NFE is a computationally intensive metric, requiring cross-tabulating a list of engagers with a list of subscribers. It is an important metric to gain insight into whether recommendation algorithms are being gamed to enhance distribution.

Consistency of labelling — Some platforms have introduced labels to tag accounts and content, for instance as state media, or to apply fact-checks or warnings. These labels have been used as criteria for demotion and/or increasing friction between users and the potentially harmful content, e.g. in the form of a confirmation popup. This metric can be calculated as the *percentage of accounts or posts correctly labelled*, as determined by a manual review of a sample of accounts.

Responsiveness to user notifications — Most platforms have established mechanisms for users to report content that appears to violate platform policies. This metric can be calculated as a *percentage of user-flags (notice and action) that receive a confirmation, response, or action* from a sample of accounts or posts. Additional metrics include median time between flagging and response, action (e.g. labelling) or removal.

Redress of denial of service — Platforms are required to provide redress measures and to protect fundamental rights as part of their mitigation measures, for instance the freedom of expression. This metric measures *how long it takes for incorrectly banned users to regain access* to their accounts, and for stolen accounts to be returned to their rightful owners.

Restrictions on Inauthentic Behaviour — Platforms have policies forbidding the use of fake accounts to undermine platform integrity. Platforms attempt to prevent users from creating fake accounts, prevent coordinated posting by those accounts, and prevent their ability to game algorithms, for instance to fake trending topics. It is hard to assess the platforms' successes in these areas, as they report blocking huge numbers of accounts, yet researchers continue to report a problem with fake accounts and their activity. In practice, experimental research using commercial social media manipulation providers may offer the best approach for researchers external to the platforms to understand how effective platform defences are. Using such a method, one can track a sample of fake accounts and calculate the *percentage of accounts removed, and the percentage of fake activity removed.*

Restrictions on Algorithmic Exploitation – Platforms have policies that prohibit the coordination of posting and engagement to artificially boost a particular post, topic or trend. Using methods such as mapping identical, copy-pasted posts, the trending of spam hashtags, and highly disproportionate ratio of engagements to exposures, researchers can approximate the degree to which manipulation was reduced on the platform.

Denylisting URLs – Some platforms have policies that disable linking in posts to domains known to host content prohibited on the service. This metric measures the *percentage of denied URLs which were blocked.*

This approach allows researchers to identify areas where platform mitigation measures were effective and where they were not. Once again, the framework is modular by design. It is not necessary to evaluate all mitigation measures across each area of systemic risk on all platforms. Specific assessments of particular variables are equally valid. Our framework has the potential to reveal gaps in existing platform policies, enforcement processes, transparency reports and the kind of data that is available to researchers and regulators.

4. Applying the framework: Kremlin disinformation surrounding Russia's war in Ukraine

a. Context

In February 2022, in response to Russia's full-scale invasion of Ukraine, Reset began a rapid response operation to monitor potential Kremlin disinformation campaigns and document platforms' responses to them. This effort was conducted in partnership with civil society researchers in Ukraine and adjacent Central and Eastern European countries who worked within the Disinformation Situation Center (DSC). The DSC was a network of over a dozen international civil society organisations tracking Russia's information war and monitoring platform mitigation. Between March and December 2022, the coalition produced 29 analytical reports and 16 newsletters addressing specific cases of pro-Kremlin online activity in the EU. These reports and newsletters tracked the activities of pro-Kremlin social media accounts, their engagement with target audiences, and the evolution of their tactics. These reports – conducted in near real-time – trialled methods for analysing Russian information operations and identified key behaviours and content types that were common in these operations. Key findings from the DSC research are included throughout this case study alongside assessments conducted months later using the same data.

Leveraging the large and diverse evidence base of data about the actors, behaviours and content types involved in the Kremlin's disinformation campaign, we began this case study. We set out to retroactively apply our Baseline Framework to this rich evidence base in order to support the European Commission in scoping possible standards for evaluating upcoming Risk Assessments submitted by VLOPs and VLOSEs. To begin, we used a combination of qualitative and quantitative measures to conduct a risk assessment of Russian disinformation, identifying numerous areas in which particular content and behaviour related to Article 34 risk categories rose to a severity level sufficient to identify systemic risk. The risk assessment section explains how the assessment metrics were applied in the context of disinformation operations during the war, based on real evidence of risks perpetuated by Kremlin operatives and their proxies.

The mitigation analysis demonstrates how we designed a systematic investigation to monitor Kremlin-backed and Kremlin-aligned actors, and the degree to which the online platforms mitigated the risks posed by these actors. We applied our mitigation metrics and evaluated first whether platforms had a policy to address these risks and then whether or not they applied it effectively.

In an ideal world, research would begin with a structural framework that precedes data collection and analysis. But this study began as a rapid response to Russia's military assault and the accompanying information warfare operations. Consequently, we have not tried to force the data into a pristine parallel construction – where, for example, we apply a metric to each specific risk factor assessed and then apply the reverse metric in a perfectly mirrored evaluation of mitigation. Instead, we have let the data speak for itself, organised inside the structure to show how it works (and can work in future studies) without being overly prescriptive. Our framework is capable of providing a deep understanding of both the risks

posed by the Kremlin's disinformation campaigns as well as the effectiveness of platform mitigation measures, including systemic gaps and lack of enforcement across an ecosystem of online platforms. The extended case study that follows not only provides deep insight into the Kremlin disinformation campaign, it offers a demonstration of the Baseline Framework in practice, however imperfectly on this first attempt, and indicates what is possible for future replication using this model for any given topic, language, platform and time-period with sufficient preparation, resources, and capacity.

b. Data & Sources

For this study, we analysed content posted by more than 2200 accounts on Facebook, Instagram, Twitter, YouTube, TikTok, and Telegram. This list was created over almost two years of intensive monitoring research. The criteria for inclusion on this manually curated list of sources were as follows:

- **Direct links with the Russian state** – Accounts in this category explicitly state that they represent a Russian government institution or state media outlet. Our dataset for such directly affiliated accounts is representative. These accounts are defined as Kremlin-backed accounts.
- **Proximity to the Russian state** – Accounts in this category are not directly connected with the Russian state, but the individual behind the account has been associated with a Russian government institution or state media outlet, be it through current or former employment. The proximity parameter is used to designate accounts of individuals that may retain some level of independence from the Kremlin, but whose posting activity is aligned with the more explicit Kremlin disinformation campaigns. Depending on the level of direct connection with the Kremlin, accounts are either defined Kremlin-backed or Kremlin-aligned.
- **Ideological alignment with the Russian state** – Accounts in this category post content that is often similar or identical to the content posted by accounts affiliated with or in close proximity to Russian government institutions or state media. In many cases, those accounts directly share links to affiliated accounts. Other accounts parrot the Kremlin's narratives through originally produced content or by spreading Kremlin-aligned narratives to different target audiences and languages. The accounts are defined as Kremlin-aligned accounts.

From this curation process, we assess with high confidence that all accounts in our sample are **"pro-Kremlin"** – meaning either that they are controlled by the Kremlin or highly aligned with state sponsored information operations. In many cases in the analysis below, we subdivide this list into explicitly **"Kremlin-backed"** accounts and the less-overt **"Kremlin-aligned"** accounts (further defined in the Appendix). We analysed close to 7 million pieces of content posted between December 2021 and December 2022 across 11 official EU languages[9] as well as Russian. Our datasets across the sample contain a higher percentage of Facebook and Telegram users relative to other platforms – which reflects the popularity of these services in Central and Eastern Europe. We evaluated a significant sample of more than two thousand

[9] In order of the number of posts collected: German, Romanian, Hungarian, Slovakian, Czech, French, Polish, English, Spanish, Bulgarian, Swedish, Portuguese, and Italian. Additionally we included accounts belonging to RT, Sputnik, and Russian Embassies operating in Greek, Lithuanian, Macedonian, Latvian, and Norwegian.

accounts across all platforms which between them published 50,000 YouTube videos, 17,500 TikTok videos, 1.7m tweets, and 1m Facebook posts. The YouTube videos in our sample totalled more than 3 billion views, while the Facebook posts racked up at least 227 million engagements.

To supplement data gathered during specific investigations during the months of rapid response, we collected more comprehensive, historical data over a period of one week using platforms' official APIs where possible. We chose a 12-month observation period. This includes a time period of 3 months prior to the full-scale invasion of Ukraine as a reference point for comparison. The period of analysis continued for 10 months after the invasion to assess the impact of platform policy changes, sanctions and regulatory actions after the initial burst of war-related content had settled down. In the case of Facebook and Instagram, we collected the data using CrowdTangle. In our Twitter and Telegram analysis, we excluded retweets and shares to prevent inflated engagement metrics. We track interactions with original posts from monitored accounts only, not their shared content from unmonitored sources.

c. Part 1: Risk Assessment

i. Risk Definition

In the following section, we show that the Kremlin's disinformation campaign affected three categories of systemic risks identified in Article 34 of the Digital Services Act. Threaded through the analysis, we examine the content of the communications in the sample through the lens of the modified Rabat proportionality test from our Baseline Framework. As a starting point, we find that the *nature of the speaker* (government controlled or aligned) and the *context* (war of aggression) distinguish anything published by these actor accounts as potentially causing systemic risk. The UN Special Rapporteur for Freedom of Expression notes that while state-sponsored disinformation is not per se unlawful under international law, it "has a potent impact on human rights, the rule of law, democratic processes, national sovereignty and geopolitical stability because of the resources and reach of States and because of their ability to simultaneously suppress independent and critical voices in the country so that there can be no challenge to the official narratives".[10]

Below, we assess the *content* itself and its *form* in each risk category in order to further evaluate severity. Paired with this qualitative evaluation, we provide quantitative data about the *prevalence* and *reach* of this content within particular audiences to determine whether it is a systemic risk. The level of severity and potential impact are inversely proportional to the scale of reach/prevalence to judge it a systemic risk.

ii. Risk Category #1 – Article 34(1)(a): Dissemination of illegal content

The Kremlin's disinformation campaign accompanying the war in Ukraine significantly increases the risk of illegal content disseminating on online platforms. In general, an act of

[10] United Nations, General Assembly. 2022. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression." A/77/288. Accessed on 2 June 2023. https://www.ohchr.org/en/documents/thematic-reports/a77288-disinformation-and-freedom-opinion-and-expression-during-armed.

violent aggression of this magnitude will inevitably also find manifestation in increased online incitement to violence. More specifically, our analysis suggests that Kremlin disinformation both directly and indirectly fuels the dissemination of illegal content.

In our taxonomy of threats in the Appendix, we have clustered types of harmful content that carry risk and rise to the level of illegal speech. We also include specific sub-categories – incitement to violence and other violent content, the glorification of war and war crimes, and discriminatory and dehumanising content. Examples of this kind of material – which could qualify as illegal under many EU Member State jurisdictions – are widespread in the sample of pro-Kremlin accounts we studied.

For instance, Member States such as Czech Republic, Estonia and Germany banned the public display of the 'Z' war propaganda symbol as an expression of support for Russia's military aggression. Nonetheless, Z themed content circulated widely on online platforms across many languages. In March 2022, we conducted a dedicated analysis of Z content to examine the scope of its distribution, looking beyond our sample of pro-Kremlin accounts to a much broader set of accounts in the first month after the full-scale invasion when Z propaganda was at its apex. We found that posts with Z propaganda content received 1.2 billion views on TikTok by mid-March 2022. In March 2022, there were over 1 million Instagram posts with hashtags referring to military symbolism or glorifying Russia's war (e.g. the letters #Z and #V, as well as different war-related phrases).[11] Our analysis of 1,200 Facebook pages and YouTube channels found that only 2.5% and 1.3% of posts, respectively, had been taken down—and that only 5.5% of individual posts and videos had been removed. On YouTube, 20 videos containing Z war propaganda in the early days of the war generated over 2 million views and 45,000 likes and comments. As of May 2023, 13 of these videos are still available.

Other examples include more direct physical threats against individuals that are alarmingly common on the platforms. The example below is from a German Telegram account:

[11] #своихнебросаем ("We don't leave our people"), #Zамир ("For peace"), #Zанаших ("For our people"), #зароссию ("For Russia"), #силавправде (Power in truth).

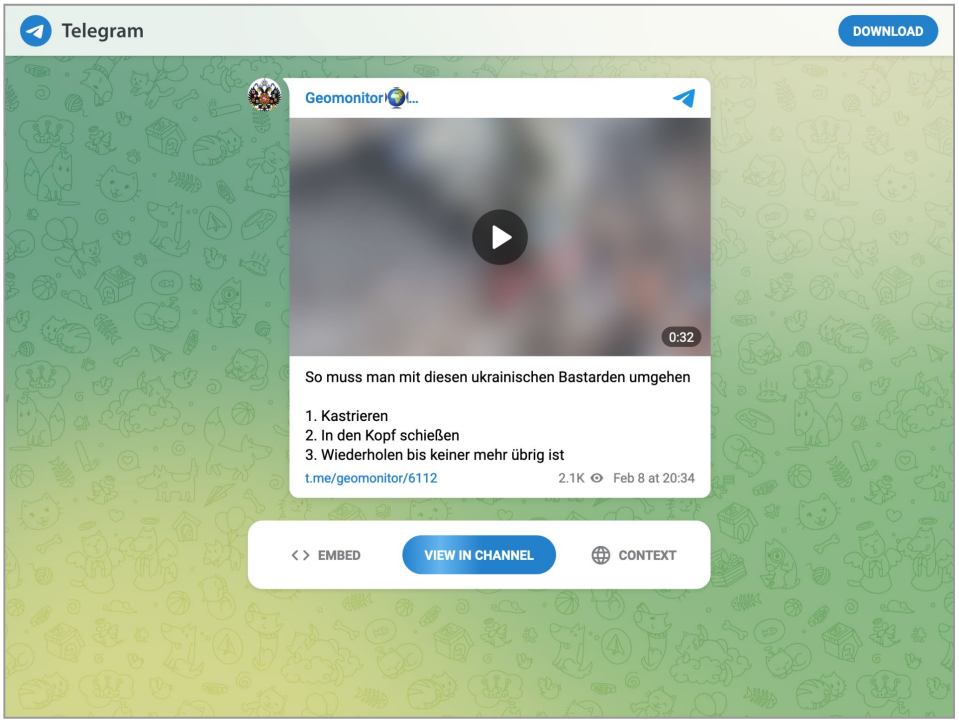

Figure 1: Screenshot of a post shared on Telegram. Translation below.

"That's the way to deal with those Ukrainian bastards

1. Castrate.
2. Shoot in the head.
3. Repeat until there are no more left."

We investigated ethnic slurs against Ukrainians, a common practice in Russian-language social media posts, and found that such content is rarely addressed by the platforms. The word "hohol", for example, appears in 3,100 posts from our monitored list of pro-Kremlin accounts and those posts are still available across all platforms. Similarly, content aimed at denigrating and dehumanising Ukrainians propagates freely on Twitter: since March 2022, 210,000 English-language tweets include the term 'ukronazis'. The example below is particularly extreme but not atypical of this type of incitement to violence against Ukrainians. At such an extreme level of toxicity (and illegality in some jurisdictions), these kinds of posts need to find only a small audience in order to reach a severity level that meets the standards for systemic risk.

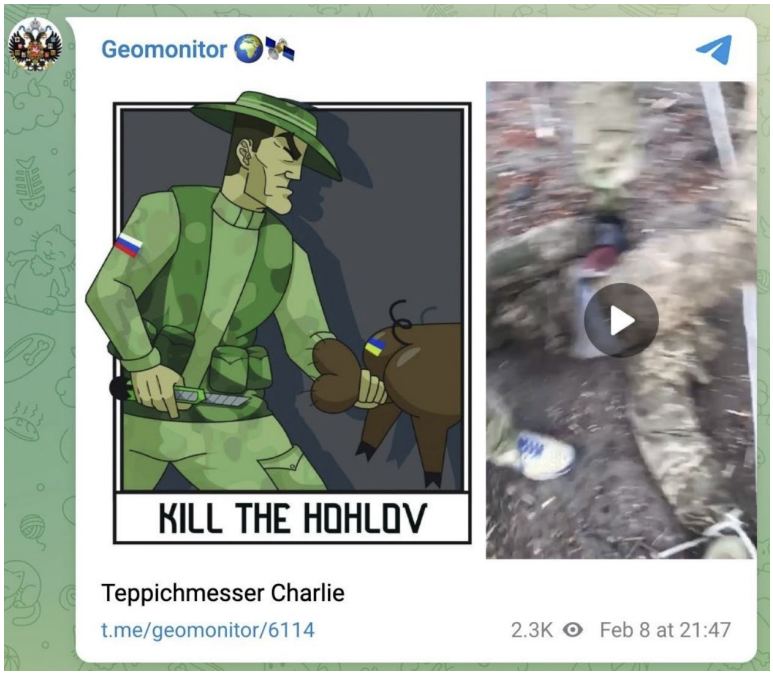

Figure 2: Screenshot of a post on a Kremlin-aligned Telegram channel inciting violence against Ukrainians.

