Code of Practice on Disinformation

A Comparative Analysis of the Prevalence and Sources of Disinformation across Major Social Media Platforms in Poland, Slovakia, and Spain

SEPTEMBER 2023

SEMI-ANNUAL REPORT

TrustLab

Code of Practice on Disinformation

The Code of Practice on Disinformation was established following the European Commission's guidance by major online platforms, emerging and specialized platforms, players in the advertising industry, fact-checkers, research, and civil society organizations to deliver a strengthened Code of Practice on Disinformation.

This measurement study was commissioned by Meta, YouTube, TikTok, and LinkedIn as part of the European Commission's Code of Practice.

You can read more about the Code of Practice framework here.

Table of **Contents**

Executive Summary

The European Union's (EU) 2022 Code of Practice on Disinformation represents a milestone in the global fight against online disinformation. Online disinformation is an ambiguous and fast-changing phenomenon, and measuring disinformation is challenging. As the first empirical application of the Code, this study set out to evaluate the prevalence and sources of disinformation across six major social media platforms (Facebook, Instagram, LinkedIn, TikTok, Twitter (now known as X),[1] and YouTube) in three countries: Poland, Slovakia, and Spain.

A total of 6,155 unique social media posts and 4,460 unique accounts were sampled by searching popular disinformation keywords using a platforms' native search functionality (Table 4). The key metrics examined are discoverability, relative post engagement, absolute post engagement, and properties about disinformation actors, including ratio of disinformation actors, their account activities, and engagement with other users.

- Discoverability refers to the ratio of mis/disinformation posts among sensitive content. The platform with the largest discoverability was Twitter (0.428), followed by Facebook (0.313). YouTube had the lowest ratio of discoverability (0.082) (Figure 1-Discoverability).

- Absolute post engagement is defined as the absolute amount of engagement that mis/disinformation posts obtained on average. Relative post engagement is defined as the ratio of average absolute engagement with mis/disinformation content over average absolute engagement with non-mis/disinformation content. High relative engagement and high absolute engagement are related, but distinct measures of audience's exposure and potential for harm (Figure 1-Relative Post Engagement and Absolute Post Engagement). Twitter had the largest relative engagement ratio of 1.977, but relatively low absolute engagement, while the opposite effect was observed on TikTok.

[1] Twitter was rebranded to X during the course of this research, but the authors decided to use the former for consistency for the remainder of the report.

- Ratio of disinformation actors refers to the proportion of disinformation actors relative to the total accounts sampled on a platform. In our study, platforms with a larger fraction of mis/disinformation content have a larger fraction of disinformation actors as well. The ratios of disinformation actors on Twitter and Facebook are the largest and of similar size (8-9 percent), whereas YouTube had the smallest ratio at 0.8 percent (Figure 1-Ratio of Disinformation Actors). Study findings about the characteristics of disinformation actors are limited due to the amount of data, but disinformation actors were found to follow more users than their non-disinformation counterparts and also tend to have joined the platform more recently than non-disinformation users.

This study establishes an initial benchmark for the implementation of the Code of Practice, and paves the way for further discussion and advances in the measurement of disinformation. Disinformation measurement is known to be a hard problem, and accordingly our metrics and methodology can, and should be, improved. More time and budget, better access to platform data, and broader agreement among stakeholders on detailed specifications of mis/disinformation as well as metric definitions that better normalise platform differences can strengthen future measurements. We also hope to broaden the scope beyond three countries and six platforms, allowing disinformation measurements to be carried out across many more platforms and countries.

Executive Summary

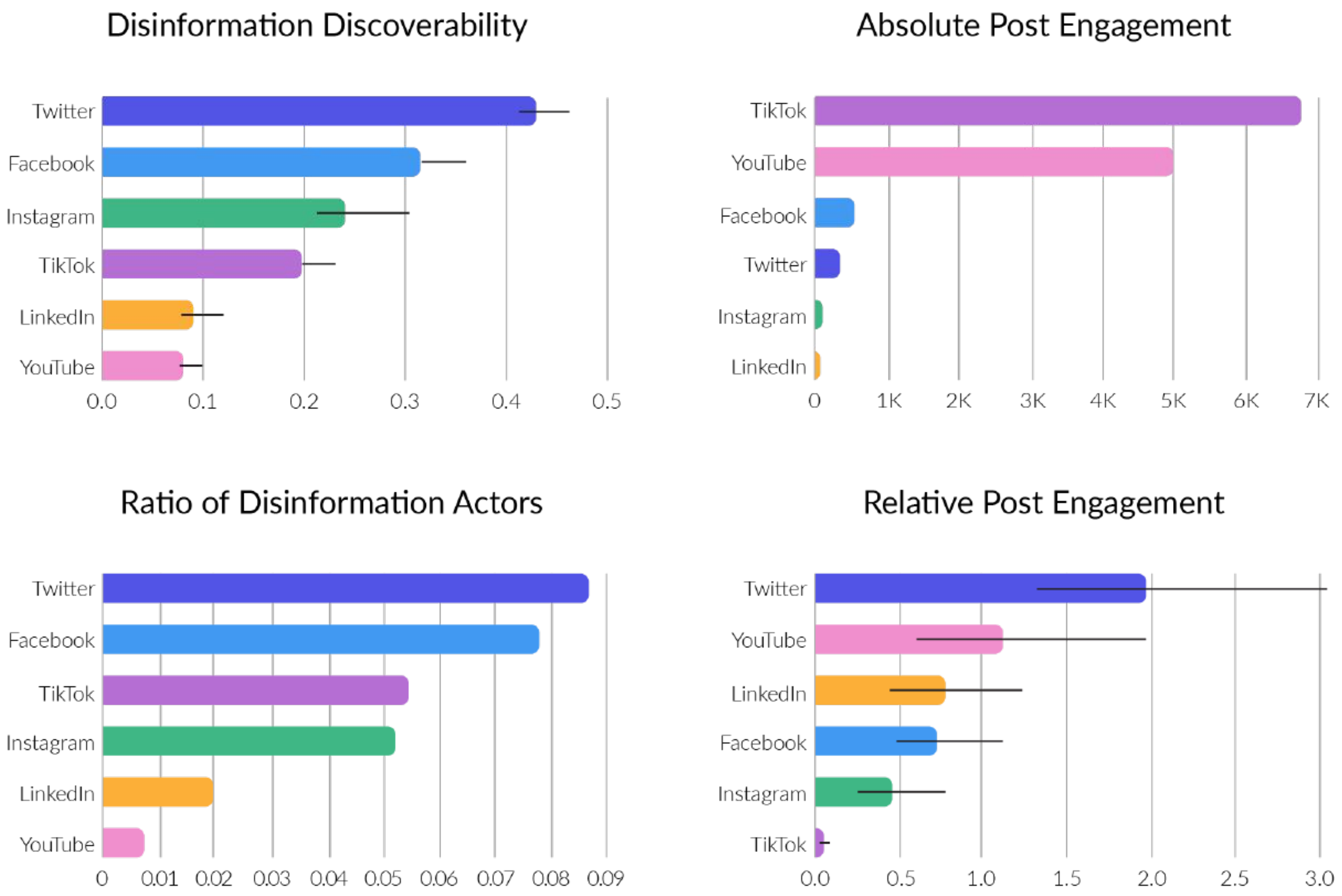


Figure 1: Platform Performance in Discoverability, Relative Post Engagement, Absolute Post Engagement, and Ratio of Disinformation Actors.
See Terminology for metric definitions.

About the Author

TrustLab's mission is to build a safer internet for everyone.

It is a global leader in disinformation monitoring and analytics, and serves as a trusted third-party solution for detecting and mitigating critical safety threats on the internet. Its three founders have over forty years of combined Trust & Safety experience from companies including Google, YouTube, Reddit, and TikTok.

TrustLab has worked with many social media platforms, messaging companies, government bodies, and online marketplaces to deliver its innovative, independent, and unbiased measurement solution. Leveraging state-of-the-art, patent-protected technology, it is able to accurately and rapidly identify harmful content at scale across multiple languages, sources and abuse verticals.

TrustLab's customers range from small companies building out their internal teams and policies to large enterprises with complex Trust & Safety needs. By providing its cutting-edge software and expertise, TrustLab helps its clients protect their users against harmful content and in doing so, serve its mission to make the internet safer.

About the Participants and Partners

The participants in the pilot study were six major social media platforms (Facebook, Instagram, LinkedIn, TikTok, Twitter (now known as X), and YouTube). Initially, all platforms were voluntarily signatories to the Code of Practice on disinformation. Partway through the study, X withdrew from the Code of Practice on disinformation.

Facebook is an online social media and social networking service owned by Meta Platforms.

Instagram is a photo and video-sharing social networking service owned by Meta Platforms.

LinkedIn is a business and employment-focused social media platform owned by Microsoft.

TikTok is a short-form video hosting service owned by ByteDance.

X is an online social media and social networking service owned by Elon Musk.

YouTube is an online video-sharing and social media platform owned by Google.

The partners of the pilot study were the Permanent Taskforce of the Code of Practice, in particular the European Commission, Avaaz, European Regulators Group for Audiovisual Media Services (ERGA), and European Digital Media Observatory (EDMO).

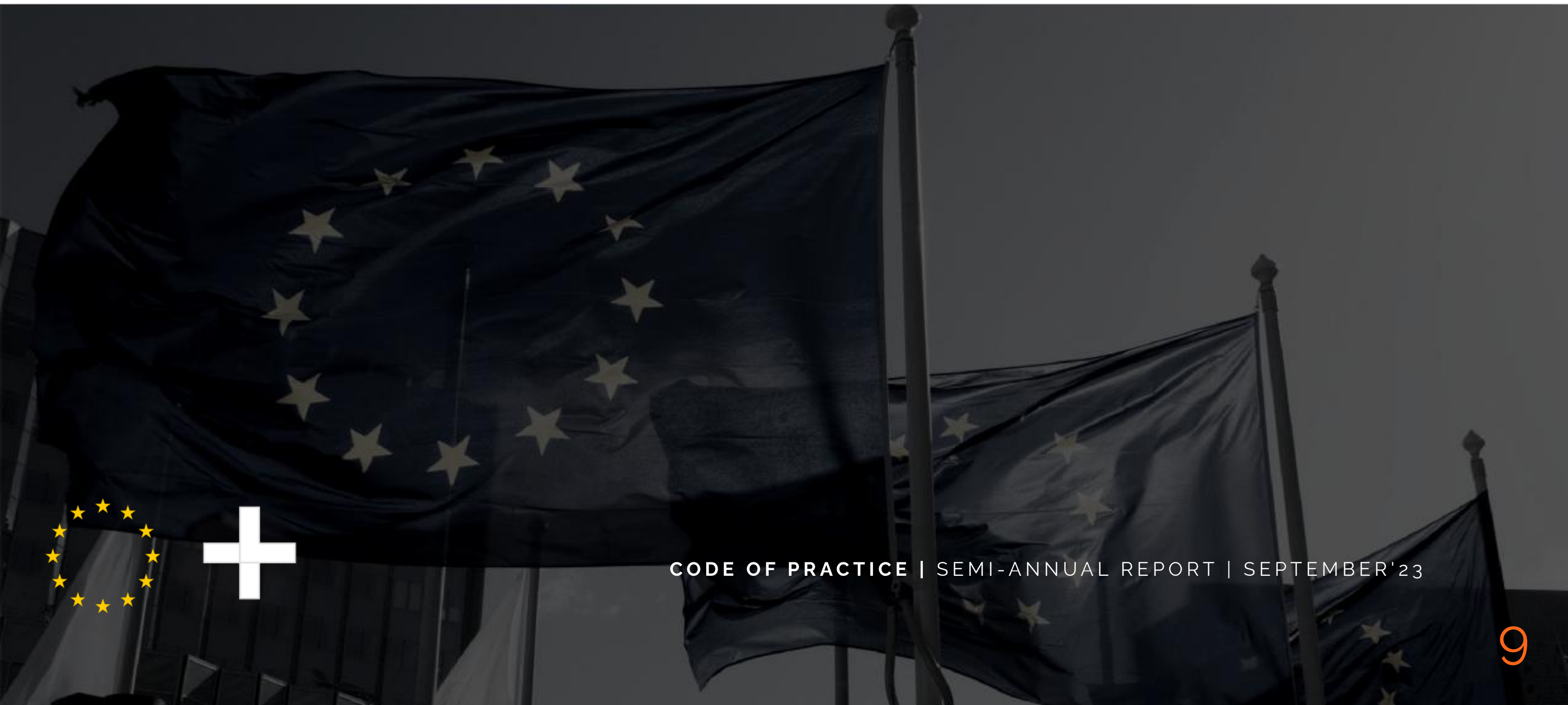
The **European Commission** helps to shape the European Unions' overall strategy, proposes new EU laws and policies, monitors their implementation and manages the EU budget. It also plays a significant role in supporting international development and delivering aid.

About the Participants and Partners

Avaaz is a global web movement to bring people-powered politics to decision-making everywhere.

The **European Regulators Group for Audiovisual Media Services (ERGA)** brings together heads or high level representatives of national independent regulatory bodies in the field of audiovisual services, to advise the Commission on the implementation of the EU's Audiovisual Media Services Directive (AVMSD).

The **European Digital Media Observatory (EDMO)** brings together fact-checkers, media literacy experts, and academic researchers to understand and analyse disinformation, in collaboration with media organisations, online platforms and media literacy practitioners.



About the Structural Indicators

Under Commitment 41 of the Code, signatories commit to working within the Task-force to develop Structural Indicators designed to assess the effectiveness of the Code in reducing the spread of online disinformation for each relevant signatory and for the entire online ecosystem in the EU and at the Member State level.

To achieve this, signatories established a Working Group in June 2022 following the launch of the Strengthened Code and the European Commission requested EDMO to create a first proposal for Structural Indicators to initiate discussions within the Working Group. EDMO presented a proposal at the beginning of September 2022, encompassing six different areas: prevalence, sources, audience, demonetisation of disinformation, collaboration, and investments in fact-checking and Code implementation. Due to the comprehensiveness of EDMO's proposal and the limited time available, the Working Group and EDMO agreed to focus on the prevalence, sources and audience of disinformation as the initial set of Indicators for the 2023 reporting.

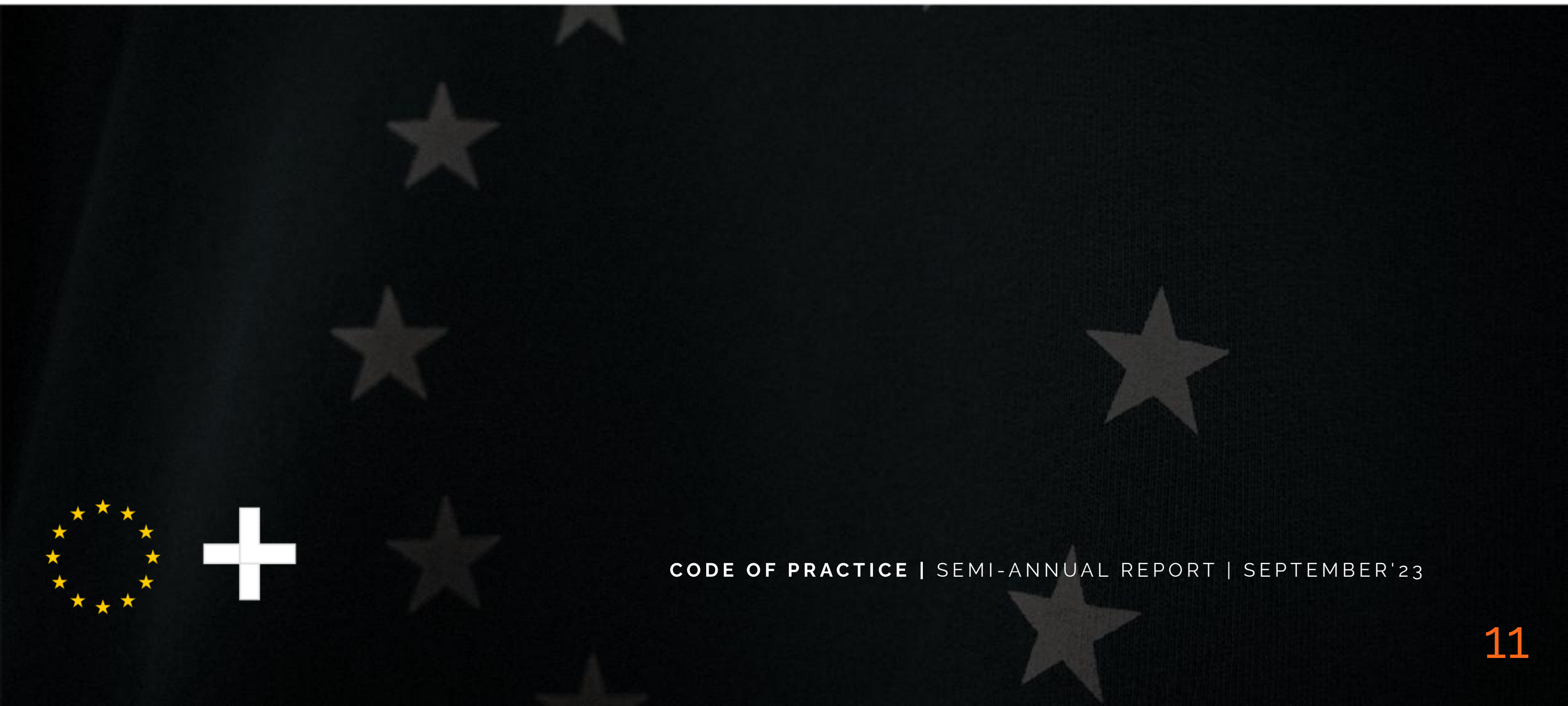
While in the course of autumn of 2022, several platform signatories had worked towards significantly increasing their data point availability on the prevalence and sources of disinformation, the tabled datasets and data points did not allow for satisfactory cross-platform Structural Indicators. Platform signatories noted that they had done their utmost to meet the Working Group's timelines and accommodate said data requests, taking into account (legal) constraints.

