A platform policy implementation audit of actions against Russia's state-controlled media

Sofya Glazunova, Anna Ryzhova, Axel Bruns, Silvia Ximena Montaña-Niño, Arista Beseler, Ehsan Dehghan.

Short bio(s): **Dr Sofya Glazunova** is a postdoctoral research fellow at the Digital Media Research Centre, Queensland University of Technology, Australia. **Anna Ryzhova** is a PhD candidate within the RUSINFORM research group in the University of Passau. **Professor Axel Bruns** is an Australian Laureate Fellow and professor in the Digital Media Research Centre, Queensland University of Technology, Australia. **Dr Silvia Ximena Montaña-Niño** is a postdoctoral research fellow at the Queensland University of Technology's Node of the Automated Decision Making + Society Centre. **Arista Beseler** is a PhD candidate within the RUSINFORM research group in the University of Passau. **Dr Ehsan Dehghan** is a senior lecturer in the Digital Media Research Centre, Queensland University of Technology, Australia.

Abstract (150-200 words): The information influence of Russia's state-controlled media outlets such as *RT* and *Sputnik* on global multimillion audiences has been one of the major concerns for Western democracies and Ukraine in the last decade. With the start of the 2022 full-scale invasion of Ukraine by Russia, they were recognised as a threat to international security and several major bans were implemented towards *RT* and *Sputnik*, including their ban in the EU and its member states, and non-EU countries such as Canada, the UK and Australia, and reinforced content moderation by digital platforms globally. Digital platforms had to step up as new arbitrators of digital public spheres in this crisis event, which led us to question *how major digital platforms (Twitter, YouTube, Facebook, Instagram, TikTok, and Telegram) have implemented their content moderation policies towards* RT *and* Sputnik *accounts following Russia's full-scale invasion of Ukraine in 2022, across ten countries*. We present a platform policy implementation audit method to analyse such content moderation measures, and demonstrate its implementation by six coders after two months of the full-scale invasion. Our audit shows largely inconsistent trends in platform policy implementation towards *RT* and *Sputnik*, as well as a wide catalogue of measures taken by tech giants. We conclude with a discussion of the further implications and effectiveness of such content moderation measures for global digital audiences.

Teaser (one-sentence description): A platform policy implementation audit of how major digital platforms implemented their content moderation policies towards *RT* and *Sputnik* accounts at the beginning of *Russia's full-scale invasion of Ukraine in February 2022* shows a wide, yet inconsistent range of measures taken by tech giants.

Keywords (5): platform policies, content moderation, state-controlled media, Russia Today, policy audit.

Publication fractions: Sofya Glazunova, Queensland University of Technology: 0.250 fraction; Anna Ryzhova, University of Passau: 0.200 fraction; Axel Bruns, Queensland University of Technology: 0.200; Silvia Ximena Montaña-Niño, Queensland University of Technology: 0.150; Arista Beseler, University of Passau: 0.100; Ehsan Dehghan, Queensland University of Technology: 0.100.

1. Introduction

The full-scale invasion of Ukraine by Russia on 24 February 2022 brought immense human loss and devastation to Ukraine, as well as affecting European and global constituencies that are now dealing with the political, societal, and other implications of the invasion. Russia's information influence on global multimillion broadcast and online audiences has been a major concern for Ukraine, Western democracies, and their public spheres in the last decade. In an unprecedented move, the major mouthpieces of Kremlin politics abroad, Russia's state-controlled media for international audiences *RT* and *Sputnik,* were banned by the EU within its 27 member states following the full-scale invasion. Countries including Canada, the UK, and Australia have implemented broadcast bans on *RT* and *Sputnik* or asked major tech giants to restrict their content on their platforms. Digital platforms, in their turn, had to step up as arbitrators of digital public spheres during the invasion and reinforced content moderation of Russia's state-controlled media, following the EU and national legislations and their own content moderation policies.

However, to date there is no systematic analysis of *how major digital platforms have implemented their content moderation policies towards* RT *and* Sputnik *accounts*

after the first two months of Russia's full-scale invasion on 24 February. Indeed, as yet there are no well-established approaches to producing such an analysis, even though the EU's actions against Russian state-controlled media, as well as other, less centrally coordinated regulatory actions against mis- and disinformation in other critical contexts (such as that of the COVID-19 pandemic), have increasingly highlighted the need for methodological frameworks that verify whether, how, and to what extent such regulatory steps have been implemented by the various platform operators.

To address this twofold gap in scholarship, this article a) introduces a new research approach which we describe as a *platform policy implementation audit*, designed to systematically observe and document the various measures against identified problematic content that have been implemented by selected platforms, and b) demonstrates the utility of this approach by undertaking such a policy implementation audit for six platforms in ten countries (including EU and non-EU countries) to observe what content moderation measures these platforms have taken towards *RT* and *Sputnik*'s digital branches in response to the EU bans. Notably, our analysis finds largely inconsistent trends in policy implementation among platforms towards Russia's state-controlled media, and documents a wide and diverse catalogue of measures taken by tech giants in this unprecedented case. We conclude with a discussion of the further implications and effectiveness of such measures by platforms for global digital audiences.

The structure of the article is as follows: after the introduction, we proceed with context on Russia's war on Ukraine since 2014, the more recent full-scale invasion in 2022, and on Russia's information influence in Ukraine and the EU. We then provide background information on Russia's state-controlled media outlets, *RT* and *Sputnik*, and

their social media audiences. We introduce a classification of different types of content moderation measures across platforms, supported by actual examples of how *RT* and *Sputnik* content was moderated by platforms in the first two months of the full-scale invasion. The central part of our article introduces the platform policy implementation audit approach, its application across platforms, and its limitations. Demonstrating the utility of this approach, we then show the results of our audit across countries and platforms and, finally, provide concluding remarks.

2. The war on Ukraine, Russia's state-sponsored media, and content moderation measures

2.1. The war on Ukraine

In this article, we analyse regulatory actions taken within two months of a new phase of Russia's war on Ukraine: specifically, the beginning of the full-scale invasion by Russia that began on 24 February 2022[1]. Here and throughout this article, we refer to this phase as the "full-scale invasion". This phase is a part of the longer war between Russia and Ukraine that started in 2014 following the Euromaidan protests in Ukraine (2013-2014). The protests were sparked by Russian-aligned then-president Victor Yanukovich's decision, against the Ukrainian Parliament's vote, not to sign the European Union–Ukraine Association Agreement and to change the country's course towards greater cooperation with Russia (Harding, 2017). The scope of the protests widened, triggered among other factors by the violence of Ukrainian law enforcement against protesters.

The events of Maidan were followed by the ousting of Yanukovich, pro-Russian

[1] President Vladimir Putin on this day labelled this phase as a 'Special Military Operation' in his address to the nation. Since then, it has been imperative in Russian public discourse to use this label instead 'war' or 'invasion', subject to administrative and criminal punishment in the country.

unrest in eastern and southern Ukraine, the annexation of Crimea by Russia in March 2014, and the war in Donbas (Ukraine's Eastern Donetsk and Luhansk oblasts) between the Ukrainian military and pro-Russian separatists since April 2014. Many early separatist leaders were reported to hold Russian passports and have long-standing ties to Russia's security services (Wilson, 2016), while multiple reports from the Organization for Security and Co-operation in Europe's (OSCE) missions to Donbas from 2014 to 2022 also suggested Russia had provided military aid[2] (e.g., Organization for Security and Co-operation in Europe, 2014, 2015, 2018), with such aid ranging from material supplies to the direct involvement of Russian mercenary as well as regular forces in the war in Donbas (Wilson, 2016). Putin admitted the presence of Russian military officers in eastern Ukraine in 2015 (Walker, 2015).