However, in the majority of cases, the pro-Kremlin accounts in our source list post content that is inflammatory but remains beneath the illegality threshold. For instance, in one Telegram post in April 2022, Deputy Chairman of Russia's Security Council Dimitry Medvedev referred to Russia's enemies as "bastards and scum", claiming he would do "anything to make them disappear"[12]. While most readers of this post would infer that Medvedev directed this specific threat to the Ukrainian people, "Russia's enemies" is not specific enough to qualify as a distinct group. The post would therefore likely not qualify as illegal content in most Member States.

While the Kremlin's disinformation content often appears designed to exploit legal grey areas, it is apparently tailored and targeted so as to mobilise both offline and online violence, and thus poses a strong indirect risk of increasing the dissemination of illegal content by others. This form of incitement not only risks radicalising audiences sympathetic to the Kremlin but also those in vehement opposition. This cycle of polarisation – action and reaction that ratchets up the level of severity – is a significant contributor to increased toxicity in social media ecosystems. The DSA requires risk assessments to account for the indirect ways in which disinformation may contribute to the spread of illegal content. Recital 84 states that VLOPs must "also focus on the information which is not illegal, but contributes to the systemic risks identified in this Regulation." Even if itself not unlawful, content by pro-Kremlin accounts incited hatred and violence among target audiences on the platforms and thus carried an intrinsic risk

[12] In many Member States, this would likely not qualify as incitement to violence as their criminal codes stipulate that statements only constitute incitement to violence where they make explicit reference to a particular societal group or a predefined set of individuals. In Germany, for instance, Section 130 of the Criminal Code only prohibits incitement to violence against segments of the population based on their national, racial, religious or ethnic background.

of contributing to the dissemination of illegal content such as illegal hate speech or incitement of violence.

For instance, we compared the *comments* responding to the posts of pro-Kremlin accounts across platforms in the months before and after the February 2022 full-scale invasion of Ukraine. We used the Perspective API[13] to assign a toxicity score – which is a probability figure determined through automated content analysis that estimates the propensity to provoke or cause offence. The proportion of toxic, and potentially illegal, comments to the posts from pro-Kremlin accounts increased sharply and immediately between February and April of 2022. Our data show a 120% increase in toxic posts on Twitter and a 70% increase on YouTube.[14]

iii. Risk Category #2 – Article 34(1)(b): Negative impact on the exercise of fundamental rights

The Kremlin's disinformation campaign poses significant risks to the exercise of fundamental rights. It propagates discriminatory content at scale, denigrating and often dehumanising particular groups or individuals on the basis of protected characteristics such as nationality, sex, gender or religion.[15]

For example, an underreported aspect of pro-Kremlin disinformation campaigns focused on gender-based attacks. Analysts at the Ukrainian research organisation Detector media performed a dedicated study of pro-Kremlin social media posts across platforms in Ukrainian between February and August 2022.[16] They found the denigration of Ukrainian women as a widespread theme. It took many forms, but in particular, the pro-Kremlin narrative alleged female Ukrainian refugees were prostituting themselves to Europe.

Women leaders prominent in the opposition to the Russian war of aggression were particular targets – including the First Lady of Ukraine, Olena Zelenska, the President of Moldova, Maia Sandu, and the former Prime Minister of Moldova, Natalia Gavrilița. We found numerous instances of direct attacks against women on the channels in our sample. The screenshot below is taken from a well-produced cartoon video that circulated across all major social media platforms. It depicts the First Lady of Ukraine as a prostitute for the leaders of NATO.

---

[13] https://perspectiveapi.com/.

[14] We define toxic here to mean comments with a Perspective API toxicity score above 0.8. The periods of comparison are 1 December 2021 – 20 February 2022 and 1 April 2022 – 30 November 2022.

[15] Strand, Cecilia, Svensson, Jakob. 2021. "Disinformation campaigns about LGBTI+ people in the EU and foreign influence." Briefing requested by the INGE committee. Policy Department for External Relations Directorate General for External Policies of the Union, July 2. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/653644/EXPO_BRI(2021)653644_EN.pdf; Bilousenko, Olha. 2023. "Ethnocide of Hungarians and Jewish Conspiracy. Russian Disinformation on Ethnic Groups on Social Media." Detector Media. Accessed 2 June 2023. https://en.detector.media/post/ethnocide-of-hungarians-and-jewish-conspiracy-russian-disinformation-on-ethnic-groups-on-social-media?fbclid=IwAR3Wm5SVZZvcwIFWrfMJy643PW2gf67SpSbrM0_m6JPL4gWqqHxH-47Y4PA; Polianska, Inna. 2022. "A History of Defamation: Key Russian Narratives on Ukrainian Sovereignty." EU vs Disinfo. Accessed June 2 2023. https://euvsdisinfo.eu/a-history-of-defamation-key-russian-narratives-on-ukrainian-sovereignty-2/.

[16] Bilousenko et al. 2022. "'Prostitution will save Ukraine from the default.' Investigating Russian gender disinformation in social networks." Dectetor Media. Accessed 2 June 2023. https://detector.media/propahanda_vplyvy/article/203226/2022-09-28-prostitution-will-save-ukraine-from-the-default-investigating-russian-gender-disinformation-in-social-networks/.

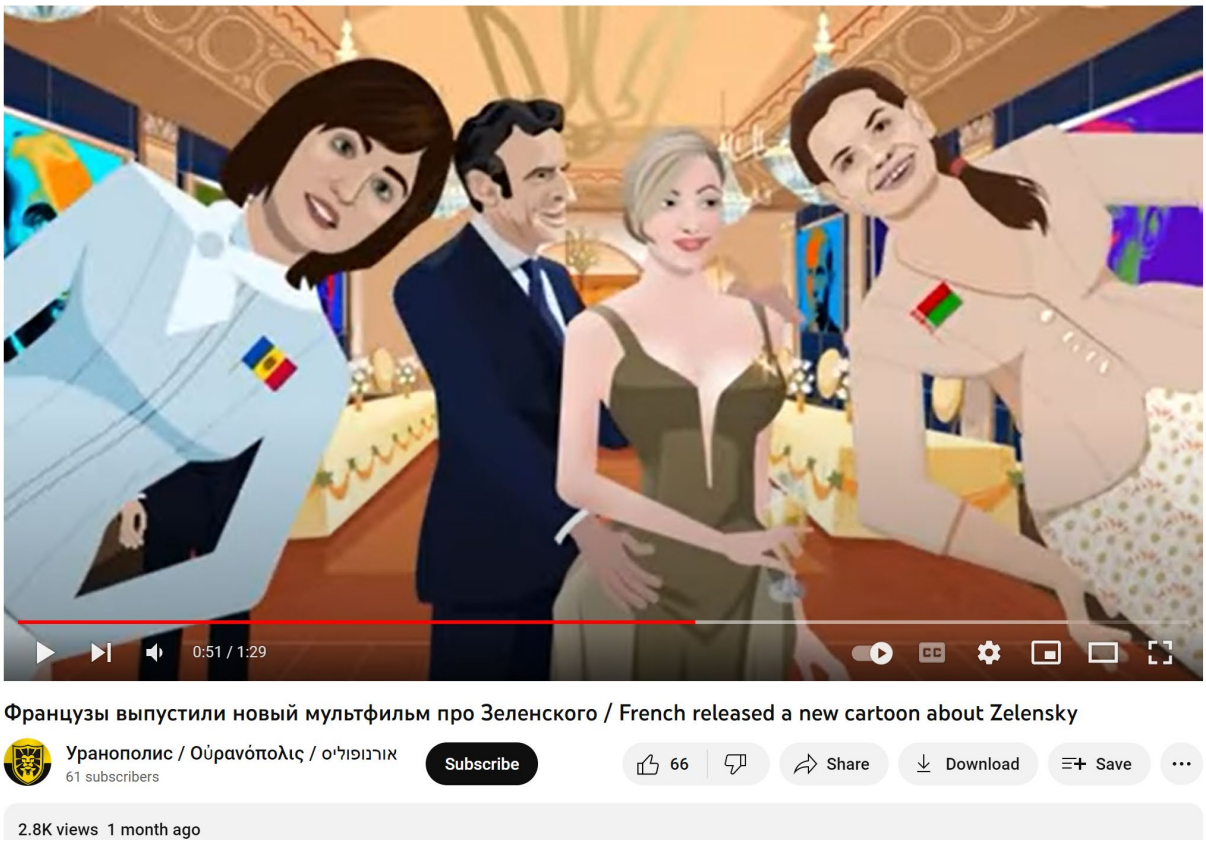

Figure 3: Screenshot of a YouTube video displaying gender-based violence against Olena Zelenska, First Lady of Ukraine.

Besides the right to non-discrimination, the Kremlin's disinformation campaign also affects the ability of citizens both in Ukraine and in EU Member States to exercise their right to freedom of expression. Relevant behaviours associated with the Kremlin, which we have labelled Suppression Behaviours in the taxonomy provided in the Appendix, encompass actions such as sending abusive notifications, engaging in impersonation or coordinated harassment, and disseminating discriminatory or hateful content. These behaviours result in the silencing of individuals who become targets of such operations.

The Kremlin's disinformation campaign on online platforms also causes clear risks for the freedom and pluralism of the media. We have collected evidence of abusive notification campaigns on online platforms that resulted in the shutting down of social media accounts of independent media outlets in Ukraine and other countries. This happens when malicious actors bombard legitimate accounts with user-notifications falsely flagging the account as spam or as engaged in activity that violates the platform's Terms and Conditions. Ukrainian media outlets have lost significant shares of their Facebook traffic since the beginning of the war, thus exacerbating economic pressure on Ukrainian publishers and reducing access to reliable information for Ukrainian users, including in areas of active conflict. In late 2022, *Ukrainska Pravda*, Ukraine's largest independent newsroom, coordinated a survey among Ukrainian news outlets that use Facebook for distribution to document their challenges. Out of 57 independent local news outlets, 37% reported major reductions in organic reach. Their Facebook pages were given 'red' or 'orange' status for alleged violations of Meta's content policies. Moreover, affected outlets lost access to Meta's monetization tools. The Ukrainians

were not alone as targets of Kremlin suppression campaigns. The Facebook accounts of influential public figures in Bulgaria expressing overt pro-Ukrainian views have also been subject to inexplicable blocking or temporary suspensions, likely as a result of mass reporting of those accounts in 2022.

Along with these suppression behaviours, large-scale dissemination of disinformation and state propaganda about the war narrowed the space for independent and pluralistic news media content on the platforms. More generally, the harassment, intimidation and suppression of independent journalism has observable effects in EU Member States as well. Germany, for instance, recently dropped five ranks in the Media Freedom Index because of growing online and offline attacks against journalists in the context of conspiracy ideologies and disinformation that incite hostility towards independent media.[17]

Our investigation of pro-Kremlin accounts shows significant amounts of "coordinated harassment" that targets particular individuals and institutions. For example, the organisation known as Cyber Front Z — described[18] in a UK Foreign Office report as a "sick Russian troll factory" — has since its founding in mid-March of 2022 been among the most visible groups coordinating pro-Kremlin information operations against Ukraine. Trolls, paid and unpaid, make up the core of the group. However, the group seeks to mobilise 'patriotic' Russians to assist in the trolling. Coordination of the trolling primarily happens through a network of Telegram channels, of which Cyber Front Z[19] is the most prominent known example. The trolling is a cross-platform endeavour. Coordinated on Telegram, the majority of posts call on supporters to move to other platforms such as YouTube or Instagram and share specific propaganda memes, up- or down-vote specific posts, report users for violating platform Terms and Conditions, and hound opponents in the comments section.

The coordination was often highly structured, including specific instructions to participating "cyber soldiers." The channels regularly posted content that falls into two categories: pro-Kremlin propaganda for dissemination ('mediabombs') and targets for harassment ('traitors'). Material targeting enemies or traitors was usually very specific – as demonstrated in the screenshots of Telegram posts below. (We have provided translations from the original Russian into English for ease of review.) These posts contained links to the target's social media platforms and directions to perform particular actions. In many cases, links to individual posts were provided to demonstrate which channels to target (Figure 4) as well how the targeting should be performed (Figure 5). The accounts targeted span from a wide variety of perceived opponents such as the social media accounts of foreign and domestic politicians (Figure 6), Ukrainian soldiers, Russian citizens in opposition to Russia,[20] Ukrainian government and ministries,[21] foreign governments and representatives, social media companies as well as foreign press. The coordination process also includes the recruitment of

[17] Reporters Without Borders RSF. 2023. "World Press Freedom Index." Accessed 2 June 2023. https://rsf.org/en/index.
[18] UK Foreign, Commonwealth & Development Office. 2022. "UK exposes sick Russian troll factory plaguing social media with Kremlin propaganda." Accessed 2 June 2023. https://www.gov.uk/government/news/uk-exposes-sick-russian-troll-factory-plaguing-social-media-with-kremlin-propaganda.
[19] https://t.me/cyber_frontZ, accessed 2 June 2023.
[20] https://www.youtube.com/watch?v=aDgJS1p1I9M&t=2s, accessed 2 June 2023.
[21] https://www.youtube.com/watch?v=v5I5nQIcLsw, accessed 2 June 2023.

new cyber soldiers where specific information is communicated on what type of competencies the Cyber Front Z are looking for as well as information about working conditions (Figure 7).

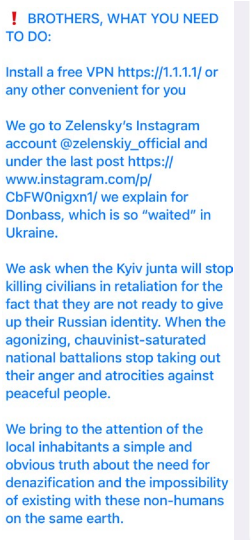

Figure 4

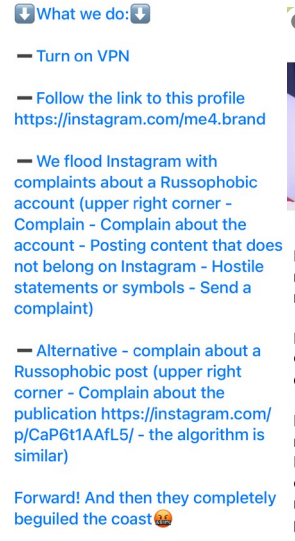

Figure 5

Figure 6

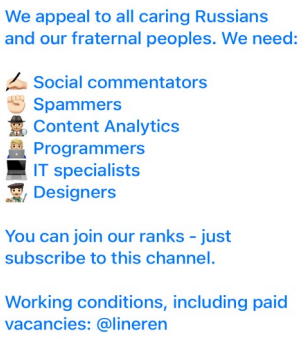

Figure 7

Lastly, doxxing is a behaviour Kremlin-aligned accounts frequently apply on online platforms to suppress opposing viewpoints, thereby undermining the ability of their targets to exercise their right to data protection. 'Project Nemesis' is a pro-Kremlin mass doxxing operation aiming to hunt down and expose Ukrainian "Nazis" and "those who help them". Using a website and a Telegram channel with over 50k subscribers, activists published photographs and personal details of hundreds of individuals fighting on behalf of Ukraine (including birth dates, addresses, telephone numbers, passport numbers, social media profiles, etc), risking the safety of individuals. While doxxing is often coordinated between Telegram and VKontakte, it targets accounts on most of the large platforms.

iv. Risk Category #3 – Article 34(1)(c): Negative effect on electoral processes, civic discourse and public security

The Kremlin's disinformation campaign accompanying the illegal war in Ukraine constitutes a risk to public security *per se*. Kremlin operatives have repeatedly and explicitly positioned disinformation as a weapon in the Kremlin's arsenal along with other hybrid, as well as conventional, military capabilities. It is the declared objective of the Russian state propaganda apparatus to "conduct information war against the whole Western world", and to "conquer" and "grow" audiences in order to access those audiences in "critical moments."[22] Key operatives behind Russia's disinformation apparatus are known to be interlinked with the Ministry of Defence, Russian Armed Forces and its intelligence arm, the GRU. The documented aim of the Russian disinformation campaign is to legitimise and promote violence, and to weaken public institutions inside the European Union.[23]

[22] Eu vs Disinfo. 2022. "The Kremlin's weapons of deception: 7 things you need to know about RT and Sputnik." Accessed 2 June 2023. https://euvsdisinfo.eu/the-kremlins-weapons-of-deception-7-things-you-need-to-know-about-rt-sputnik/.
[23] Organisation for Economic Co-operation and Development. 2022. "Disinformation and Russia's war of aggression against Ukraine: Threats and governance responses." Accessed 2 June 2023. https://www.oecd.org/ukraine-hub/policy-responses/disinformation-and-russia-s-war-of-aggression-against-ukraine-37186bde/.