In January 2023, platforms committed to evaluating whether one or more third parties should be selected to assist in delivering the first set of Structural Indicators, either independently or with the support of EDMO, by the first reporting period. To ensure a harmonised approach across the main platforms (Facebook, Instagram, LinkedIn, TikTok, and YouTube) and to adhere to their Terms of Service, the Working Group issued a call for proposals and decided to contract TrustLab for an independent analysis of the selected indicators.

About the Structural Indicators

Due to the complexity of the task and given the limited time and resources available, the Working Group agreed to focus initially on a pilot analysis, comprising a smaller set of indicators (prevalence and sources of disinformation) and covering only three EU Member States: Poland, Spain, and Slovakia.

Looking ahead, the Working Group continues to explore ways to expand the scope and methodology of the Structural Indicators. In the interim, we are pleased to present TrustLab's pilot analysis of Structural Indicators.



Terminology

The following definitions are based on an amalgamation of peer-reviewed studies that TrustLab considers to be broadly aligned with industry standards that were then adopted for TrustLab's policies. These definitions are aligned with the European Union's 2022 Code of Practice on Disinformation; however, "foreign interference in the information space" is outside the scope of the current study. Better access to platform data and alignment on operational definitions can enable future measurements to address this limitation.

Term	Definition
Misinformation	Misinformation is false or misleading content shared without harmful intent though the effects can be still harmful.
Disinformation	False or misleading information that is spread with an intention to deceive or secure some advantage. This is a simplification of the EC definition ("false or misleading content that is spread with an intention to deceive or secure economic or political gain, and which may cause public harm") and reflects what was operationalized in the study. The intent of the actor, nature of gain they hope to receive, and how much public harm can be (or is intended to be) caused are not readily visible from the content itself, and need to be inferred, with ample room for subjectivity. To operationalise the measurement of disinformation, we focused on visible signs from the user who posted the content such as (but not limited to) repeat activity, size of the follower network, manipulation of images, video, or audio clips, the deliberate use of misleading headlines, or clickbait as a way to attract attention and promote false narratives.
Mis/Disinformation	A term intended to include both misinformation and disinformation.
Disinformation Actors	Accounts actively posting disinformation.

Terminology

Structural Indicator: Prevalence and Sources of Disinformation[2]

The following definitions relate to the metrics used by TrustLab in this pilot study.

Term	Definition
Discoverability	The percentage of content returned from searching disinformation keywords which is mis/disinformation. It captures how easily a platform surfaces mis/disinformation content to a user searching for sensitive topics.
Relative Post Engagement	The ratio of mis/disinformation engagement (where the underlying content is mis/disinformation) to non-mis/disinformation engagement (where the underlying content is non-mis/disinformation content).
Absolute Post Enagement	The magnitude, in absolute terms, of engagement with mis/disinformation content (with the caveat that the underlying data availability and nature across platforms can affect the magnitude of the metric).
Ratio of Disinformation Actors By Platform	The proportion of disinformation actors relative to the total accounts sampled on a platform.
Engagement With Disinformation Actors	The ratio of the engagement of disinformation actors with other users over the engagement of non-disinformation actors with other users on the platform. Absolute comparison is also provided. This sheds light on the influence that disinformation actors may exert on other users.
Disinformation Actor Account Activities	The group differences between disinformation and non-disinformation actors in post frequency and network size.

[2] Two sources metrics in the original proposal, share of disinformation actors by trends and disinformation actor demographics, are unavailable due to data limitations.

Introduction

Disinformation is a global issue that poses a threat to democracy and puts the health, security, and environment of (EU) citizens at risk. We define disinformation as false or misleading content that is spread with an intention to deceive, or secure some gain. Such content is typically spread through strategic campaigns, often targeting specific individuals or groups, aiming to mislead or distort public perception. Its existence predates the digital era. Disinformation is often perpetrated by stakeholders occupying spaces of power (politics, health authorities, culture, and arts) by means of traditional media (television, radio, and written press). However, the rise of new technologies at the beginning of the 21st century is considered to be, along with high speed propagation, its main catalysts today. For as long as individuals have attempted to manipulate information through disinformation and other tactics, others have tried to detect and counter it.

The implementation of the European Union's 2022 Code of Practice on Disinformation was transformative for the global fight against online disinformation. In order to assess the effectiveness of the Code, six structural indicators were proposed by the European Digital Media Observatory (EDMO)[3] to measure adherence to the Code for the leading large social media platforms hosting user generated content. In this pilot study, a comparative analysis was conducted to examine the first two structural indicators, namely the prevalence of disinformation and sources of disinformation. These indicators were measured across six major social media platforms - Facebook, Instagram, LinkedIn, TikTok, Twitter, and YouTube - across three countries: Poland, Slovakia, and Spain. The chosen countries represent both large and small EU Member States, each characterised by distinct population size, language, geography, and political ideologies.

[3] Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation. https://cadmus.eui.eu/handle/1814/75558

Furthermore, these countries were assessed to have a higher likelihood of being targets of disinformation during the pilot period (particularly due to upcoming elections or proximity to ongoing conflict between Russia and Ukraine). Taking a keyword-based approach, this study sampled posts and accounts that surfaced by searching sensitive topics on platforms. It evaluated the degree of mis/disinformation content and actors on each platform, and analysed the characteristics of these content and users, such as engagement and account activity patterns. Through this systematic cross-platform assessment, the study aims to establish a benchmark for implementing the Code of Practice, collaborate with the academic community to enhance methodologies, and lay the groundwork for forthcoming measurements in a wider array of countries and platforms.

Countries Overview

It is important to understand disinformation in the context of each country, such as the population size, social media usage, trust in media and institutions, and any existing measurements of disinformation. Slovakia’s country population, as well as internet user population is significantly smaller than that of Poland and Spain (Table 1). The internet adoption rate, however, is similar across the three countries (over 85 percent). YouTube and Facebook are consistently the two most widely used social media platforms in all three countries, with their particular dominance in Poland and Slovakia (much more popular than the third platform Instagram) (Figure 2).[4] Across these three countries, disinformation actors, fake accounts, and political bots all contribute to the historically-rooted and growing disinformation landscapes, as disinformation is increasingly being deployed to influence politics and public life. [5] [6] [7]

Table 1: Population Size and Internet Users by Country

	Poland	Slovakia	Spain
Population Size [8]	36,624,748	5,322,188	47,569,620
Internet Users [9]	32,300,000	4,806,000	44,180,000

[4] Google Trends, ‘Facebook, Twitter, Instagram, YouTube, Linkedin’, 2023, https://trends.google.com/trends/explore?geo=ES&q=Facebook,Twitter,Instagram,YouTube,LinkedIn&hl=en

[5] Gorwa, R. 'Poland: Unpacking the Ecosystem of Social Media Manipulation', in Samuel C. Woolley, and Philip N. Howard (eds), Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media, Oxford Studies in Digital Politics (New York, 2018; online edn, Oxford Academic, 22 Nov. 2018), https://doi.org/10.1093/oso/9780190931407.003.0005

[6] Miller, L., ‘Polarisation in Spain: more divided by ideology and identity than by public policies’, 2020, https://www.esade.edu/ecpol/en/publications/polarisation-spain/

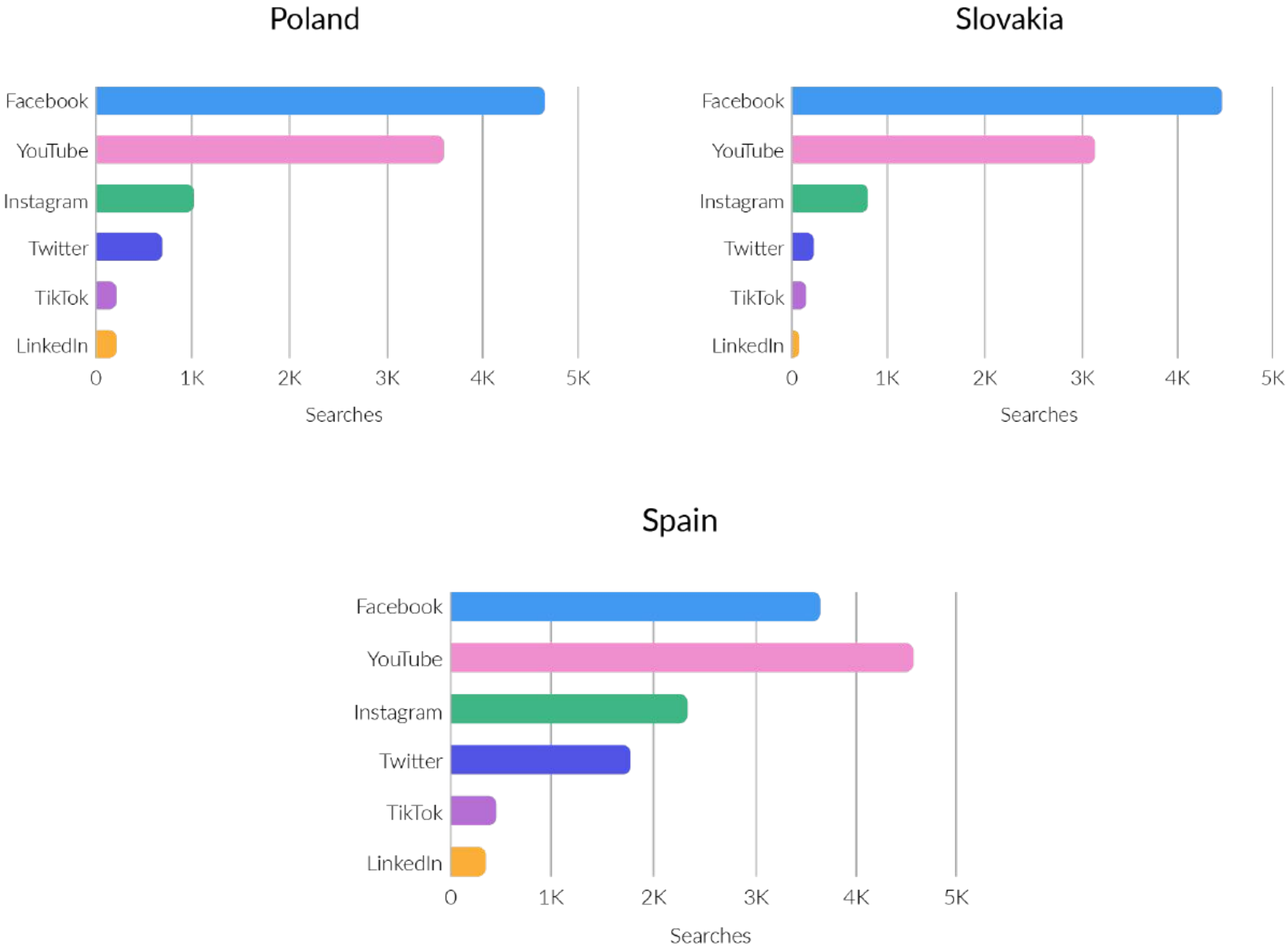
[7] Mrvová, I., ‘Skúma medziľudskú dôveru Vzorce spoločenského rozkladu dnes vidíme už aj na Slovensku’, 2023, https://www.postoj.sk/132944/vzorce-spolocenskeho-rozkladu-dnes-vidime-uz-aj-na-slovensku?fbclid=IwAR2ztLR7D37SgdeTkA2bLTRCHjmBOe2RzWLqfUlhYx8NttdcuaOUMoMh060

[8] The World Bank, ‘Data Catalog’, 2023, https://databank.worldbank.org/source/population-estimates-and-projections

[9] Central Intelligence Agency, ‘Country Comparisons Internet Users’, 2021, https://www.cia.gov/the-world-factbook/field/internet-users/country-comparison/

Figure 2: Platform Popularity on Google Trends.

Platform popularity is a normalised measure of total searches of the platform names on Google Search. Time period: last twelve months. Data downloaded on August 17, 2023. Note that the data in this chart is only meant to provide anecdotal indication of platform popularity in these regions and is not used as the basis for any of the metrics or conclusions in this study. A thorough review of various data sources to estimate platform popularity in a rigorous way is outside the scope of the current study.



Countries Overview

POLAND

In Poland, over half of internet users report encountering disinformation, and over a third report encountering false information online on a weekly basis.[10] [11] Yet, as many as 19 percent of Polish internet users state they did not check the credibility of online information or its sources.[12] Dominant disinformation narrative themes in Poland include socioeconomics, war reporting, politics, and conspiracy theories.[13] In addition, it is crucial to underline that since the outbreak of war in Ukraine, Poland has faced a surge in Russian disinformation since 2022, following the outbreak of the war in Ukraine with campaigns such as the one that accused Ukrainian refugees of assaults and rapes in Przemyśl and its neighbouring countries.[14] Although not within the scope of this pilot study, Telegram was a key social media platform in which disinformation relating to the Russo-Ukrainian war spread in Poland.[15]

SLOVAKIA

Slovakia appeared to be more vulnerable to large-scale disinformation campaigns compared to other nations in Central and Eastern Europe.[16] The vulnerability is rooted in the persistent pro-Russian sentiments that promoted its Communist past and spread anti-NATO/EU discourse.[17] A lack of trust in the government and news organisations has created fertile ground for disinformation campaigns to grow.[18] [19] A 2022 poll found that only 26 percent of Slovakian respondents trusted mainstream news, the lowest of 46 countries in the study; not many Slovakians believed the media was free from political (16 percent) or business (15 percent) influence.[20] This poll also confirmed that Facebook and YouTube were top platforms for messaging and social media in Slovakia.

The top disinformation narratives in Slovakia from 2018 to 2020 include migration, kidnapping of children, alternative medicine, and moon-landings.[21] Following events related to the Russo-Ukrainian War in 2022, Russian propaganda and disinformation about the war has surged in Slovakia. The Slovakian government has recognised disinformation as an issue and has taken recent regulatory actions against its spread, amending its cybersecurity law to block websites publishing harmful content.[22]

SPAIN

Spaniards widely recognize disinformation as a problem facing the country. A 2018 Eurobarometer report showed that 83 percent of respondents perceived fake news as a danger to Spanish democracy. Meanwhile, 53 percent of Spainards reported to have encountered fake news daily or almost daily,[23] and 57 percent of Spainards admitted to believing non-factual events had taken place.[24] At the platform level, a survey of the Spanish adult population found that the respondents perceived Facebook, Twitter, and WhatsApp to be the platforms with the highest rates of fake news.[25] Another COVID-19 misinformation study similarly reported highest rates of misinformation spread via WhatsApp and Twitter.[26]

Dominant Spanish disinformation narratives themes include politics (political and economic polarisation weakening democracy), identity attacks (Islamophobia, anti-Moroccan sentiment and gender and identity-based disinformation), health, and climate change.[27] Among its efforts to combat disinformation, in 2022, the Spanish government sentenced a person for spreading 'fake news' that stigmatised unaccompanied migrant minors, one of many recent regulatory actions that Spain has taken against disinformation.[28]

Countries Overview

Given each country's historical predisposition and current development of mis/disinformation spread, high rates of mis/disinformation content and actors are expected in all countries. There are likely to be country variations in metrics due to the differences in population size, geopolitical climate, and social media usage and norms.

