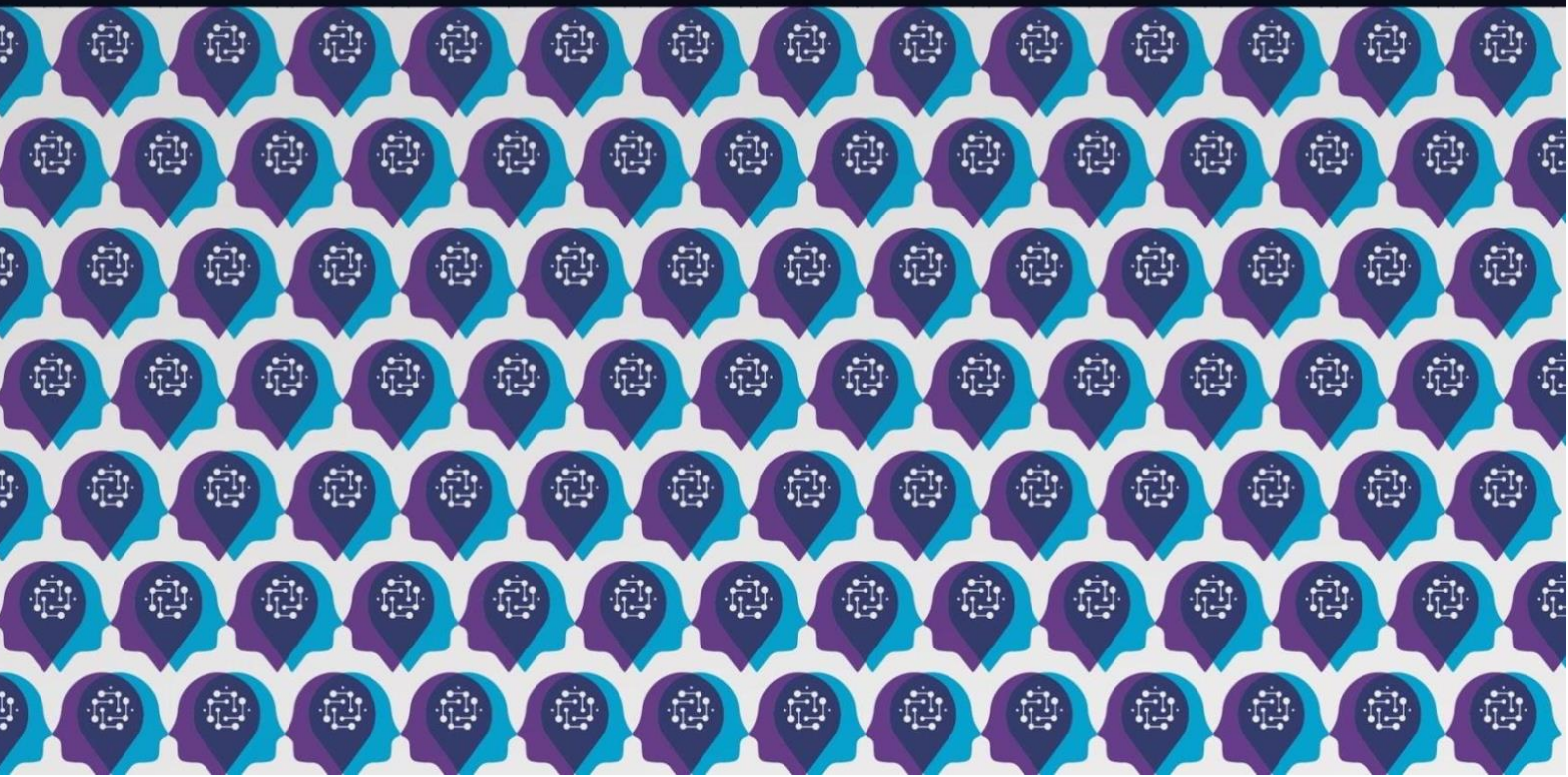
AI4Debunk

D12.1 : Possible impacts of the tool on the perceptions of the citizens and the social media users- January 2025

Funded by
the European Union



Grant Agreement No.: 101135757
Call: HORIZON-CL4-2023-HUMAN-01-CNECT
Topic: HORIZON-CL4-2023-HUMAN-01-05
Type of action: HORIZON Innovation Actions

D.12.1. POSSIBLE IMPACTS OF THE TOOL ON THE PERCEPTIONS OF THE CITIZENS AND THE SOCIAL MEDIA USERS

Project Acronym	AI4Debunk
Project Number	101135757
Project Full Title	Participative Assistive AI-powered Tools for Supporting Trustworthy Online Activity of Citizens and Debunking Disinformation
Work package	WP 12
Task	Task 12.1
Due date	31/03/2025
Submission date	31/03/2025
Deliverable lead	Partner P4D
Version	V1.0
Authors	Pascaline Gaborit (Pilot4dev), Joen Martinsen (Pilot4dev)
Contributors	Vishnu Rao (Pilot4dev)
Reviewers	Zaneta Ozolina (University of Latvia), Álvaro Parafita (BSC)
Abstract	This report presents the first results of the desk research and of the online Survey of WP12. It focuses on the spread of misinformation and disinformation online, but also on the citizens' perceptions, and also on the regulation and moderation of social media to counter the circulation of fake news.
Keywords	Social Media Platforms, Online Survey, Citizens' Perceptions