These efforts must be understood not only as a wartime action but as a premeditated strategy. Since early 2020, there is clear evidence that pro-Kremlin sources built large social media audiences in languages such as German, French and English by tailoring and targeting extreme content to cultivate susceptible audiences. They were frequent purveyors of vaccine disinformation during the COVID-19 pandemic, for instance. Social media algorithms recommend new conspiratorial content to audiences that have previously consumed similar content. This kind of intentional manipulation of civic discourse in the EU enabled the Kremlin to develop networks of content distribution through which to roll out war propaganda to significant portions of the online populations in many EU countries in late February 2022.

We assess therefore that the Kremlin's disinformation campaign carries high risk of negative effects not only to public security but also to electoral processes and civic discourse. International human rights law suggests that both types of risk can emerge when deceptive means are employed to prevent citizens from fully exercising their rights, for example the rights to vote, to freedom of expression and information, or the right to freedom of opinion. In its General Comment No. 25 on Article 25 of the International Covenant on Civil and Political Rights, the UN Human Rights Committee specifies that "voters should be able to form opinions independently, free of violence or threat of violence, compulsion, inducement or manipulative interference of any kind."[24] In her Report on Disinformation and Freedom of Opinion and Expression during Armed Conflicts, the UN Special Rapporteur for Freedom of Expression similarly notes that "access to diverse, verifiable sources of information is a fundamental human right" and an "essential necessity for people in conflict-affected societies". She further notes that

> *"coercive, involuntary or non-consensual manipulation of the thinking process, such as indoctrination or "brainwashing" by State or non-State actors, violates freedom of opinion. Content curation through powerful platform recommendations or microtargeting, which plays a key element in amplifying disinformation and aggravating political tensions, is non-consensual manipulation of users' innermost thinking processes in digitized form. As such, it amounts to a violation of the right to freedom of opinion."*[25]

The Kremlin's disinformation campaign employs a wide range of behaviours and content types on online platforms that deceive users on matters of vital public relevance in the context of the war, and artificially inflate the reach of Kremlin disinformation at the expense of diverse and verifiable sources of information. Unless mitigated, the Kremlin's disinformation campaign thus arguably poses significant risks to electoral processes and civic discourse. Summarised here is an overview of the types of behaviour and content used by the pro-Kremlin accounts that generate the potential for systemic risks to both.

---

[24] UN Human Rights Committee. 1996. "CCPR General Comment No. 25: Article 25 (Participation in Public Affairs and the Right to Vote), The Right to Participate in Public Affairs, Voting Rights and the Right of Equal Access to Public Service." CCPR/C/21/Rev.1/Add.7. Accessed on 2 June 2023. https://www.refworld.org/docid/453883fc22.html.

[25] United Nations, General Assembly. 2022. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression." A/77/288. Accessed on 2 June 2023. https://www.ohchr.org/en/documents/thematic-reports/a77288-disinformation-and-freedom-opinion-and-expression-during-armed.

The first category of activity we documented was Circumvention Behaviours – detailed more thoroughly in the taxonomy in the Appendix – which Kremlin-aligned actors employed to bypass detection and mitigation by online platforms. This category of behaviours includes the use of deceptive identities and rebranding. For instance, Reset and other research organisations identified networks of thousands of social media accounts that deceptively posed as "news media" on online platforms, but were set up with the sole intention of repurposing and disseminating Russian disinformation and war propaganda content. In many cases, these accounts were set up specifically to continue distributing content by Russian state media accounts that had been banned or geoblocked in response to EU sanctions.

"Clones" of authentic media often use domain names and designs similar to legitimate and well-known news sources to target users with disinformation. A prominent case documented by EU DisinfoLab includes at least 17 cloned media providers, including Bild, 20minutes, Ansa, The Guardian or RBC Ukraine, which depicted Ukraine as a Nazi-governed state and denied the Bucha massacre.[26] More recently, we have identified a network of at least 30 coordinated, inauthentic accounts on Facebook with a collective audience of 1.6m followers. These accounts masquerade as local media outlets while promoting content produced by Russia state media and targeting French-language audiences in Europe and Africa.

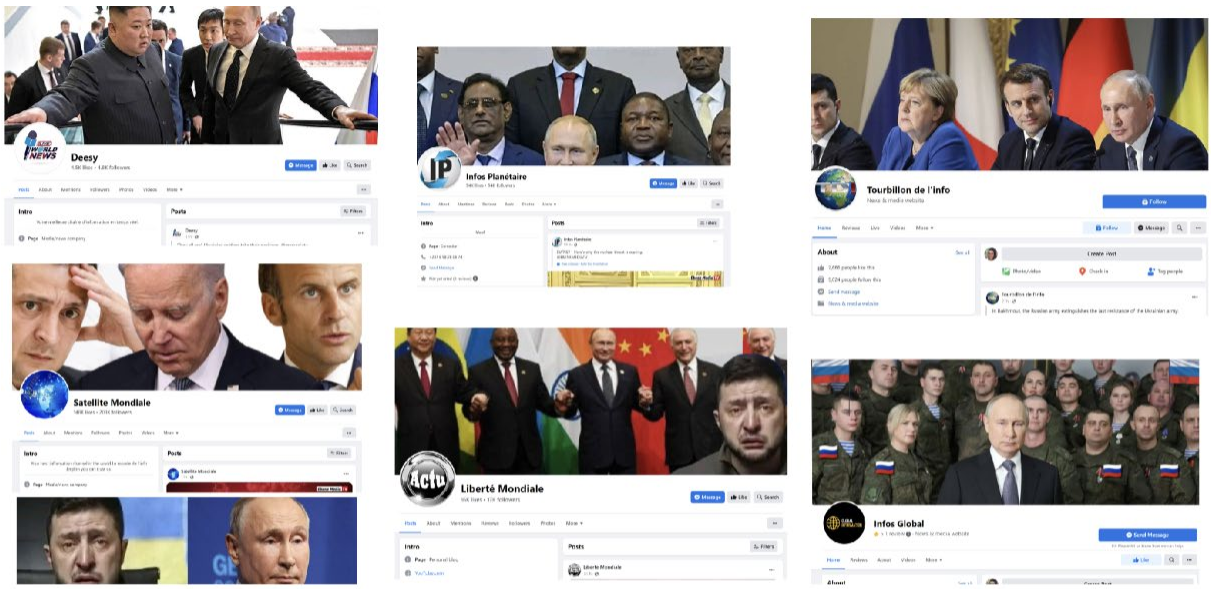

Figure 8: Screenshots of "cloned" authentic media imitating legitimate media.

In addition, a network of seven Telegram channels dubbed NSDV (a Russian acronym from "Na samom dele v…", referring to "current news from..") pretends to broadcast local news from Ukrainian cities but also promotes pro-Kremlin narratives to its audience of 100,000 subscribers. Another network of 25+ channels posing as media outlets and branded under the name "InfoDefence", promotes pro-Kremlin narratives in many European languages and shares content from Russian state-affiliated media to a total audience of 175,000 subscribers.

[26] Alaphilippe et al. 2022. "Doppelganger – Media clones serving Russian propaganda." EU Disinfo Lab. Accessed 2 June 2023. https://www.disinfo.eu/doppelganger/.

Related behaviours designed to spread deceptive content about the war included the cultivation of back-up accounts (in case the primary accounts were banned by a platform) and the “re-channeling” of social media audiences to external, unmoderated platforms and domains. After the Russian Federation blocked Facebook, Instagram and Twitter from use within its territory, many Russian government accounts started using their accounts as “landing pages" by including links to Russian platforms in the biography or description sections of their accounts. Between March and April 2022, more than 40 embassies, consulates, delegations, and missions created Telegram channels, and three quarters of the Russian government accounts that we monitored directed their followers to Telegram – indicating a concerted exodus to pre-empt interventions by other platforms. We also identified 2,100 tweets by Kremlin state media, Russian state-affiliated journalists and Russian diplomatic accounts directing users to follow RT and Sputnik on Odysee.

We identified a wide variety of content promoted by pro-Kremlin accounts carrying potential systemic risk to public security, electoral processes and civic discourse through intentional disinformation. This type of narrative has a particularly high potential of increasing risk to civic discourse and elections, in that it is designed to deceive and “brainwash” audiences on matters of critical public importance, including matters of life and death.

The example below is a video that circulated widely on German language YouTube – generating hundreds of thousands of views. It is one of many similar posts spread across European digital media platforms by pro-Kremlin actors alleging that the Ukrainians were responsible for a nuclear cloud blowing into the EU. Other examples of attempts to pollute the public debate in European countries with hyperbole and falsehood included claims that Europe would run out of gas in the winter or that Putin had won the war and NATO was abandoning Ukraine.

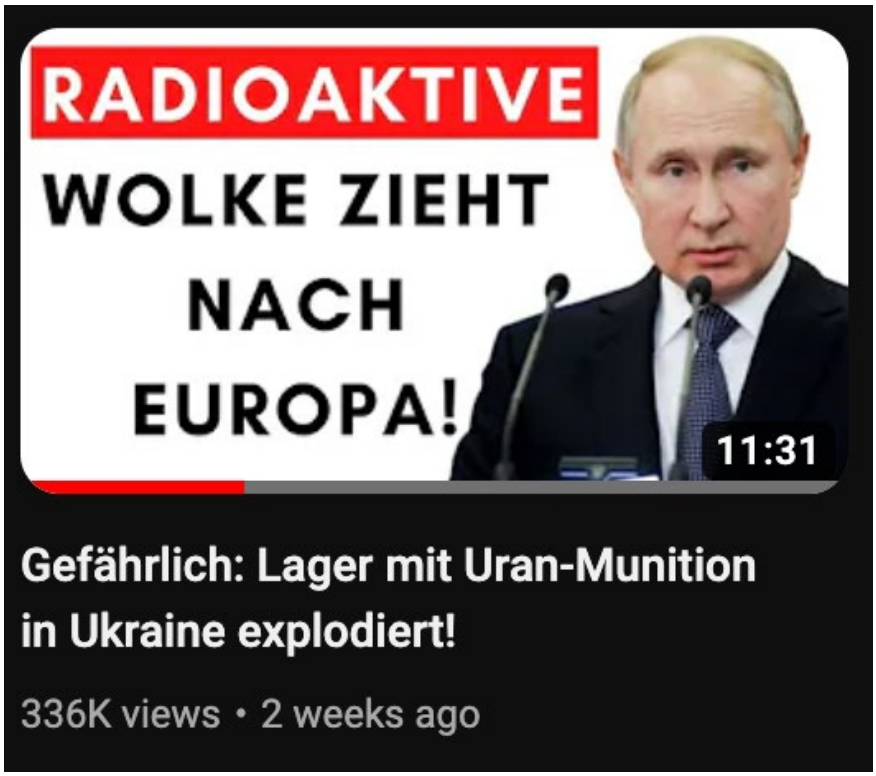

Figure 9: Screenshot of a German-language YouTube video amplifying Kremlin disinformation about an alleged "nuclear cloud" moving towards Europe from Ukraine.

This risk category also includes content that denies war crimes, or likely war crimes, such as the Bucha massacre. After Russian forces retreated from the Ukrainian town of Bucha, northwest of Kyiv, on March 31, 2022, the Ukrainian military re-entered the town to discover numerous sites of mass murder that happened during Russian occupation. After the crimes attracted global attention, the Kremlin denied Russia's involvement and introduced narratives online which blamed Ukrainians. We performed an analysis of the 100 highest-performing posts from our sample of pro-Kremlin accounts that mentioned Bucha on Facebook, Instagram, Telegram and TikTok during this period. It showed that the vast majority (from 85% on Facebook to 100% on TikTok) of all the top-performing content contained disinformation about the massacre, including denial that it had even taken place. For example, a video published on the German YouTube channel COMPACT TV titled "Bucha: Facts versus propaganda" suggests that photos and videos from Bucha could have been manipulated or staged by Western media.[27]

[27] https://archive.ph/CeH9D, accessed 2 June 2023.

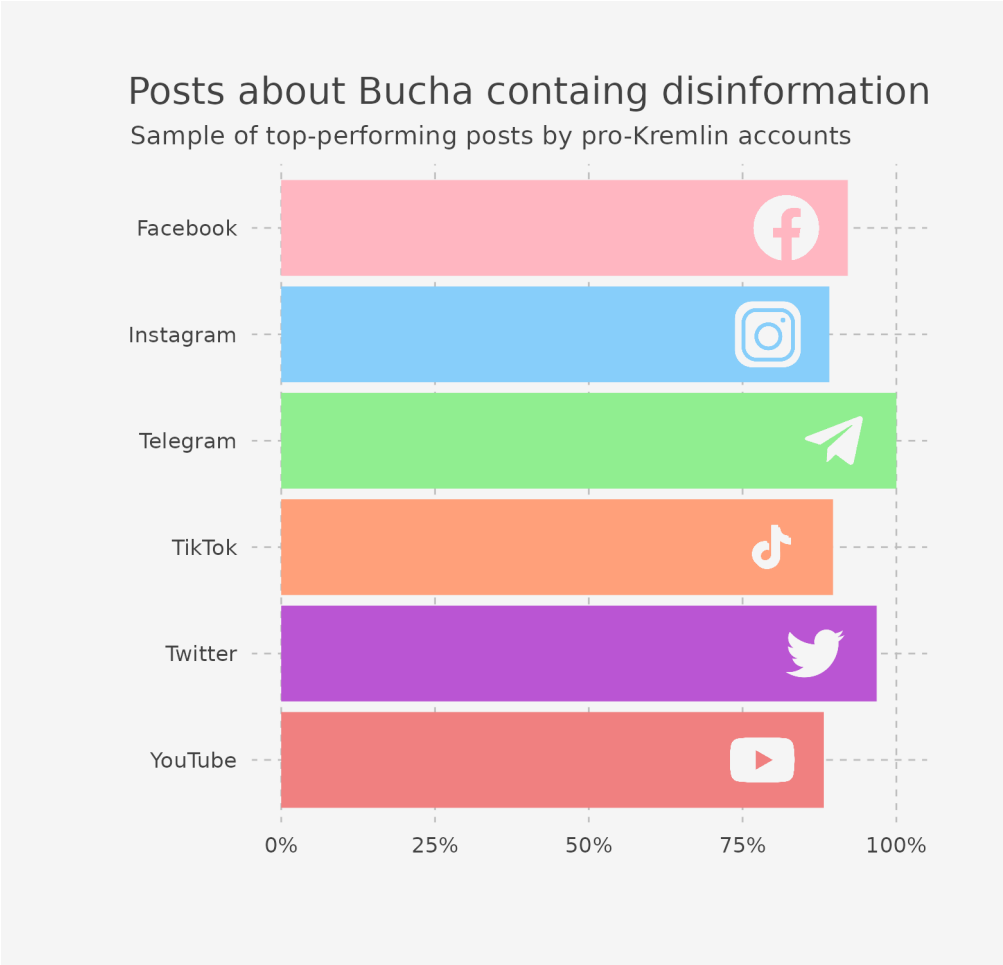

Figure 10: Percentage of posts about Bucha containing disinformation, from a sample of top-performing posts by pro-Kremlin accounts per platform.

For every high profile example of war crime denial or threats of nuclear fallout, there are dozens and dozens of posts circulating on the platforms that carry lower-risk disinformation, the everyday falsehoods that comprise the standard fare of many of the pro-Kremlin channels in our sample. These may be evaluated for risk in their own right, but they should also be seen as a cumulative phenomenon. The more audiences are exposed to intentional disinformation, the more the narratives it carries are normalised. Most people are conditioned as information consumers to equate frequency of exposure with probability of veracity. Even if audiences never fully believe the falsehoods, the disinformation succeeds if all it does is call into question the facts. Below are a common set of examples of disinformation in different languages and on different topics. The number of engagements and exposures varies, but this type of content is persistently shared across platforms to a growing audience. The effect in aggregate poses a significant risk to civic discourse over time.

Figure 11: Pro-Kremlin content claiming Hungarian Prime Minister Victor Orban is warning the EU is sending troops to Ukraine and will start a third world war (Facebook).

Figure 12: RIA accusing the Ukrainians of bombing a theatre full of civilians in Mariupol that was destroyed by Russian shelling (Twitter).

Figure 13: War propaganda on Russian diplomatic account (Twitter).

Figure 14: Pro-Kremlin content claiming a nuclear cloud in Ukraine (YouTube).

v. Scale

Operating within the Baseline Framework presented here, the next step after the qualitative assessment of risk within particular types of content that match the categories in Article 34 is to evaluate the scale of reach using risk metrics. Under the modified Rabat structure, this sequence permits a clearer finding as to the severity of the risk and whether it meets the threshold of systemic risk under the Regulation. Severity here is a function of the qualitative risk assessment paired with the quantitative reach calculations where the higher the assessed risk the lower the required reach in order to meet a standard of systemic risk. This section presents the quantitative analysis of reach and engagement of the pro-Kremlin (Kremlin-backed and Kremlin-aligned) accounts in our case study in order to put the qualitative assessment of risk into the proper context.