[10] Naukowa i Akademicka Sieć Komputerowa (NASK), 'Badania Nask: Ponad Połowa Polskich Internautów Styka Się Z Manipulacją I Dezinformacją W Internecie', 2019, https://www.nask.pl/pl/aktualnosci/2249,Badania-NASK-ponad-polowa-polskich-internautow-styka-sie-z-manipulacja-i-dezinfo.html

[11] DigitalPoland, 'Dezinformacja oczami Polaków', 2022, https://digitalpoland.org/publikacje/pobierz?id=4f2e2116-82a6-47b5-a984-801b5e704b56

[12] Naukowa i Akademicka Sieć Komputerowa (NASK), 'Badania Nask: Ponad Połowa Polskich Internautów Styka Się Z Manipulacją I Dezinformacją W Internecie', 2019, https://www.nask.pl/pl/aktualnosci/2249,Badania-NASK-ponad-polowa-polskich-internautow-styka-sie-z-manipulacja-i-dezinfo.html

[13] Thompson, S. 'Fake News in the Polish Information Sphere Following the Russian Invasion of Ukraine in February 2022', 2022, https://www.diva-portal.org/smash/get/diva2:1697514/FULLTEXT02.pdf

[14] VoxUkraine, 'Russian disinformation in Poland: policy brief within Kremlin Watchers Movement project', 2023, https://voxukraine.org/en/russian-disinformation-in-poland-policy-brief-within-kremlin-watchers-movement-project

[15] European External Action Service, '1st EEAS Report on Foreign Information Manipulation and Interference Threats Towards a framework for networked defence', 2023, https://euvsdisinfo.eu/uploads/2023/02/EEAS-ThreatReport-February2023-02.pdf

[16] GLOBSEC, 'Which Slovaks Believe in Conspiracy Theories?', 2019, https://www.globsec.org/what-we-do/commentaries/which-slovaks-believe-conspiracy-theories

[17] Reporting Democracy, 'With Tiktoks And School Trips, Activists Take On Slovakia's Disinformation Ecosystem', 2023, https://balkaninsight.com/2023/02/28/with-tiktoks-and-school-trips-activists-take-on-slovakias-disinformation-ecosystem/

[18] Mrvová, I., 'Skúma medziľudskú dôveru Vzorce spoločenského rozkladu dnes vidíme už aj na Slovensku', 2023, https://www.postoj.sk/132944/vzorce-spolocenskeho-rozkladu-dnes-vidime-uz-aj-na-slovensku?fbclid=IwAR2ztLR7D37SgdeTkA2bLTRCHjmBOe2RzWLqfUlhYx8NttdcuaOUMoMh060

[19] GLOBSEC, 'Nový prieskum GLOBSEC-u: Slovensko zaznamenalo výrazný prepad v prozápadných postojoch', 2023, https://www.globsec.org/what-we-do/press-releases/novy-prieskum-globsec-u-slovensko-zaznamenalo-vyrazny-prepad-v

[20] Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., & Nielsen, R. K., 'Reuters Institute Digital News Report 2022', 2022, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

[21] Policajný Zbor, 'Správa Policajného Zboru O Dezinformáciách Na Slovensku V Roku 2022', 2022, https://www.minv.sk/swift_data/source/images/sprava-o-dezinformaciach-sr-2022.pdf

[22] Hečková, A. C. & Smith, S., ‘Slovakia’, 2022 https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/slovakia

[23] European Union, ‘Fake news and disinformation online’, 2018, https://europa.eu/eurobarometer/surveys/detail/2183

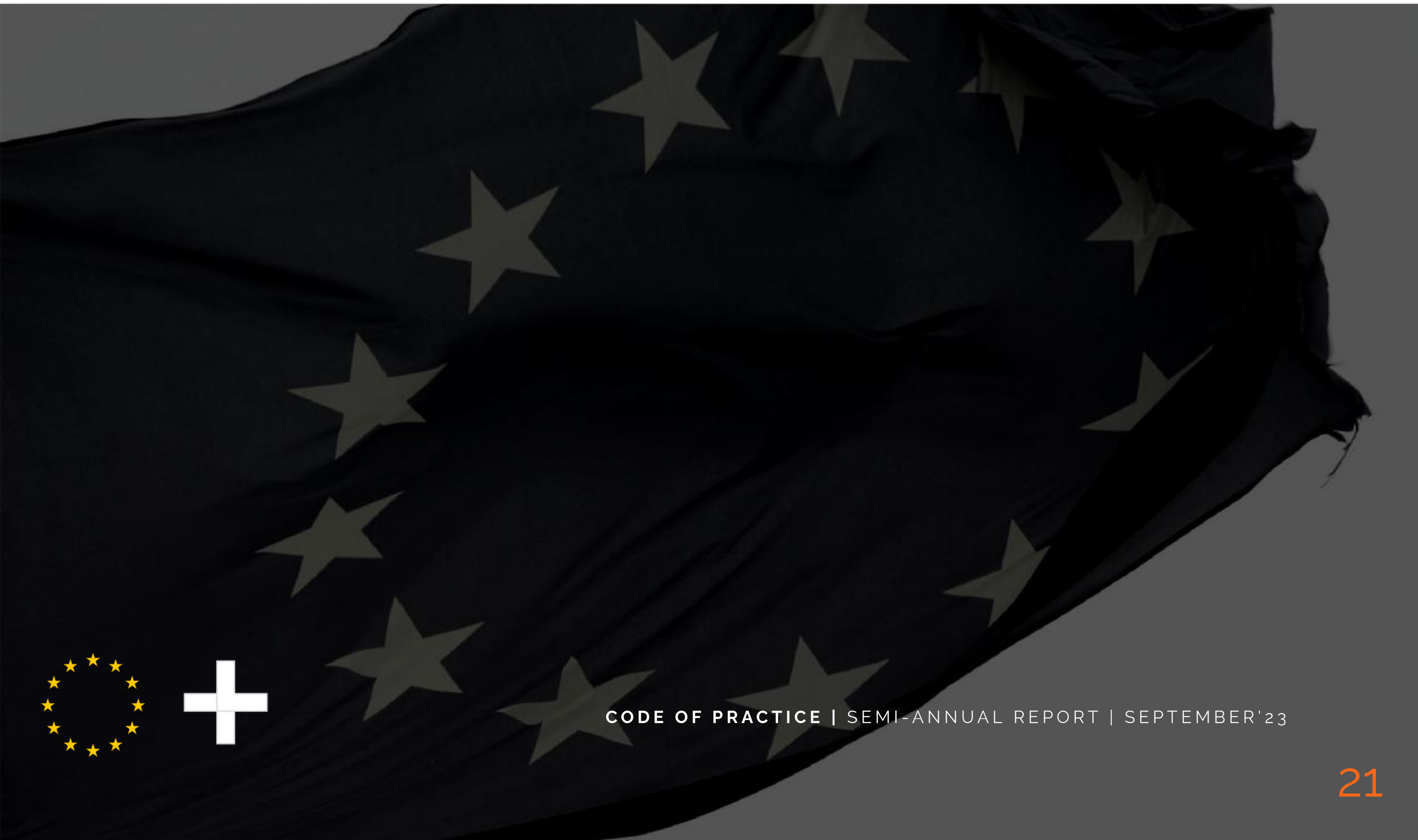
[24] Ipsos,’Fake News, Filter Bubbles, Post-truth and Trust: A Study Across 27 Countries’, 2018

[25] Blanco-Herrero, D., Amores, J.J. and Sánchez-Holgado, P., ‘Citizen perceptions of fake news in Spain: socioeconomic, demographic, and ideological differences’, 2021, *Publications*, *9*(3), p.35.

[26] Almansa-Martínez, A., Fernández-Torres, M.J. and Rodríguez-Fernández, L., ‘Disinformation in Spain one year after COVID-19. Analysis of the Newtral and Maldita verifications’, 2022, *Revista Latina de Comunicación Social*, (80), pp.183-199.

[27] EU DisinfoLab, ‘Disinformation Landscape in Spain’, 2023, https://www.disinfo.eu/wp-content/uploads/2023/03/20230224_SP_DisinfoFS.pdf

[28] El País, ‘First conviction in Spain for spreading 'fake news' about migrant minors’, 2022, https://elpais.com/espana/catalunya/2022-11-08/primera-condena-en-espana-por-difundir-fake-news-sobre-los-menores-migrantes.html



Methodology

The structural indicators proposed by EDMO[29] ideally require collecting random samples of all content and accounts on platforms. This study, however, took a different sampling approach due to a lack of access to internal platform data. Specifically, the sampling approach in this study was to search for specific keywords related to current mis/disinformation topics on each platform's native search engine, and collect the posts and accounts from the search results. The collected data does not represent a random sample of all content and accounts on that platform, but instead represents the content and accounts encountered by users of the platform who are searching for keywords related to mis/disinformation topics.[30]

To understand a key difference between the two sampling approaches, consider that when normalising the amount of vaccine misinformation, the denominator can either be all content on the platform or only vaccine related content. Both measures are meaningful, and offer different perspectives on the prevalence of misinformation. A recent Stanford study[31] shows that extremist YouTube content is mainly viewed by those seeking it out, indicating the need for sampling from the perspective of users actively seeking related content.[32]

The data collection process consisted of compiling the latest popular mis/disinformation claims, extracting keywords from them, and then sampling content and accounts using these keywords, identifying positive cases (mis/disinformation content or actors) and neutral or negative cases (non-disinformation content or users). It was carried out in three measurements over a period of six weeks. For many metrics, we looked at the relative proportions of good-to-bad (non-mis/disinformation to mis/disinformation) content, or good-to-bad actors behind the content, in order to draw cross-platform conclusions.

[29] Structural indicators to assess effectiveness of the EU's Code of Practice on Disinformation. https://cadmus.eui.eu/handle/1814/75558

[30] By broadening the set of keywords and applying other techniques from the literature, we can approach a random sample, but this was outside the scope of the current study due to time and budget constraints.

[31] Study Finds Extremist YouTube Content Mainly Viewed by Those Seeking it Out. https://cyber.fsi.stanford.edu/news/study-finds-extremist-youtube-content-mainly-viewed-those-seeking-it-out

[32] While the current study only employs keyword search based sampling for finding related content, the methodology can also be extended to the feed as outlined in the Limitations and Implications section.

Mis/Disinformation Claim Collection

To generate keywords, a list of popular mis/disinformation claims on the internet was first curated for each of the local markets: Poland, Spain, and Slovakia. A qualified mis/disinformation claim is a claim that:

- References a mis/disinformation narrative (factually inaccurate claims based on current fact check articles or other strong evidence presented by a trusted source)
- Is harmful (focused on mis/disinformation claims in critical topics such as elections, politics, COVID-19, and the Russo-Ukrainian war)

In each country, a team of three disinformation analysts worked together to collect disinformation claims. They first identified certified trustworthy fact-check websites, then searched for other sources beyond fact-checks based on fact-check articles, research papers, newspaper articles, and review websites that curate lists of disinformation sources such as Media Bias Fact Check and the Global Information Index. The fact-check websites and supplemental materials were evaluated based on the timeliness of claims, the popularity on social media with tools like CrowdTangle, and high potential for impact (high viewership or measured popularity). The saturation principle[33] was used to collect an exhaustive set of claims within the bounds of time and resource budgets for the current project.

The same groups of analysts then selected the most recent and relevant mis/disinformation claims from fact-checking websites, disinformation websites, and disinformation users. (Note the terms "disinformation websites" and "disinformation users" in this section refer to the sources that provided disinformation claims for the study.)

[33] To be more precise, our approach to claim collection adhered to the principle of saturation, a widely used concept in qualitative research for choosing the appropriate sample size and data gathering procedure. According to this principle, the data collection process continues until redundancy becomes evident in the gathered information, making it unlikely that further data collection would yield significantly distinct findings. In other words, saturation is achieved. See more details in Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. 'Saturation in qualitative research: exploring its conceptualization and operationalization', 2018. Qual Quant. 52(4):1893-1907. doi: 10.1007/s11135-017-0574-8. Epub 2017 Sep 14. PMID: 29937585; PMCID: PMC5993836.

Each claim was fact-checked as mis/disinformation in the same thorough fact-checking process as described above. Popular claims (i.e. with high user engagement) were prioritised, while ensuring that claims from a variety of topics were included. In each country, the final collection of claims included a list of 100 mis/disinformation claims during the first round of data collection and 30 new claims in each of the next two rounds of data collection The final claims cover a broad range of topics, such as politics, healthcare, and the Russo-Ukrainian war.[34]

Mis/Disinformation Keyword Generation

During each measurement, we randomly sampled 40 disinformation claims per country,[35] from which we extracted disinformation keywords. This number (along with the decision to collect five posts per keyword in the next section) was chosen based on a combination of power calculations,[36] expertise from the team based on previous studies that TrustLab has performed, and budget and time restrictions of this project. Disinformation specialists summarised these selected claims as commonly used words and phrases known as keywords.

A qualified keyword must accurately represent its corresponding claim and must be precise, which means that a Google search for the keyword should yield at least one of the top three search results related to the same topic as the original disinformation claim. Note that agents were not instructed to select keywords based on whether mis/disinformation was found in the top three results, but only based on whether related content was found. The content could be mis/disinfo, or a fact-check article or a neutral opinion related to the claim.

[34] In the first measurement, the analysts were instructed to collect recent disinformation claims published from the past month. In the subsequent measurements, they focused on finding fresh claims published since the last measurement (i.e. in the past two weeks).

[35] In the first period, a sample of 40 claims was randomly chosen from a pool of 100 available claims. In the second period, another 40 claims were selected from a pool of 130 available claims. Finally, during the third measurement period, a sample of 40 claims was drawn from a total pool of 160 available claims.

[36] Power calculations represent a frequently used statistical technique for evaluating whether a study possesses sufficient statistical power to derive meaningful findings. In the context of this study, we explored different combinations of values for keyword counts and posts per keyword in order to find the appropriate sample size capable of attaining the predefined standard errors.

Similarly, agents are asked to not issue queries to social media sources during the keyword selection process to avoid bias in favour of or against one or more social platforms. At each measurement, mis/disinformation claims collected from the previous measurements remained relevant and were re-used in the sample to draw keywords, so there exists some keyword overlap between measurements per country.[37]

Social Media Data Collection

Local analysts living in Poland, Spain, and Slovakia collected social media data in their respective country to ensure that the data collected reflected typical online search results in the specified country. Each analyst registered a new social media account on each of the six platforms being measured to remove bias from previous account history. These analysts searched posts using the platform's native search functionality. For each keyword search, a maximum of five posts were collected.[38] For each user identified from keyword searches, up to 50 posts were collected from the profile of the user.

A few measures were taken to ensure data collection quality. First, after the initial training, agent performance was continuously monitored. Second, a random sample of 1,000 (<1 percent) data points, such as the text of a post or the number of followers of a user, were audited for data accuracy.

[37] During the first measurement, a random sample of 40 disinformation claims out of 100 total claims was chosen for keyword generation. In the second measurement, the 100 initial disinformation claims were pooled together with the 30 new claims, and a weighted sample of 40 claims were drawn for keyword generation (new claims got double the weights than the first-round claims to increase the likelihood of new claims being selected). By the third measurement, a total of 160 disinformation claims had been gathered (100 initial claims and 30 new claims from each of the subsequent measurements). Again, a weighted sample of 40 disinformation claims were selected for keyword extraction.

[38] The post collection process was altered for Instagram because the typical search keyword format (multiple words separated by space) failed to yield many search results on Instagram (Appendix A1).

Mis/Disinformation Content Labelling

Mis/disinformation identification had three dimensions: dubiousness,[39] fact checkability,[40] and harmfulness.[41] Analysts reviewed the original posts and answered a series of questions about the three components. The final mis/disinformation label was then derived from the scores on the three components.[42]

The labelling process was multi-tiered. Tier-1 analysts are frontline content moderators responsible for handling the initial review of a piece of content. Tier-2 analysts are more experienced and specialised content moderators who deal with complex or nuanced issues that require expertise beyond the scope of Tier-1 analysts. After Tier-1 analysts reviewed the content, tier-2 analysts re-reviewed the positive cases flagged by the Tier-1 analysts as a quality assurance. All analysts that worked on mis/disinformation labelling had completed pre-project training. Further quality assurance measures included continuous agent performance monitoring, feedback sessions, and daily updates of the precision metric based on the tier-2 reviews. Moreover, tier-2 analysts reviewed a random sample of 10 percent (n = 600) negative cases and calculated the ratio of true negative ratings (Appendix A2).

Disinformation Actor Identification

The analysts that performed the mis/disinformation content labelling also performed the disinformation actor labelling. The analysts received rigorous training, regular debriefing sessions, and full-time support from TrustLab's policy team throughout the project.

[39] Content that raises significant doubts about its accuracy, truthfulness, or credibility. It implies that the information is questionable or suspicious and may not be trustworthy, potentially containing elements of misinformation or disinformation.

[40] The degree to which a statement, claim, or piece of information can be verified or corroborated by reliable and objective sources, and it's accuracy can be proven or disproven.

[41] The negative impact and consequences that false or misleading information can have on individuals, communities, societies.

[42] These were assessed on a 1-5 scale encompassing the least risky content to most risky content, whether the content is true and can be explicitly fact-checked to content is false and can be explicitly proven as false, and no likelihood of serious physical harm to high likelihood for serious physical harm or death.

Analysts reviewed (up to) 15 posts per user, among the (up to) 50 that were collected, which included all of the known mis/disinformation posts by the user and the most recent posts of the user.[43] Each account was evaluated for the following criteria (Table 2): (1) if the account posted three or more misinformation posts (if the answer was true, the account proceeded to the next check); (2) if the account had a large following (5,000[44] followers or more); and, (3) if the account posted frequently (3+ times a month) about a specific topic, or both. If the account met the criteria again, it was sent to a secondary review. Secondary review consisted of an assessment of disinformation account status (i.e. adherence to the disinformation actor definition) via examination of an account's content topics, post engagement, post frequency, follower count, and review of additional pieces of content beyond the 15 pieces of content. The secondary review process reassessed the content to determine if the analysts were correct in their evaluation of said content. At the same time, analysts were given the opportunity to escalate accounts that passed the first check, but not the second check for a holistic secondary review if the account seemed suspicious. Accounts that passed secondary review (either the escalated accounts or accounts that satisfy all three criteria) make up the sample of disinformation actors and the basis to calculate sources-related metrics.

[43] Users differ in their number of identified mis/disinformation posts. Therefore, to show analysts 15 posts per user, the number of most recent posts shown to analysts also differs by user. If a user had n identified mis/disinformation posts, 15-n of their most recent posts would be included to compose the 15-post sample. For example, if a user had 3 identified mis/disinformation posts in our database, 12 of their most recent posts would be included in the 15-post sample for analyst review.

[44] An industry standard for a large following is approximately 20k followers. However, this threshold needed to be adjusted due to the relative population of these countries so a following of 10k was determined to be a large following for the purposes of this project. Upon first review, agents were finding the 10k threshold was hard to meet based on the population of these countries so it was lowered to 5k to be more reasonable of a threshold to meet when determining disinformation account status.

Methodology

The secondary review contains some degree of subjectivity (a holistic review of multiple account characteristics), so we consider two alternative quantifiable definitions of disinformation actors as sensitivity checks. One alternative definition treats accounts that pass the first criterion (i.e. posting three or more pieces of misinformation recently), regardless of the second criteria or secondary review, as disinformation actors. We call this group of accounts Level 1 disinformation actors. The second alternative definition requires accounts to pass both of the first two criteria to qualify as disinformation actors. These accounts are referred to as Level 2 disinformation actors. By definition, both the main sample and the Level 2 disinformation actors are a subset of Level 1 disinformation actors. The main sample is also largely a subset of Label 2 disinformation actors (except for the escalated cases). Measurement of disinformation sources will be applied to the main sample, as well as the two alternative samples as robustness checks. (Appendix A7)

Table 2: Disinformation Actor Labelling Procedure

Condition 1	Did the account post three or more misinformation posts recently?
Condition 2	(1) Did the account have a large following (5,000 followers or more), (2) posted frequently (3+ times a month) about a specific topic, (3) or both? *(An account may be escalated to bypass Condition 2 to proceed to Condition 3)*
Condition 2	Did the account pass secondary review (examination of an account's content topics, post engagement, posting frequency, and follower count)?