The further full-scale invasion of Ukraine by Russia in February 2022 has led to severe casualties. From 24 February 2022 to 13 March 2023, the Office of the UN High Commissioner for Human Rights (OHCHR) has recorded 21,965 civilian casualties in the country: 8,231 killed and 13,734 injured (United Nations, 2023)[3]. Given the lack of definitive information from the areas most affected by the war, this is likely to be a substantial underestimation; it also does not account for military losses, which to date have been estimated at 120,000 for Ukraine and 160,000 for Russia in western media reports (Khurshudyan et al., 2023; Cole, 2023). The full, disastrous implications of the invasion of Ukraine are hard to estimate, as it is still unfolding at the time of writing.

[2] OSCE's Special Monitoring Mission in Ukraine (2014-2022) describes itself as ''an unarmed, civilian mission, operating on the ground 24/7 Ukraine. Its main tasks were to observe and report in an impartial and objective manner on the security situation in Ukraine; and to facilitate dialogue among all parties to the conflict" (Organization for Security and Co-operation in Europe, 2022). During its presence in the region, the mission has been publicly criticised and "occasionally been simultaneously considered to be an ally of the confronting sides" (see more in Härtel et al., 2021).
[3] OHCHR estimated the total number of conflict-related casualties in Ukraine before 2022, from 14 April 2014 to 31 December 2021, to be 51,000–54,000. (United Nations, 2022).

Russia's information influence since the Euromaidan events has become a source of threat and concern not only in Ukraine but also in Europe. In Ukraine, a series of actions limiting Russia's information influence were taken. They include but are not limited to a broadcast ban on 14 television channels in 2014 including *RT* (Reuters, 19 August 2014), and a ban of the Russian VK (former VKontakte) social media platform in the country in 2017. In 2021, Ukrainian President Volodymyr Zelenskyy banned three television channels – 112, NewsOne, and ZIK – for spreading Kremlin propaganda, a move that EU foreign policy spokesperson Josep Borell criticised at the time as limiting the freedom of the media (Mirovalev, 5 February 2021).

The EU has also addressed threats from Russia's strategic communication abroad, and established the East StratCom Task Force to counter Russian disinformation in 2015, which *inter alia* set up the *EU vs Disinfo* site to identify and debunk such disinformation. Further, several EU countries have banned or limited the operation of Russia's state-sponsored media *RT* and *Sputnik* in their territory; for instance, Germany in 2021, Estonia in 2020, Latvia in 2016 and 2020, the UK in 2019, and France in 2017. However, extraordinary and crisis circumstances, such as the full-scale invasion of Ukraine by Russia in 2022 and the related information warfare directed at Ukraine and the broader European continent, led the EU to take more decisive steps to protect European audiences from the Kremlin's information influence. These included total bans on *RT* and *Sputnik*.

It is outside the scope of our article to discuss the justifiability of such sanctions. However, we note that in Europe and beyond they have been subject to intense debate amongst and between politicians, policy-makers, scholars, and civil society organisations, weighing the implications of these state media bans as potential

precedents for restrictions on freedom of the press, freedom of expression, and overall human rights on the one hand against the consequences of continuing to allow Russian influence and disinformation campaigns designed to circulate unchecked amongst European and world audiences on the other hand (for a valuable introduction to the current debate, see Helberger & Schulz, 2022; for earlier discussion, see Richter, 2015). While the individual scholars contributing to the present article represent a range of views on this question, as a collective effort our article takes no sides on this issue; instead, we proceed from the reality that these regulatory actions against *RT* and *Sputnik* have now been taken, and that their implementation by the various platforms and across the countries within which they operate should be reviewed.

2.2. RT and Sputnik

Many states rely on and support international broadcasters to promote their political agendas across the world and have influence on global public opinion in critical geopolitical events (Sheafer & Gabay, 2009). In this sense, both democratic and authoritarian political regimes try to "sway international policy-making and gain political control" through state-sponsored media (2009, p. 447). These efforts in public diplomacy of democratic countries have been often labelled as 'soft power' (Nye, 2004), meaning the promotion of a country's positive image through cultural, information, and educational practices. But in authoritarian countries state-sponsored media and their content are commonly associated with the distortion of information or promotion of unattractive authoritarian values or 'sharp power' (International Forum for Democratic Studies, 2017). In reality, this distinction between 'sharp' and 'soft power' within the work and content of one state-sponsored media outlet is hard to establish or measure,

as such outlets criticise their enemies and promote a positive image of a country at the same time. Recent studies exploring this duality of 'soft' and 'sharp power' influence on the audiences of *RT* on Facebook have demonstrated the dominance of the latter form of power (Glazunova et al., 2022). Moreover, the balance of these public diplomacy powers can change over time. *RT* and *Sputnik* as Russia's international broadcasters were initially established in the early 2000s to promote Russia's positive image in the world (Elswah & Howard, 2022), but gradually their role has changed due to Russia's involvement in multiple wars (e.g., the Russo-Georgian war in 2008); since 2022, *RT* and *Sputnik* have engaged in blatant disinformation and directed extreme and belligerent rhetoric towards Ukraine and the West. This shift has influenced not only the content strategies of *RT* and *Sputnik*, but also their audiences and organisational practices. In this paper, we refer to this 'dark' side of *RT* as an instrument of 'sharp power' which became a matter of concern and a threat for Ukraine, the EU states, and beyond following the full-scale invasion.

RT (formerly *Russia Today*) was launched in 2005 but rose to prominence after the five-day Russo-Georgian war and its related information warfare in 2008. At that time, Russia did not possess a proper arsenal for informational counterattacks; in addition to journalistic work, *RT* became an information warfare tool in the following wars (Sukhodolov, 2015), and a diplomacy tool in Russia's international relations spreading strategic Kremlin narratives. The outlet has been found to disseminate conspiracy theories (Yablokov, 2015), mis- and disinformation (Cull et al., 2017), antisemitism (Rosenberg, 2015), Islamophobia (Lytvynenko & Silverman, 2019), denial of war crimes, and other problematic content. From an organisational point of view, interviews with former *RT* employees have shown that the channel ignores professional journalistic

conventions, that its editors are appointed by the government, that it promotes anti-Western ideologies, and that its operations are driven by objectives other than profit, creating influential propaganda instead (Elswah & Howard, 2022, p. 625).

At the same time, Russia's international broadcasters do not always employ propaganda strategies such as distributing misinformation and conspiracy theories, as described above. On certain issues, they use more subtle techniques to convince audiences of their credibility (Crilley et al., 2022): for instance, during the 2018 World Cup in Russia, *RT*'s usual tendency of focussing on negative reporting on Western institutions was replaced with a strong focus on promoting the sports megaevent in Russia and promoting Russia's positive image (Crilley et al., 2021, p. 137).

Beyond TV broadcasts, such content is further amplified through the *RT* website and the extensive use of digital platforms. *RT* itself has reported a substantial growth of broadcast and online audiences over the years; however, scholarly estimates of *RT* online audiences are fragmented and incomplete (Crilley et al., 2022; Nielsen, 2022). According to the recent study of *RT* and *Sputnik* audiences in 21 countries (Kling et al., 2022), the reach of these outlets via their apps and website is relatively small, i.e. it did not exceed 5% of the digital population in any of the countries between 2019 and 2021. *RT* and *Sputnik* managed to gather their biggest monthly audiences in Spain, Germany, and France (Kling et al., 2022, p. 3). At the same time, in terms of reach on Facebook, in Germany and France *RT* audiences are not much smaller than national newspapers' audiences (e.g., *Les Echoes* in France has only 2.8 times the audience of RT, and *Der Spiegel* in Germany has just 1.6 times RT's audience), due to "fractured audiences" in these countries (Nielsen, 2022).

The demographics and political identification of *RT*'s online audiences is understudied too. On Twitter, *RT* audiences are diverse and fragmented rather than politically extreme, far more likely to be male, and/or slightly older than an average Twitter user (Crilley et al., 2022). On Facebook, *RT* content across its six languages of coverage is consumed by disenfranchised ideological communities, both from the far right and the far left, who criticise the weakness of mainstream liberal-democratic leadership and the oppressive nature of global Western hegemony (Glazunova et al., 2022). However, *RT* is not the only instrument of such information influence on global audiences.