Actors		Number of accounts	Volume	Audience Size	Exposure*	Engagement	Audience Size (change)	Mean Exposure (change)	Mean Engagement (change)
Kremlin-backed	State media (affected by EU sanctions)	168	1,390,000	64,400,000	3,060,000,000	125,000,000	4%	-81%	-42%
	State media (global)	170	1,590,000	95,600,000	4,550,000,000	178,000,000	8%	-87%	-72%
	Official government	518	213,000	21,300,000	812,000,000	28,500,000	21%	58%	-2%
	Kremlin staffers	244	335,000	48,300,000	11,100,000,000	55,900,000	65%	301%	-37%
Kremlin-aligned	Media outlets	330	1,890,000	29,100,000	13,700,000,000	103,000,000	63%	209%	42%
	Influencers	807	2,840,000	67,600,000	56,200,000,000	218,000,000	126%	129%	49%
Total		2237	8,258,000	326,300,000	89,422,000,000	708,400,000	31%	58%	-24%

Figure 15: Overview of the main metrics collected grouped by actor category, including both raw numbers and the changes observed in the period 1 April 2022 – 30 November 2022, compared to the pre-war baseline (1 December 2021 – 20 February 2022). Darker cells indicate an increase, bright yellow indicates a large reduction, showing a trend where other Kremlin-linked actors increasingly replaced state media.

Overall, accounts involved in the Kremlin's disinformation campaign continue to reach vast audiences across online platforms. Accounts affiliated with Russian state media maintained an aggregate audience of over 160 million across platforms in December 2022. Additionally, other Kremlin-backed accounts, such as government accounts and the personal accounts of Kremlin staffers, had an audience of at least 60 million. Overall, Kremlin-backed accounts that regularly disseminate disinformation or propaganda thus have an aggregate audience of at least 220 million across online platforms.[28] Despite the different restrictive measures introduced by governments and platforms, the reach of Kremlin-backed accounts in terms of audience size grew by 31 percent compared to the pre-war period. In total, content by the approximately 1110 Kremlin-backed accounts in our sample was viewed at least 19 billion times during our monitoring period, and accumulated almost 400 million engagements.

Pro-Kremlin accounts continue to reach the largest audience on the Meta platforms Facebook and Instagram with a combined audience size of 142 million for the 719 accounts in our sample. At the same time, we observed strong reductions on other risk metrics for Facebook and Instagram: posting volume (-35 and -45 percent respectively), average engagement (-37

[28] Note: it is not possible to calculate the number of unique users in light of data access limitations.

and -72 percent) and average exposure (-84 and -65 percent) declined strongly. On other platforms we observed a similar downward trend only for TikTok, where the audience was only 1/6th the size of that on Facebook. Meanwhile, the audience size for pro-Kremlin accounts instead increased by 88 percent on YouTube and by as much as 308 percent on Telegram. Similarly, exposure and engagement grew by 60 and 84 percent on YouTube, engagement by 54 percent on Twitter, and exposure by 194 percent on Telegram.

Our analysis suggests that government interventions were the main factor driving down exposure and engagement for Kremlin-backed accounts on the Meta platforms. Meta's own mitigation measures also played a role, but the causal effect is less straightforward (see section "Algorithmic Recommender Systems" that is part of the Mitigation Analysis).[29] We observed that the decreases in engagement were sharper for Russian-language Kremlin-backed accounts than for accounts operated in EU languages. The explanation for these differences is domestic censorship efforts in Russia. In mid-March 2022, the Kremlin labelled Meta an extremist organisation. Russian Internet service providers thereafter systematically blocked access to Facebook and Instagram, cutting off a significant share of the potential audience for Russian-language content on the platforms.

For Kremlin-backed accounts operating in EU-languages, government measures were also a large causal factor. Among the different types of Kremlin-backed accounts, engagement dropped the most for state media accounts, many of which were geo-blocked inside the EU in response to several rounds of sanctions by the Council of the EU. At the same time, engagement for official government accounts operating in EU languages on Facebook, which were not subject to any restrictions by EU governments or Meta itself, in fact grew by more than 50 percent. On Twitter and YouTube, engagement on Russian official accounts posting in EU languages quadrupled and tripled respectively. By the end of our monitoring period, Kremlin-backed accounts other than state media had a combined audience of 9.2 million on Facebook and Instagram. This constitutes an 8% rise on the pre-February 2022 figure, and an absolute increase of 850 000 subscribers.

In our assessment, restrictive measures on social media platforms were effective tools to decrease exposure for Kremlin-backed accounts. However they were used rarely, allowing the Kremlin to compensate by leveraging its other assets, in particular official government accounts and the personal accounts of Kremlin staffers. Notably, these accounts remained unbanned across all platforms, enabling the continued dissemination of targeted disinformation and war propaganda to both domestic and international audiences. The movement of audiences to accounts that evaded restrictions to a considerable degree neutralised the effect of restricting access to Russian state-media accounts. YouTube's binary policy of outright banning accounts or allowing them to operate apparently without restrictions similarly enabled initially smaller channels to rapidly grow and fill the vacated space.

This within-platform movement of audiences happened on an even larger scale between platforms. Telegram emerged as a crucial vector for the Kremlin to reach audiences in Russia and abroad. For every metric considered, the significance of Telegram increased by multiples. We were unable to collect engagement data for Telegram as those metrics are only visible in

[29] Meanwhile, it seems unlikely that an "organic" decline in user interest drove decreasing engagement rates, given that Russia's war in Ukraine has dominated global news cycles since February 2022.

its mobile app – the web or desktop versions merely indicating view counts. The reduction in engagement numbers in the table above reflects an overall drop on the other platforms, whereas exposure— which includes Telegram—shows a 58% overall increase.

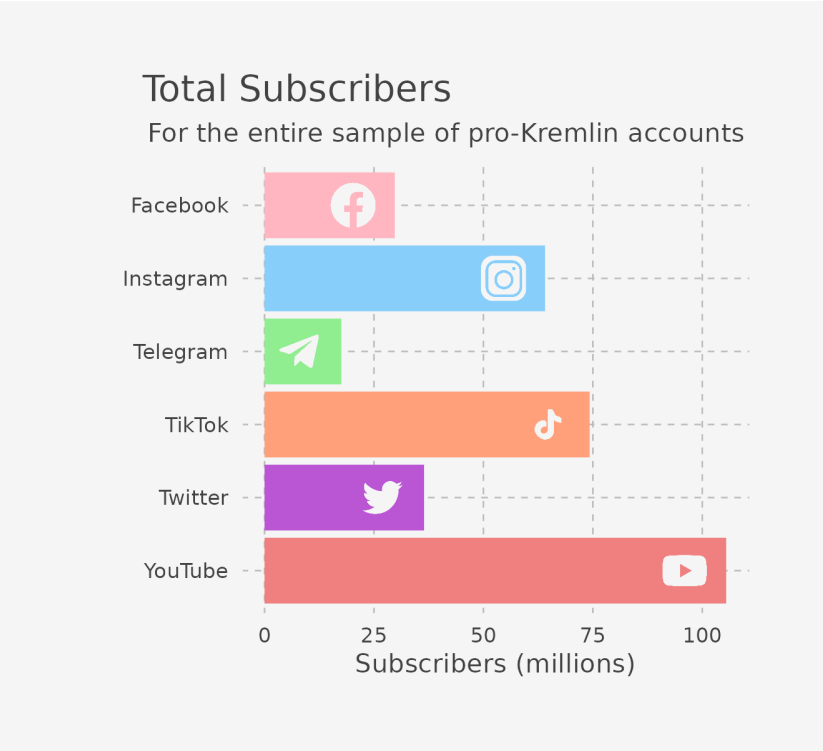

Figure 16: Total number of subscribers for the entire sample of pro-Kremlin accounts by platform as of December 2022.

We conclude based on the qualitative and quantitative evidence presented here that Kremlin disinformation campaigns between February and November of 2022 reached a level of severity more than sufficient to qualify as systemic risk across all of the platforms. The proportionality test of the modified Rabat framework can be robustly assessed. We find that the actors (state-backed and state-affiliated accounts) and context (information operations supporting a war of aggression) are sufficient to judge both the intent of the speakers and the expectations of the audiences. Intent is clear from the Kremlin's own statements about their activities, such as when RT head Margarita Simonyan described relying on guerilla tactics to circumvent YouTube's content moderation.[30] In addition, both the circumvention of platform policies and repeated violations of platforms' Terms and Conditions serve as indication of the intent of Kremlin-backed actors. Further, we can see from the nature of the content itself that it contains pervasive examples of incitement to violence, intentional attempts at deception, organised efforts to harass others and suppress freedom of expression. The form of the content and the activities of its publishers and distributors also meets the standard of severe risk – often containing false statements, manipulated images, deceptive branding and attempts to distort the public sphere through artificial amplification of messages.

Finally, we see that this content across the sample of accounts studied has reached very large audiences across multiple platforms, languages and countries. Despite the efforts of

[30] https://twitter.com/JuliaDavisNews/status/1514040178591117317, accessed 2 June 2023.

governments and platforms, high risk content reached very large audiences measured in the hundreds of millions. Over the period of study, these audiences grew and the severity level of the risk rose with repeat exposure and engagement. The analysis provided here to assess the severity threshold of systemic risk – a combination of qualitative review of the content and quantitative study of the audience size – demonstrates that it is clearly met.

d. Part 2: Mitigation Analysis

Having established through the Risk Assessment that the conditions for systemic risk are clearly associated with the Kremlin disinformation campaigns under analysis, we move now to evaluate the mitigation measures applied by the platforms. The mitigation analysis below focuses primarily on the **actor** dimension of our risk assessment. It is designed to comprehensively assess how well online platforms mitigated risks associated with Kremlin-backed and Kremlin-aligned actors through their stated policies and their application via content moderation processes and recommender systems. The objective is not to enumerate all aspects of potential risk mitigation analysis across all aspects of the risks assessed above. That is beyond the scope of this paper and publicly available data. Rather, our purpose is to offer a proof-of-concept for structuring replicable investigations using a set of standardised metrics that could be applied to each aspect of the risk assessment.

Following the Baseline Framework and Article 35 of the Regulation, our mitigation analysis is split into two main parts: 1) A *qualitative* evaluation of the policies that platforms either introduced or already had in place to address the systemic risks identified in the risk assessment above ("Terms and Conditions"). Here we are able to provide an assessment of Actors, Behaviours and Content risk categories; and 2) A series of *quantitative* analyses measuring how effectively platforms applied and enforced these policies, sub-categorised into two segments – "Content Moderation Measures" and "Algorithmic Recommender Systems."

i. Terms and Conditions

We analysed the Terms and Conditions ("policies") of six platforms: Facebook and Instagram (collectively referred to as Meta), YouTube, TikTok, Twitter and Telegram to determine the extent to which they cover the actors, behaviours and content associated with the Kremlin's disinformation campaign. We began by conducting desk research to identify relevant policies and created a structured table (the "platform policy tracker") for documenting and standardising these policies across the platform ecosystem. This table includes the platforms' written Terms and Conditions and public announcements regarding changes that they were making in response to Russia's full scale invasion of Ukraine. However, we also include the numerous *pre-existing* policies that apply to the evidence we detected in our sample, such as rules against doxxing or incitement to violence. These standards should apply to the Kremlin's disinformation campaign even though they were not developed specifically in response to it. Changes that were made to platforms' Terms and Conditions in response to the invasion are indicated by an *asterisk* in the tables below. We rely only on public information that has been released by companies, not on leaked materials or inferences about possible internal policies.

(1) Actors

Actors		Meta	YouTube	TikTok	Twitter	Telegram
Kremlin-backed	State media (affected by EU sanctions)	geoblock*	geoblock*	geoblock*	geoblock*	geoblock*
	State media (global)	label*	ban*	label*	label	allow
		demote*			demote	
	Official government	allow	allow	allow	label	allow
					demote*	
	Kremlin staffers	allow	allow	allow	label	allow
					demote*	
Kremlin-aligned	Media outlets	allow	allow	allow	allow	allow
	Influencers	allow	allow	allow	allow	allow

Kremlin-backed accounts

Twitter was the only platform that introduced policies covering all the types of Kremlin-backed accounts. Twitter's policy response thus was the most comprehensive in scope. By contrast, Telegram did not introduce any policy against Kremlin-backed accounts, except for legally mandated geoblocks on the accounts of Russian state media as required by EU sanctions. Notably, under the new leadership of Elon Musk, Twitter has recently (and publicly) reversed many of these policies.[31]

Whereas Twitter's policy response was the widest in scope, YouTube's was arguably the most invasive: Unlike any other platform, YouTube banned all Russian state media accounts globally. However, the real scope of YouTube's policy was impossible to verify as the platform did not disclose a list of the outlets or accounts to which it applied the ban.

[31] https://twitter.com/elonmusk/status/1645177202961534977?s=20, accessed 2 June 2023.

None of the major platforms we examined banned all Kremlin-backed accounts, including official government accounts, such as the accounts of Russian Embassies, or the personal accounts of Kremlin staffers. As a result, the Kremlin was able to continue targeting disinformation and war propaganda at growing Russian-language and international audiences. The following chart contrasts the weekly exposure of Kremlin state media and government accounts with those of Kremlin staffers for the duration of our monitoring period. As the data show, the Kremlin was able to recuperate much of the social media reach it had lost as a result of restrictive policies against state media accounts, generating up to 300 million weekly views through the personal accounts of Kremlin staffers. For instance, the Telegram channel of Dmitry Medvedev has accumulated more than 1 million subscribers since its creation in March 2022 and established itself as a regular distributor of dehumanising and violence inciting messaging.

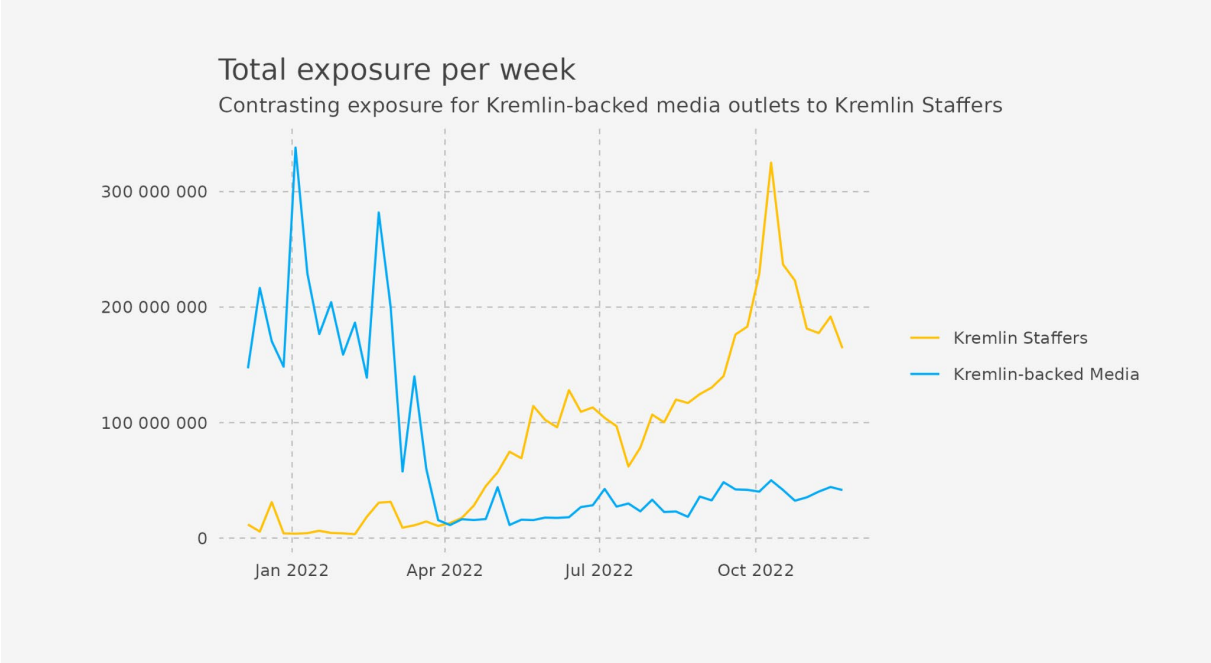

Figure 17: Total exposure to accounts by Kremlin-backed media contrasted to total exposure to accounts by Kremlin staffers across all platforms.

Notably, there is precedent for all-encompassing bans on state-backed accounts in the context of military violence. In February 2021, Facebook banned all accounts linked to the Myanmar military after the latter had seized control of the state, with Facebook citing "exceptionally severe human rights abuses and the clear risk of future military-initiated violence."[32]

Kremlin-aligned accounts

The sustained overall reach of Kremlin disinformation was reinforced by the absence of policies applied to the broader ecosystem of Kremlin-aligned accounts such as "influencers" and self-proclaimed media that persistently propagate the Kremlin's disinformation, but do not have direct, public links to the Kremlin itself. In the period following Russia's attack on Ukraine, the aggregate audience for the Kremlin-aligned accounts in our sample more than doubled to over 100 million across platforms. These accounts were prolific producers. Users were exposed to their content at least 80 billion times.