Metrics

The search-based sampling approach described in the Methodology section underpins all of the metrics calculated in this study. The metrics are described in more detail below.

Discoverability

The percentage of misinformation content is a common measure of misinformation prevalence on social media platforms.[45,46] This study calculates a variation of prevalence called discoverability. Discoverability represents the proportion of search results from the study sample that are labelled as mis/disinformation content. A higher discoverability implies that a user can more easily find mis/disinformation content on a platform when they search for keywords related to popular mis/disinformation narratives.

The formula of discoverability is: $\boldsymbol{discoverability} = \frac{N_{mis/disinfo}}{N}$

where $N_{mis/disinfo}$ represents the number of mis/disinformation posts in the search results, and N represents the total number of posts in the search results. The standard error of the discoverability metric follows the formula $SE = \sqrt{\frac{d(1-d)}{N}}$ (where d refers to the value of discoverability), and 90 percent confidence intervals were reported.

[45] Do Nascimento, I.J.B., Pizarro, A.B., Almeida, J.M., Azzopardi-Muscat, N., Gonçalves, M.A., Björklund, M. & Novillo-Ortiz, D., 'Infodemics and health misinformation: a systematic review of reviews', 2022. Bulletin of the World Health Organization, 100(9), p.544.

[46] Sell, T.K., Hosangadi, D. & Trotochaud, M., 'Misinformation and the US Ebola communication crisis: analyzing the veracity and content of social media messages related to a fear-inducing infectious disease outbreak', 2020, BMC Public Health, 20(1), pp.1-10.

Relative Post Engagement

Relative post engagement is operationalized as a simple ratio of average active engagement with mis/disinformation posts over average active engagement with non-mis/disinformation posts among all the posts from searching disinformation queries. In other words, it is the ratio of “bad-to-good” engagement (where bad or good engagement is determined by whether the underlying content is mis/disinformation) collected from search results. Note that although the posts were collected using keyword search, engagement data on these posts is a reflection of all ways in which users of the platform find this content, through search, through feed, through external referrals and so on. Bootstrap confidence interval was calculated.[47]

$$Relative\ Post\ Engagement\ = \frac{Average\ (Engagement\ with\ Mis/disinfo\ Posts)}{Average\ (Engagement\ with\ Non{-}Mis/disinfo\ Posts)}$$

Active post engagement is measured as the sum of the volume of three subtypes of engagement: reactions, comments, and shares whenever available, which is consistent with past studies.[48,49,50] Reactions, comments, and shares represent a kind of active user interactions, as opposed to views. Views as engagement are not within the scope of the current study. Posts that have existed on platforms longer are more likely to have larger engagement. Although the main model could not account for time, we were able to control for time in the robustness checks.

[47] Bootstrap standard errors were calculated by drawing 2,000 repeated samples with replacement from the initial posts, calculating the relative engagement metrics in each sample, and selecting the 5th and 95th percentiles (the lower and upper bounds of the 90 percent confidence interval respectively).

[48] Allcott, H., Gentzkow, M. & Yu, C., ‘Trends in the diffusion of misinformation on social media’, 2019,. Research & Politics, 6(2), p.2053168019848554.

[49] Fletcher, R., Cornia, A., Graves, L. & Nielsen, R.K., ‘Measuring the reach of "fake news" and online disinformation in Europe’, 2018. Australasian Policing, 10(2).

[50] Marchal, N., Kollanyi, B., Neudert, L. M., Au, H., & Howard, P. N. ‘Junk News & Information Sharing During the 2019 UK General Election’, 2020, arXiv preprint arXiv:2002.12069.

Table 3: Disinformation Actor Labelling Procedure[51]

Platform	Type of Engagement
Facebook	Reactions + Comments + Shares
Instagram	Reactions
LinkedIn	Reactions + Comments + Shares
Tik Tok	Reactions + Comments + Shares
Twitter	Reactions + Comments + Shares
YouTube	Reactions + Comments

Two sets of robustness checks were conducted.[52] First, relative post engagement was estimated using a Poisson regression model. The unit of analysis is posts. Post engagement, which is a count variable, is the dependent variable (DV). Whether a post contains mis/disinformation is the independent variable (IV). The control variable is the number of days between the time a post was created and the time the post was collected. The coefficient of the IV represents the log change in post engagement between mis/disinformation posts and non-mis/disinformation posts. The exponent of the coefficient represents the ratio of relative engagement.[53]

[51] Numbers of comments and shares are unavailable on Instagram. Number of shares is unavailable on YouTube.

[52] Duplicate posts were dropped to prevent correlated errors in the regression models. They were also dropped in the ratio measurement.

[53] The exponent of the coefficient represents the ratio of the expected count for mis/disinformation engagement divided by the expected count for non-mis/disinformation engagement in both Poisson and negative binomial regression models.

Absolute Post Engagement

The absolute number of engagement counts is another way to examine the reach and influence of a post. Suppose two hypothetical platforms have the same relative engagement ratio of 0.1. On one platform, the ratio is derived from dividing 1 unit of “bad” engagement (with mis/disinformation) over 10 units of “good” engagement (with non-mis/disinformation) (1 / 10 = 0.1). On the other platform, the ratio is derived from dividing 1 million units of “bad” engagement by 10 million units of “good” engagement (1 million / 10 millions = 0.1). While the relative post engagement makes platform-wise comparison easier, it does not capture the difference in the scope of engagement. Absolute engagement, however, should be interpreted with the caveat that availability and nature of the underlying engagement data may differ across platforms.

Absolute engagement is measured as the average “bad” engagement and average “good” engagement in absolute numbers (where “bad” or “good” is determined by whether the underlying content is mis/disinformation or not), and the share of absolute engagement of the top N posts over the total engagement of all posts.

Absolute engagement is measured as the average “bad” engagement and average “good” engagement in absolute numbers (where “bad” or “good” is determined by whether the underlying content is mis/disinformation or not), and the share of absolute engagement of the top N posts over the total engagement of all posts.

Distribution of Disinformation Actors by Platform

This metric refers to the proportion of disinformation accounts relative to the total accounts sampled on the platform in this study. (As discussed in the Methodology section, content and accounts were sampled using keyword searches.)

Disinformation Actor Account Activity

Disinformation actor account activity is measured by the posting frequency and number of followers and following.

Engagement with Disinformation Actors

This metric is built on account-level engagement, which describes the amount of interaction that a disinformation actor had with other users. Account-level engagement is an aggregate measure of engagement with posts made by a user. Post engagement is calculated in the same way as the sum of reactions, comments, and shares (Table 3). Up to 50 of the most recent posts were collected for each user. The content size of 50 was chosen to maximise the number of posts per user and user activity history within feasible resources. All (up to) 50 posts or a subset of them were used to calculate account-level engagement variables. For each user, we calculated the maximal post engagement, which is the largest post engagement among all the posts by the user. The maximum may be subject to outliers, so in the main analysis a more robust measure of the 90th percentile was adopted instead. We computed the average level of engagement across all the posts gathered from each single user. The last account-level engagement measure is the total post engagement within a week, which is the sum of engagement of posts that the user published in the week prior to the time of data collection (see Appendix A2 for more details). The group means and ratio of the 90th percentile engagement, average post engagement, and total weekly engagement were calculated for disinformation and non-disinformation actors respectively.

Descriptive Statistics

By searching disinformation keywords on platforms, a total of 6,155 unique social media posts were collected across all three Member States. They were then reviewed by analysts for misinformation content and used to calculate prevalence metrics (Table 4). The accounts behind these posts represent the sample for the sources metrics. There were a total of 4,460 unique accounts. User profile information and up to 50 most recent posts were collected from each account.

The collected data was distributed unevenly across the various platforms (Table 4). Keyword searches found the most posts on YouTube (1,777 posts) and TikTok (1,730 posts), and the least posts on LinkedIn (609 posts) and Instagram (296 posts). While the same keywords and methodology were applied to each platform, the number of posts differs because of variations in keyword popularity and platform usage in each local market. YouTube had the largest number of posts, but its number of accounts was not the largest, which implies that some posts in the sample were created by the same users. The most accounts were collected on TikTok (1,197), Facebook (977), and Twitter (804).

Table 4: Sample Size

	Data for Prevalence	Data for Source
Platform	Posts	Accounts
Facebook	1,457	977
Instagram	281	229
LinkedIn	508	462
Tik Tok	1,463	1,197
Twitter	966	804
YouTube	1,480	791
Total	**6,155**	**4,460**

Prevalence of Disinformation

Finding 1: Twitter has the highest discoverability, while YouTube has the lowest.

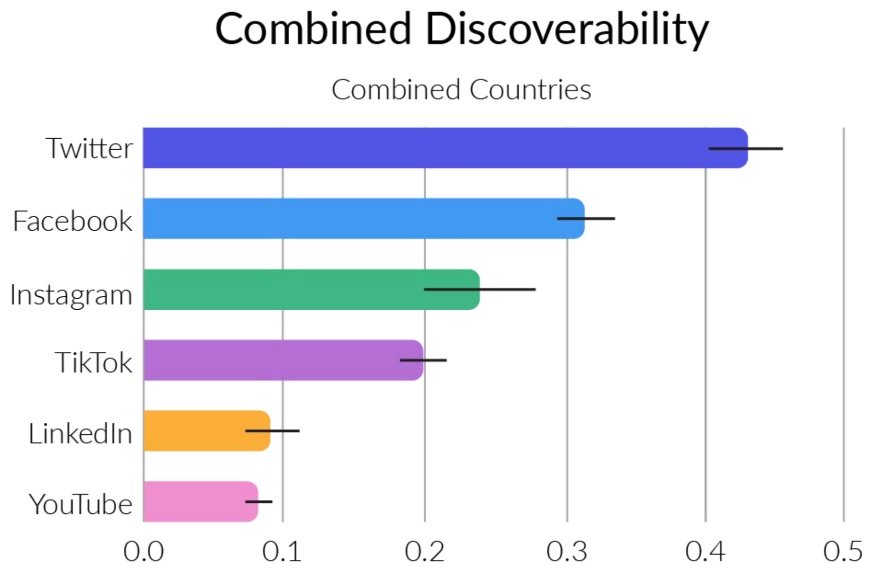
We begin by addressing the overall analysis that consolidated data from all countries and measurements and has the most extensive sample size. A fraction of 0.428 of the posts collected via disinformation search queries are mis/disinformation on Twitter, the highest overall mis/disinformation discoverability (Figure 3a; Table 5). Facebook has the second highest mis/disinformation ratio of 0.313. LinkedIn (0.092) and YouTube (0.082) have the lowest ratios of mis/disinformation, both of which are under 10 percent.[55]

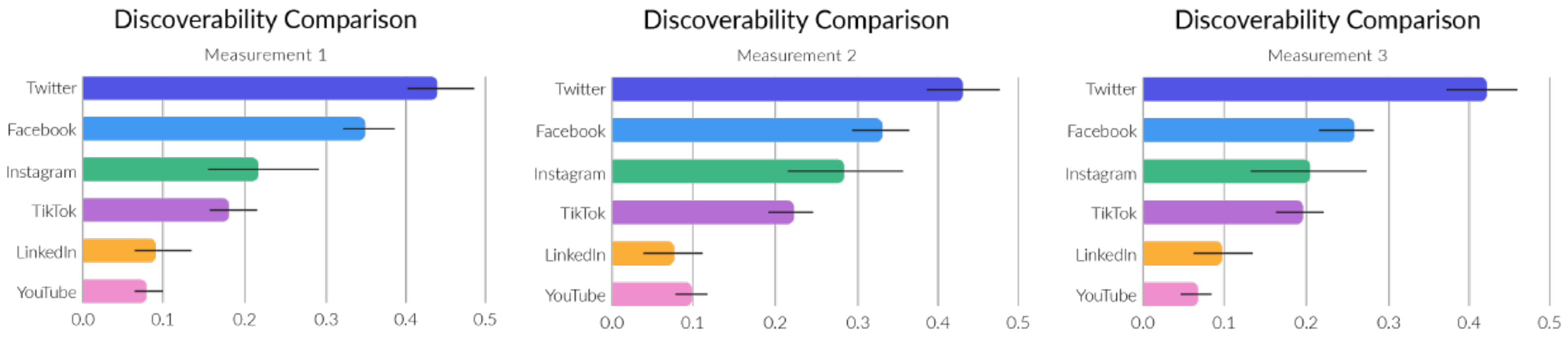
The platform order remains stable between measurements (Figure 3b). The top four platforms stay the same at each measurement, yet the last two platforms, YouTube and LinkedIn, switched places between measurements, but their difference in discoverability was never statistically significant (Appendix Tables A4.2-4).

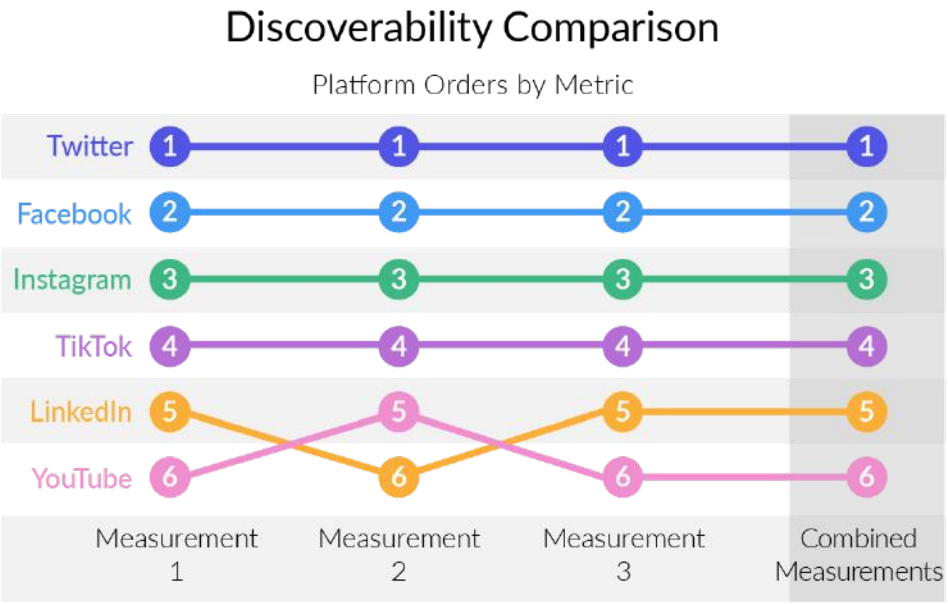
In contrast, the platform's discoverability exhibits greater variability and is less consistent across countries (Figure 3C). This variance may reflect differences in local contexts or the roles platforms play in various countries. However, it is important to acknowledge that certain estimates, especially for specific platforms in Slovakia, may be less precise due to the limited size of the sample.

[55] See Appendix A4 for the pairwise t-tests for platform differences in discoverability.

Findings







	Measurement 1	Measurement 2	Measurement 3	Combined Measurements
Twitter	1	1	1	1
Facebook	2	2	2	2
Instagram	3	3	3	3
TikTok	4	4	4	4
LinkedIn	5	6	5	5
YouTube	6	5	6	6

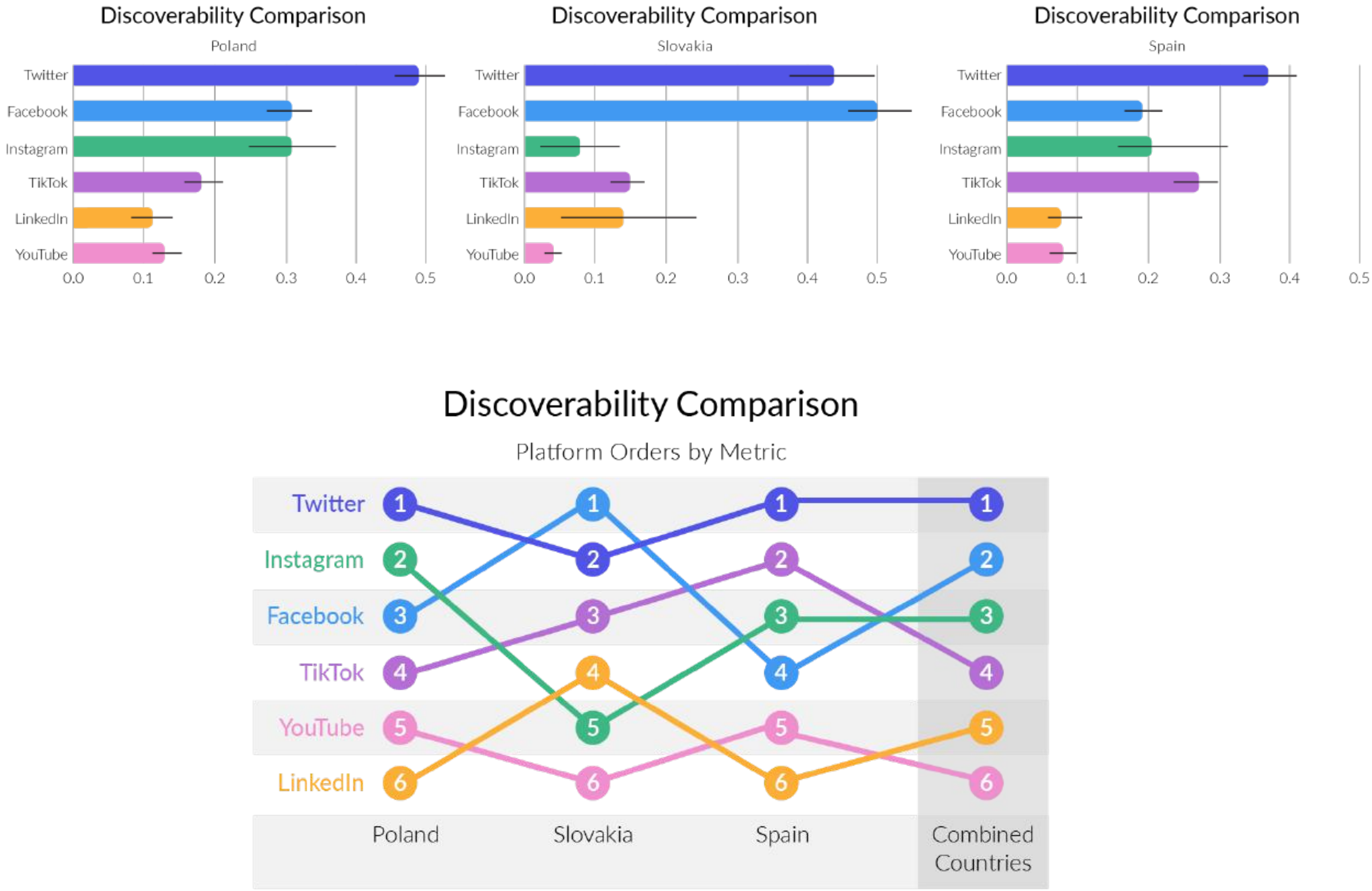


Figure 3: Misinformation Discoverability by Platform, Country, and Measurement. In the bar charts, the bars represent the point estimates of discoverability, and the black error bars represent the 90 percent confidence intervals. Note that the Instagram estimates have slightly wider confidence intervals due to its smaller sample size relative to other platforms. The line chart aims to assist in discerning changes in platform positions. A straight line indicates a consistent platform position between countries or measurements, while an intersection denotes a shift in the platform's position.