Funded by the European Union

DOCUMENT DISSEMINATION LEVEL

Dissemination level	
X	PU – Public
	SEN – Sensitive

DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributor(s)
0.1	27/11/2024	Changes from lead partner	UL
0.2	16/12/2024	Reviewed version with comments	BSC
0.3	18/12/2024	Revised version	P4D
0.4	23/12/2024	Project Coordinator Review	UL
1.0	30/01/2025	Final version ready for submission	P4D

STATEMENT ON MAINSTREAMING GENDER

The AI4Debunk consortium is committed to including gender and intersectionality as a transversal aspect in the project's activities. In line with EU guidelines and objectives, all partners – including the authors of this deliverable – recognise the importance of advancing gender analysis and sex-disaggregated data collection in the development of scientific research. Therefore, we commit to paying particular attention to including, monitoring, and periodically evaluating the participation of different genders in all activities developed within the project, including workshops, webinars and events but also surveys, interviews and research, in general. While applying a non-binary approach to data collection and promoting the participation of all genders in the activities, the partners will periodically reflect and inform about the limitations of their approach. Through an iterative learning process, they commit to plan and implement strategies that maximise the inclusion of more and more intersectional perspectives in their activities.

Funded by the European Union

DISCLAIMER

The AI4Debunk project has received funding from the European Union's Horizon Europe Programme under the Grant Agreement No. 101135757.

Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

COPYRIGHT NOTICE

How to cite this report: AI4Debunk (2025) : **Gaborit P., Martinsen J.,'Possible impacts of the tool on the perceptions of the citizens and the social media users'**.

The AI4Debunk consortium is the following:

Participant number	Participant organisation name	Short name	Country
1	LATVIJAS UNIVERSITATE	UL	LV
2	FREE MEDIA BULGARIA	EURACTIV	BE
3	PILOT4DEV	P4D	BE
4	INTERNEWS UKRAINE	IUA	UA
5	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR-IRPPS	IT
6	UNIVERSITA DEGLI STUDI DI FIRENZE	MICC/UNIFI	IT
6.1	CONSORZIO NAZIONALE INTERUNIVERSITARIO PER LE TELECOMUNICAZIONI	CNIT	IT
7	BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION	BSC	ES
8	DOTSOFT OLOKLIROMENES EFARMOGES DIADIKTIOY KAI VASEON DEDOMENON AE	DOTSOFT	EL
9	UNIVERSITE DE MONS	UMONS	BE
10	NATIONAL UNIVERSITY OF IRELAND GALWAY	NUIG	IE
11	STICHTING HOGESCHOOL UTRECHT	HU	NL
12	STICHTING INNOVATIVE POWER	IP	NL
13	F6S NETWORK IRELAND LIMITED	F6S	IE

TABLE OF CONTENTS

Contents

Funded by the European Union

LIST OF FIGURES

Funded by the European Union

ABBREVIATIONS

WP	Work Package
EC	European Commission
DSA	Digital Services Act
MAU	Monthly Active Users
VLOP	Very Large Online Platform

Funded by the European Union

EXECUTIVE SUMMARY

The spread of misinformation and disinformation online is a recent but well-researched topic, though it is rapidly evolving with many aspects still undocumented. Studies demonstrate that social media platforms significantly amplify fake news through their structural design and user engagement patterns. On these platforms, fake news can spread instantly without verification. Such content often travels faster than factual news because of its engaging nature, which increases the likelihood of sharing. Social media algorithms compound this issue by prioritizing engaging content—typically favoring sensational information—which can further amplify fake news through the platforms' own mechanisms.

The research team analyzed the social media platform content moderation, but also set up an online survey for citizens. The report provides a comprehensive analysis of this AI4DEBUNK online survey, a detailed questionnaire comprising 15 carefully crafted questions that garnered meaningful responses from 329 participants across various demographics and user groups (Sections II and III). The analysis then delves into an extensive examination of content moderation practices implemented by major social media platforms, alongside a thorough review of current and upcoming EU regulations that shape the digital landscape (Section IV). The final section presents a detailed overview of existing technological tools and solutions, thoroughly analyzing their potential applications and specific contributions that could enhance the effectiveness of the AI4DEBUNK project's objectives (Section V).

Funded by the European Union

1 INTRODUCTION

The spread of misinformation and disinformation online is a recent but well-researched topic, though it is rapidly evolving with many aspects still undocumented. Studies demonstrate that social media platforms significantly amplify fake news through their structural design and user engagement patterns. On these platforms, fake news can spread instantly without verification. Such content often travels faster than factual news because of its engaging nature, which increases the likelihood of sharing. Social media algorithms compound this issue by prioritizing engaging content—typically favoring sensational information—which can further amplify fake news through the platforms' own mechanisms.

The NATO StratCom social media manipulation experiment reveals platform vulnerabilities and highlights efforts to identify and counter commercial manipulation and bot-generated AI content (Bay et al., 2023).

AI-powered algorithms are also becoming increasingly present in citizens' everyday lives. These algorithms are defined as systems capable of "interpreting external data accurately, learning from it, and using those insights to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019: 15).

Research shows that in decisions involving mechanical tasks, people perceive algorithmic and human-made decisions as equally fair and trustworthy, with similar emotional responses. Algorithms' perceived fairness and trustworthiness stemmed from their efficiency and objectivity. (Lee, 2018:8). The entities responsible for developing and regulating AI have a substantial impact on public trust. People have the highest confidence in national universities, research institutions, and defense organizations to develop, implement, and govern AI in the public interest, with 76 to 82 percent expressing trust in these institutions. In contrast, trust is markedly lower for governments and commercial organizations, with a third of respondents expressing limited confidence in these entities' ability to responsibly develop, utilize, and oversee AI systems (Gillespie et al., 2023:.9).