[32] Reuters. 2021. "Facebook bans Myanmar military from its platforms with immediate effect." Accessed 2 June 2023. https://www.reuters.com/article/uk-myanmar-politics-facebook-idUKKBN2AP0BO.

While it is important to document the ecosystem-level effects of platform policies, we do not mean to suggest that companies *necessarily* should have imposed bans or demotions on all Kremlin-aligned accounts. However, in the context of systemic risk mitigation – as defined under Article 35 of the DSA – it is clear that mitigation requirements are not defined by the actor but rather by the severity of risk. As we have demonstrated in the Risk Assessment, all of the sub-categories of accounts in our sample – Kremlin-backed and Kremlin-aligned – met the standard of severe and systemic risk. Arguably, the ease at which the Kremlin's disinformation campaign was able to simply "switch" channels was facilitated by a lack of horizontal policies covering all Kremlin-backed accounts, as well as a lack of horizontal policies covering Kremlin-aligned accounts engaging in the Kremlin's disinformation campaign.

(2) Behaviours

Behaviours		Meta	YouTube	TikTok	Twitter	Telegram
Circumvention	Deceptive identities and Rebranding	ban	ban	ban	ban	allow
	Back-up accounts	ban	ban	ban	ban	allow
	Re-channeling audiences	may label* may demote*	ban	may ban	may ban	allow
	Republishing state media content	may label* may demote*	ban	allow	may label* may demote*	allow
Amplification	Mass posting	ban	ban	may ban	ban	allow
	Cross-platform coordination	allow	allow	allow	allow	allow
	Manipulating	ban	ban	may ban	ban	allow

	algorithmic reach					
	Operating networks of accounts	ban	may ban	ban	ban	allow
Suppression	Abusive notifications (brigading)	ban	ban	ban	ban	allow
	Impersonation and identity theft	ban	ban	ban	ban	allow
	Bullying and harassment	ban	ban	ban	ban	allow
	Doxxing	ban*	ban	ban	ban	allow

Circumvention

With the exception of Telegram, the other major platforms had pre-existing policies addressing several of the most common circumvention behaviours. Some added new ones after the invasion. However, they were notably unable to address the re-channeling of audiences. In anticipation of an increasingly restrictive regulatory environment, pro-Kremlin accounts undertook significant efforts to move their audiences from VLOPs to unregulated platforms: 55% of Russian-language accounts and 28% of EU-language accounts in our source list featured outlinks to “alternative” platforms such as Telegram, VKontakte or RuTube in their biography or description. None of the VLOPs had effective or comprehensive policies to mitigate this circumvention behaviour, which we have labelled “re-channeling”.

Specifically in response to the war, Meta and Twitter introduced new policies to label and demote posts that contained links to Russian state media websites. However, these policies seemingly did not apply when an outlink instead sent a user to the same state media outlet’s account on another online platform, such as Telegram. By the same token, YouTube, TikTok and Twitter all have more general policies that ban users from redirecting their audiences to some types of external content if it violates the VLOP’s own policies. For instance, YouTube prohibits users from posting “links to content that would violate our hate or harassment policies if uploaded to YouTube.” In addition, all of the VLOPs have a policy that bans users from setting up back-up accounts on the same platform if the goal is to bypass a policy or enforcement action. However, no platform keeps users from linking to back-up accounts on other platforms, even if the aim is to bypass the company’s policy. In sum, none of the

aforementioned policies prevent users from re-channeling their audiences to Kremlin-backed accounts on other platforms. This policy gap has likely contributed to the rapid growth of Telegram as a key distribution hub for Kremlin disinformation: The audience of pro-Kremlin accounts on Telegram grew more than threefold after the invasion of February 2022 to 73 million total subscribers.

Meta's and Twitter's tailored policies on republishing Russian state media content had an additional scope limitation: They only covered the posting of URLs, but did not set out to limit the re-publication of state media content in other formats, such as by re-uploading audiovisual content from RT or Sputnik or simply copy-pasting their textual content into new posts. In addition, the effectiveness of these policies was difficult to assess because neither platform disclosed lists of the state media outlets to which they applied their demotion and labelling policies.

Amplification

Except Telegram, all platforms have policies that partly or fully ban mass posting, a behaviour that pro-Kremlin accounts frequently employed to artificially inflate the visibility of pro-Russian content. Besides Telegram, all platforms have policies addressing other coordinated behaviours that pro-Kremlin accounts frequently employed to boost their content, including a range of tactics to manipulate algorithmic reach,[33] and the operation of networks of accounts. However, we noted that relevant policies only addressed behaviours where coordination occurred *on* the platform – for example, where a network of several Facebook accounts was run by the same account owner to drive up the engagement metrics on pro-Kremlin posts. Meta's recent adversarial threat reports have noted the trend of increased cross-platform coordination,[34] yet no platform had a policy to address off-platform coordination, where large numbers of pro-Kremlin accounts used a joint hub, such as a Telegram channel CyberFront Z, to coordinate attacks on the accounts of pro-Ukrainian voices on VLOPs.

Suppression

All platforms but Telegram have policies that ban the behaviours of pro-Kremlin accounts employed to suppress opposing voices or perspectives. Meta even extended its policy on doxxing in May 2022 by prohibiting content that reveals the identity or location of a Prisoner of War in the context of an armed conflict. In September 2022, the policy was further extended to include content that puts a defector at risk by outing the individual with personally identifiable information when the content is reported by credible government channels.

[33] Notably, Telegram is the only platform in our sample that is not based on algorithmic content curation.
[34] https://about.fb.com/wp-content/uploads/2023/05/Meta-Quarterly-Adversarial-Threat-Report-Q1-2023.pdf, accessed 2 June 2023.

(3) Content

Content		Meta	YouTube	TikTok	Twitter	Telegram
Promotion of hate and violence	Glorification of the war and war crimes	may ban	may ban	may ban	may ban	allow
	Violent content	may ban	may ban	ban	may ban	allow
	Discriminatory or dehumanising content	ban	ban	ban	ban	allow
Deception	Denial of war crimes	ban*	ban	ban	ban	allow
	Political disinformation	may ban	may ban	may ban	may ban	allow
	Cheap fakes	may ban	may ban	ban	may ban	allow
	Deep fakes	ban	ban	ban	ban	allow

Promotion of hate and violence

We noted above that platforms have appeared reluctant to introduce account-level policies targeting all Russian government actors. A similar pattern emerges in relation to content policies. While several platforms have rules that explicitly prohibit the praise or promotion of extremist or terrorist organisations, no policies exist to limit the praise or promotion of a government even when it is killing civilians. Similarly, while platforms prohibit "statements that advocate for high-severity violence" (Meta), no explicit policies exist to dissuade users from advocating for a war or – as Kremlin propaganda calls it – a "special operation". Twitter prohibits praise for "individual perpetrators of violent attacks" – however, it seems unlikely that an entire government would fall under this definition. In short, the policies against promoting

and inciting violence that are common among major platforms were not applied consistently to these behaviours by Russian government actors or in praise of Russian government activities. Again, this contrasts pointedly with the posture taken by Meta vis-a-vis the military government in Myanmar.

Deception

As opposed to misinformation, disinformation is defined by the speaker's intention – an intention to deceive and cause harm. A review of relevant content policies suggests that platforms are reluctant to evaluate the status and intent of the speaker in assessing content risks – contrary to what a human rights based approach such as the Rabat proportionality test would recommend. In fact, not a single platform has any explicit policy on disinformation or even state-backed disinformation in their Terms and Conditions despite most of them being signatories to the EU's Code of Practice against Disinformation.[35] However all platforms have rules against *mis*information, and thus on specific types of false content, independently of the actor who is propagating it. The result is a permissive posture towards persistently misleading and conspiratorial narratives promoted by pro-Kremlin accounts to very large audiences. This suited Kremlin propagandists, whose first priority of disinforming the public is followed closely by the objective of casting a shadow of confusion and doubt over widely reported facts.

In general, Meta was the only platform to change a content policy in response to the war: It specified its policy on denial of war crimes by prohibiting "governments that have instituted sustained blocks of social media to use their official departments, agencies, and embassies to deny the use of force or violent events in the context of an attack against the territorial integrity of another state in violation of Article 2(4) of the UN charter."[36] At the time of introduction, the Russian government was the only government worldwide to which this specification applied. While this policy change sets Meta favourably apart from the other platforms in terms of the scope of its mitigation measures, it is worth noting that the additional restriction applied only to a very specific disinformation narrative – the denial of the use of force. Conspiracies that celebrated the Russian use of force under false pretences were not similarly restricted.

This lack of horizontal policies, and the gaps within existing policies, have allowed for actors, behaviours, and content that are part of the Kremlin's disinformation campaigns to spread across the platform ecosystem. This suggests that the platforms' Terms and Conditions have not accounted for, or appropriately addressed, the systemic risks outlined in Article 34(1). They therefore likely do not serve as reasonable, proportionate and effective mitigation against them.

ii. Preparedness and Transparency

Before we look at platform policies in practice, it is worth spending a moment on the questions of preparedness and transparency. Platforms were not unaware that the Russian government was engaged in state-sponsored information operations prior to the February 2022 invasion. The Kremlin's disinformation campaign targeting Ukraine, the EU and its partners has been

[35] "The 2022 Code of Practice on Disinformation." Accessed on 2 June 2023. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.
[36] Facebook's Policy on "Inauthentic Behaviour." Accessed on 2 June 2023. https://transparency.fb.com/en-gb/policies/community-standards/inauthentic-behavior/.

documented on online platforms at least since 2014, when Russia illegally annexed Crimea. What's more, governments and experts alike had been calling on online platforms to increase their content moderation efforts months before Russia attacked all of Ukraine in February 2022. In September 2021, Ukraine's Minister for Digital Transformation Mykhailo Fedorov travelled to the United States to personally ask Google to open offices in Ukraine, following reports that the company had been conducting much of its Ukrainian content moderation from its Moscow offices.[37]

It would therefore seem reasonable to expect that online platforms had increased their content moderation capacities and fine tuned their internal processes to prepare for a possible full-blown war. However, we see little evidence that the platforms were prepared to deploy the necessary resources to apply mitigation measures effectively in Central and Eastern European languages. Therefore, despite the transparency of the policies described in the preceding sections, the companies did not have the resources (technical, human or financial) deployed to apply them consistently, even when they were present.

On the question of transparency, we can say on the one hand that policies are described in Terms and Conditions and public statements. But, on the other hand, it is a monumental task simply to compile all of the policies and to get clarity on what they mean in practice. The companies' Transparency Reports should add significant detail and provide a more comprehensive picture. In most cases, these are incomplete at best. Transparency reports focused on enforcement of platform rules are published quarterly by Meta, YouTube, and TikTok, and twice a year by Twitter.[38] These reports provide general statistics on platform moderation actions and provide no details about particular content or accounts. The reports do not include metrics on the speed of content moderation (apart from TikTok, which provides a share of removal within 24 hours only) or the processing speed of appeals. Though the timeliness of moderation efforts has a huge impact on whether risk persists or is averted, the absence of these metrics by platforms makes temporal evaluations nearly impossible using these reports.

The adequacy of these platform transparency reports for the measurement of actual risk mitigation fails on other metrics as well. By and large, the platform transparency reports do not provide geographic breakdowns of their moderation efforts. Meta does not cover this aspect at all, and TikTok and YouTube only provide broad numbers on volume of video removals across 50 and 30 countries, respectively, without providing any data about the proportion these videos represent compared to the overall number of videos published on the platform. No platform provides data about appeals, which is an important metric that could shed light on false-positive removals—and thus the accuracy and precision of platforms' content moderation processes and systems. As of February 2023, Twitter—which used to publish transparency reports approximately every six months—has not released a report covering any period of time

[37] Promote Ukraine. 2021. "Opening of Google, YouTube Offices in Ukraine to Help Fight Russian Propaganda." Accessed 2 June 2023. https://www.promoteukraine.org/opening-of-google-youtube-offices-in-ukraine-to-help-fight-russian-propaganda/.

[38] https://transparency.fb.com/data/community-standards-enforcement/; https://transparencyreport.google.com/youtube-policy/removals?hl=en; https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2022-3/; https://transparency.twitter.com/en/reports/rules-enforcement.html#2019-jul-dec, all accessed 2 June 2023.

since December 2021. It is unclear whether this is a result of recent company layoffs impacting key staffers responsible for transparency reporting.

To piece together the details of policy implementation that platforms do not share, we were compelled to engage in painstaking, manual investigation. In many cases, affected accounts were identified through reverse engineering, by observing sudden drops in engagement on particular channels and tracing them back to respective policy announcements. Additionally, none of the platforms provide a list of accounts that have been attributed specific labels, such as "Russian state media", nor a list of countries where this labelling is being applied. Platforms do not offer a straightforward way to check when an account was first labelled, except by relying on archived links from public domains, such as the Internet Archive's Wayback Machine, or by deduction from media articles. This lack of transparency complicates empirical analyses of the platforms' policy interventions and their effectiveness in reducing the reach of harmful content and state media actors.

Particularly on YouTube, prominent Russian-language channels disappear altogether without any accompanying communication by the platform, leaving questions as to why a given channel was suspended, when it happened and what content was available on the channel previously. From our sample, 75 out of a list of monitored 329 YouTube channels are currently either unavailable or terminated, without clear notices as to why. While YouTube's account suspensions seem at first sight desirable, this lack of transparency makes it difficult to determine the reasons why accounts were closed and therefore to assess whether the actions taken fulfil the threshold of Article 35(1) or whether the policies were applied consistently.

iii. Content Moderation Measures

In almost all cases, the policies described in the preceding section that platforms impose through Terms and Conditions are either some form of content moderation (removal or labelling) or a change to algorithmic recommendation that reduces visibility of content on the platform (demotion, removal from search, removal from recommendations). Adjustments at the level of Terms and Conditions can be a necessary, but never a sufficient, mitigation measure in response to any systemic risk. Only if policies are applied swiftly and consistently can they fulfil their stated objective. We begin here with an evaluation of content moderation measures.

To assess how effectively platforms applied their Terms and Conditions across geographies and language areas, ranging from Ukraine, to Bulgaria, Czech Republic to Germany or the English-speaking world, we designed and conducted standardised investigations across all platforms. To do so, we turned the risk metrics, which we applied in our risk assessment, into risk mitigation metrics. Thus, instead of audience size, engagement or exposure, we would measure mean changes in audience size, engagement or exposure. Rather than looking at the prevalence of content, we look at the speed and consistency of content moderation.

Applying risk mitigation metrics to content moderation

Looking at our pro-Kremlin sources, we compared how these accounts performed between two periods of time: Phase 1 (1st December 2021 and 20th February 2022) and Phase 2 (1st April 2022 to 30th November 2022) – before and after the full-scale invasion of Ukraine. Essentially, we analysed whether and how the platforms have put into place measures to

mitigate against the systemic risks outlined in Article 34(1). We examined specific mitigation measures described in Article 35(1). This allowed us to determine to what extent the platforms actually applied their own policies, and whether their actions in relation to content moderation and algorithmic recommender systems thereby constitute reasonable, proportionate, and effective measures (Article 35(1)) to mitigate against the systemic risks described in Article 34(1).

platform	Number of Accounts	Volume	Audience size	Exposure*	Engagement	VolumeΔ	Mean ExposureΔ	Mean EngagementΔ	SubscriberΔ
Facebook	430	1,080,000	105,600,000	3,320,000,000	227,000,000	-35%	-84%	-37%	2%
Instagram	289	117,000	36,400,000	728,000,000	100,000,000	-47%	-65%	-72%	4%
Telegram	744	5,460,000	72,800,000	78,100,000,000		96%	194%		308%
TikTok	138	17,500	17,300,000	4,120,000,000	130,000,000	-6%	-73%	-68%	128%
Twitter	388	1,740,000	64,800,000		101,000,000	-28%		54%	3%
YouTube	251	52,600	29,600,000	3,130,000,000	150,000,000	4%	60%	84%	88%

Figure 18: Number of accounts per platform and select risk metrics, including both raw numbers and the changes observed between the periods 1 December 2021 – 20 February 2022 and 1 April 2022 – 30 November 2022.

The most important measure of whether a content moderation policy was effective comes down to the changes in exposure and engagement levels. This is simply because of the importance of scale in the proportionality test that determines severity of risk. The table below provides a top-line summary of the performance of the pro-Kremlin accounts that we tracked in our sample throughout the period of research.

We observe consistently across this study that policies effectively applied to the narrow category of Kremlin-backed media did *not* lead to significant overall reductions in the audience size or engagement levels of pro-Kremlin narratives on the platforms. Within platforms, the audiences shifted to the channels to which no mitigation measures were applied, including networks of smaller accounts, which grew steadily in size and relevance as a result: In December 2021-February 2022, accounts with 300,000 or more followers accounted for 70% of all engagement. From March onwards, this proportion more than halved. While platform policies disproportionately targeted the largest accounts, content from small and medium sized accounts filled the emerging vacuum, quickly increasing the audiences and engagement that these accounts received.