Table 5: Mis/Disinformation Discoverability by Platform, Measurement, and Country

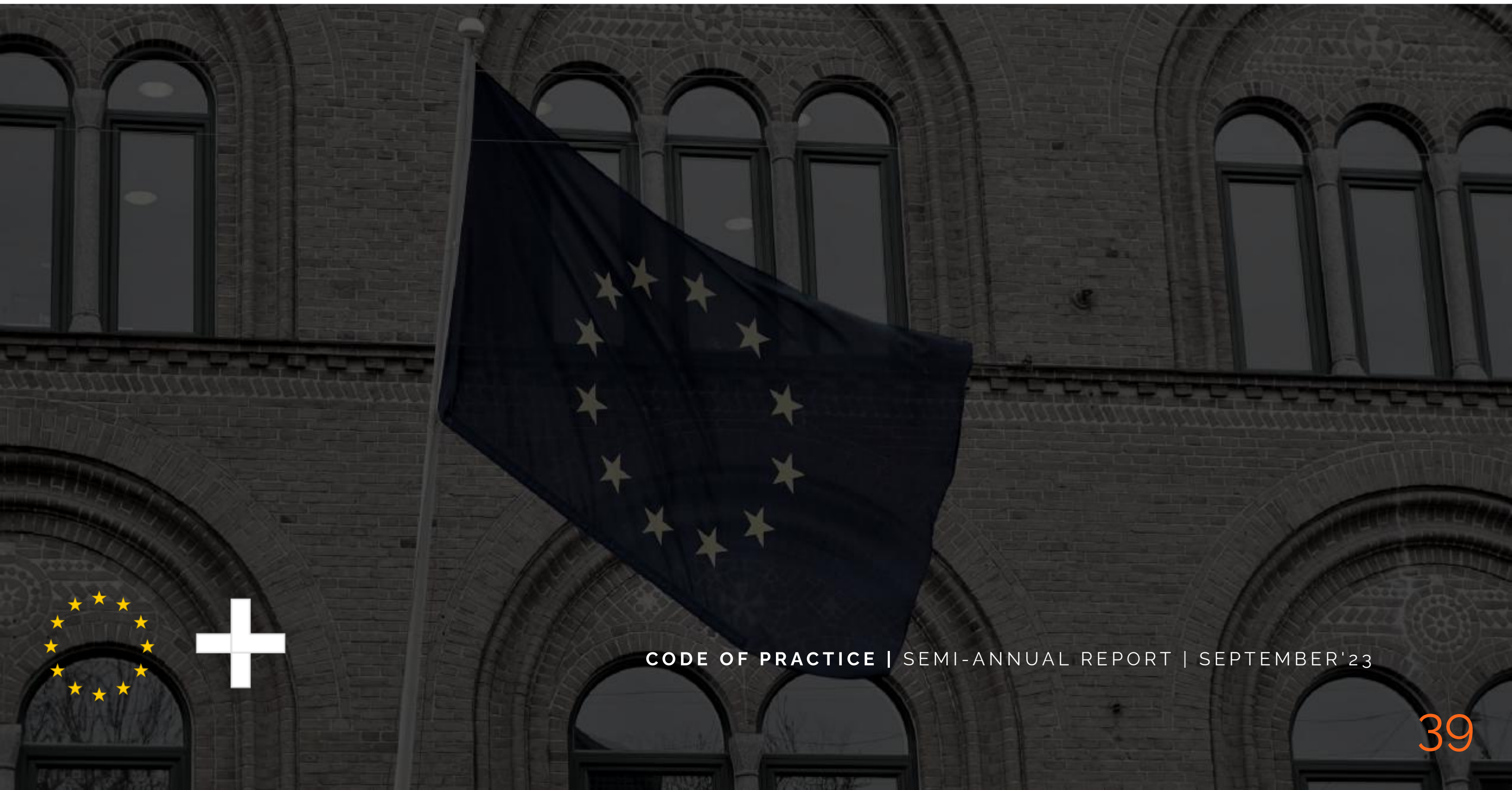
Overall Sample

Platform	Sample Size	Discoverability *
Facebook	1,503	0.313 (0.294, 0.333)
Instagram	296	0.240 (0.199, 0.281)
LinkedIn	609	0.092 (0.073, 0.111)
TikTok	1,730	0.199 (0.183, 0.215)
Twitter	1,075	0.428 (0.403, 0.453)
YouTube	1,777	0.082 (0.071, 0.093)

* Estimate and 90% confidence interval

	Measurement 1		Measurement 2		Measurement 3	
Platform	Sample Size	Discoverability	Sample Size	Discoverability	Sample Size	Discoverability
Facebook	532	0.353 (0.319, 0.387)	496	0.327 (0.292, 0.361)	475	0.255 (0.222, 0.288)
Instagram	98	0.224 (0.155, 0.294)	110	0.282 (0.211, 0.352)	88	0.205 (0.134, 0.275)
LinkedIn	231	0.1 (0.067, 0.132)	177	0.073 (0.041, 0.106)	201	0.1 (0.065, 0.134)
TikTok	585	0.186 (0.16, 0.213)	550	0.216 (0.187, 0.245)	595	0.195 (0.168, 0.222)
Twitter	395	0.438 (0.397, 0.479)	348	0.425 (0.382, 0.469)	332	0.419 (0.374, 0.463)
YouTube	580	0.084 (0.065, 0.103)	599	0.097 (0.077, 0.117)	598	0.065 (0.049, 0.082)

	Poland		Slovakia		Spain	
Platform	Sample Size	Discoverability	Sample Size	Discoverability	Sample Size	Discoverability
Facebook	523	0.306 (0.273, 0.339)	404	0.502 (0.462, 0.543)	576	0.188 (0.161, 0.214)
Instagram	156	0.308 (0.247, 0.368)	61	0.082 (0.024, 0.14)	79	0.228 (0.15, 0.305)
LinkedIn	229	0.109 (0.075, 0.143)	35	0.143 (0.046, 0.24)	345	0.075 (0.052, 0.099)
TikTok	570	0.182 (0.156, 0.209)	585	0.149 (0.125, 0.173)	575	0.266 (0.236, 0.296)
Twitter	444	0.486 (0.447, 0.526)	179	0.436 (0.375, 0.497)	452	0.367 (0.33, 0.405)
YouTube	600	0.128 (0.106, 0.151)	598	0.042 (0.028, 0.055)	579	0.076 (0.058, 0.094)



Finding 2: Mis/disinformation content received more engagement than non-mis/disinformation content on Twitter. The opposite is true on TikTok.

Relative post engagement measures the ratio of the average level of active engagement with mis/disinformation posts compared to that with non-mis/disinformation posts. When this ratio exceeds one, it implies that the average level of active engagement with mis/disinformation posts is higher than that with non-mis/disinformation posts. Conversely, a ratio below one indicates the opposite. Twitter showed the highest relative post engagement. The average engagement with mis/disinformation content found on Twitter is 1.977 times as high as the average engagement with non-mis/disinformation. This effect is statistically significant (p<0.10, Table 6). However, further analysis at the country level later revealed that this effect is mainly attributable to Twitter's activities in Spain (Figure 4a; Table 6). .

YouTube is the only other platform where mis/disinformation received more engagement than non-mis/disinformation, but the relative engagement ratio is closer to 1.0 (1.114) and is not statistically significant (p>0.10). On the rest of the platforms, mis/disinformation posts, on average, received less engagement than non-misinformation or disinformation posts, with the difference being statistically significant only on TikTok and Instagram.On Instagram, mis/disinformation on average got less than half of the engagement of non-mis/disinformation did. The relative engagement ratio on TikTok is much smaller (0.048), likely influenced by the few popular non-mis/disinformation post outliers on TikTok in Slovakia (Figure 4c; Table 6).

Breakdown by measurement shows a similar pattern (Figure 4b; Table 6). The countries with the largest relative engagement (Twitter) and the smallest relative engagement (Instagram and TikTok) remain the same. The relative positions of the other countries shifted, but their effects are largely neutral or insignificant. At the country level, the platform order in Spain is similar to the overall pattern, whereas Poland and Slovakia show great divergence (Figure 4c; Table 6). The results remain robust when tested using poisson regression models and negative binomial regression models (Appendix A5).

There has been mixed empirical evidence about the relationship between misinformation and user engagement.[56,57,58]

Past research has found that on Facebook the ratio of engagement between junk news and professional news in the English language was 4:1, supporting that junk news received more engagement, but 1:3 in the Italian language, 2:3 in Polish, and 1:1 in Spanish, showing junk news getting equal or less engagement.[59] While centre and left-leaning sources of misinformation experienced a decline in engagement, right-leaning sources did not.[60] This project complements the literature by coming to a similar conclusion, while employing different samples and methodologies (such as a wider range of platforms and keyword-based sampling method). The mixed findings may result from the properties of misinformation messages (such as novelty, polarisation, niche target audience, or lack of credibility) or platform policies and interventions.[61, 62]

[56] Cybersecurity for Democracy. Far-right news sources on Facebook more engaging. 2021. https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90

[57] Dwoskin, E. Misinformation on Facebook got six times more clicks than factual news during the 2020 election, study says. 2021. The Washington Post. https://www.washingtonpost.com/technology/2021/09/03/facebook-misinformation-nyu-study/

[58] Hutchinson, A. New Study Shows that Misinformation Sees Significantly More Engagement than Real News on Facebook. 2019. *SocialMediaToday*. https://www.socialmediatoday.com/news/new-study-shows-that-misinformation-sees-significantly-more-engagement-than/555286/

[59] Hutchinson, A. New Study Shows that Misinformation Sees Significantly More Engagement than Real News on Facebook. 2019. *SocialMediaToday*. https://www.socialmediatoday.com/news/new-study-shows-that-misinformation-sees-significantly-more-engagement-than/555286/

[60] Cybersecurity for Democracy. Far-right news sources on Facebook more engaging. 2021. https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90

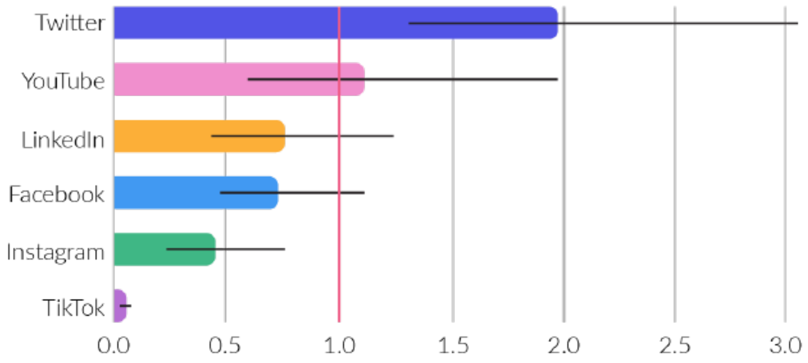
[61] Vosoughi, S., Roy, D., & Aral, S. The spread of true and false news online. 2018. *Science*. 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[62] Ceylan, G., Anderson, I. A., & Wood, W. Sharing of misinformation is habitual, not just lazy or biased. 2023. *Proceedings of the National Academy of Sciences*, *120*(4). https://doi.org/10.1073/pnas.2216614120

Findings

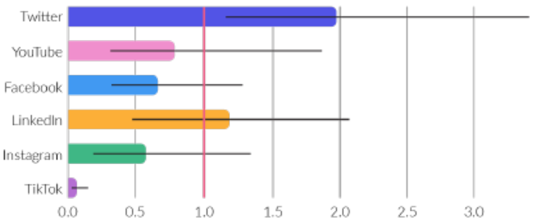
Overall Relative Post Engagement

Combined Countries



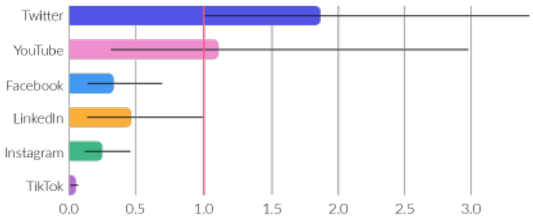
Relative Post Engagement Comparison

Measurement 1



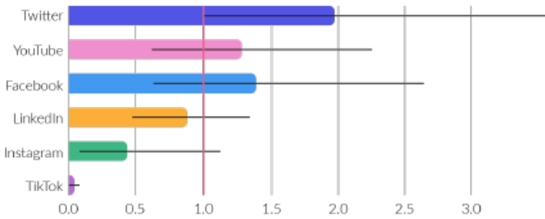
Relative Post Engagement Comparison

Measurement 2



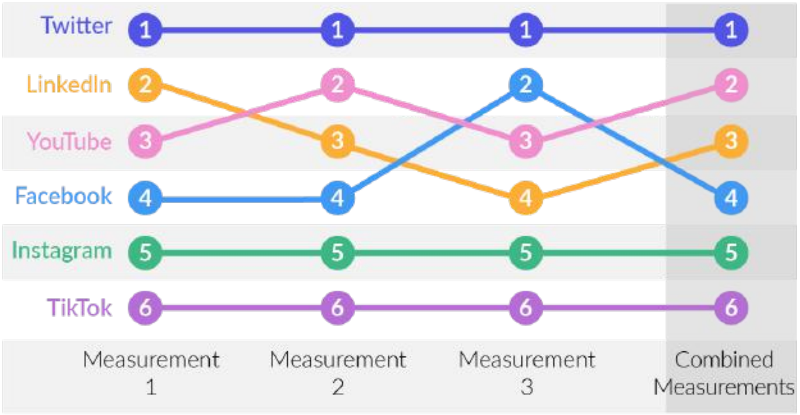
Relative Post Engagement Comparison

Measurement 3



Relative Post Engagement Comparison

Platform Orders by Metric



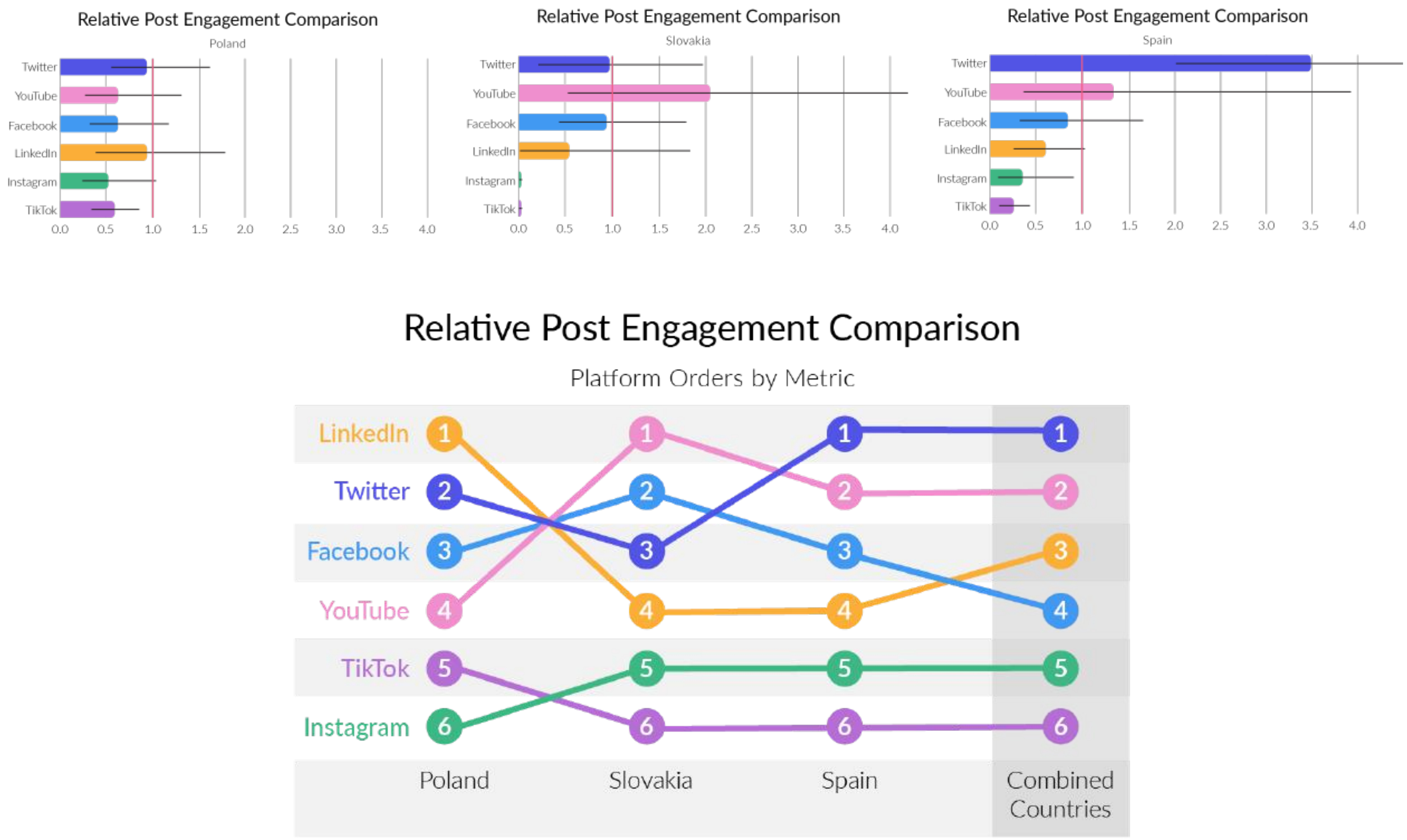


Figure 4: Relative Post Engagement by Platform, Measurement, and Country.