The results from our online survey show distrust in social media as a source for information online. Although it is frequently used as a source for news consumption as our results show, it is still perceived as the most likely source of disinformation and misinformation.

While social media is frequently used for news consumption, our findings reveal it is still perceived as the most likely source of disinformation and misinformation. Regarding measures to combat disinformation, the survey underscores the importance of regulations, particularly the regulation of social media platforms and of Media Literacy.

In 2018, the EU Commission created a High-Level Expert Group on Fake News and Online Disinformation, which outlined five key areas to tackle disinformation: (i) enhancing transparency within the digital information ecosystem, (ii) promoting media and information literacy, (iii) developing tools to empower users and journalists while encouraging positive engagement, (iv) ensuring the diversity and sustainability of the European news media landscape, and (v) conducting ongoing research on the effects of disinformation in Europe (European Commission: CNECT, 2019: 35). The report also highlights the

importance of fact-checking, advocating for increased visibility of fact-checking organizations to reach a broader audience (European Commission: CNECT, 2019:15).

In December 2018, the European Commission published an Action Plan Against Disinformation, outlining measures to strengthen the EU's capabilities to counter disinformation campaigns (EC, 2018 d). The plan includes initiatives to improve detection, analysis, and response to disinformation, enhance coordination among EU institutions and member states, and promote media literacy and critical thinking.

The European Democracy Action Plan was proposed in December 2020, to safeguard the integrity of elections and democratic processes in the EU. It includes measures to address disinformation, improve transparency of political advertising, support quality journalism, and strengthen media literacy (EC, 2020, b). The Digital Services Act (DSA) (EC, 2023)**:** was also proposed by the European Commission to update and harmonize rules for digital services in the EU. It includes provisions to tackle illegal content, including disinformation, by imposing obligations on online platforms to take measures to prevent the dissemination of harmful content while respecting fundamental rights. With the digital Service Act, the European Union transfers responsibility and accountability of the moderation to the online platforms themselves.

Our survey shows that most of the respondents consider Media Literacy as an important tool. From a prevention angle, the EU promotes media literacy initiatives to empower citizens with the skills to critically assess information and recognize disinformation. Funding programs support projects that enhance media literacy and promote quality journalism. The EU also established a Rapid Alert System in 2019 to facilitate the exchange of information among member states on disinformation campaigns targeting EU elections and other critical events. The system enables timely detection and response to disinformation threats. The European Digital Media Observatory (EDMO), launched in June 2020, was set up as a network of fact-checkers, researchers, and academics across Europe working to combat disinformation. It supports fact-checking activities, conducts research on disinformation trends, and provides analysis to policymakers and the public.

Finally, the AI Act (EC, 2021, b) is the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a more visible role globally.

The idea of media literacy education is to teach individuals how to think and not what to think (Art, 2018: 66). There are mainly two disciplines when it comes to how critical thinking skills should be taught. One is a general approach that aims to teach critical thinking as a stand-alone skill, while the second is a discipline-embedded approach that still teaches critical thinking within a specific context of one discipline, for example media literacy (Tiruneh et al., 2014: 3). While the general approach supports the idea of "teaching how to think instead of what to think," it has a key limitation. If critical thinking is taught only within the framework of a standard subject matter course, students may struggle to identify and apply thinking skills outside that context. As a result, they may not transfer what they have learned to other situations effectively (Tiruneh et al., 2014: 3). Therefore, a general approach to teaching critical thinking might not be effective to build resilience against misinformation.

Beyond media literacy in public education, video games such as the “Bad News Game” have been utilized to improve people’s media literacy and psychological resistance against online misinformation. Educational gaming can be a fun and visually appealing way to learn new concepts (Squire & Steinkuehler, 2011). Roozenbeek & Van der Linden (2019) provided evidence that the game “Bad News Game” improved the players ability to detect and resist a whole range of misinformation in the form of deceptive Twitter posts, and this result was consistent regardless of age, gender, and political standings (liberal or conservative). The intervention with the Bad News game also improved people's abilities to detect the tactics used to deceive in twitter posts with misinformation (Roozenbeek et al., 2019; Basol et al. 2020).

Structure of the report:

The report provides a comprehensive analysis of the AI4DEBUNK online survey, a detailed questionnaire comprising 15 carefully crafted questions that garnered meaningful responses from 329 participants across various demographics and user groups (Sections II and III). The analysis then delves into an extensive examination of content moderation practices implemented by major social media platforms, alongside a thorough review of current and upcoming EU regulations that shape the digital landscape (Section IV). The final section presents a detailed overview of existing technological tools and solutions, thoroughly analyzing their potential applications and specific contributions that could enhance the effectiveness of the AI4DEBUNK project's objectives (Section V).

References

Art, S. (2018). Media literacy and critical thinking. *International Journal of Media and Information Literacy*, *3*(2), 66-71. URL: https://cyberleninka.ru/article/n/media-literacy-and-critical-thinking

Basol, M., Roozenbeek, J., & Van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition*, *3*(1). doi: 10.5334/joc.91. PMID: 31934684

Bay, S., Dek, A., Fredheim, R., Haiduchyk, T., Stolze, M. Social Media Manipulation (2023). Assessing the Ability of Social Media Companies to Combat Platform Manipulation. Riga: NATO Strategic Communications Centre of Excellence. URL: https://stratcomcoe.org/publications/social-media-manipulation-20222023-assessing-the-ability-of-social-media-companies-to-combat-platform-manipulation/272

Bostrom, A., Demuth, J. L., Wirz, C. D., Cains, M. G., Schumacher, A., Madlambayan, D., ... & Williams, J. K. (2024). Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Analysis*, *44*(6), 1498-1513.