Along with *RT*, *Sputnik* served similar geopolitical purposes: to "secur[e] the national interests of the Russian Federation in the information sphere" (Bradshaw et al., 2022, p. 6). Established in 2014 by the Russian media group *Rossiya Segodnya* (*Russia Today*), *Sputnik* operates in more than 30 languages. In contrast to *RT*, *Sputnik* has received considerably less scholarly attention, and is often studied alongside *RT*. Some recent studies examined the narratives it promotes (Deverell et al., 2021; Wagnsson et al., 2017; Wagnsson & Barzanje, 2019). Deverell et al. (2021) found that *Sputnik*'s narratives about the Nordic countries focussed primarily on portraying them 1) as politically dysfunctional, and 2) as Russophobic and in perpetual conflict with Russia. In the Swedish case, *Sputnik* supported and further promoted already existing nationalist and anti-liberal narratives about Swedish society, rather than developing novel opposing ideologies (Wagnsson 2021, p. 12). As for the differences between *RT* and *Sputnik*, on average, *Sputnik* publishes significantly more content, by translating and reprinting articles from its non-English outlets into English (Ramsay & Robertshaw, 2019, p. 18).

Given the history of *RT* and Sputnik's spread of misleading information and being the information instrument of Russia's diplomacy and influence abroad, in this article, we scrutinise in detail how the leading social media platforms reacted to Russia's full-scale invasion of Ukraine in 2022 in terms of restricting the *RT* and *Sputnik* posts, links and accounts.

2.3. A catalogue of content moderation measures

Since the mid-2010s, the growing concerns about 'fake news', or more properly mis-, dis- and malinformation (Wardle & Derakhshan, 2017) in political, commercial, personal, or medical contexts have encouraged the development and diversification of an arsenal of possible interventions at the platform level. While considering diverging affordances of different social media platforms, it is nonetheless possible to identify broad categories of intervention, directed either at the accounts with problematic content; at the content; or at the platform users who seek to engage with such content. As we will show, all three types of intervention were used in restricting the dissemination of *RT* and *Sputnik* content following the 2022 Russian full-scale invasion of Ukraine.

The most powerful measure to target a publisher account is to remove it from the platform altogether. Such removal can be temporary or permanent, as implemented most famously by Twitter against former US President Donald Trump in response to his call for an armed insurrection against Congress in early 2021 (Twitter, 2021). Further, where the underlying infrastructure permits, it may also be implemented within specific geographic areas, especially if accounts are found to be in violation of local laws on free

speech, decency, or blasphemy (BBC News, 2012). At its core, the term 'deplatforming' describes these forms of intervention.

Platforms may also choose more subtle interventions: while allowing the account to continue to operate, they may exclude it from on-site advertising and other monetisation functionality, and thereby remove significant incentives for further activity (Caplan & Gillespie, 2020). Finally, accounts may also simply be tagged with warning labels, indicating that they have been identified as sources of misinformation, or are sponsored by specific state or commercial interests.

A second category of interventions focusses on problematic content per se. Here, individual content posts may be taken down permanently or suspended temporarily (to argue for posts' reinstatement) or made unavailable in specific countries or regions – for instance, in response to national laws preventing the display of Nazi insignia or other hateful content (Shih, 2012). Such actions may be taken via automated content flagging, human content moderation, reports by the user community, or a combination of all three; they may be triggered if reports through one or more of these mechanisms pass a preset threshold; and they are therefore largely reactive and *a posteriori* rather than proactive and *a priori*.

Further, warning labels of various forms can be applied at the level of the post itself. The post may be labelled as originating from a state-sponsored media outlet or some other problematic category of account. Additionally, it may also receive labels that mark it as containing confronting language or imagery, dealing with controversial topics, or containing media objects that are presented out of their original context (Kraus, 2020). Such labelling is most likely to result from a combination of automated and human intervention.

Finally, a third category of interventions may become fully visible only as ordinary platform users attempt to interact with problematic content and accounts. Users may encounter click-through warnings as they seek to access the images, videos, or external links embedded in a problematic post: these warnings may again alert them to the problematic origins, the confronting or controversial subject matter, or the uncertain provenance of this material.

Subsequently, ordinary users might then also be presented with such warnings if they attempt to disseminate such posts, for example by retweeting them on Twitter or on-sharing them in equivalent ways on other platforms. These on-sharing warnings can take various forms, from a simple click-through that asks the user to reflect on their choice before confirming the on-sharing action, to an outright ban on further on-sharing on the platform. Depending on the nature of this implementation, such on-sharing interventions have the potential to considerably impede the further dissemination of problematic content beyond its point of origin.

Adjusted for the specific affordances of each platform, all such interventions are available in principle for the *RT* and *Sputnik* accounts and content on the six platforms we examine in this article. Which *levels* of intervention have been taken then, reveals the extent to which these platforms have chosen to respond to the political and societal concerns to act against Kremlin propaganda following the 2022 invasion of Ukraine, and points to a considerable divergence in their evaluation of the balance between corporate social responsibility and adherence to freedom of expression ideals.

It is important to note in this context that the implementation of these measures does indeed predominantly reflect the platform operators' choice and is not generally hindered by technological limitations. Different affordances notwithstanding, all six

platforms could take down accounts or their posts, and to implement other less drastic measures; Twitter, as we have noted, even banned a former US President from its platform, while in response to a regulatory disagreement Facebook notoriously even removed all news content from its Australian operation for over a week in February 2021 (Leaver, 2021). If such measures are available to defend a platform's own commercial and political interests, they are also available at least in principle in response to political and societal concerns to prevent the spread of state propaganda.

3. Sanctions towards *RT* and *Sputnik*

Before implementing our audit, we first built a database of all reported bans and platform content moderation measures towards *RT* and *Sputnik* in the first two months of the invasion. We relied on an extensive search of their mentions in international and Russian news media, but also on self-reporting by *RT* in their social media accounts. We also monitored the platforms' official blogs and accounts for statements on content moderation measures towards Russia's state-sponsored media, or state-sponsored media in general. Lastly, we also checked official legal documents that identified policies towards *RT* and *Sputnik*; for instance, the *Official Journal of the European Union* documented the EU's actions towards *RT* and *Sputnik* on social media.

This produced an extensive database and timeline of content moderation measures towards *RT* and *Sputnik* during the two months of our analysis. Our database contains several examples of post takedowns, demonetisation, and account suspension actions by platforms. However, for future extensions of this research a more systematic approach to track these activities by platforms in real time in the media and other sources will be needed. This could include live content gathering from *RT* and *Sputnik*

social media channels, and tracking of self-reporting by *RT* and *Sputnik* journalists and chief editors like Margarita Simonyan, who usually report on such actions by the platforms.

Various governments had already applied multiple sanctions to *RT* and *Sputnik* before the full-scale invasion in 2022, in order to prevent the outlets from influencing national audiences. But most of these sanctions concerned national broadcast channels, with limited bans on social media content. In the European context, several restrictive actions were implemented towards *RT* or *Sputnik* before the full-scale invasion, as described in Section 2.1.. These past measures had a fragmented character and targeted their broadcasts and websites for national constituencies.

After the Euromaidan protests, Russia's annexation of Crimea, and the war in the Donbas (2014-2022), Ukraine restricted the reach of the Kremlin's broadcast media channels including *RT* in 2014 (Moscow Times, 2021a); however, these measures were aimed mostly towards Russian-language media and their broadcasts, rather than social media accounts. Until May 2022, the only measure implemented by Ukraine concerning the *RT* and *Sputnik* social media accounts was to ban downloads of *RT*'s mobile app from the Android app store in Ukraine (Reuters, 2022).