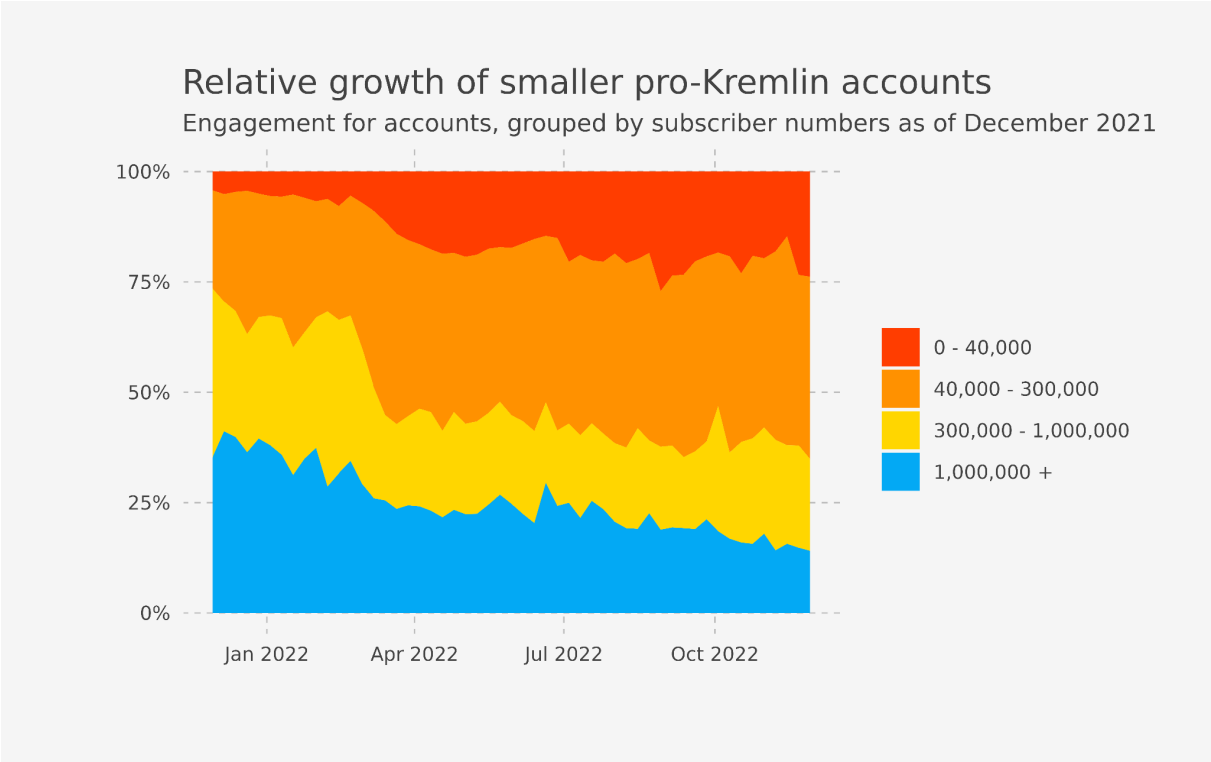

Figure 19: Proportion of engagement for accounts ranked according to the number of followers in December 2021.

Across the platform ecosystem, audiences gravitated to the least restrictive spaces. Telegram stands out as the main beneficiary from movements away from other platforms. There has been a dramatic increase in attention paid to pro-Kremlin content on that smaller platform. Exposure overall has more than doubled, and subscriber counts of Kremlin-backed accounts increased by roughly 450% and of Kremlin-aligned accounts by more than 250%. YouTube also enabled pro-Kremlin actors to propagate disinformation about the war: For both Russian and English, they have dramatically increased both their average exposure and subscriber counts across almost all categories of Kremlin-backed and Kremlin-aligned YouTube accounts.

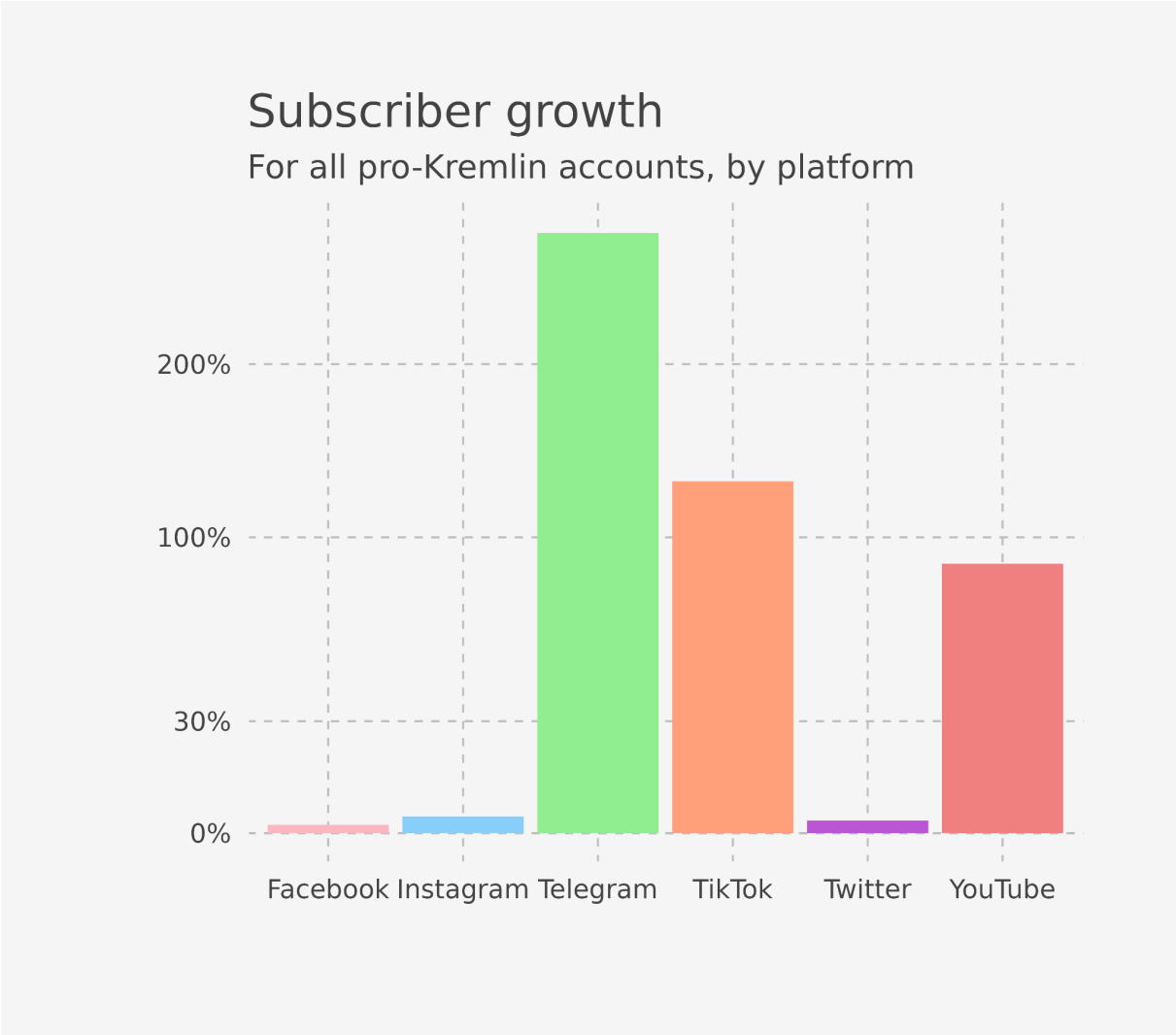

Figure 20: Subscriber growth for all pro-Kremlin accounts between February 2022 and December 2022 by platform.

These numbers paint a picture of outcomes. But the levers pulled by the platforms to affect these outcomes represent a variety of techniques of content moderation and changes to algorithmic recommendation. We discuss some of these in the subsections that follow.

Reactive content moderation (notice and action)

There are two forms of content moderation undertaken by platforms – proactive and reactive. Proactive moderation means the platform applied its policy unilaterally to remove or label content. This type of moderation is profiled in the geo-blocking and labelling analyses below. Reactive moderation means that the platform applied its policy (action) in response to a flag initiated by a user (notice) through the application's interface. In the case of reactive moderation, the key metrics are these: the speed and probability of response, the speed and probability of action taken on the content, the consistency of the actions taken on similar content (including across languages), and the impact on the overall level of exposure and engagement on this content.

To test the performance of major platforms in reactive moderation, we conducted an experiment on Facebook, YouTube and Twitter across multiple languages – Czech, Slovak,

Hungarian, Bulgarian and Ukrainian – in the summer of 2022. First, we ran the Perspective API across a large volume of posts from our source list. We selected a subset of content that scored above 95% probability for containing an intention to inflict pain, injury or violence against an individual or group – a violation of Terms and Conditions for all platforms. Posts confirmed by local experts as violating Terms and Conditions were reported using user-interface flags.

The results demonstrated wide variation in the application of reactive moderation policies across languages and platforms. To take Twitter as an example in the chart below, moderation was almost non-existent in Czech. Reported tweets in other languages on average received at least an acknowledgement in three quarters of cases. These experiments based on limited samples of user flags show significant shortcomings in policy enforcement that indicate the need for more systemic research in the future with larger data sets.

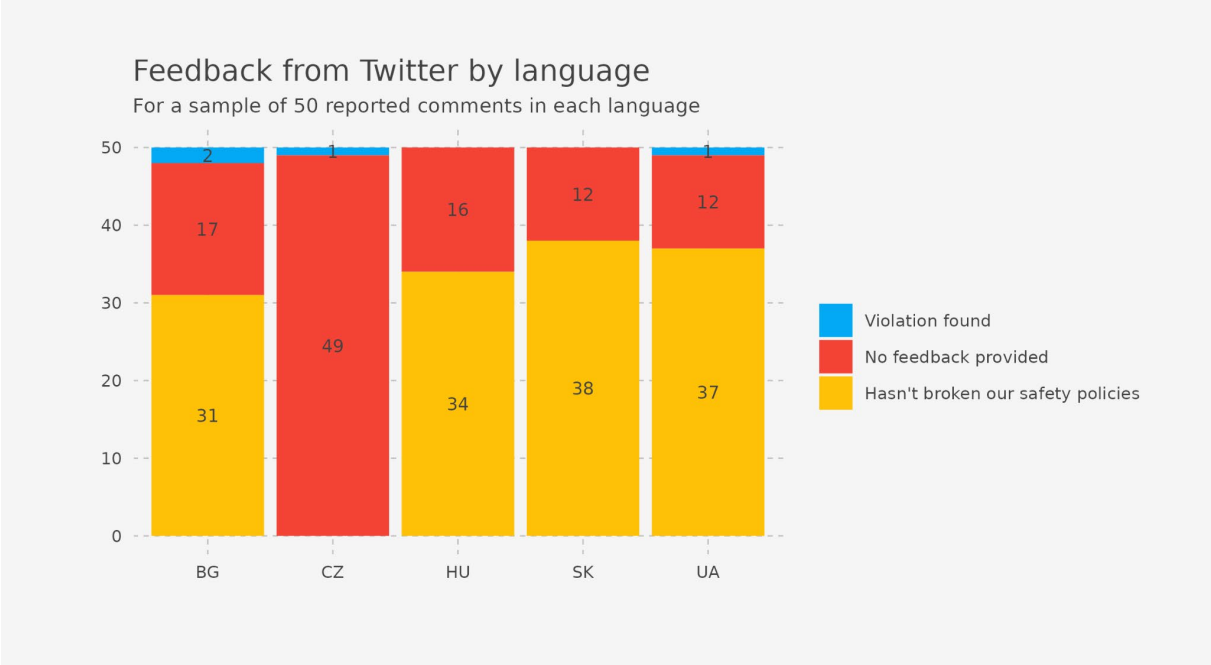

Figure 21: Feedback from Twitter on reactive content moderation pertaining to a sample of 50 reported comments, broken down by language: Czech, Slovak, Hungarian, Bulgarian and Ukrainian.

The **quality and adequacy of feedback** from the platforms also varied considerably. Facebook removed approximately 20% of the reported content within seven days. Nonetheless, the most common response received from the platform was either no response at all, or that they were unable to review the content. In Hungarian, only 8% of violative comments were reviewed and removed, compared to 28% for Bulgarian and Ukrainian. Even these low levels of reactive content moderation on Facebook compared favourably with YouTube. YouTube offered zero feedback on any flagged comment in any language. 99% of the accounts that had posted violent comments remained active in January 2023.

Geo-blocking:

Platform	March package			June & December packages			Together		
	Un-available	Total	Geo-blocked	Un-available	Total	Geo-blocked	Un-available	Available	Failed to geo-block
Facebook	47	50	94%	5	7	71%	52	5	9%
Instagram	32	38	84%	4	6	67%	36	8	18%
Twitter	17	19	89%	4	7	57%	21	5	19%
YouTube	39	39	100%	5	5	100%	44	0	0%
Telegram	37	40	93%	0	9	0%	37	12	24%
TikTok	12	16	75%	1	7	14%	13	10	43%

Figure 22: Total number and percentage of geo-blocked accounts belonging to sanctioned media outlets (on 23 February 2023).

Following Russia's full-scale invasion of Ukraine, the EU introduced several rounds of sanctions that included measures to suspend the broadcasting activities of Russian state media outlets in the EU (we refer to these sanctions by the month they were introduced, e.g. “March sanctions”).[39] In response, all platforms adapted their policies to geoblock accounts operated by Russian state media. Our research suggests that platforms implemented their geoblocking policies with different levels of consistency: By 23 February 2023, Facebook had failed to geoblock 9% of the state media accounts impacted by the three rounds of EU sanctions, whereas TikTok had failed to geoblock 43% of the affected accounts.

We were able to analyse the effectiveness of geoblocking by comparing mitigation metrics for the RT and Sputnik accounts in different languages. The chart below demonstrates the reduction in engagement across platforms. The reduction in exposure, engagement, and audience growth were particularly marked for state media accounts that had the majority of their audiences inside the EU, such as RT Deutsch and RT France (*rtenfrancais*). For these accounts, geoblocking led to near complete disruption of global traffic. The average engagement on tweets by RT France collapsed by more than 80%, and RT Deutsch on Instagram saw the average number of weekly engagements fall from 1720 to 32.

[39] Council Decision (CFSP) 2022/351 of 1 March 2022 amending Decision 2014/512/CFSP concerning restrictive measures in view of Russia’s actions destabilising the situation in Ukraine *OJ L 65, 2.3.2022, p. 5-7*.

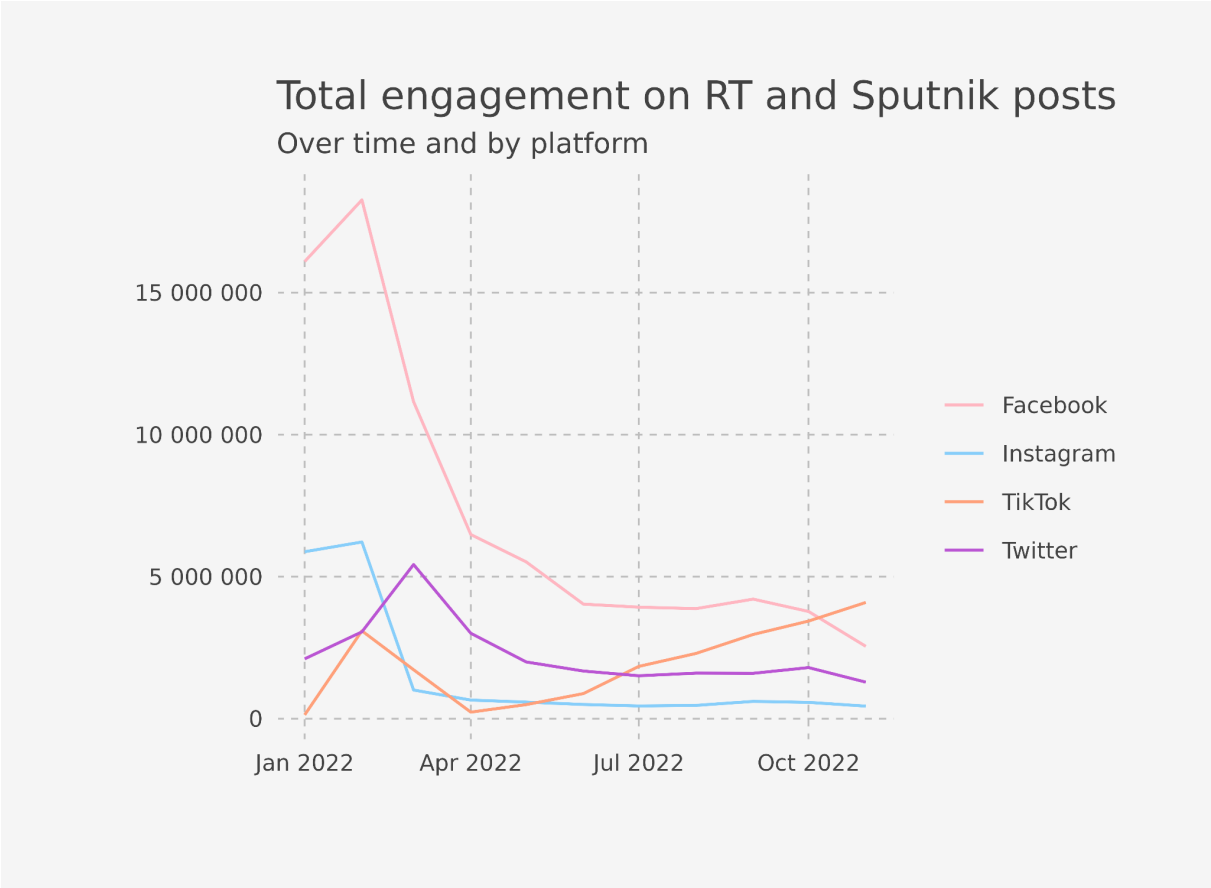

Figure 23: Total worldwide engagement with content by RT and Sputnik over time and by platform.