The red line at 1 indicates null effect, which means that mis/disinformation and non-mis/disinformation on average received the same amount of engagement. In the bar charts, the bars represent the point estimates of relative post engagement. The black error bars indicate the 90 percent confidence intervals. The line chart demonstrates the changes in the platform positions across countries or measurements.

Table 6: Relative Post Engagement Ratio by Platform, Country, and Measurement

Overall Sample

Platform	Relative Post Engagement *
Facebook	0.734 (0.471, 1.124)
Instagram	0.454 (0.233, 0.774)
LinkedIn	0.766 (0.43, 1.236)
TikTok	0.048 (0.03, 0.078)
Twitter	1.977 (1.319, 3.057)
YouTube	1.114 (0.596, 1.979)

* Estimate and 90% confidence interval

By Measurement	Measurement 1	Measurement 2	Measurement 3
Platform	Relative Post Engagement	Relative Post Engagement	Relative Post Engagement
Facebook	0.679 (0.324, 1.306)	0.34 (0.157, 0.693)	1.418 (0.612, 2.674)
Instagram	0.584 (0.174, 1.364)	0.239 (0.119, 0.452)	0.439 (0.063, 1.16)
LinkedIn	1.167 (0.466, 2.085)	0.438 (0.141, 0.989)	0.888 (0.486, 1.364)
TikTok	0.063 (0.029, 0.14)	0.039 (0.021, 0.082)	0.022 (0.01, 0.045)
Twitter	1.994 (1.128, 3.432)	1.886 (1.001, 3.456)	2.002 (1.026, 3.836)
YouTube	0.763 (0.306, 1.862)	1.079 (0.317, 2.988)	1.321 (0.617, 2.273)

By Measurement	Poland	Slovakia	Spain
Platform	Relative Post Engagement	Relative Post Engagement	Relative Post Engagement
Facebook	0.616 (0.316, 1.152)	0.914 (0.41, 1.754)	0.833 (0.301, 1.689)
Instagram	0.516 (0.222, 1.026)	NA[63]	0.343 (0.084, 0.914)
LinkedIn	0.931 (0.374, 1.777)	NA	0.575 (0.253, 1.044)
TikTok	0.559 (0.328, 0.85)	0.002 (0.001, 0.005)	0.216 (0.106, 0.437)
Twitter	0.912 (0.524, 1.603)	0.824 (0.204, 1.969)	3.536 (2.053, 5.993)
YouTube	0.615 (0.268, 1.301)	2.007 (0.5, 4.512)	1.349 (0.336, 3.91)

[63] Relative post engagement ratio is unavailable on Instagram and LinkedIn in Slovakia due to small sample sizes.

Finding 3: A large relative engagement ratio on some platforms like Twitter may be a result of moderate absolute engagement (and vice versa on TikTok).

Relative engagement does not reflect the volume of engagement in absolute number, which we examine using the average absolute engagement per post. The largest relative engagement ratio of 1.977 was identified on Twitter. It was derived from dividing the average of “bad” engagement of 361.7 (with mis/disinformation) over the average of “good” engagement of 183.0 (with non-mis/disinformation) (Figure 5; Table 7). By contrast, the very small relative engagement on TikTok is derived from a larger average of “bad” engagement of 7429.7 and an average of “good” engagement of 155220.7 (Table 7). *t*-tests showed that the average engagement is significantly different between mis/disinformation and non-mis/disinformation content only on Instagram, TikTok, and Twitter.[64] Absolute engagement should be interpreted with the consideration that it is affected by the nature of the underlying engagement data and the sampling methodology. High relative engagement or high absolute engagement can both be indicators of audience exposure and potential for harm and can be examined together.

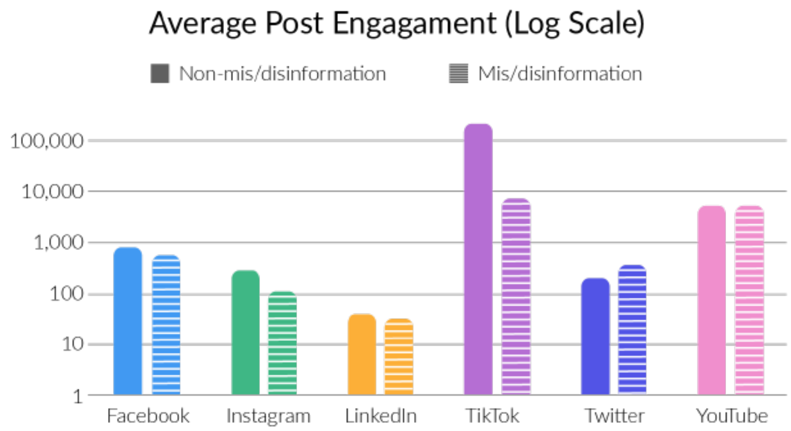


Figure 5: Average (Log) Post Engagement by Post Type.

The lined bars represent average engagement with mis/disinformation posts, and the solid bars present average engagement with non-mis/disinformation posts.

[64] This is consistent with the findings from the relative engagement section (Table 7). The group means being significantly different between mis/disinformation and non-mis/disinformation posts implies that the ratio of the group means (i.e. relative engagement) is statistically significant.

Table 7: Average Post Engagement by Post Type

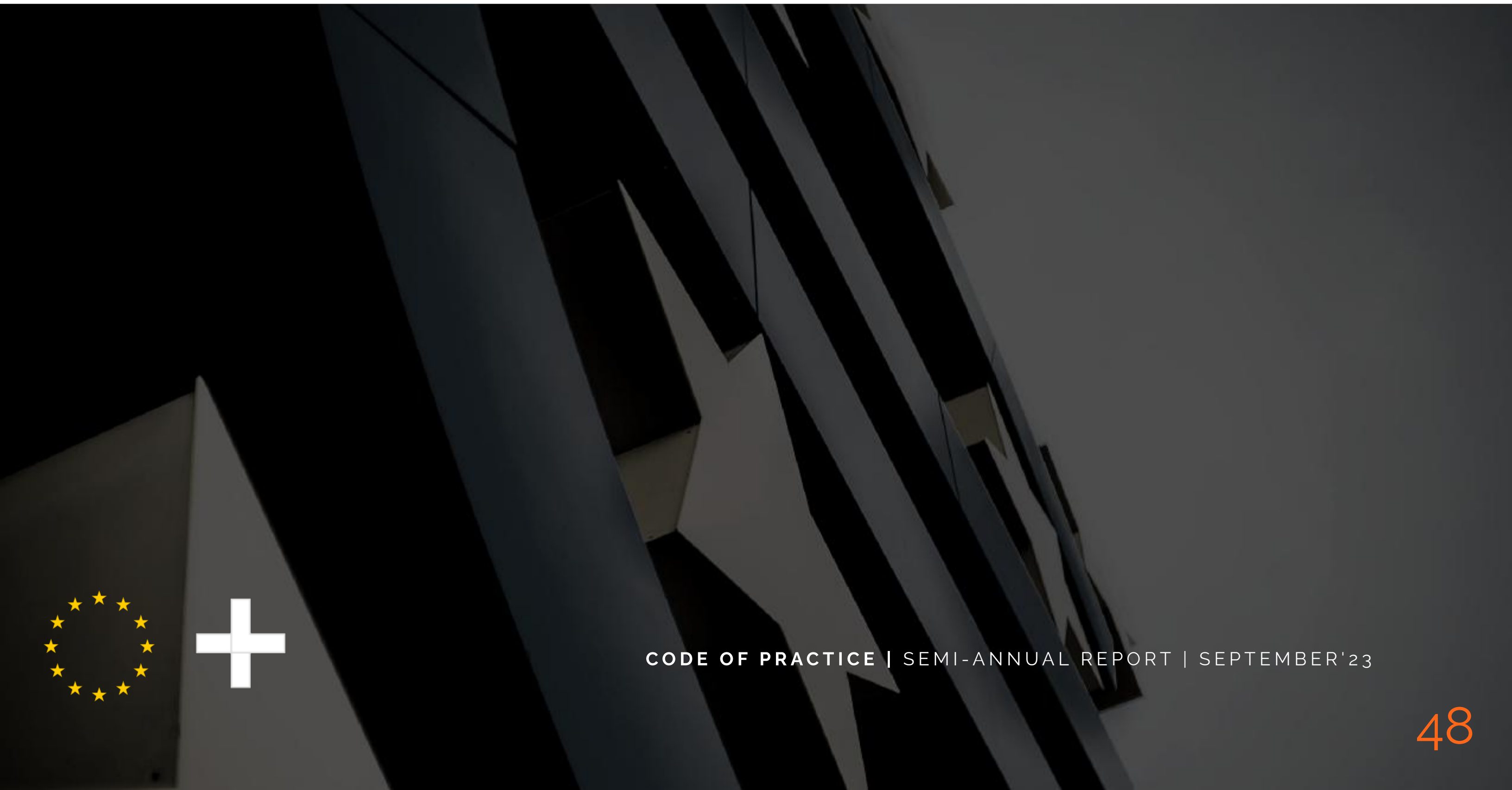
Platform	Average Mis/Dis	Average Non-Mis/Dis	Ratio
Facebook	558.2	760.6	0.734
Instagram	113.7	250.5 *	0.454
LinkedIn	30.3	39.5	0.766
TikTok	7429.7	155220.7 **	0.048
Twitter	361.7	183.0 ***	1.977
YouTube	5719.3	5135.8	1.114

	Poland			Slovakia			Spain		
Platform	Average Mis/Dis	Average Non-Mis/Dis	Ratio	Average Mis/Dis	Average Non-Mis/Dis	Ratio	Average Mis/Dis	Average Non-Mis/Dis	Ratio
Facebook	465.6	755.5	0.616	559.4	612	0.914	688.9	827.5	0.833
Instagram	134.4	260.2	0.516	1.8	235.2	0.007	84.6	246.8	0.343
LinkedIn	46.3	49.7	0.931	12.8	23.4	0.545	20.5	35.6	0.575
TikTok	13969.0	24977 *	0.559	889.3	393960.3 *	0.002	6072.1	28170.8	0.216
Twitter	214.2	234.9 ***	0.912	40.6	49.3	0.824	666.3	188.4	3.536
YouTube	4583.0	7453.3	0.615	5377.8	2680.0	2.007	7612.6	5643.1	1.349

* indicates the *p-value of t*-test is < 0.10. ** indicates *p*-value < 0.05. *** indicates *p*-value < 0.01.

Findings

There is an unequal distribution of engagement across posts, especially among mis/disinformation posts. Figure 6a shows that the top five most popular mis/disinformation posts accounted for between one third and two thirds of total engagement with all the collected mis/disinformation posts. The share of the top 20 posts reached as high as 97.3 percent on LinkedIn (Figure 6c; Table 8). The same number of top non-mis/disinformation posts generally occupied a smaller share of total non-mis/disinformation engagement (Table 8). The concentration of engagement among a small number of posts or users may present opportunities to minimise the adverse impact of disinformation.



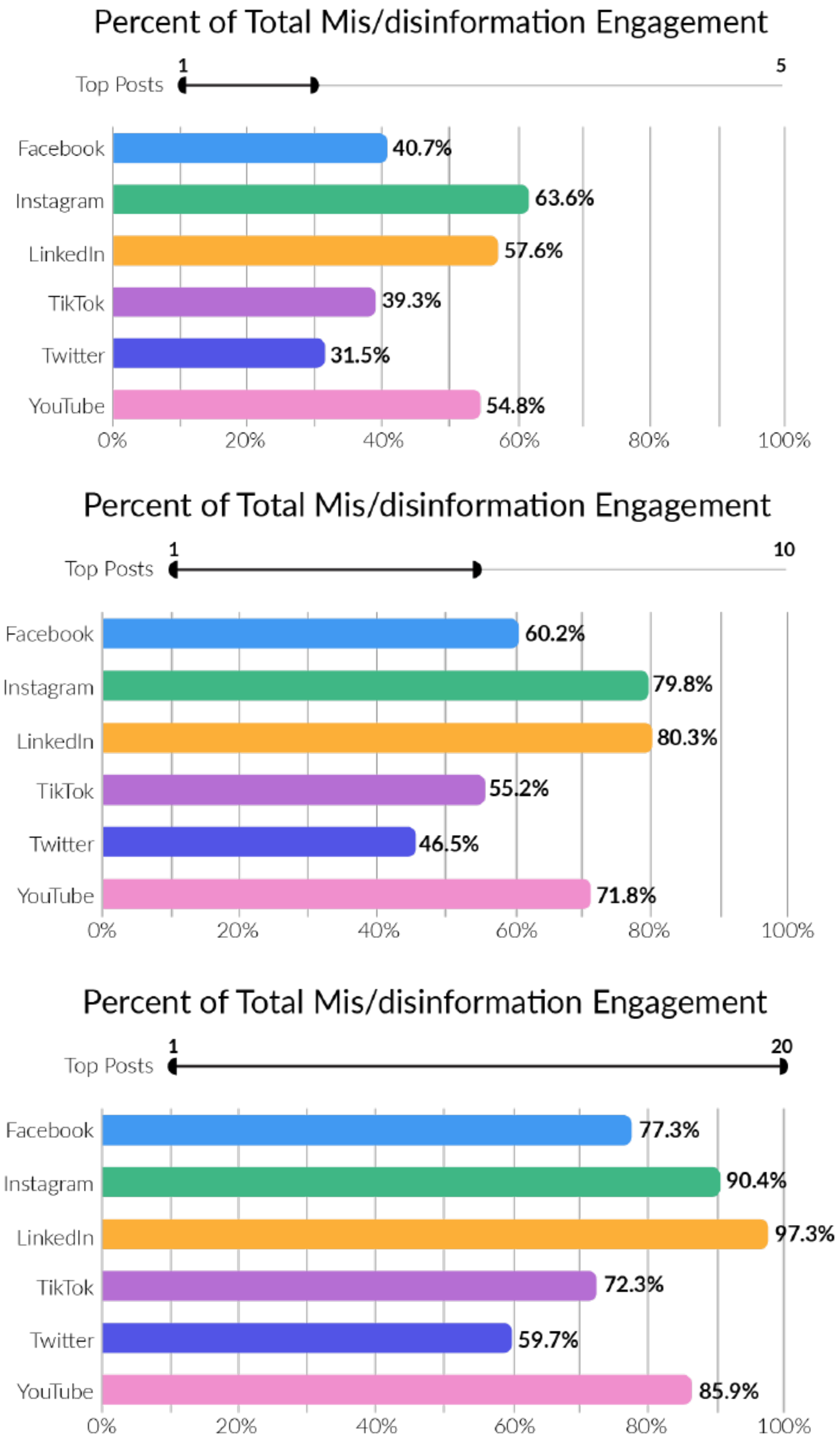


Figure 6: Share of Engagement with Top N Mis/Disinformation Posts among Total Mis/Disinformation Engagement (N = 5, 10, 20).

	Mis/disinformation			Non Mis/disinformation		
Platform	Top 5 Posts	Top 10 Posts	Top 20 Posts	Top 5 Posts	Top 10 Posts	Top 20 Posts
Facebook	0.399	0.589	0.757	0.264	0.396	0.555
Instagram	0.634	0.795	0.900	0.288	0.434	0.608
LinkedIn	0.460	0.695	0.918	0.199	0.274	0.383
TikTok	0.357	0.501	0.663	0.355	0.586	0.729
Twitter	0.297	0.439	0.568	0.313	0.409	0.510
YouTube	0.499	0.673	0.826	0.354	0.432	0.524

Summary of Prevalence Metrics

Synthesising the findings from prevalence metrics reveals three distinct platform patterns, summarised along two dimensions: discoverability and engagement (average engagement per mis/disinformation posts) (Figure 7). Twitter and Facebook are high-discoverability, medium-engagement platforms indicated by the pink background. The rate of mis/disinformation and relative engagement with such content is high, although the absolute engagement is moderate. YouTube and TikTok are low-discoverability, high-engagement platforms indicated by the blue background. Note that both platforms are video-sharing platforms, and that their incentive mechanisms of user engagement may differ from the other types of platforms. Instagram and LinkedIn are low-discoverability, low-engagement platforms indicated by the green background. Their findings may be related to the smaller sample size and smaller user base of these platforms.