Bulger, Monica, and Patrick Davison. (2018). The promises, challenges, and futures of media literacy. *Journal of Media Literacy Education* 10.1: 1-21. Retrieved from: https://doi.org/10.23860/JMLE-2018-10-1-1

Funded by the European Union

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European management journal*, *40*(5), 685-706.

European Commission. (2018, September 26). Code of practice on disinformation. Digital Strategy. https://digital-strategy.ec.europa.eu/en/news/code-practice-disinformation

European Commission, 2018a, A Multi-dimensional Approach to Disinformation: Report of the Independent High-Level Group on Fake News and Online Disinformation. Directorate-General for Communication Networks, Content and Technology. Available at https:// ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformatio

European Commission, 2018b, Code of Practice on Disinformation. Available at https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

European Commission, 2018c, Synopsis Report of the Public Consultation on Fake News and Online Disinformation. https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation

European Commission (2018d) Tackling Online Disinformation: A European Approach (Communication) COM(2018) 236 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236

European Commission (2020a) Assessment of the Code of Practice on Disinformation —Achievements and areas for further improvement. Commission Staff working document (SWD(2020) 180 final).

European Commission (2020b) European Democracy Action Plan (Communication) COM(2020) 790 final.https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

European Commission: Directorate-General for Communications Networks, Content and Technology. (2018). *A multi-dimensional approach to disinformation : report of the independent High level Group on fake news and online disinformation*. Publications Office. **https://data.europa.eu/doi/10.2759/739290**.

European Commission. (2019, February 26). Antitrust: Commission fines Google €1.49 billion for abusive practices in online advertising [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_19_746

European Commission (2021) Guidance on Strengthening the Code of Practice on Disinformation (COM(2021) 262 final). https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation

European Digital Media Observatory. (2024). BECID D3.4 Report. Retrieved from https://edmo.eu/wp-content/uploads/2024/06/BECID-D3.4._report.pdf

European Digital Media Observatory. (2022). Policies to tackle disinformation in EU member states – Part II. https://edmo.eu/wp-content/uploads/2022/07/Policies-to-tackle-disinformation-in-EU-member-states-%E2%80%93-Part-II.pdf

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). Trust in Artificial Intelligence: A Global Study. The University of Queensland and KPMG Australia. doi: 10.14264/00d3c94

Handschuh, K. (2023). Critical Thinking and Media Literacy in an Age of Misinformation. Retrieved from: http://doi.org/10.33774/apsa-2023-l43bk

Haenlein, M. and Kaplan, A. (2019). A Brief History of AI: On the Past, Present,and Future of Artificial Intelligence. California Management Review, 61(4),5-14

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684.

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1-10. https://doi.org/10.1057/s41599-019-0279-9

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. In Higher Education Studies (Vol. 4, Number 1, pp. 1–17). Canadian Center of Science and Education. https://doi.org/10.5539/hes.v4n1p1

Ventsel, A., Hansson, S., Rickberg, M., & Madisson, M. L. (2023). Building Resilience Against Hostile Information

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146-1151. DOI:10.1126/science.aap9559

Yee, Amy (2022, January 31). Subjected to repeated disinformation campaigns, the tiny Baltic country of Estonia sees media literacy education as part of its digital-first culture and national security. BBC. Retrieved from https://www.bbc.com/future/article/20220128-thecountry-inoculating-against-disinformation

Zhang, L., Zhang, H., & Wang, K. (2020). Media literacy education and curriculum integration: A literature review. *International Journal of Contemporary Education*, *3*(1), 55-64. https://doi.org/10.11114/ijce.v3i1.4769

Funded by the European Union

2 PART II. AI4DEBUNK ONLINE SURVEY

The team of Pilot4dev created an online survey in June 2024 in the form of an online questionnaire with a mix of multiple-choice questions and open-ended questions. The questionnaire is available online[1]. It is anonymous and the names and email addresses of participants have not been collected. The questionnaire was translated into French, Italian, Dutch, Latvian, Bulgarian, Ukrainian, German, Norwegian and Greek. It was published on social media, the project website and an online poll platform. It was also disseminated to students, for instance, in Belgium and the Netherlands. The poll is open until the end of 2024, and we are regularly collecting and analyzing answers.

2.1 SURVEY DESIGN

The study utilized an online survey consisting of 15 questions (+ two questions about age and primary occupation) designed to gather insights into citizens' perspectives on disinformation detection and their preferences for the use of artificial intelligence (AI) tools to debunk disinformation. The survey included a combination of multiple-choice questions, Likert scale items, and open-ended questions. The multiple-choice and Likert scale questions provided quantitative data on participants' media consumption habits, exposure to disinformation, and opinions on the effectiveness of AI in combating false information. Open-ended questions allowed participants to express qualitative feedback regarding their preferences for AI tool design and functionality.

The online survey was chosen as the preferred method for this research over other methods, such as interviews, for several key reasons. First and foremost, the online survey format allowed us to consult a larger number of respondents across multiple countries and languages, ensuring a broader and more representative sample of citizens' perspectives on disinformation detection and the use of AI tools. This reach was particularly important for a project aimed at understanding how diverse groups of citizens spread across different geographical and linguistic contexts, interact with and perceive disinformation.