Labelling:

Platform	March package			June & December packages			Together		
	Labelled	Total	%	Labelled	Total	%	Labelled	Failed to label	Failed to label
Facebook	49	50	98%	6	7	86%	55	2	4%
Instagram	31	38	82%	5	6	83%	36	8	18%
Twitter	18	19	95%	5	7	71%	23	3	12%
TikTok	12	16	75%	2	7	29%	14	9	39%

Figure 24: Total number and percentage of labelled accounts belonging to sanctioned media outlets (on 23 February 2023).

The geoblocking of state media only applied to users inside the EU. At global level, all platforms—with the exception of YouTube and Telegram—additionally began applying labels to state media accounts. Meta, Twitter and TikTok handled labelling in different ways, with similar levels of scope but varying levels of consistency. Firstly, all four platforms defined state media as media outlets that are under the editorial control of the state, excluding outlets that are publicly funded, but editorially independent.

No platform disclosed lists of the accounts they considered to be within the scope of their definitions. In order to assess the consistency of labelling, we therefore collected a list of accounts affiliated with the Russian state media outlets the EU targeted with sanctions.[40]As regards the consistency of labelling, our analysis suggests that Facebook performed best, whereas TikTok performed significantly worse than other platforms. Subsequent analysis revealed that Facebook failed to label 4 % of the accounts affiliated with Russian state media, whereas TikTok failed to label 39 %.

iv. Algorithmic Recommender Systems

In practice, changes in quantitative metrics such as audience growth rates, exposure or engagement are driven by an interplay of content moderation and algorithmic recommender systems. Arguably, platform curation systems that are tuned primarily for attention capture and maximum engagement levels override narrowly focused content moderation.[41] However, the conditions for studying recommender systems are unfavourable for independent research. Without access to the methods of algorithmic ranking proprietary within each platform, it is difficult to measure the impact of changes to recommendation systems. It must be done with proxy variables. In order to isolate the effect of recommendation systems on content distribution more accurately, we have developed the **non-follower engagement (NFE) metric**.

Our NFE metric captures a measure of how many users interacted with material as a result of content recommendations or algorithmic sorting, rather than subscriptions (i.e. a user choice to follow a particular account). This allows for an approximation of the extent to which non-subscribers interact with content on the bases of algorithmic cues. NFE is a computationally intensive metric that requires cross-tabulating a list of engagers with a list of subscribers. This is challenging for outside researchers, as no platforms currently subdivide engagement metrics into subscribers and non-subscribers. To analyse the effects, and shortcomings, of the platforms' demotion measures, we have closely reviewed the platforms' de-amplification of Russian government accounts and state-media accounts – which are subsets of our Kremlin-backed source list.

By studying NFE, we found that official Russian government accounts – such as embassy accounts – seemed to enjoy a very considerable increase in algorithmic promotion in the period after the invasion of Ukraine and the geo-blocking of Russian state media. Take for example the Twitter account of the Russian Embassy in Germany. The Embassy's posts went from averaging less than 30 engagements to more than 300 from mid-March 2022 onwards – with a very large percentage coming from non-followers. In general, accounts that received the highest non-follower engagements tended to be for small accounts without a large, pre-existing subscriber base.

In aggregate, Russian embassy accounts from our sample show a marked increase in NFE during February and March 2022. The average engagement for these accounts increased by a factor of 5 on Twitter and 2.5 on Facebook. By contrast in the period before the February

[40] Council Decision (CFSP) 2022/351 of 1 March 2022 amending Decision 2014/512/CFSP concerning restrictive measures in view of Russia's actions destabilising the situation in Ukraine *OJ L 65, 2.3.2022, p. 5-7*.
[41] See, e.g. Allen, Jeff. 2022. "Misinformation Amplification Analysis and Tracking Dashboard." Integrity Institute. Accessed 2 June 2023. https://integrityinstitute.org/blog/misinformation-amplification-tracking-dashboard.

invasion, Russian state institution accounts received only slightly more non-follower engagements than engagements by followers. It is very unusual to see the levels of NFE spike like this. By comparison, roughly 1 in 4 engagements for influencer accounts and 1 in 3 for media accounts were by non followers. The conclusion we draw is either that one or more influence factors are at play: 1) Russian operators gamed the system with fake engagements to drive up these numbers; 2) inflammatory content promoted by the Embassy accounts in the wake of the geo-blocks on Russian state media were rewarded by algorithmic recommendation systems with greater distribution and hence greater engagement; 3) one or both of the previous factors were operating and served to raise algorithmic weighting of the accounts into a cycle of virality. It is possible all of these things happened at once and were mutually reinforcing.

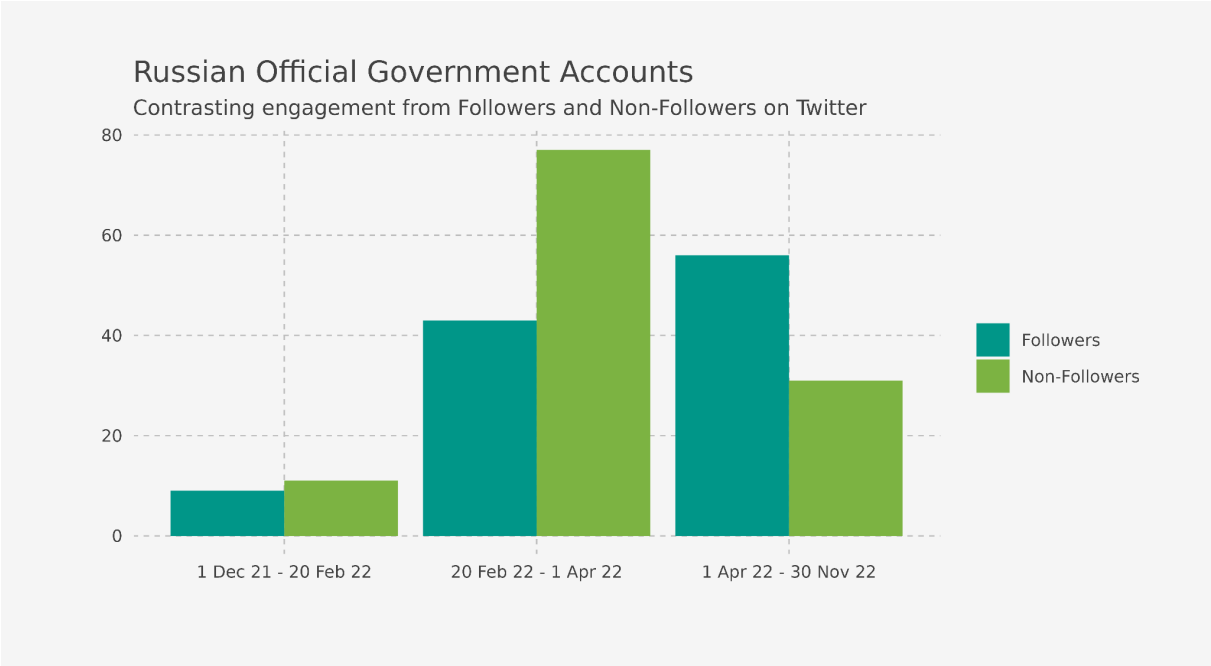

Figure 25: Engagement from followers contrasted to engagement from non-followers on Twitter during the periods 1 December 2021 – 20 February 2022, 20 February 2022 – 1 April 2022, and 1 April 2022 – 30 November 2022.

We can also see the phenomenon of re-channeling appear in the NFE data whereby non-moderated channels not only saw increased follower numbers but also increases in non-follower engagement. For example, the graph below tracking NFE on YouTube at first suggests that the platform enforced its policies pertaining to Russian state media effectively. However, much as we have seen across many platform mitigation measures, because its policies did not cover official government accounts, Kremlin staffers, and Kremlin-aligned influencers, NFE for content posted by these other account types increased. This increase offsets the declining engagement for Russian state media accounts in particular. In the case of YouTube, the channels that picked up the traffic moving away from the blocked state media outlets doubled in audience size (subscribers) on average and exposure increased by 70%. This appears to be again a mutually reinforcing virality cycle where increases in audience size and inflammatory content from rechanneling yields a benefit in algorithmic recommendation, which in turn promotes NFE, and so on. In some cases, the content on these channels is simply re-packaged material that would otherwise be viewed on the state media channel. Viewed from

this perspective, it appears far less clear that YouTube reasonably, proportionably, and effectively mitigated the systemic risks outlined in Article 34(1).

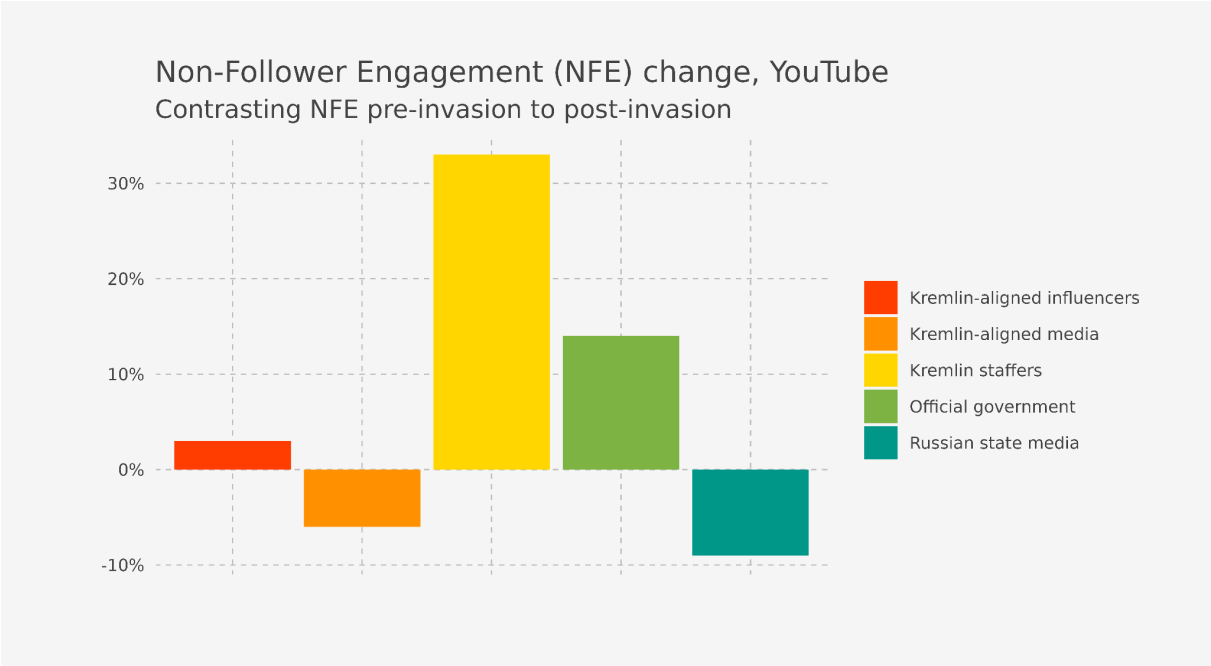

Figure 26: NFE pre-invasion (1 December 2021 – 20 February 2022) contrasted to NFE post-invasion (1 April 2022 – 30 November 2022) on YouTube.

The centrality of algorithmic recommendation as a variable for understanding changes in audience and engagement is especially important for YouTube. According to numbers from 2018, 70% of YouTube views are recommended by the platform's algorithm.[42] YouTube announced in March 2022 that it would uprank authoritative sources about the war, but there is little information about what else the platform has done to affect how content recommendations are made. The data from our modest sample shows that 90% of comments under videos posted by Kremlin-aligned influencers did not come from subscribers to the channel, meaning that users had arrived at this content by some other means, be that through a recommendation external to the platform (search engine, group chats, etc), through internal YouTube search, or content algorithmically tailored to the user and recommended on the platform's landing page, sidebar, or at the end of videos. In the context of ongoing Russian information operations and coordinated trolling, a high proportion of non-follower comments may also be an indicator of mass posting campaigns. Because the systemic risks described in Article 34(1) do not only emanate from Kremlin-backed but also from Kremlin-aligned accounts, our data therefore suggests that YouTube may not have done enough to fulfil the threshold of Article 35(1)(c) to adapt its algorithmic systems, including its recommender systems.

[42] Rodriguez, Ashley. 2018. "YouTube's recommendations drive 70% of what we watch." Quartz. Accessed 2 June 2023. https://qz.com/1178125/youtubes-recommendations-drive-70-of-what-we-watch.

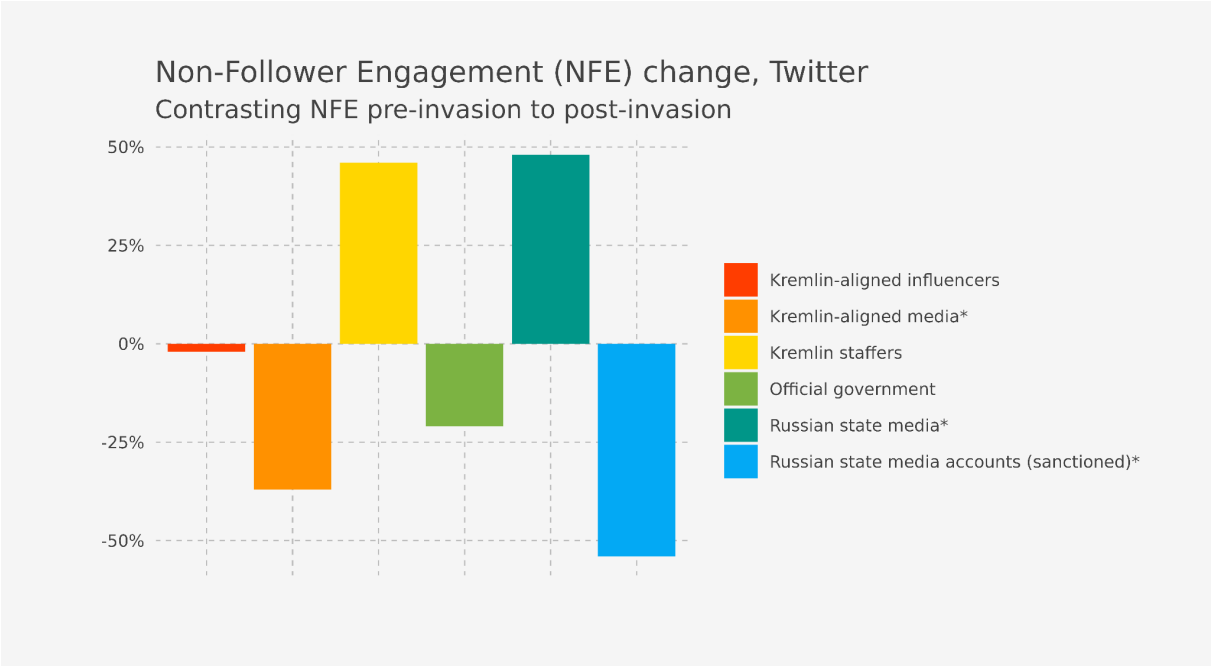

Figure 27: NFE pre-invasion (1 December 2021 – 20 February 2022) contrasted to NFE post-invasion (1 April 2022 – 30 November 2022) on Twitter.