The full-scale invasion triggered a chain reaction of more significant bans, as well as greater concern with the social media accounts of Russian state-controlled media. These bans were initiated by multiple stakeholders, including the EU, social media platforms, public and private national broadcasters, and other actors. We describe the major milestones and policies that were introduced from the beginning of the Russian full-scale invasion on 24 February to 12 May 2022, and that concerned or influenced the social media accounts of *RT* and *Sputnik*.

3.1. The EU ban

On 1 March 2022, the Council of the EU adopted a decision that prohibits Russian state-sponsored outlets in Europe to "broadcast or to enable, facilitate or otherwise contribute to broadcast … any content" with the purpose of disseminating propaganda directed against Ukraine and the European Union (Art. 2f. Regulation (EU) No 833/2014, as amended on 1 March 2022 by Council Regulation (EU) 2022/350). Distribution by any means was banned, including "cable, satellite, IP-TV, internet service providers, internet video-sharing platforms or applications, whether new or pre-installed" (Official Journal of the European Union 2022, 65/2). These measures officially came into effect on 2 March 2022.

3.2. Country bans

Non-EU countries like Canada, Australia, and the UK also imposed various restricting measures towards *RT*, *Sputnik*, and other Russian state media, mostly towards their broadcasts. In the US, *RT America* ceased production on its own accord, citing "unforeseen business interruption events" (Darcy, 2022). Some of the countries took public steps and appealed to digital platforms, too.

On 3 March 2022, Nadine Dorries, the UK Secretary for Digital, Culture, Media, and Sport, published an official letter asking Meta, Twitter and TikTok to restrict access to *Sputnik* and *RT* pages on their platforms; in response, Meta restricted such access (Martin, 2022). On the same day, the Australian Communications Minister, Paul Fletcher, asked digital platforms inclusive of Meta, TikTok, Twitter and Google to block content from Russian state-sponsored media in Australia (Hurst & Butler, 2022). On 8

March, Marise Payne, the Australian Minister for Foreign Affairs (2022), stated that the Australian government was working with Facebook, Twitter and Google to suspend the dissemination of content by Russia-affiliated outlets in Australia. However, as the findings of our audit will show, both in the UK and Australia, the bans did not take place during the period of our analysis.

3.3. Measures by platforms

The platforms like Meta, YouTube and TikTok announced they conformed with the wide EU ban of *RT* and *Sputnik* (Clegg, 2022; YouTubeInsider, 2022; Bell, 2022). In addition, some platforms started to implement other restrictions towards *RT*, *Sputnik*, and their digital branches at the beginning of the full-scale invasion. We outline the most significant steps taken by platforms Google, YouTube, Facebook, Twitter, Telegram, and TikTok by May 2022.

3.4. Flagging content

In the first days of the full-scale invasion, Instagram, Facebook and Twitter reportedly started to flag the accounts of *RT* and *Sputnik* as Russian 'state-sponsored', 'state-affiliated' or 'state-controlled' media referring to their policies (Instagram Help Center, 2022; Twitter Help Center, 2022). Similar policy statements were not visible on YouTube, even though YouTube provides information about publisher context (YouTube Help, 2022). However, YouTube had banned most Russian state-sponsored media outright later. TikTok, by contrast, only started to develop and implement policies to label state media outlets almost two months after the start of the full-scale invasion (TikTok, 2022).

3.5. Demonetisation

By 27 February 2022, Facebook, Google, YouTube, and Twitter were all understood to prohibit Russian state-sponsored media including *RT* from advertising or monetising their content on the platforms (Bond, 2022). Twitter had already banned advertising from state-controlled media in 2019, but in addition now paused all ads from Russia and Ukraine during the 2022 full-scale invasion.

3.6. Post takedowns

According to *RT in Russian* (2022a), on 26 February 2022 Facebook took down a post about the Ukrainian soldiers who "gave up without a fight" on Snake Island, due to it being "false". Facebook collaborates with multiple third-party fact-checkers, including the Ukrainian *StopFake*, which checked the post and stated that the soldiers had been fighting "till the end". While at first Ukrainian President Zelensky had publicly stated that all 13 soldiers on Snake Island had died heroically, the soldiers were later believed to be "alive and well" (Shukla & Kolirin, 2022).

3.7. Temporary suspensions

On 11 March 2022, according to *RT in Russian*, Twitter limited access to the *RT* account for 12 hours due to the publication of a tweet on the alleged bombardment of the Mariupol maternity hospital, as its content was believed to violate the platform's rules against abuse and harassment (RT in Russian, 2022b). On 8 April, Twitter reportedly again suspended the *RT in Russian* account, preventing *RT* from publishing any new posts, because of a previous tweet about a captioned Ukrainian soldier.

3.8. Platform-wide bans

On 12 March 2022, YouTube announced that it had banned all Russian state-sponsored media from the platform, "citing a policy barring content that denies, minimizes or trivializes well-documented violent events" (Guardian, 2022).

4. Method

4.1. Platform Policy Implementation Audit

To analyse these measure implementations consistently, we introduce a *Platform Policy Implementation Audit* method to examine how the mainstream digital platforms have implemented their content moderation policies towards *RT* and *Sputnik*, review possible inconsistencies in that implementation, and potentially also detect how *RT* and *Sputnik* have sought to circumvent these measures. In doing so, we draw on previous studies of online censorship (Aceto & Pescapé, 2015), *hard* and *soft* platform moderation measures (York & Zuckerman, 2019), and commercial content moderation (Roberts, 2018).

Drawing from Aceto and Pescapé's (2015) work, our method was designed to systematically observe the differences in content flagging, suspending, removing, and other common moderation measures associated with *RT* and *Sputnik* as targets, while also considering the apparent geographical location of the user as a potential trigger for differentiated restriction measures. The specific targets comprise the official *RT* and *Sputnik* accounts, pages, and channels, and their posts, in the six major languages covered by both outlets, as well as the video-focussed *RT* spin-off *Ruptly* (a total of 13 accounts).

These are:

RT main branches	Sputnik main branches
RT English	Sputnik News
RT Deutsch	Sputnik Mundo
RT France	Sputnik France
RT Arabic	Sputnik Arabic
RT Español	Радио Sputnik
RT Russian	SNA (German)
Ruptly	

We examine these for the six major social media platforms Facebook, Twitter, Instagram, TikTok, Telegram and YouTube. We further test for differences in the restrictive actions implemented against our targets that may be triggered by the user's apparent geographical location, by attempting to access and engage with these accounts and their content from a total of ten countries, using a combination of genuine in-country browsing and access via a VPN service that simulates access from these countries.

In addition to Germany, Australia and Spain, where the coders were located, we included seven further countries in our audit. To explore the implementation of restrictions there, we utilised the VPN software ExpressVPN. We based our selection of countries on the public announcements and actions about restrictions and bans against *RT* and *Sputnik* and sought to include countries that at the time of the audit represented a range of official political positions towards Russia, from friendly (e.g., Hungary) through ambivalent (e.g., Germany) to hostile (e.g., Poland and, of course, Ukraine). Our list includes the following countries:

- Germany (EU) *
- Poland (EU)
- Spain (EU) *
- Lithuania (EU)

- Hungary (EU)
- United States of America
- Australia *
- United Kingdom
- Canada
- Ukraine

(*: direct access; access from all other countries simulated via VPN software)

By systematically accessing *targets,* 13 outlets, six platforms, and 10 countries, we *trigger* the platforms' restrictive *actions*, and record the resulting content restriction *symptoms* that an ordinary user would experience; we further note any evidence of *circumvention* mechanisms that *RT* or *Sputnik* may have implemented in response.

We note that in doing so, we can only speculate about the exact nature of the *surveillance* or *censoring devices* that the platforms may have implemented (i.e., about how they identify the triggers that activate the censoring process, and how they implement their restrictions at the technical level), since these technical details relate to the human and algorithmic practices internal to these companies. Such devices are likely to draw on a combination of human and algorithmic actions, however: broader content restrictions (such as the permanent flagging, suspension, or removal of *RT* or *Sputnik* accounts) may be implemented by human operators, while more targeted restrictions (e.g., takedowns or post flags) may result from the algorithmic review of content; similarly, algorithms may detect the user's location and implement geographically differentiated actions on that basis, but the selection of which restrictions to implement in which location is a human policy decision.