The survey was created using Google Forms to ensure ease of distribution and data collection. Using Google Forms was preferred because it was a cost-effective alternative which was considered beneficial and convenient to produce the survey in different languages. Moreover, Google Form is easy and user friendly to the responders. The poll was translated into 11 languages—English, German, French, Italian, Dutch, Greek, Norwegian, Latvian, Bulgarian, Ukrainian—to accommodate diverse participants across different countries and linguistic backgrounds. The intention of translating the survey in different languages was to make the survey more accessible to non-English speakers. These languages were not intended for country-specific comparative analysis but rather to broaden participation by including individuals who are either uncomfortable with English or prefer their native language. However, without specific cultural or regional targeting, the results cannot be used for comparing disinformation patterns

[1] https://docs.google.com/forms/d/e/1FAIpQLSfJ6RAs1makx1Y23CqKg2HZi5BuVtymJuiGvQ_ApO8jqJOwzQ/viewform?usp=sf_link

or AI tool preferences across different countries. Therefore, the language diversity served more as a way to gather broader perspectives but does not support any formal cross-country comparisons.

2.2 TARGET POPULATION AND SAMPLE

The target population for the survey was broad, aiming to capture the views of adult citizens across various European countries, who engage with online media and may have encountered disinformation. Recruitment was conducted through convenience sampling, with the survey distributed by various project partners across their social media networks and shared on the project's own social media channels and website. To reach a wider audience, the survey was also made available on a website called PollPool, a platform designed for gathering responses to public surveys. The use of multiple distribution channels helped to increase participation across diverse demographics.

The survey collected a total of 329 respondents that participated in the survey between all the languages combined. It was distributed between the surveys in different languages accordingly; English (60), German (52), Greek (51), Dutch (39), Norwegian (29), Bulgarian (25), Ukrainian (24), French (22) Lavtian (18), Italian (9).

The survey response rate was highest among young adults, with 80 respondents aged 18-24 and 85 aged 25-34. Response numbers then declined steadily by age group, with 57 respondents aged 35-44, 51 aged 45-54, and 36 aged 55-64. The 65-and-older group contributed 14 responses, while those younger than 18 were the fewest, with only 6 responses. This age distribution highlights a stronger engagement from younger age groups, tapering off significantly among older participants.

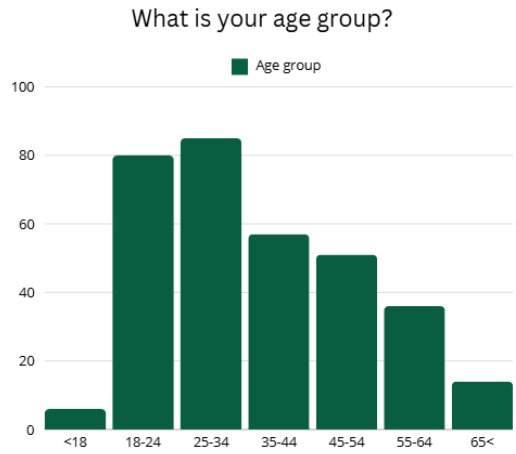


FIGURE 1: ILLUSTRATION OF AGE DISTRUBTION FROM SURVEY PARTICIPANTS

2.3 DATA COLLECTION PROCESS

Data collection occurred entirely online through Google Forms. The survey was open for participation over a specified period and was promoted through the networks of project partners, project-specific social media accounts, and on the project's website. Additional exposure was achieved by posting the survey on PollPool to attract more respondents from a broader audience.

In consideration of privacy concerns, the survey was designed to be fully anonymous. No personally identifiable information, such as names or email addresses, was collected from respondents. Participants were informed that their responses would be confidential and used solely for research purposes.

2.4 SURVEY INSTRUMENT

The survey was structured into three key sections:

Media Consumption and Disinformation Exposure: This section aimed to gather information about participants' habits in consuming news and media, their experiences with encountering disinformation, and their ability to recognize misleading or false information. Multiple-choice questions and Likert scales were used to assess these aspects.

Perspectives on AI for Debunking Disinformation: Respondents were asked about their familiarity with AI and whether they would consider using an AI tool to help detect disinformation. Likert scale questions were used to gauge the level of trust in such tools and their perceived usefulness.

AI Tool Design Preferences: Open-ended questions were included to allow participants to express their opinions on what features they would like to see in an AI tool designed to combat disinformation. This qualitative feedback was crucial for identifying user preferences in terms of interface, usability, and specific functionalities (e.g., real-time fact-checking, visual alerts).

2.5 ETHICAL CONSIDERATIONS

The survey adhered to ethical guidelines regarding privacy and anonymity. The survey was made compliant with General Data Protection Regulation (GDPR) policy standards for privacy concerns, meaning no names, email addresses, or other personally identifiable information collected, and participants were assured of their anonymity throughout the process. Additionally, there was no form for tracking tools used to identify participants. Participation was entirely voluntary, and no incentives were offered. Meaning the participants could stop the survey at any time they wanted, and were not forced or pressured to finish the survey if they didn't want to.

2.6 LIMITATIONS OF THE SURVEY

While the online survey provided valuable insights into citizens' perspectives on disinformation and preferences for AI tools designed to combat it, there are several limitations that should be acknowledged regarding the sample size, distribution method, and demographic representation.

One significant limitation of the poll is that it collected responses from only 329 participants. While this provides a useful snapshot of public opinion, it is not sufficient to claim full representativeness of the broader population, particularly when dealing with a complex issue like disinformation that affects people across various age groups, regions, and socio-economic backgrounds. This small sample also increases the margin of error, which could lead to inaccurate estimations of public opinion. A larger sample size would have allowed for more robust statistical analysis and the ability to draw more generalized conclusions from the data.