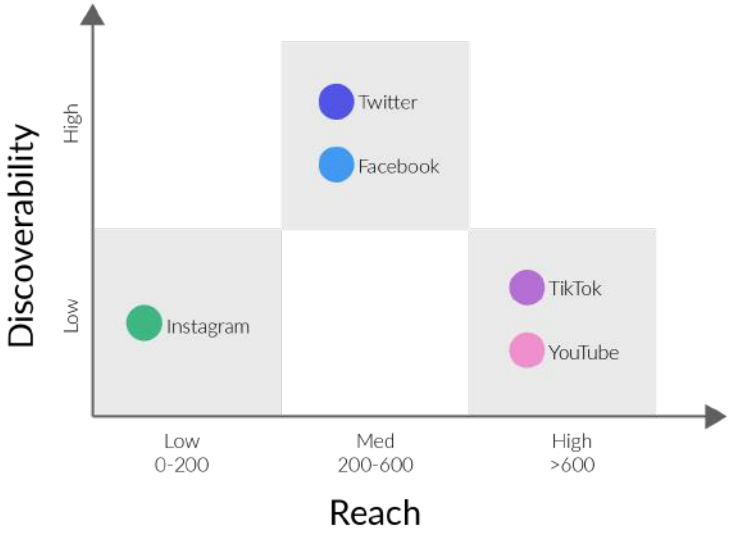


Figure 7: Prevalence-Based Platform Typology.

The X-axis represents the level of absolute engagement with misinformation or disinformation, categorised into three groups: low, medium, and high. The Y-axis represents discoverability, which is divided into two categories: low (below the mean) and high (above the mean). [65]

Sources of Disinformation

Finding 4: Twitter has the largest ratio of disinformation actors, while YouTube has the smallest.

Accounts After accounts were collected from disinformation search queries, they were reviewed forto identify disinformation actors, and per the percentageDisinformation Actor Identification section above. The number of disinformation actorsaccounts was calculatedthen divided by the total number of accounts sampled on the platform (through keyword searches per the Methodology section earlier) to obtain the ratio of disinformation accounts. Twitter and Facebook have the largest ratios of disinformation actors (between 8-9 percent; Figure 8; Table 9). The difference between the two platforms is not statistically significant (Appendix A6). YouTube had the lowest ratio of disinformation actors at 0.8 percent. Given that our disinformation actor definition includes a subjective secondary review, sensitivityRobustness checks using two more objective alternative definitions were conducted, and they yieldedof disinformation actors found the same platform order (Appendix A7).

[65] Instagram has a discoverability value of 0.24, slightly above the sample average of 0.22.

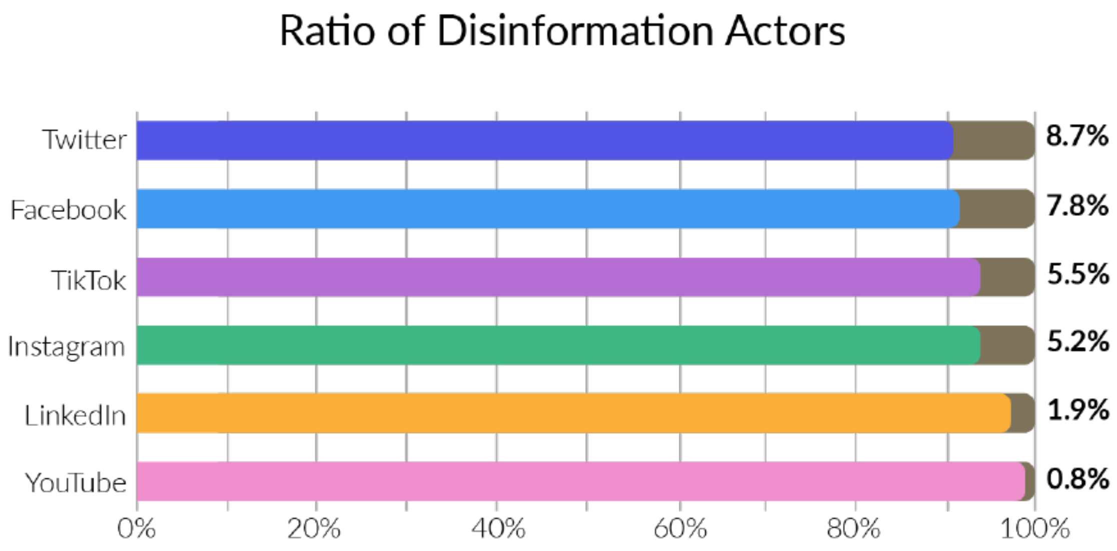


Figure 8: Ratio of Disinformation Actors by Platform. The figures displayed are the percentage of disinformation actors.

Table 9: Ratios of Disinformation Actors and Non-Disinformation Actors

Platform	Ratio of non-Disinfo. Actors	Ratio of Disinfo. Actors
Facebook	0.922	0.078
Instagram	0.948	0.052
LinkedIn	0.981	0.019
TikTok	0.945	0.055
Twitter	0.913	0.087
YouTube	0.992	0.008

Finding 5: Disinformation actors tend to follow more users, but have fewer followers compared to non-disinformation actors. Disinformation actors are also more likely to have joined the platforms more recently.

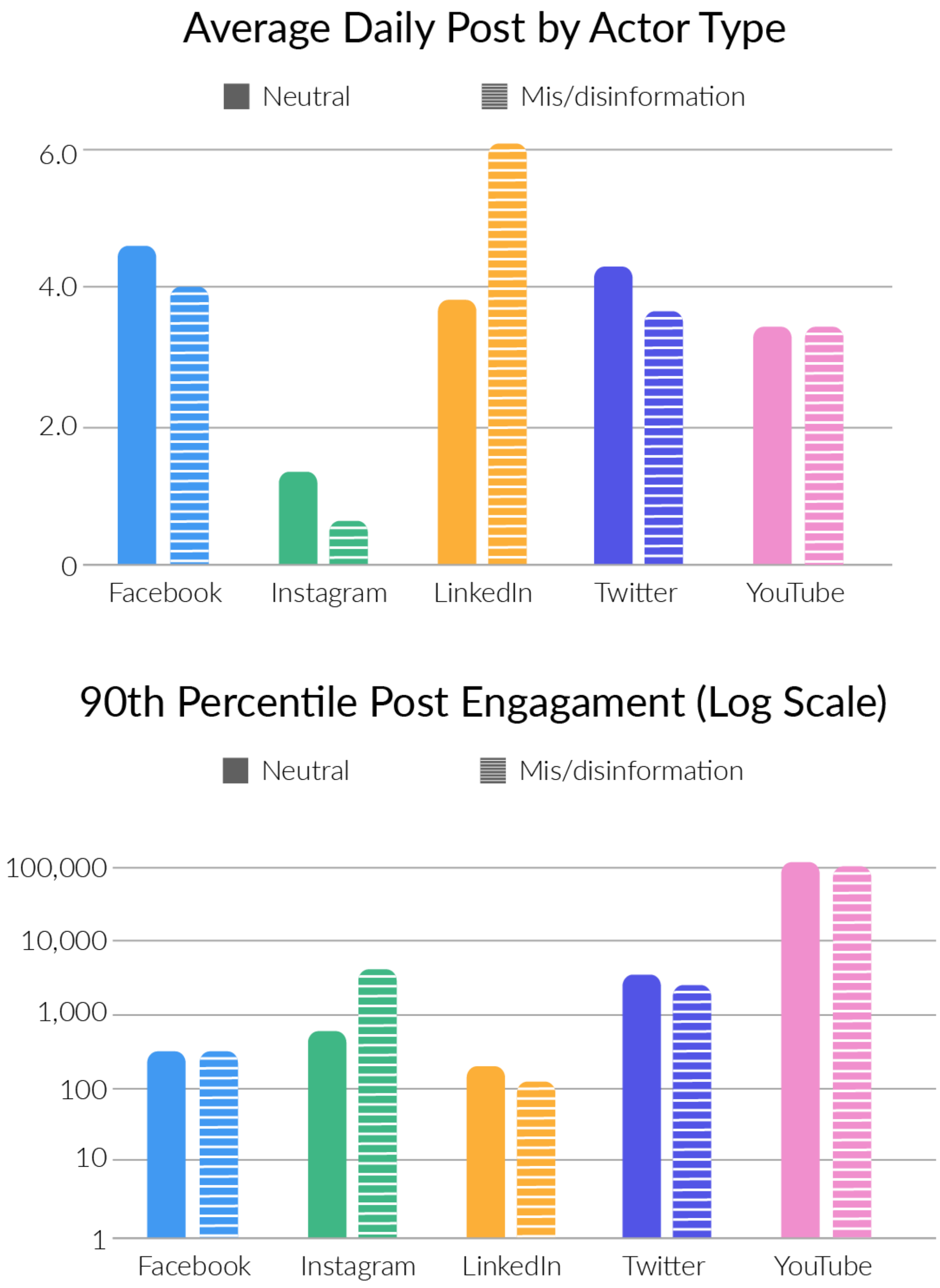
Disinformation actors are significantly more likely to follow more users (i.e. have a larger number of followings) than non-disinformation actors do on Instagram, TikTok, and Twitter.[66] For instance, disinformation actors on TikTok on average follow approximately 1,017 users, whereas non-disinformation actors on TikTok on average follow 582.4 users, resulting in a ratio of 1.7 disinformation actors for every non-disinformation actor on TikTok (Table 10). Disinformation actors, however, are generally followed by fewer users, although the results are largely not statistically significant. On all platforms but Instagram (a small sample with possible outliers), the ratio of average followers between disinformation actors and non-disinformation actors ranges between 0.1 and 0.3 (Table 10).[67]

Table 10: Average Number of Followers and Followings By Actor Type

	Number of Followings			Number of Followers		
Platform	Average of Disinfo Actors	Average of Non-Disinfo Actors	Ratio	Average of Disinfo Actors	Average of Non-Disinfo Actors	Ratio
Facebook		N/A		10824.8	71165.5	0.2
Instagram	1514.1	651.0 **	2.3	32899.5	21660.4	1.5
LinkedIn		N/A		1284.1	4104.5	0.3
TikTok	1017.0	582.4 **	1.7	15625.9	210795.8	0.1
Twitter	3259.5	1612.8 ***	2.0	14679.3	87962.7	0.2
YouTube		N/A		172216.7	621951.1	0.3

[66] Number of followings is not supported on Facebook, LinkedIn, and YouTube.

[67] This is consistent with collective first-hand experience from TrustLab's policy team that disinformation actors tend to spread misleading content by following other users, but they are less likely to gain followers themselves.



Average Daily Post by Actor Type

90th Percentile Post Engagament (Log Scale)

Figure 9: Group Averages in Time on Platform, Number of Daily Posts, and 90th Percentile Post Engagement by Actor Type

Table 11: Average Number of Daily Posts by Actor Type

	Average Daily Posts		
Platform	Average of Disinfo Actors	Average of Non-Disinfo Actors	Ratio
Facebook	4.3	5.2	0.8
Instagram	0.9	1.6	0.5
LinkedIn	6.6	4.0	1.6
TikTok	N/A *	N/A	N/A
Twitter	3.8	4.6	0.8
YouTube	3.9	3.6	1.1

Table 12: Average Actor Engagement by Actor Type

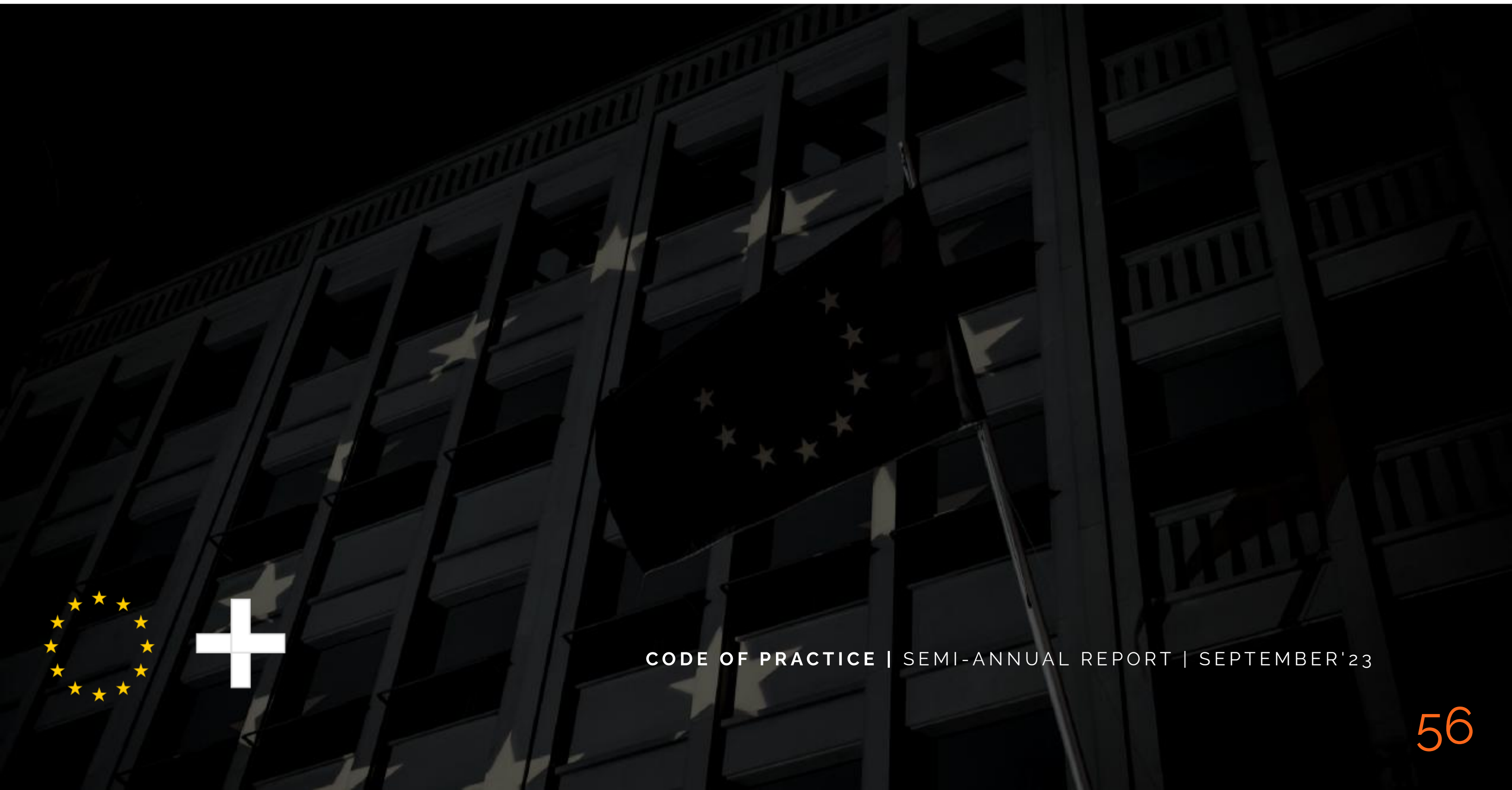
	90th Percentile of Post Engagement Per Account			Total Weekly Engagement Per Account		
Platform	Average of Disinfo Actors	Average of Non-Disinfo Actors	Ratio	Average of Disinfo Actors	Average of Non-Disinfo Actors	Ratio
Facebook	338.9	301.4	1.1	1846.4	3370.1	0.5
Instagram	3969.5	582.9 ***	6.8	9251.2	4104.7	2.3
LinkedIn	109.9	187.0	0.6	1353.8	1598.8	0.8
TikTok	N/A *	N/A	N/A	N/A	N/A	N/A
Twitter	2786.5	3181.1	0.9	36746.3	27323.1	1.3
YouTube	102577.8	116415.1	0.9	118797.3	348487.0	0.3

* average daily posts unavailable for TikTok due to data collection issues

*** indicates the *p-value of t*-test is < 0.01.

Summary of Sources Metrics

Platforms with higher discoverability also have higher rates of disinformation actors. Twitter scored the highest in both metrics, and YouTube the lowest. With the caveat that our sample of disinformation actors is small, and that the conclusions we could draw about disinformation actors were limited, disinformation actors were found to tend to follow more users, but have fewer followers for themselves compared to non-disinformation actors.



Limitations and Implications

As the first empirical study under the Code of Practice, this project measured the prevalence and sources of online disinformation and provided a benchmark for policy evaluation and for the monitoring of disinformation over time.[67] This study was carried out by TrustLab, a third-party provider of online trust and safety services. Independent research conducted by third parties, such as TrustLab, offers valuable external insights that complement the platforms' self-reported measurements. Furthermore, applying a consistent methodology and metrics across all platforms leads to fresh comparative insights which cannot be found in platform transparency reports. We also note that consistency of methodology and metrics will further improve as the discussions among stakeholders about detailed definitions of misinformation, disinformation, engagement, and other factors continue.

The insights from the study should be interpreted with the caveat that budget and time constraints have led to small sample sizes and imprecise estimates for some metrics. The lack of direct access to platform data in the current study limits its capacity to measure broader dimensions of online disinformation. Additionally, the manual labelling of mis/disinformation content and disinformation actors may be subject to human error. Despite two tiers of analysts, sufficient pre-training, and ongoing feedback during data collection, a small number of labelling errors are possible. Future projects may address this issue by further minimising the possibility of human error, and including the possibility of human errors as an uncertainty element in the metrics.

The platforms in this study ranged from text-based to video-based, with distinctive user dynamics and platform structures. To draw cross-platform conclusions, we standardised metrics and employed relative measures. The relative measures include the disinformation-to-non-disinformation content ratio and the comparison of characteristics between disinformation actors and non-disinformation actors.