The codebook for our audit, then, distinguishes 11 types of actions (or inaction), the list of which is based on our combined inductive observations, and existing literature

on the most common content moderation practices (York & Zuckerman, 2019). Table 1 depicts the codes and their descriptions.

Table 1: Platform Policy Implementation Audit Codebook

Code	**Description**
No measures	No observable measures applied to the *RT* or *Sputnik* social media account or its content.
On-sharing ban	Users are prevented from sharing news stories from *RT* or *Sputnik* accounts or their domains on their own social media accounts.
On-sharing flagging	When attempting to share *RT* or *Sputnik* content (using the relevant platform features, e.g., retweeting on Twitter, sharing a post on Facebook), users are warned that they are sharing content from Russian state-sponsored media, or containing mis- or disinformation, and must acknowledge that warning before proceeding with their on-sharing action.
Click-through flagging	When clicking on links to *RT* or *Sputnik* content in social media posts, users are warned that the links may lead to potentially misleading content, or content from Russian state-sponsored media, and must acknowledge that warning before they can proceed to this content.
Flagging	The *RT* or *Sputnik* account in general, or a piece of *RT* or *Sputnik* content in particular, is flagged on the platform as originating from Russian state-controlled media or as containing mis- or disinformation.
Demonetisation	The platform bars *RT* or *Sputnik* from receiving income for advertising attached to their accounts or content.
Post takedown	The platform has removed one or more posts from the *RT* or *Sputnik* account's content feed, due to a violation of the platform's policies or rules.
Temporary suspension	Temporary suspension of the social media account from the platform for a violation of platform rules and policies.
Country block	The account or content of *RT* or *Sputnik* is not available in a particular country.
Banning from the platform	Permanent suspension of an *RT* or *Sputnik* social media account from the platform.
The page of *RT*/*Sputnik* on the platform does not exist or could not be found.	The page of *RT/Sputnik* on the platform does not exist or could not be found. (Note: this may also indicate that an account for the specific *RT* or *Sputnik* branch was never created on a particular platform.)

To test our assumptions, we accessed our targets and engaged with their posts from 6 to 12 May 2022, with our team divided into three groups of observers in different physical locations (in Australia, Germany, and Spain). We further simulated accessing these targets from additional geographical locations by the VPN software. In each country, we tested policy implementations for the desktop browser versions of the six platforms. Overall, this audit produces a comprehensive snapshot of the

implementation of access restriction policies towards *RT* and *Sputnik* (and the target outlets' attempts to circumvent such restrictions) as they were in place on our audit date.

4.2. Limitations

This approach introduces several unavoidable limitations. Firstly, we simulate users based in seven countries by using a VPN service. In doing so, our study does not account for measures implemented by individual internet service providers (ISPs) in these countries, which are potentially also powerful actors in banning *RT* and *Sputnik*. The EU ban also concerns ISPs, which may have had to set up network blocks to implement the provisions listed in the ban directive (Biselli et al., 2022). However, such ISP-level network bans would directly target the websites of *RT* and *Sputnik*, while our study focusses on the social media platforms where their content may also circulate (Ó Fathaigh, 2022).

Secondly, we tested only for those measures that platforms and governments said they would implement and did not investigate 'shadow bans' and other less obvious measures on digital platforms. Such adjustments to recommendation and ranking algorithms on platforms are difficult to test for systematically, however, and are therefore not included in our approach.

Finally, we performed our audit only for the desktop browser versions of these platforms. Mobile versions were checked only for TikTok, to examine the discrepancies on its desktop version that our audit revealed, which hinted that the platform might display *Sputnik* and *RT* accounts differently on its mobile version. This assumption was confirmed.

5. Findings

5.1. YouTube

YouTube has implemented the most consistent and homogenous policy towards the *RT* and *Sputnik* accounts across studied platforms. All the *RT* and *Sputnik* accounts investigated were first demonetised (see section 2.4) and then banned entirely from the platform following the beginning of the full-scale invasion, with no *RT* or *Sputnik* channels or videos accessible at all at the moment of analysis. The only such content that still circulates on YouTube therefore results from individual users re-uploading it; such attempts to help *RT* and *Sputnik* circumvent YouTube's blanket ban on these channels are likely to reach a much smaller audience than the original channels themselves, however.

5.2. Facebook

Facebook had demonetised *RT* and *Sputnik*'s branches and taken down posts by *RT Russian* at the beginning of Russia's full-scale invasion in February 2022 (see section 3.5). In addition, Facebook seems to comply with EU directives at least partially, and has blocked *RT* and *Sputnik* in most of the EU countries on our list: Hungary, Poland, Germany and Spain; they were also unavailable in Ukraine. However, we could still access these pages in Lithuania, which points to an unequal and incomplete implementation of the EU-wide ban. They also remained available in the US, UK, Australia and Canada.

In those five countries where *RT* and *Sputnik* pages were still available, Facebook flagged any *posts* by *RT* and *Sputnik* as "Russia state-controlled media"; however, the

pages themselves were not flagged as such (either in the search function of Facebook, or in the headers of *RT* and *Sputnik* Facebook pages). We also observed inconsistent patterns in the implementation of on-sharing flagging and click-through flagging. In Australia, the US, the UK, Canada, and Lithuania, click-through flagging was not applied to *RT Arabic*, *RT Español*, *Ruptly*, *Ραδιο Sputnik* (In English: *Radio Sputnik*) and *SNA*; while on-sharing flagging was not implemented for *Sputnik Arabic*. We note that except for *Ruptly*, which mostly posts video content, these are all non-English pages. Therefore, it is possible that, for reasons that we can only speculate about, Facebook's implementation of such flagging (in countries where it was not compelled to block the pages of *RT* and *Sputnik* altogether) has focussed predominantly on the English-language versions of these pages. Previous Facebook document leaks in 2021 revealed that "Facebook under-invests in content safety systems for non-English languages" (Duffy et al., 2021)

5.3. Instagram

The Instagram pages of *RT* and *Sputnik* were inaccessible in all the EU countries on our list (including Lithuania); they remained available in the US, UK, Australia, Canada, and Ukraine. The latter is somewhat surprising as the other major platform owned by Meta, Facebook, had banned both outlets' pages in Ukraine: in their handling of these outlets, Facebook and Instagram thus take divergent approaches even despite their common ownership by Meta.

In those countries where the Instagram pages remained available, the platform systematically flagged all *RT* and *Sputnik* accounts and posts as "Russia state-controlled media". On-sharing flagging was applied to the posts of most of these Instagram

accounts, except *Sputnik France*, which received no on-sharing warning in any of the five countries, and *Sputnik Mundo* and *Радио Sputnik*, which received no on-sharing warning in Ukraine but did so in the four other countries. Further, for *Sputnik News* we encountered on-sharing warnings only some of the time.

Our exploration of click-through flagging measures on Instagram was limited by the fact that not every *Sputnik* or *RT* page posted stories during the audit timeframe (where these links could be clicked). We could observe such click-through flagging for *RT Arabic*, *Sputnik Mundo*, and *Sputnik Arabic* in all countries except Ukraine, and for *Радио Sputnik* in the US. This may again point to an incomplete implementation of such measures across the different language versions of *RT* and *Sputnik*. Additionally, we noticed inconsistencies in the focus of these click-through warnings: Instagram generally warned users clicking on such links about the fact that the content was originating from Russian state media; however, if the story contained information about COVID-19, it instead warned about potential COVID-19 misinformation and did not highlight the nature of the outlet itself.