The way in which the survey was distributed also presents challenges to the representativeness of the sample. Recruitment relied on convenience sampling, with the survey being distributed through the social media networks of project partners, the project's own social media channels and website, and PollPool, a public survey website. While these channels helped reach a broad audience, they may have attracted respondents who are already engaged in the issue of disinformation or who are more digitally literate. This could lead to sample bias, as those who are less active online, such as older adults or individuals with limited digital literacy, may be underrepresented.

Furthermore, the use of online platforms for survey distribution likely skewed the sample toward a younger, more tech-savvy demographic. This is evident in the fact that the majority of respondents were aged between 18 and 34. While this demographic is certainly affected by disinformation, older adults—who tend to be more vulnerable to fake news—were underrepresented in the survey. Older individuals may lack the technological skills or the motivation to participate in an online survey, making it less likely that their perspectives were captured.

This overrepresentation of young people brings us to a significant limitation of the survey, which is the underrepresentation of vulnerable groups. The small sample makes the results vulnerable to underrepresent sub-populations who are arguably most in need of tools to help debunk disinformation—particularly older adults. The results have very few participants from the age group 65 and older. Research consistently shows that older adults are more susceptible to disinformation, partly due to lower levels of digital literacy and a higher likelihood of encountering misleading information on social media platforms. The fact that the majority of respondents were younger (aged 18-34) suggests that the poll may not have fully captured the perspectives and needs of older adults who might benefit the most from an AI tool designed to detect fake news.

As a result, the feedback obtained may be biased toward younger users' preferences for AI tool design, which might not align with the needs of older individuals, who may require different interfaces or levels

of support when using such tools. This presents a challenge for ensuring that the final AI tool effectively serves the populations most vulnerable to disinformation.

The survey design may also present limitations in capturing important nuances because of the design of the questions. Many of the questions use multiple-choice formats with predefined options to streamline the survey experience and reduce the risk of participant fatigue. However, these fixed options might not fully represent the diversity of respondent experiences or perspectives, and they may be too broad or generalized to yield meaningful insights. For instance, categories like "Social media," "News organizations," and "Independent media" are quite vague and can carry different interpretations depending on cultural or regional contexts. This lack of specificity could obscure critical nuances in how respondents perceive and interact with disinformation, making it challenging to draw actionable conclusions from the data. A more refined approach, incorporating precise categories or opened for more opportunities for open-ended responses, could help capture the depth and complexity of participants' views more effectively.

Finally, the voluntary nature of participation introduces the possibility of self-selection bias. Individuals who chose to complete the survey might have done so because they have a particular interest or awareness of disinformation, which could result in a skewed representation of public opinion. Those who are less informed or concerned about fake news may not have been as motivated to participate, further limiting the generalizability of the findings.

Funded by the European Union

3 PART III RESULTS OF ONLINE SURVEY

3.1 CONCERNS ABOUT FAKE NEWS AND SOCIAL MEDIA PLATFORMS

The first question in the survey asked the respondents "**How confident are you in your ability to Identify fake news?"**. The question was multiple choice, and the respondent had to put themselves on a scale, from 5 being "very confident" in their abilities, to 1 "very unconfident". The results gave us 39 responses

FIGURE 3: RESULTS FROM QUESTION "HOW CONFIDENT ARE YOU IN YOUR ABILITY TO IDENTIFY FAKE NEWS?"

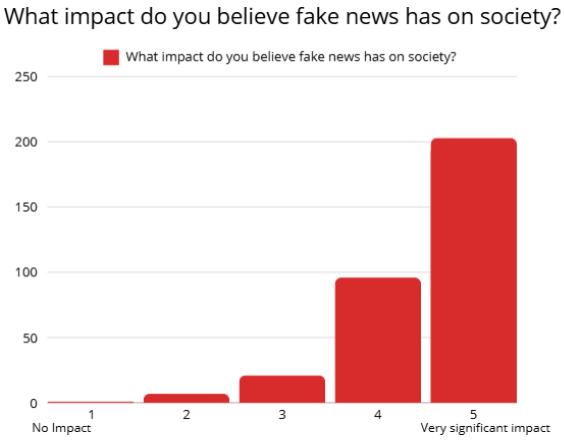


FIGURE 2: RESULTS FROM QUESTION "WHAT IMPACT DO YOU BELIEVE FAKE NEWS HAS ON SOCIETY

for very confident, while the most common response was "4", with 150 respondents being confident in their abilities. Then 103 responses for "3", 27 respondents reported themselves to "2" on the scale and lastly 9 respondents reported to be "very unconfident" in their ability. The results show very few on both extremes, with most responses in the middle leaning towards the confident side.

The second question then asked the respondents, "**What impact do you believe fake news has on society?"**. There was a strong consensus among the respondents that it has a highly significant impact. In fact, 91% of all respondents indicated either a "significant impact" or "very significant impact" of fake news on society, where "Very significant" was the most common answer with 203 responders choosing this option. This concern was shared across all countries, age groups, and genders. Followed by the second most common response which was "4" on the scale, indicating that respondents perceived it to have a significant impact, 21 responders answered "3" on the scale, and 7 put "2". Only one single respondent

Funded by the European Union

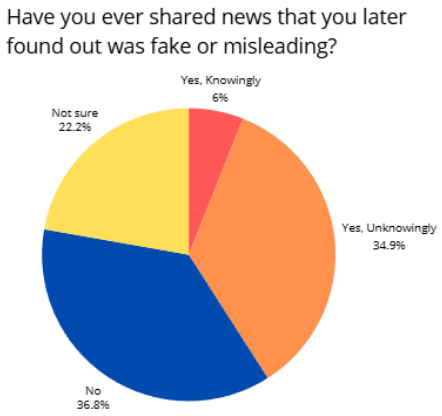


FIGURE 4: RESULTS FROM QUESTION "HAVE YOU EVER SHARED NEWS THAT YOU LATER FOUND OUT WAS FAKE OR MISLEADING?"

out of 329 reported to perceive that fake news had "no impact at all" highlighting the broad recognition of fake news as a serious societal issue from a citizen's perspective.