[67] Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D. and Schudson, M., 2018. The science of fake news. *Science*, *359*(6380), pp.1094-1096.

Nonetheless, platform heterogeneity can still present challenges. An example is the construction of the post engagement variable. Not all types of engagement are accessible across all platforms, and even the same type of engagement may carry different meanings and implications depending on the specific content being engaged with and the functionality and established norms of the platform.

More resources (time, budget and data access) can help make several enhancements to this study. For example, this study uses platform search engines to surface mis/disinformation content and accounts. Many social media platforms nowadays allow users to discover content through a feed. TrustLab has previously developed a methodology for measuring harmful content recommended on the feed to assess the impact of algorithmic amplification. It has applied this methodology in an EU project measuring Terrorism and Violent Extremism content.[69] Although the feed-based approach was proposed for the current study, it was deemed infeasible due to budget constraints; we strongly recommend including this approach in future measurements.

With respect to disinformation actor classification, greater access to platform data and increased time and budget can help improve the methodology by more comprehensively accounting for the actor network, evidence of coordinated activity and foreign influence. We can also measure a platform's active efforts in combating disinformation from how the platform addresses user-flagged content, such as the rate in which the platform responds and the timeliness in the platform actions (see an example in a past TrustLab project).[70] Note that platform actions for mis/disinformation will not necessarily be content removal but could include demonetization, applying warning labels or adding pointers to fact-check articles.

[69] TrustLab, 'Study to Inform the EU Internet Forum with Measurements on the Impact of the Misuse of Algorithmic Amplification of Terrorist and Violent Extremist Content', 2023, Request 000005 – DG HOME

[70] TrustLab (2023).

Limitations and Implications

The current study provided three biweekly measurements of disinformation, which does not inform meaningful time trends. Longitudinal analysis that tracks the same metrics over a long period of time can yield more robust platform measures and better capture time trends. While this study takes a macro approach to understand online disinformation, future research may also benefit from studying user interaction patterns at the micro level or with a qualitative or mixed-methods approach.[71, 72]

This study measured the first two structural indicators (prevalence and sources of disinformation) of the Code. With more structural indicators implemented, more insights about the trends and characteristics about disinformation will be uncovered to assist policy makers, platforms, and trust and safety teams with the identification of high-risk areas and communities that are prone to mis/disinformation and strengthening content moderation practices. Experimental or quasi-experimental research designs, coupled with the measurement methodology in this study, are useful tools for evaluating the effectiveness of policy interventions and establishing causal relationships.

[70] Correia, R. P., Silva, B. M. C., Jerónimo, P., & Garcia, N. A Micro-interaction Tool for Online Text Analysis. In T. Guarda, F. Portela, & M. F. Augusto (Eds.). 2022. Advanced Research in Technologies, Information, Innovation and Sustainability (pp. 511–523). Springer Nature Switzerland.

[71] Meredith, J. Conversation Analysis and Online Interaction. 2019. Research on Language and Social Interaction, 52(3), 241–256.

Conclusion

As the first empirical study under the Code of Practice, this study implemented a robust cross-platform measurement of the prevalence and sources of disinformation across Poland, Spain, and Slovakia. Based on this study, searching mis/disinformation-related queries on social media platforms leads to a sizable share of mis/disinformation and disinformation actors, which suggests that additional steps are needed to reduce the visibility of mis/disinformation through the search functionality. Furthermore, engagement data on the found mis/disinformation content indicates that at least some of this content has high levels of engagement. The findings from this study provide a benchmark for the implementation of the Code of Practice and the evaluation of disinformation on social media platforms. Future studies can replicate our methodology to monitor the long-term trends of disinformation. At the same time, alternative methods, such as feed-based data collection and tracking platform responses to user-flagged harmful content, can be applied in future work and complement our existing findings. In addition, the scope of the study can be expanded to cover many more countries and platforms in future measurements.

Conflicts and Acknowledgements

Conflicts of Interest

Although all measured platforms excluding Twitter funded the study, TrustLab independently developed measurement methodology and conducted post-measurement analysis, without platform input.

Acknowledgments

We would like to thank the Permanent Task-force of the Code of Practice, in particular the European Commission, Avaaz, European Regulators Group for Audiovisual Media Services (ERGA), European Digital Media Observatory (EDMO), and the platform signatories.

Appendix

A1: Disinformation Keyword Search on Instagram

The same disinformation keywords were searched across platforms, but they yielded very few search results on Instagram compared to the other platforms. Based on our past data collection experience, single words and hashtags tend to perform better in keyword searches on Instagram than keyphrases made up of several words and space in-between. Thus two changes were applied to keyword search on Instagram. First, alternative shorter keywords or hashtags were generated for the same disinformation claims. Analysts explored alternative keywords by changing verb tenses, paraphrasing, or shortening the words, while meeting the same precision requirement that at least one out of the top three search results returned from a keyword had the same disinformation content as the keyword. The second change was that keyword searches were conducted on Instagram mobile app, whereas the other platforms were accessed through the web browser. The change was due to the observation that searching the same keyword on Instagram mobile app tended to yield many more posts than searching in the web version of Instagram. The altered way of keyword search on Instagram resulted in more posts and more robust estimates, but it had also potentially introduced bias and inconsistent comparison between platforms.

A2: Mis/Disinformation Content Labelling Quality Metrics

Table **Mis/Disinformation Post Labelling Precision and True Negative Ratio**

	Poland	Slovakia	Spain
Precision	0.72	0.75	0.79
True Negative Ratio	0.81	0.87	0.97

Each country team achieved a precision score above 0.70, which means that out of ten mis/disinformation posts labelled by the Tier-1 analysts, on average more than seven of the ten posts are correctly labelled (true positives).

The ratio of true negative labels among all negative labels is consistently above 0.80. This means that on average at least four out of five posts that Tier-1 analysts labelled as non-mis/disinformation were indeed non-mis/disinformation.

A3: Construction of Disinformation Actor Engagement Variables

Account-level engagement is an aggregate measure of engagement of posts published by the account. We collected up to 50 most recent posts on each user's profile page. The size 50 was chosen to maximise the number of posts per user and user activity history within feasible resources.

To demonstrate the construction of the variables, suppose we collected n posts made by user u. We refer to the posts as $\{p_1, p_2, ..., p_n\}$,1<=n<=50. We represent the engagement of post p_i as E_{p_i} so the engagement of the n posts becomes $\left\{E_{p_1}, E_{p_2}, ..., E_{p_n}\right\}$.

For each user, we calculate three aggregate measures:

1. The maximal post engagement (maximum of engagement of the n individual posts by the user; the most robust measure of 90th percentile of engagement of n posts is adopted in the main analysis):

Maximal post engagement for user u = $max(\{E_{p_i}\}),\ 1 <= i <= n$

Robust "maximal" post engagement for user u = $90th\ percentile(\{E_{p_i}\}),\ 1 <= i <= n$

2. Average post engagement (average of the n individual posts' engagement)

$$\text{Average post engagement} = \frac{1}{n}\Sigma\,(E_{p_i}),\ 1 <= i <= n$$

3. Total post engagement from the most recent week (sum of engagement of posts published within the week prior to the time of data collection)

$$\text{Total post engagement from last week} = \Sigma\,(E_{p_j}),$$

where $\{p_j\}$ are posts published by user u in the past week when data were collected.

Next, we compare the group averages between disinformation actors and non-disinformation actors in these three account-level engagement variables, both in terms of absolute numbers and ratios.

A4: Pairwise *t*-tests of Discoverability

Table A4.1: *p*-values of Pairwise Platform Differences in Overall Discoverability

Platform	Facebook	Twitter	Instagram	TikTok	YouTube	LinkedIn
Facebook	NA	<0.001	0.176	<0.001	<0.001	<0.001
Twitter	<0.001	NA	<0.001	<0.001	<0.001	<0.001
Instagram	0.176	<0.001	NA	1	<0.001	<0.001
TikTok	<0.001	<0.001	1	NA	<0.001	<0.001
YouTube	<0.001	<0.001	<0.001	<0.001	NA	1
LinkedIn	<0.001	<0.001	<0.001	<0.001	1	NA

p-values were corrected for multiple tests using the Bonferroni method (alpha = 0.1, corrected alpha = 0.0067)

Table A4.2: *p*-values of Pairwise Platform Differences in Discoverability in Measurement 1

Platform	Facebook	Twitter	Instagram	TikTok	YouTube	LinkedIn
Facebook	NA	0.135	0.192	<0.001	<0.001	<0.001
Twitter	0.135	NA	0.002	<0.001	<0.001	<0.001
Instagram	0.192	0.002	NA	1	<0.001	0.037
TikTok	<0.001	<0.001	1	NA	<0.001	0.036
YouTube	<0.001	<0.001	<0.001	<0.001	NA	1
LinkedIn	<0.001	<0.001	0.037	0.036	1	NA

p-values were corrected for multiple tests using the Bonferroni method (alpha = 0.1, corrected alpha = 0.0067)

Table A4.3: *p*-values of Pairwise Platform Differences in Discoverability in Measurement 2

Platform	Facebook	Twitter	Instagram	TikTok	YouTube	LinkedIn
Facebook	NA	0.051	1	<0.001	<0.001	<0.001
Twitter	0.051	NA	0.107	<0.001	<0.001	<0.001
Instagram	1.000	0.107	NA	1	<0.001	<0.001
TikTok	<0.001	<0.001	1	NA	<0.001	<0.001
YouTube	<0.001	<0.001	<0.001	<0.001	NA	1
LinkedIn	<0.001	<0.001	<0.001	<0.001	1	NA

p-values were corrected for multiple tests using the Bonferroni method (alpha = 0.1, corrected alpha = 0.0067)

Table A4.4: *p*-values of Pairwise Platform Differences in Discoverability in Measurement 3

Platform	Facebook	Twitter	Instagram	TikTok	YouTube	LinkedIn
Facebook	NA	0.135	0.192	0.289	<0.001	<0.001
Twitter	<0.001	NA	0.003	<0.001	<0.001	<0.001
Instagram	1.000	0.003	NA	1	<0.001	0.224
TikTok	0.289	<0.001	1	NA	<0.001	0.028
YouTube	<0.001	<0.001	<0.001	<0.001	NA	1
LinkedIn	<0.001	<0.001	0.224	0.028	1	NA

p-values were corrected for multiple tests using the Bonferroni method
(alpha = 0.1, corrected alpha = 0.0067)

Table A4.5: *p*-values of Pairwise Platform Differences in Discoverability in Poland

Platform	Facebook	Twitter	Instagram	TikTok	YouTube	LinkedIn
Facebook	NA	<0.001	1	<0.001	<0.001	<0.001
Twitter	<0.001	NA	0.002	<0.001	<0.001	<0.001
Instagram	1.000	0.002	NA	0.01	<0.001	<0.001
TikTok	<0.001	<0.001	0.01	NA	0.157	0.163
YouTube	<0.001	<0.001	<0.001	0.157	NA	1
LinkedIn	<0.001	<0.001	<0.001	0.163	1	NA

p-values were corrected for multiple tests using the Bonferroni method
(alpha = 0.1, corrected alpha = 0.0067)

Table A4.6: *p*-values of Pairwise Platform Differences in Discoverability in Slovakia

Platform	Facebook	Twitter	Instagram	TikTok	YouTube	LinkedIn
Facebook	NA	0.135	<0.001	<0.001	<0.001	<0.001
Twitter	1	NA	<0.001	<0.001	<0.001	0.016
Instagram	<0.001	<0.001	NA	1	1	1
TikTok	<0.001	<0.001	1	NA	<0.001	1
YouTube	<0.001	<0.001	1	<0.001	NA	0.093
LinkedIn	<0.001	0.016	1	1	0.093	NA

p-values were corrected for multiple tests using the Bonferroni method
(alpha = 0.1, corrected alpha = 0.0067)

Table A4.7: *p*-values of Pairwise Platform Differences in Discoverability in Spain

Platform	Facebook	Twitter	Instagram	TikTok	YouTube	LinkedIn
Facebook	NA	<0.001	1	0.022	<0.001	<0.001
Twitter	<0.001	NA	0.244	0.007	<0.001	<0.001
Instagram	1.000	0.244	NA	1	<0.001	<0.001
TikTok	0.022	0.007	1	NA	<0.001	<0.001
YouTube	<0.001	<0.001	<0.001	<0.001	NA	1
LinkedIn	<0.001	<0.001	<0.001	<0.001	1	NA

p-values were corrected for multiple tests using the Bonferroni method
(alpha = 0.1, corrected alpha = 0.0067)

A5: Robustness Checks of Relative Post Engagement Estimates

Table A5.1: Poisson Regression Model Coefficients and 90th Percent Confidence Intervals

Platform	Estimate
Facebook	0.772 (0.51, 1.169)
Instagram	0.448 (0.244, 0.821)
LinkedIn	0.771 (0.451, 1.317)
TikTok	0.044 (0.027, 0.073)
Twitter	1.898 (1.235, 2.918)
YouTube	1.129 (0.622, 2.049)

Platform	Measurement 1	Measurement 2	Measurement 3
Facebook	0.691 (0.362, 1.319)	0.378 (0.189, 0.753)	1.488 (0.759, 2.919)
Instagram	0.576 (0.219, 1.515)	0.255 (0.132, 0.492)	0.432 (0.115, 1.624)
LinkedIn	1.157 (0.571, 2.348)	0.453 (0.18, 1.144)	0.899 (0.527, 1.531)
TikTok	0.059 (0.026, 0.132)	0.038 (0.02, 0.074)	0.02 (0.01, 0.04)
Twitter	1.798 (1.038, 3.112)	1.855 (0.977, 3.524)	2.012 (1.03, 3.93)
YouTube	0.793 (0.332, 1.896)	0.973 (0.332, 2.854)	1.332 (0.703, 2.524)

Platform	Poland	Slovakia	Spain
Facebook	0.609 (0.312, 1.189)	0.893 (0.446, 1.786)	0.905 (0.418, 1.957)
Instagram	0.507 (0.238, 1.082)	0.008 (0.001, 0.052)	0.329 (0.103, 1.05)
LinkedIn	0.955 (0.45, 2.028)	0.5 (0.073, 3.402)	0.566 (0.283, 1.132)
TikTok	0.573 (0.36, 0.913)	0.002 (0.001, 0.005)	0.21 (0.102, 0.432)
Twitter	0.85 (0.464, 1.56)	0.961 (0.362, 2.555)	3.405 (1.897, 6.11)
YouTube	0.62 (0.287, 1.339)	2.164 (0.769, 6.092)	1.36 (0.437, 4.236)

Table A5.2: Poisson Regression Model Coefficients and 90th Percent Confidence Intervals

Platform	Estimate
Facebook	0.858 (0.699, 1.058)
Instagram	0.454 (0.298, 0.71)
LinkedIn	0.773 (0.523, 1.192)
TikTok	0.053 (0.041, 0.071)
Twitter	1.977 (1.558, 2.517)
YouTube	1.144 (0.819, 1.655)

Appendix

Platform	Measurement 1	Measurement 2	Measurement 3
Facebook	0.714 (0.503, 1.022)	0.485 (0.347, 0.685)	1.518 (1.068, 2.203)
Instagram	0.652 (0.293, 1.616)	0.276 (0.153, 0.525)	0.707 (0.316, 1.753)
LinkedIn	1.149 (0.693, 2.045)	0.462 (0.219, 1.171)	0.897 (0.523, 1.674)
TikTok	0.066 (0.042, 0.108)	0.037 (0.025, 0.057)	0.021 (0.014, 0.033)
Twitter	1.548 (1.063, 2.257)	1.669 (1.082, 2.585)	2.107 (1.424, 3.142)
YouTube	0.932 (0.55, 1.714)	1.055 (0.661, 1.802)	1.488 (0.879, 2.76)

Platform	Poland	Slovakia	Spain
Facebook	0.692 (0.514, 0.944)	0.887 (0.603, 1.302)	1.224 (0.801, 1.927)
Instagram	0.528 (0.316, 0.91)	0.007 (0.002, 0.036)	0.344 (0.135, 1.042)
LinkedIn	0.976 (0.558, 1.862)	0.144 (0.028, 0.875)	0.559 (0.32, 1.068)
TikTok	0.566 (0.394, 0.838)	0.002 (0.001, 0.004)	0.205 (0.138, 0.313)
Twitter	0.775 (0.554, 1.083)	0.981 (0.489, 2.061)	3.785 (2.623, 5.535)
YouTube	0.629 (0.403, 1.036)	2.206 (1.065, 5.607)	1.372 (0.768, 2.745)

A6: Pairwise t-tests of Ratios of Disinformation Actors

Table A6: p-values of Pairwise Platform Differences in Ratio of Disinformation Actors

Platform	Facebook	Twitter	Instagram	TikTok	YouTube	LinkedIn
Facebook	NA	1	<0.001	0.502	1	<0.001
Twitter	1	NA	0.582	1	0.797	<0.001
Instagram	<0.001	0.582	NA	0.026	<0.001	0.928
TikTok	0.502	1	0.026	NA	0.214	<0.001
YouTube	1	0.797	<0.001	0.214	NA	<0.001
LinkedIn	<0.001	<0.001	0.928	<0.001	<0.001	NA

p-values were corrected for multiple tests using the Bonferroni method
(alpha = 0.1, corrected alpha = 0.0067)

A7: Alternative Definitions of Disinformation Actors

Table A7: Ratio of Disinformation Actors with Alternative Definitions

Platform	Level 1	Level 2	Main Sample
Facebook	0.21	0.17	0.08
Instagram	0.09	0.09	0.05
LinkedIn	0.04	0.04	0.02
TikTok	0.15	0.14	0.06
Twitter	0.23	0.20	0.09
YouTube	0.05	0.05	0.01

Thank you.

For more information please contact
info@trustlab.com