Finally, our audit of the implementation of these restrictive measures on Instagram also produced evidence of *RT*'s efforts to circumvent such measures: notably, stories posted by *RT Deutsch* on Instagram led to an *RT* mirror domain: *test.rtde.me*. This alternative domain is likely to have been introduced to evade content, on-sharing and click-through flagging as well as outright content bans.

5.4. Telegram

Telegram channels for *Sputnik France* and *SNA* were not available or did not exist at the time of our audit. The remaining Telegram channels of *RT* and *Sputnik* in our audit were

not accessible in most EU countries: Germany, Poland, Hungary, and Lithuania. In Spain, where one of our coders was located, we could access these *RT* and *Sputnik* channels; however, when trying to access them through a VPN from Germany, the channels were not available. It is unclear what would have caused this variable accessibility.

Except where it was required to conform with EU directives, in the rest of the countries Telegram did not implement any restrictions towards *RT* or *Sputnik*. This might be due to the well-known libertarian stance of its founder Pavel Durov and his opposition to government intervention in platform policies.

5.5. TikTok

There are no TikTok accounts for *RT France*, *Sputnik France*, *Sputnik Arabic*, and *SNA*. We could not establish whether *RT Deutsch* was an authentic TikTok account of *RT*, as it did not receive any flagging, and there are no links from the *RT* website to it. For the others, despite TikTok's announcement of a ban of *RT* and *Sputnik* accounts in the EU, the outlets' accounts themselves remained available, but their videos were hidden when tested from EU countries (Germany, Poland, Lithuania, and Hungary; Spain was an exception).

Moreover, this pattern of active accounts but unavailable videos was observable only in the desktop version, while in the mobile version of TikTok all the videos posted by *RT* and *Sputnik* accounts remained visible. In the UK, which had asked TikTok to restrict access to the *RT* and *Sputnik* accounts, they were similarly blocked in the desktop version, while their videos remained available in the mobile version; we observed the same in Canada, which has not explicitly announced any sanctions against the social media accounts of Russian state-sponsored media.

In Australia, the United States, and Ukraine, *Sputnik* and *RT* accounts were consistently flagged as Russian state-sponsored media, but the accounts and their content remained accessible to users. The most heterogeneous case in terms of inconsistent measures implemented against *RT* and *Sputnik* on TikTok was Spain. Tested from Spain directly, without using a VPN, we found that the main *RT* accounts, such as *RT*, *RT Arabic*, and *RT Russian*, were not available from the desktop version but available from the mobile version, while other accounts, such as *RT Español*, *Ruptly*, *Sputnik News*, *Sputnik Mundo*, and *Paðuo Sputnik* were only flagged as Russian state-sponsored media. There is no obvious explanation for this inconsistency.

5.6. Twitter

The implementation of restrictive measures on Twitter was inconsistent. In Germany, Poland, Lithuania, and Hungary, almost all of *RT* and *Sputnik* accounts were blocked in compliance with EU directives. The exceptions were the accounts of *Ruptly* and *Paðuo Sputnik,* which, when checked from these countries, were only flagged. In terms of the country block enforcement in the EU, Spain was an exemption, as when accessed from it, the accounts of *RT* and *Sputnik* were flagged, and *RT Deutsch* had an on-sharing warning, however, no country-wide block was enforced.

In the remaining countries, we found that all *RT* and *Sputnik* accounts were flagged as "Russia state-affiliated media" in Australia, the United States, the United Kingdom, Canada and Ukraine. The *RT Deutsch* account also received an on-sharing warning. Twitter's differential treatment of *RT Deutsch* is especially puzzling. While our audit does not enable us to establish causality, we do note that, instead of linking to *rt.com*, *RT Deutsch*'s tweets linked to a new domain *pressefreiheit.rtde.live* (which translates as

'*pressfreedom.rtde.live*'); it is possible that this domain was adopted only after the initial implementation of content and on-sharing warnings in order to circumvent or at least protest what *RT* would have perceived as censorship of its content, or that the special treatment of *RT Deutsch* on Twitter has resulted precisely from the use of such blatant evasion techniques in the first place. (At the time of our audit, other *RT* accounts continued to link to standard *rt.com* URLs.)

6. Conclusion

Our systematic audit of the availability and treatment of 13 *RT* and *Sputnik* accounts on six platforms in ten countries points both to substantial differences between countries and platforms, and to significant inconsistencies in the implementation of restrictive measures even within the same platforms. As we have noted, YouTube took the strongest action against these outlets by removing their channels entirely and globally from its platform since the beginning of the full-scale invasion. The other five platforms largely implemented the EU's directive to ban these outlets at least in the countries where that directive applied with a few exceptions.

This indicates that even the comprehensive global bans of *RT* and *Sputnik* content that sites like YouTube have implemented can only have a limited (if substantial) effect on the circulation of such content: given sufficient effort and motivation, new websites can always be created to publish such content, and new accounts can always be set up to disseminate it on leading social media platforms. However, this does not render the bans and other measures whose implementation we have audited here ineffectual or meaningless: while they cannot fully and permanently protect vulnerable populations from exposure to state-sponsored propaganda and disinformation, or prevent state-

sponsored propaganda and disinformation from affecting the processes of opinion formation by democratic polities, they can significantly delay and diminish the circulation of such content.

Further, while we did not explicitly seek to identify such efforts systematically, we also noted clear evidence of *RT* attempts to evade and circumvent such restrictions. While our audit could not explicitly explore such circumvention measures, the fact that we noted them for both Twitter and Instagram points to an ongoing struggle between governments, platform providers, and these state media outlets about the unfettered circulation of *RT* and *Sputnik* content. A recent study by the Institute for Strategic Dialogue (see Balint et al., 2022) explores such circumvention strategies in more detail. This, in our view, is a prospective topic for future research.

While *RT* and *Sputnik* are important tools in the Kremlin's information warfare arsenal, increasing attention is now also being directed to the activities of the accounts of Russian embassies, ambassadors, and other official government accounts on various social media platforms. While in democratic systems there is usually a clear distinction between government, political, and media actors, and the protection of press freedom means that press outlets are largely free from such restrictions, in autocracies such as Russia the same distinction cannot be made, and 'press' accounts such as those of *RT* and *Sputnik* merely represent another government function.

Should western governments proceed to implementing such bans or restrictions on the social media accounts of Russian state officials, then the audit approach we have utilised here will again be able to produce systematic insights into the implementation of such measures across platforms and countries. We note here, as above, that several limitations still apply to this approach.

Our audit offers a systematic snapshot of the state of restrictions against *RT* and *Sputnik* from February to May 2022, but it is likely that the situation will have changed again since then, both as the outlets have attempted to further evade and circumvent these measures, and as governments and platforms have sought to strengthen their restrictions against the propaganda and disinformation published by *RT* and *Sputnik*. Regular audits of the implementation of such measures, and further research into the outlets' efforts to disseminate their content through other efforts, would be valuable, therefore. Such audits could be used both to improve the effectiveness of existing measures, and to verify that platform providers have continued to implement them.