For the third question, the survey asks the respondents on their own history of sharing false or misleading information. The question was worded as “Have you ever shared news that you later found out was fake or misleading?” and the responders could either answer “Yes, Knowingly”, “Yes, Unknowingly”, “No” or “Not sure”. The Most common response was “No” with 36.8% of the respondents, then closely after “Yes, unknowingly” with 34.9% of the responses, then “Not sure” with 22.2% and finally 6% of the respondents reported “Yes, Knowingly”. So, although not many have shared fake news knowingly, it is still not an insignificant group.

In the next question, respondents were asked “**How familiar are you with the concept of Deepfakes?**” followed by “How concerned are you of the potential impact of deepfake on society”. Both questions respondents were overall less familiar with the concept of Deepfakes than Fake News.

39 responders reported to be “Not familiar at all” with deep fakes. Then less than 31 respondents reported

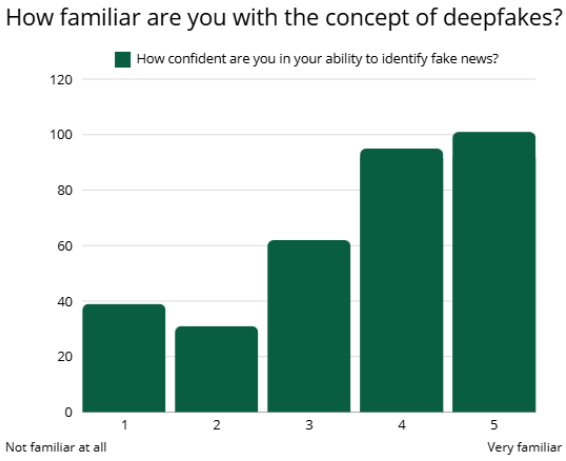


FIGURE 6: RESULTS FROM QUESTION "HOW FAMILIAR ARE YOU WITH THE CONCEPT OF DEEPFAKES?"

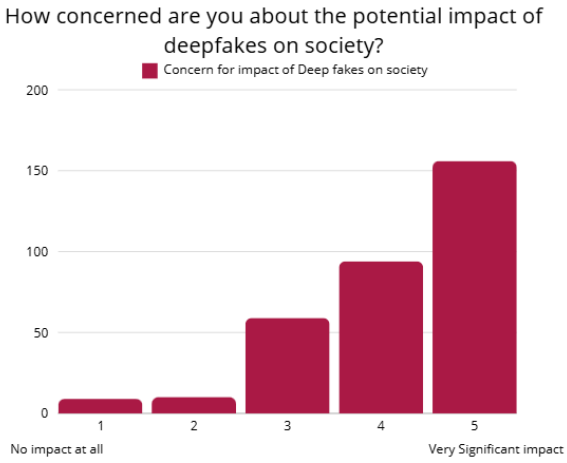


FIGURE 5: RESULTS FROM QUESTION "HOW CONCERNED ARE YOU ABOUT THE POTENTIAL IMPACT OF DEEPFAKES ON SOCIETY?"

to be not very familiar with the concept (2 on the scale), 62 respondents answered 2-3, 95 answered that they were familiar, and 101 reported to be “very familiar” with the concept of deep fakes. So, still a majority of the respondents were familiar with the concept, but the graph shows a much more complex picture, with a good part of the population being familiar with the concept, but still a significant portion not feeling familiar with it at all. Then for how concerned the respondents were for the impact of deep fakes, it shows a similar trend as the previous question on the impact of fake news on society. Most of

Funded by the European Union

the respondents answered "Very significant impact" with 156 responses, and 94 responses for "4" significant impact. Then 59 responses show for "3" in the middle, 10 responses for "2" and only 9 of the respondents reported to perceive deep fakes to have "No impact at all".

For the following question, respondents were asked about their news consumption "Where do you consume your news? (Select all that apply)". This was also a multiple-choice question where the respondents could choose between "Online news websites", "TV", "Newspapers", "Radio", "Social Media" or Others, and they were asked to apply all options that applied to them. This resulted in 265 responses for "Online news websites", followed by "Social media" with 235 responses. Then TV received 152 responses, 97 responders reported listening to Radio for news consumption, 86 responders, 19 responses were also given to the "others" category. Here, respondents reported "WhatsApp" which arguably also belongs to the social media category, others wrote "podcast", which could be put under radio category. Also, one respondent in this other category mentioned "Various expert groups" as a source for news consumption. Moreover, three responses reported some version of "I don't read news".