The NFE data for Twitter also demonstrates that algorithmic de-amplification had a significant impact not only on sanctioned Russian state media inside Europe but also on Russian government accounts. However, again, it appears that the global state media accounts and those of Kremlin staffers picked up some of this slack. In April 2022, in reaction to Russia's war against Ukraine, Twitter announced it had taken measures to restrict the reach of Russian state-backed accounts. According to our data, these actions resulted in a 28% reduction in the number of retweets for a sample of Russian official sources.[43] The number of non-followers engaging with content from the Russian MFA's Twitter account was reduced by approximately 50% compared to the number prior to demotion. These findings demonstrate not only the effectiveness of these measures but also the speed of their impact. A month later in May 2022, Twitter confirmed (with numbers similar to our findings) that its selected mitigation measure of algorithmically demoting specific accounts linked to the Russian government had been effective:

> *In April, we announced that we would not amplify or recommend government accounts of states that limit access to free information and are engaged in armed interstate conflict, beginning with Russian government accounts. Our approach has proven effective – for Russian government accounts, early results show that;*
>
> - *Engagements per Tweet decreased by approximately 25%;*
> - *The number of accounts that engaged with those Tweets decreased by 49%.*[44]

[43] Russian Embassy in London, Russian MFA (en), Russian MFA (RU), and Russian MOD.
[44] Twitter. 2022. "Our ongoing approach to the war in Ukraine." Accessed 2 June 2023. https://blog.twitter.com/en_us/topics/company/2022/our-ongoing-approach-to-the-war-in-ukraine.

In April 2023, and therefore after the period of analysis relevant for this report, Twitter decided to reverse the de-amplification of Kremlin-backed accounts.[45] Nonetheless, prior to the reversal of its decision, Twitter identified a systemic risk and applied at least partially effective mitigation measures that are likely to have fulfilled the threshold of Article 35(1)(c).

5. Conclusion

The application of the Baseline Framework for evaluating digital risk management in the case of Kremlin disinformation operations yielded clear conclusions. This contextual data delivered important insights into the standards of assessment that may be applied to Article 34 risk categories as well as the Article 35 evaluations of effective mitigation.

We applied the Rabat proportionality test to the content observed in our sample, subdividing the analysis into a focused examination of the risks of illegal content, risks to fundamental rights, and risks to electoral processes, public security and civic discourse. In each case, we demonstrated the severity level of the risk by presenting a qualitative assessment of the content in context along with a quantitative assessment of scale.

We found widespread examples of content carrying high levels of toxicity – in some instances rising to the level of illegality – that reached very large audiences. For example:

- Russia's Z propaganda campaign in March 2022 reached tens of millions in the EU with billions of exposures of videos, images and text across all major platforms.
- Incitement to violence against Ukrainians was a consistent content theme in the data, including direct and explicit calls to murder and maim.
- Dehumanising content using gender and ethnicity to radicalise audiences appeared repeatedly in the data.

We found coordinated attempts to suppress the rights of others and manipulate civic discourse.

- Pro-Kremlin actors engaged in overt, cross-platform campaigns to target and harass individuals and institutions with mass-posting.
- Organised campaigns to flag the social media pages of Ukrainian news outlets with abusive notifications of transgressive activity resulted in very significant declines in the viability of these channels.
- Pro-Kremlin actors created social media channels for fake news outlets and inauthentic clones of actual news outlets to deceive audiences and distort public information.

The scale of pro-Kremlin disinformation campaigns reached very large audiences – content from these accounts was exposed to users billions of times and drew hundreds of millions of engagements (likes, shares and comments).

[45] Kerr, Dara. 2023. "Twitter once muzzled Russian and Chinese state propaganda. That's over now." NPR. Accessed 2 June 2023. https://www.npr.org/2023/04/21/1171193551/twitter-once-muzzled-russian-and-chinese-state-propaganda-thats-over-now.

We assess that across all the risk categories identified in Article 34 of the Regulation, the proportionality test of severity and systemic risk has been met by the combination of qualitative analysis of the content and quantitative evaluation of the reach and probability of harm.

We find that the mitigation measures applied by the platforms were largely ineffective. Platforms were unprepared to meet the mitigation demands of information warfare. Their policies, though transparent, were insufficient to respond to the crisis. There were instances of effective mitigation that targeted very specific accounts and reduced the risk level for the audiences of those channels. However, at the systemic level of all accounts on the platforms engaged in Kremlin disinformation campaigns, the mitigation measures failed.

- We found that reductions in exposure and engagement for Russian state media channels that were blocked and labelled by platforms did not achieve an overall decline in the audience for pro-Kremlin disinformation.
- These restrictions were circumvented by hundreds of other channels carrying similar or identical content which attracted an even larger audience during the period when mitigation measures were applied to constrain Russian disinformation.
- The re-channeling of these audiences was achieved through a combination of direct action by pro-Kremlin actors and through algorithmic recommendation by the platforms.

We conclude that Article 35 standards of effective risk mitigation were not met in the case of Kremlin disinformation campaigns.

6. Appendix

This Appendix contributes more detail to the Baseline Framework that we used for empirical analysis of platform risk assessment and mitigation measures related to Russian information operations after the invasion of Ukraine. The following sections provide a taxonomy of the categories of Actors, Behaviours, and Content, broken down into sub-categories with additional information about definitions.

a. Actors

Since the beginning of Russia's full-scale invasion of Ukraine, we monitored accounts that were overt about their identities and under the direct control of the Russian state as well as accounts that appeared to be independent but participated in the Kremlin disinformation campaigns. We catalogued the actors in our risk assessment as either **Kremlin-backed or Kremlin-aligned.** The Kremlin-backed category features accounts that self-identify as connected to the Russian government or where we could identify a clear connection to the state using open source data. The Kremlin-aligned category features accounts that self-identify as ideologically aligned with the Kremlin. In many cases Kremlin-aligned actors may be wholly independent, but they consistently express views that mirror the Kremlin's primary narratives about the war in Ukraine or consistently engage with content produced by Kremlin-backed entities.

Actor	Definition	Risk Metrics	Available evidence
Kremlin-backed			
State media (affected by EU sanctions)	Social media accounts linked to Russian state media outlets sanctioned by the EU.	- Volume - Audience Size - Engagement - Exposure - Algorithmic Reach	Near-complete
State media (global)	Social media accounts of all Russian state media outlets (sanctioned and non-sanctioned).	- Volume - Audience Size - Engagement - Exposure - Algorithmic Reach	Near-complete
Official government	Overt-identity entities linked to the Russian state, e.g. MFA Russia, Russian embassies, Russian houses.	- Volume - Audience Size - Engagement - Exposure - Algorithmic Reach	Substantive
Kremlin staffers	Individuals working for Russian state institutions, Russian politicians, and journalists working for Kremlin-backed media.	- Volume - Audience Size - Engagement - Exposure	Partial

i. Kremlin-backed Actors

State media (affected by EU sanctions):

This category covers 168 social media accounts linked to Russian state media outlets sanctioned by the EU. Following Russia's full-scale invasion of Ukraine, the EU introduced a series of sanctions between March and December 2022, three of which covered the transmission and distribution of Russian state media content within the EU, including via online platforms.[46] The first set of sanctions covered Sputnik and RT/Russia Today (including RT English, RT UK, RT Germany, RT France, and RT Spanish) and subsequent sanctions covered seven further outlets including Rossiya RTR / RTR Planeta, Rossiya 24 / Russia 24, Rossiya 1, TV Centre International, NTV/NTV Mir, REN TV and Pervyi Kanal.

State media (global):

This category covers 338 social media accounts of all Russian state media outlets, including those that were not covered in the EU's sanctions packages. Some examples of media outlets covered by this category include Izvestiia, 5TV Channel Russia and Duma TV. Many of these accounts were directly involved in spreading Russian government narratives on social media.

Official government:

This category covers 518 social media accounts of entities that are linked to the Russian state, including the Russian Ministry of Foreign Affairs and Russian embassy accounts. Though many of these accounts were not covered by platforms' policies impacting Russian state actors, our analysis found several examples of these accounts publishing official Kremlin news and statements.

Kremlin staffers:

This category covers 244 social media accounts of individuals working for Russian state institutions, Russian politicians, and journalists working for Kremlin-backed media outlets. Many of these accounts were not covered by platforms' policies impacting Russian state actors. For example, RT Editor-in-Chief Margarita Simonyan has continued to use her personal Facebook account to broadcast official RT content to her network of 70,000 followers.

[46] Council Decision (CFSP) 2022/351 of 1 March 2022 amending Decision 2014/512/CFSP concerning restrictive measures in view of Russia's actions destabilising the situation in Ukraine *OJ L 65, 2.3.2022, p. 5-7*.

ii. Kremlin-aligned Actors

Actor	Definition	Risk Metrics	Available evidence
KREMLIN-ALIGNED			
Media outlets	Accounts of media or news websites (covert or overt identities), which have been active in the Kremlin-aligned ecosystem (i.e., engaging with or spreading pro-Kremlin narratives).	- Volume - Audience Size - Engagement - Exposure	Partial
Influencers	Individuals who share pro-Kremlin narratives (journalists, bloggers) but do not work for the Russian state or media. This category also includes anonymous influential social media channels.	- Volume - Audience Size - Engagement - Exposure	Partial

Media outlets:

This category covers 330 social media accounts of media or news websites which have actively engaged with or spread pro-Kremlin narratives. For example, DonbassItalia is an Italian media outlet with presence on multiple platforms including YouTube (@DNRLNR), Odysee (@DNRLNR), Telegram (@donbassitalia), Twitter (@DonbassItalia). Its primary aim is to distribute news content, videos and 24/7 livestream broadcasts from RT, Sputnik, Rossiya 24, Perviy kanal and more, dubbed or subtitled in Italian, while being “officially” independent of those entities. DonbassItalia effectively circumvents EU sanctions by allowing viewers from the EU easy access to sanctioned Russian state TV channels. Despite consistently spreading content from sanctioned actors, these accounts have not been blocked on any of the platforms and remain freely accessible in the EU.

Influencers:

This category covers 807 social media accounts of individuals who share pro-Kremlin narratives (e.g. journalists, bloggers) but do not (overtly) work for the Russian state or media. This category also includes anonymous influential social media channels, such as UKR Leaks. UKR Leaks, for example, is linked to Russian-Ukrainian Vasiliy Prozorov and was created in 2019 to actively fuel pro-Russian narratives in English and in French. Though it is not related to Russian state media officially, it frequently endorses and shares disinformation from channels such as ukraina.ru.

b. Behaviours

The pro-Kremlin actors we monitored engaged in a wide range of exploitative behaviours that are similar to other state and non-state driven influence operations that have been documented by researchers and platforms. We consulted past reports and worked closely with organisations and researchers on the ground to catalogue twelve specific behaviours the Kremlin used to extend their online presence and ensure broader dissemination of pro-Kremlin narratives. We sub-categorise these behaviours into Circumvention, Amplification, and Suppression tactics based on the intention and impact of these behaviours.

i. Circumvention Behaviours

Circumvention is aimed at disseminating content that violates platform policies among target audiences. We have documented how Kremlin-backed and Kremlin-aligned actors worked together to continue to spread disinformation from media outlets and other actors that had been banned from platforms. The chart below outlines the most common techniques.

CIRCUMVENTION BEHAVIOURS			
Behaviour	Definition	Risk Metrics	Available evidence
Deceptive identities and Rebranding	Tactic used to disseminate content via alternative, mostly newly opened, channels, while misleading audiences as to the real identity and affiliation of these channels.	- Exposure - Engagement - Audience Size	Substantive
Back-up accounts	Tactic used to disseminate content via back-up accounts or mirroring channels without hiding their affiliation with the sanctioned actors.	- Exposure - Engagement - Audience Size	Partial
Re-channelling audiences	Redirecting audiences to less regulated social media platforms, thereby helping promote their content on alternative online spaces.	- Exposure (on both source and target platforms) - Engagement (on both source and target platforms) - Percentage of links to other platforms - Outlinks	Near-complete
Republishing state media content	Republishing content by EU-sanctioned Kremlin-	- Exposure - Engagement	Anecdotal

	backed actors on social media, done by any social media actor regardless of their alignment with Kremlin-backed entities.	- Audience Size	

ii. Amplification Behaviours

Amplification behaviours refer to tactics for disseminating content on and across different online platforms with the intention of amplifying Kremlin-aligned narratives. Amplification *always* involves some level of inauthentic coordination (executed using automated accounts or through coordinated networks of accounts) and is *often* aimed at directly manipulating platform algorithms. Amplification may take up passive forms (liking or sharing content) and active forms (mass posting). When coordinated, amplification behaviours exhibit a number of markers that are directly quantifiable and should be closely monitored by platforms, such as posting frequency, creation date of the accounts, and "copy-pasta" messages – messages that are repeatedly copied and pasted across pages and accounts. Our data tracking amplification behaviours are limited due to the difficulty in monitoring this activity systematically without full access to platforms' data sets.

AMPLIFICATION BEHAVIOURS			
Behaviour	Definition	Risk Metrics	Available evidence
Mass posting	Posting or commenting large numbers of identical or similar content to infiltrate conversations and distort the information environment. Usually, this tactic is employed by coordinated networks of accounts.	- Exposure	Partial
Cross-platform coordination	Coordinating inauthentic posting activity on less popular platforms (e.g. chats) with the aim to infiltrate bigger platforms (public evidence required).	- Exposure	Anecdotal
Manipulating engagement metrics or other algorithm-based features	The intentional manipulation of algorithms with the aim of propelling certain pieces of content to more visible sections (feeds), e.g. by liking posts or creating content around specific hashtags (Twitter trends).	- Exposure - Algorithmic Reach	Substantive

Operating networks of accounts	Running networks of artificially coordinated accounts active on platforms.	- Exposure	Anecdotal

iii. Suppression Behaviours

Suppression behaviours aim to silence or reduce the reach of perspectives that diverge from Kremlin-aligned narratives. Suppression tactics are mostly carried out against individuals and can often take a significant psychological toll – particularly because affected individuals may be threatened and attacked both on- and offline simultaneously. Targets of these tactics are often blocked or restricted by platforms.

SUPPRESSION BEHAVIOURS			
Behaviour	Definition	Risk Metrics	Available evidence
Mass reporting (brigading)	Submitting coordinated notifications to flag invented or inflated allegations, in an attempt to get targets suspended from the platform.	- Volume of reports - Denial of service	Anecdotal
Impersonation and identity theft	Creating fake accounts or websites to impersonate reliable organisations or real people.	- Exposure - Algorithmic reach	Anecdotal
Bullying and harassment	Targeting individuals with various online attacks or threats aimed to silence or intimidate them.	- Volume and Frequency	Partial
Doxxing	Releasing the private details about an individual or organisation online.	- Exposure - Algorithmic reach	Anecdotal

c. Content

Content is the most visible vector in the ABC framework: Every user can see and form an opinion about the content of social media posts while remaining in the dark about the identity of the actor or behavioural tactics used to amplify the content. Our monitoring focuses on two types of harmful content that have been identified as problematic across all online platforms: hateful/violent content and deceptive content.

i. Hateful/Violent Content

The promotion of hatred or violence is an overarching content category addressing all attempts by Kremlin-backed or Kremlin-aligned actors to promote and justify Russia's violent and unprovoked invasion in Ukraine. This frequently includes portraying Russia's actions as moral acts and glorifying war crimes committed by the Russian army. Violent content often targets specific groups or minorities with hate speech, dehumanising or defaming messages, cyberattacks, and threats.

HATEFUL/VIOLENT CONTENT			
Content	Definition	Risk Metrics	Available evidence
Glorification of the war and war crimes	Content made to glorify war and war crimes, exonerate any war crimes or invalidate any proof of such crimes.	- Exposure - Algorithmic reach	Partial
Violent content	Content depicting violence in graphic detail (e.g., murder, rape), inciting violence against specific groups or individuals or promoting violence in general. This includes cyberstalking, public or private online threats, or the distribution of sexual images without consent.	- Exposure - Algorithmic reach	Partial
Discriminatory and dehumanising content	Content meant to amplify pre-existing racist, misogynist, xenophobic, or transphobic sentiments, including hate speech and attempts to dehumanise or discriminate against certain groups or individuals.	- Exposure - Algorithmic reach	Partial

ii. Deceptive Content

Deceptive content is an overarching content category describing all attempts by Kremlin-backed or Kremlin-aligned actors to disinform online audiences by deliberately and maliciously spreading false narratives or fabricated media. This content frequently focuses on specific events or individuals and contains false information that discredits, distorts, or manipulates.

DECEPTIVE CONTENT			
Content	Definition	Risk Metrics	Available evidence
Denial of war crimes	Content aimed to confuse and deceive audiences about war crimes. This also involves planting false evidence to shift the blame to others.	- Exposure - Algorithmic reach	Substantive
Political disinformation	Manipulative, misleading, or outright deceptive content aimed to gain political advantages, including election disinformation and disinformation about staged referenda.	- Exposure - Algorithmic reach	Anecdotal
Cheap fakes	Decontextualized audiovisual content, low-threshold manipulated content.	- Exposure	Anecdotal
Deep fakes	AI manipulated visuals that are highly deceptive and may cause serious harm.	- Exposure	Anecdotal

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

– at the following standard number: +32 22999696 or

– by email via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

Publications Office
of the European Union

doi:10.2759/764631
ISBN 978-92-68-04968-6