References

Aceto, G., & Pescapé, A. (2015). Internet censorship detection: A surve. *Computer Networks*, *83*, 381-421. https://doi.org/10.1016/j.comnet.2015.03.008

Art. 2f. Regulation (EU) No 833/2014, as amended on 1 March 2022 by Council Regulation (EU Council Regulation (EU) 2022/350. Amending Regulation (EU) No 833/2014 concerning restrictive measures in view of Russia's actions destabilising the situation in Ukraine. *Official journal of the European Union.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2022:065:TOC

Australia's Minister for Foreign Affairs Marise Payne. (2022, March 8). Further sanctions on Russia. https://www.foreignminister.gov.au/minister/marise-payne/media-release/further-sanctions-russia

Balint, K., Acrostanzo, F., Wildon, J., and Reys, K. (2022, 20 July). RT Articles Are Finding Their Way to European Audiences – But How? Institute for Strategic Dialogue. https://www.isdglobal.org/digital_dispatches/rt-articles-are-finding-their-way-to-european-audiences-but-how/

BBC News. (2012, October 18). Twitter blocks neo-Nazi account to users in Germany. https://www.bbc.com/news/technology-19988662

Bell, K. (2022, February 28). TikTok follows Facebook in blocking RT and Sputnik in the EU. *Engadget*. https://www.engadget.com/tiktok-follows-facebook-in-blocking-rt-sputnik-eu-014015312.html

Biselli, A. (2022, March 2). Sanktionen gegen Russland: EU verbietet Verbreitung von RT und Sputnik. netzpolitik.org. https://netzpolitik.org/2022/sanktionen-gegen-russland-eu-verbietet-verbreitung-von-rt-und-sputnik/

Bond, S. (2022, February 27). Facebook, Google and Twitter limit ads over Russia's invasion of Ukraine. *npr.* https://www.npr.org/2022/02/26/1083291122/russia-ukraine-facebook-google-youtube-twitter

Bradshaw, S., DiResta, R., & Miller, C. (2022). Playing Both Sides: Russian State-Backed Media Coverage of the# BlackLivesMatter Movement. *The International Journal of Press/Politics*. https://doi.org/10.1177/19401612221082052

Caplan, R., & Gillespie, T. (2020). Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society*, *6*(2). https://doi.org/10.1177/2056305120936636

Clegg, N. [*@nickclegg*]. (2022, February 28). We have received requests from a number of Governments and the EU to take further steps in relation to Russian state controlled media. Given the exceptional nature of the current situation, we will be restricting access to RT and Sputnik across the EU at this time. [Tweet] https://twitter.com/nickclegg/status/1498395147536527360?s=20&t=kJm0a5N0mCTkJrq_HLgKag

Cole, B. (2023). Russia Loses 12 Tanks, 11 Armored Vehicles and 1,040 Troops in a Day: Kyiv. *Newsweek*, 16 Mar. 2023. https://www.newsweek.com/russia-ukraine-bakhmut-losses-forces-mounting-1788274

Crilley, R., Gillespie, M., Vidgen, B. and Willis, A., (2022). Understanding RT's audiences: exposure not endorsement for twitter followers of Russian state-sponsored media. *The International Journal of Press/Politics*, *27*(1), 220-242. https://doi.org/10.1177/1940161220980692

Crilley, R., Gillespie, M., Kazakov, V., & Willis, A. (2021). 'Russia isn't a country of Putins!': How RT bridged the credibility gap in Russian public diplomacy during the 2018 FIFA World Cup. *The British Journal of Politics and International Relations*, *24*(1), 136–152. https://doi.org/10.1177/13691481211013713

Cull, N.J., V. Gatov, P. Pomerantsev, A. Applebaum, & A. Shawcross. (2017). Soviet Subversion, Disinformation and Propaganda: How the West Fought against It. An Analytic History, with Lessons for the Present. *LSE Consulting*. https://bit.ly/2L7HClk

Darcy, O. (2022, March 4). RT America ceases productions and lays off most of its staff. *CNN* https://edition.cnn.com/2022/03/03/media/rt-america-layoffs/index.html

Deverell, E., Wagnsson, C., & Olsson, E. K. (2021). Destruct, direct and suppress: Sputnik narratives on the Nordic countries. *The Journal of International Communication*, *27*(1), 15-37. https://doi.org/10.1080/13216597.2020.1817122

Duffy, C., Sangal, A., Mahtani, M. & Wagner, M. (2021, October 25). Internal Facebook documents revealed. https://edition.cnn.com/business/live-news/facebook-papers-internal-documents-10-25-21/h_99f6ed16dd82c1e8ece56b5ee5c5c3a2

Elswah, M., & Howard, P. N. (2020). "Anything that causes chaos": The organizational behavior of Russia Today (RT). *Journal of Communication*, *70*(5), 623-645.

Glazunova, S., Bruns, A., Hurcombe, E., Montaña-Niño, S. X., Coulibaly, S., & Obeid, A. K. (2022). Soft power, sharp power? Exploring RT's dual role in Russia's diplomatic toolkit. *Information, Communication & Society*, online first: https://doi.org/10.1080/1369118X.2022.2155485

Guardian. (2022, March 11). YouTube blocks Russian state-funded media channels globally. https://www.theguardian.com/technology/2022/mar/11/youtube-blocks-russian-state-funded-media

Harding, L. (2017, November 26). Viktor Yanukovych promises Ukraine will embrace Russia. *The Guardian*. https://www.theguardian.com/world/2010/mar/05/ukraine-russia-relations-viktor-yanukovych

Härtel, A., Pisarenko, A., & Umland, A. (2021). The OSCE's Special Monitoring Mission to Ukraine, *Security and Human Rights*, *31*(1-4), 121-154. doi: https://doi.org/10.1163/18750230-bja10002

Helberger, N., & Schulz, W. (2022). Understandable, but still wrong: How freedom of communication suffers in the zeal for sanctions. *LSE Blog*. https://blogs.lse.ac.uk/medialse/2022/06/10/understandable-but-still-wrong-how-freedom-of-communication-suffers-in-the-zeal-for-sanctions/

Hurst, D. & Butler, J. (2022, March 3). Morrison government asks Facebook, Twitter and Google to block Russian state media 'disinformation'. *Guardian*. https://www.theguardian.com/world/2022/mar/03/morrison-government-asks-facebook-twitter-and-google-to-block-russian-state-media-disinformation

Instagram Help Center. (2022). What is labeled state-controlled media on Instagram? https://help.instagram.com/2589432474704452

International Forum for Democratic Studies. (2017, December 5). Sharp power: Rising authoritarian influence. https://www.ned.org/sharp-power-rising-authoritarian-influence-forum-report/.

Khurshudyan, I., Sonne, P., & DeYoung, K. (2023). Ukraine short of skilled troops and munitions as losses, pessimism grow. *Washington Post*, 13 March 2023.

https://www.washingtonpost.com/world/2023/03/13/ukraine-casualties-pessimism-ammunition-shortage/

Kling, J., Toepfl, F., Thurman, N. & Fletcher, R. (2022). Mapping the website and mobile app audiences of Russia's foreign communication outlets, RT and Sputnik, across 21 countries. Harvard Kennedy School Misinformation Review, 3(6), doi: 10.37016/mr-2020- 110

Kragh, M., & Åsberg, S. (2017). Russia's strategy for influence through public diplomacy and active measures: the Swedish case. *Journal of Strategic Studies*, *40*(6), 773-816. https://doi.org/10.1080/01402390.2016.1273830

Kraus, R. (2020, November 19). Facebook labeled 180 million posts as 'false' since March. Election misinformation spread anyway. *Mashable*. https://mashable.com/article/facebook-labels-180-million-posts-false

Leaver, T. (2021). Going Dark: How Google and Facebook Fought the Australian News Media and Digital Platforms Mandatory Bargaining Code. M/C Journal, 24(2). https://doi.org/10.5204/mcj.2774 (Original work published April 26, 2021)

Lytvynenko, J., & Silverman, C. (2019, April 16). A Timeline of How the Notre Dame Fire Was Turned into an Anti-Muslim Narrative. *Buzzfeed News*. https://www.buzzfeednews.com/article/janelytvynenko/notre-dame-hoax-timeline

Martin, A. (2022, March 4). Ukraine invasion: Facebook and Instagram to block RT and Sputnik in the UK following government request. *Sky News*. https://news.sky.com/story/ukraine-invasion-facebook-and-instagram-to-block-rt-and-sputnik-in-the-uk-following-government-request-12557469

Mirovalev, M. (February 5, 2021). In risky move, Ukraine's president bans pro-Russian media. *Al Jazeera*. https://www.aljazeera.com/news/2021/2/5/ukraines-president-bans-pro-russian-networks-risking-support

Moscow Times. (2021a, August 23). Ukraine blocks access to popular Russian news sites. https://www.themoscowtimes.com/2021/08/23/ukraine-blocks-access-to-popular-russian-news-sites-a74870

Moscow Times. (2021b, September 27). Telegram Messenger Blocks Navalny Bot During Russian Election. https://www.themoscowtimes.com/2021/09/18/telegram-messenger-blocks-navalny-bot-during-russian-election-a75079

Nielsen, R.K. (2022, March 2). How many people really watch or read RT, anyway? It's hard to tell, but some of their social numbers are eye-popping. *NiemanLab*. https://www.niemanlab.org/2022/03/how-many-people-really-watch-or-read-rt-anyway-its-hard-to-tell-but-some-of-their-social-numbers-are-eye-popping/

Nye, J. S. (2004). *Soft power: The means to success in world politics*, New York: Public Affairs.

Ó Fathaigh, R. (2022). [NL] Dutch ISPS block RT and Sputnik websites. European Audiovisual Observatory. https://merlin.obs.coe.int/article/9476

Official Journal of the European Union, L 065. (2022, March 2). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2022:065:TOC

Organization for Security and Co-operation in Europe. (2015). Latest from OSCE Special Monitoring Mission (SMM) to Ukraine based on information received as of 19:30 (Kyiv time), 2 August 2015. https://www.osce.org/ukraine-smm/175736

Organization for Security and Co-operation in Europe. (2022). OSCE Special Monitoring Mission to Ukraine (closed). https://www.osce.org/special-monitoring-mission-to-ukraine-closed

OSCE Observer Mission. 2014. Weekly Update from the OSCE Observer Mission at Russian Checkpoints Gukovo and Donetsk Based on Information as of 10:00 (Moscow Time), 12 November 2014. OSCE. November 12. https://www.osce.org/om/126629.

OSCE Special Monitoring Mission to Ukraine. (2018). Latest from the OSCE Special Monitoring Mission to Ukraine (SMM), Based on Information Received as of 19:30, 8 August 2018. OSCE. August 9. https://www.osce.org/special-monitoring-mission-to-ukraine/390179.

Ramsay, G., & Robertshaw, S. (2019). Weaponising News. RT, Sputnik and Targeted Disinformation. Report, Policy Institute Center for the Study of Media, Communication and Power. London: Kings College London. https://www.kcl.ac.uk/policy-institute/assets/weaponising-news.pdf

Reuters (2014, August 19). Ukraine bans Russian TV channels for airing war 'propaganda'. https://www.reuters.com/article/us-ukraine-crisis-television-idUSKBN0GJ1QM20140819

Reuters. (2022, 27. February). Google blocks Russia's RT app downloads on Ukrainian territory. https://www.reuters.com/technology/google-blocks-russias-rt-app-downloads-ukrainian-territory-says-rt-2022-02-27/

Richter, A.G. (2015). Legal response to propaganda broadcasts related to crisis in and around Ukraine, 2014–2015. *International Journal of Communication*, 9, 3125-45. https://ijoc.org/index.php/ijoc/article/view/4149/1474

Roberts, S. T. (2018). Commercial Content Moderation And Worker Wellness: Challenges & Opportunities. https://www.techdirt.com/articles/20180206/10435939168/commercial-content-moderation-worker-wellness-challenges-opportunities.shtml

Rosenberg, Y. (2015). Russia Today Airs Bizarre Anti-Semitic Conspiracy Theory about Hillary Clinton. *Tablet Mag.* https://www.tabletmag.com/scroll/194211/russia-today-airs-bizarre-anti-semitic-conspiracy-theory-about-hillary-clinton

RT на русском. [@rt_russian]. (2022a, February 25). Нам опять Facebook влепил метку false. За что — непонятно. Предположительно, за заметку на сайте о том, что военные ВСУ добровольно сдались в районе острова Змеиный в Чёрном море… [Again, Facebook slapped us with a false label. For what is not clear. Presumably, for a note on the website that the military of the Armed Forces of Ukraine voluntarily surrendered in the area of Snake Island in the Black Sea]. [Telegram post]. https://t.me/rt_russian/94922

RT на русском. [RT in Russian]. (2022b, March 11). Twitter на 12 часов ограничил аккаунт RT за разоблачение фейка о роддоме Мариуполя. [Twitter restricted the RT account for 12 hours for exposing a fake about the Mariupol maternity hospital]. https://russian.rt.com/nopolitics/news/974455-twitter-akkaunt-rt

RT. (2022). About RT. https://www.rt.com/about-us/

Sheafer, T., & Gabay, I. (2009). Mediated public diplomacy: A strategic contest over international agenda building and frame building. *Political communication*, *26*(4), 447-467.

Shih, G. (2012, January 27). Twitter to restrict user content in some countries. *Reuters*. https://www.reuters.com/article/us-twitter-idUKTRE80P28920120127

Shukla, S., & Kolirin, L. (2022, February 28). The defiant soldiers of Snake Island are actually 'alive and well,' says Ukraine's navy. *CNN*. https://edition.cnn.com/2022/02/28/europe/snake-island-ukraine-russia-survivors-alive-intl/index.html

Sukhodolov, A. (2015). The ideological function of the media in terms relevant information wars. *Theoretical and Practical Issues of Journalism*, 4(2), 117–126

TikTok. (2022). Bringing more context to content on TikTok. https://newsroom.tiktok.com/en-us/bringing-more-context-to-content-on-tiktok

Twitter Help Center. (2022). About government and state-affiliated media account labels on Twitter. https://help.twitter.com/en/rules-and-policies/state-affiliated

Twitter. (2021, January 8). Permanent suspension of @realDonaldTrump. https://blog.twitter.com/en_us/topics/company/2020/suspension

United Nations. (2023). Ukraine: civilian casualty update 13 March 2023. *Office of the high commissioner for human rights*. 13 March. https://www.ohchr.org/en/news/2023/03/ukraine-civilian-casualty-update-13-march-2023

United Nations. (2022). Conflict-related civilian casualties in Ukraine. *Office of the high commissioner for human rights*. 27 January. https://ukraine.un.org/sites/default/files/2022-02/Conflict-related%20civilian%20casualties%20as%20of%2031%20December%202021%20%28rev%2027%20January%202022%29%20corr%20EN_0.pdf

Wagnsson, C., & Barzanje, C. (2021). A framework for analysing antagonistic narrative strategies: A Russian tale of Swedish decline. *Media, war & conflict*, *14*(2), 239-257. https://doi.org/10.1177/1750635219884343 Walker, S. (2015, December 17). Putin admits Russian military presence in Ukraine for first time. *The Guardian*. https://www.theguardian.com/world/2015/dec/17/vladimir-putin-admits-russian-military-presence-ukraine

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. http://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf

Wilson, A. (2016). The Donbas in 2014: Explaining Civil Conflict Perhaps, but not Civil War. *Europe-Asia Studies*, *68*(4), 631–652. https://doi.org/10.1080/09668136.2016.1176994

Yablokov, I. (2015). Conspiracy theories as a Russian public diplomacy tool: The case of Russia Today (RT). *Politics*, *35*(3-4), 301-315. https://doi.org/10.1111/1467-9256.12097

York, J. C., & Zuckerman, E. (2019). Moderating the public sphere. *Human rights in the age of platforms*, *137*, 143.

YouTube Help. (2022). Information panel providing publisher context. https://support.google.com/youtube/answer/7630512?hl=en

YouTubeInsider. [@YouTubeInsider]. (2022 March, 11). 2/ In line with that, we are also now blocking access to YouTube channels associated with Russian state-funded media globally, expanding from across Europe. This change is effective immediately, and we expect our systems to take time to ramp up. [Tweet]

https://twitter.com/YouTubeInsider/status/1502335085122666500?s=20&t=QggGTaBeBPPQ1JcjZRqvWQ

Zavadski, K. (2015, September 17). Putin's Propaganda TV lies about its popularity. *Daily Beast*. https://www.thedailybeast.com/putins-propaganda-tv-lies-about-its-popularity