Then, the participants were required to answer the following question: "How often do you encounter news that you believe to be fake or misleading?". The multiple-choice options were then divided into frequencies of seeing fake news either on a "Daily", "Weekly" or "Monthly"-basis, "Rarely" or "Never". The results showed that overall responses a majority of the respondents reported to encounter news they

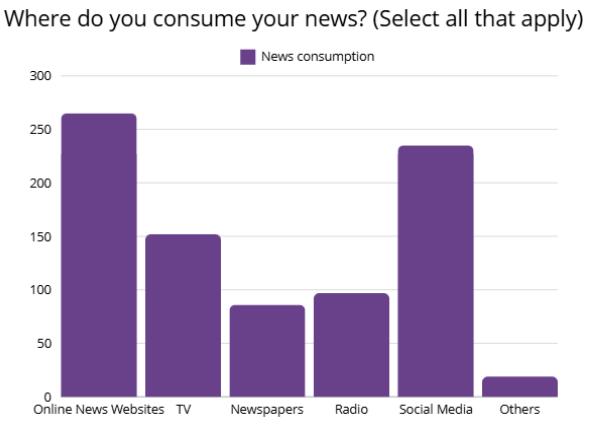


FIGURE 7: RESULTS FROM QUESTION "WHERE DO YOU CONSUME YOUR NEWS?"

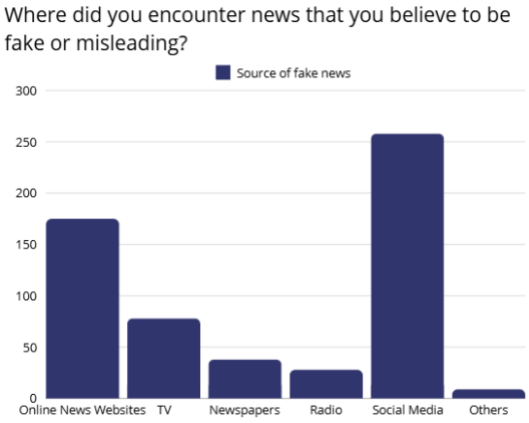


FIGURE 8: RESULTS FROM QUESTION "WHERE DID YOU ENCOUNTER NEWS THAT YOU BELIEVE TO BE FAKE OR MISLEADING?"

perceived as fake or misleading on a regular basis, with 34.6% reported to encounter it on a "Daily"-basis and 35.8% on a "Weekly"-basis. Then 15.7% percent of our respondents said they encounter it on a monthly basis, and 12.6% reported only seeing it "Rarely". Lastly, a very small portion said to "Never" encounter news they perceive as fake or misleading, and there were no demographic trends between the ones who answered "Never", there was an equal distribution of "Students" "Retired" "Full-time employed" and "other".

Funded by the European Union

From there, the questions became more specific and targeted **our two case studies, on climate change and the war in Ukraine**. First, "How often do you encounter news that you believe to be fake or misleading specifically regarding climate change?". To this question responses like "Daily" and "weekly" were less common, and "Monthly" and "Rarely", were more common than when respondents were asked just about encountering any fake news. Here 13.8% were for "Daily", and 30.6% of the responses were for "weekly". While 27.5% of the respondents answered "Monthly", and 26.6% responses were for "Rarely". So, there are some indications that climate change related fake news is less commonly encountered or at least noticed by our respondents. Also, to this question, not many reported "Never" seeing climate related news, with only 1.6% responses, mirroring the results from the previous question.

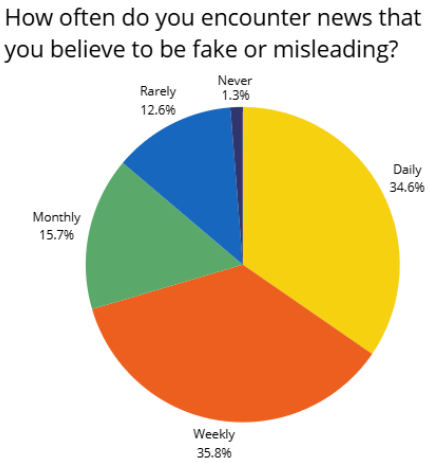


FIGURE 9: RESULTS FROM QUESTION "HOW OFTEN DO YOU ENCOUNTER NEWS THAT YOU BELIEVE TO BE FAKE OR MISLEADING?"

Meanwhile, in response to the question, "How often do you encounter news that you believe to be fake or misleading specifically regarding the war in Ukraine?" respondents reported encountering fake news about the war more frequently than about climate change. A notable 21.1% reported encountering fake news about the Ukraine war on a "daily" basis, and 35.5% on a "weekly" basis—both significantly higher rates than those for climate change. Additionally, 22.7% reported encountering such news "monthly," and 16.5% answered "rarely."

Breaking down these results by language sample, respondents in "Ukrainian," "Bulgarian," and, to a slightly lesser degree, "Latvian" and "Greek" language groups reported much higher frequencies of

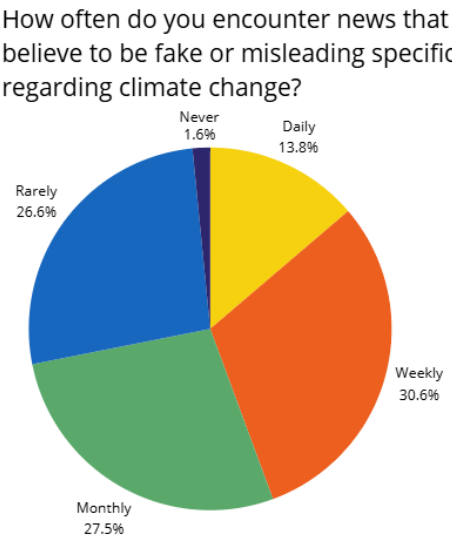


FIGURE 11: RESULTS FROM QUESTION "HOW OFTEN DO YOU ENCOUNTER NEWS THAT YOU BELIEVE TO BE FAKE OR MISLEADING SPECIALLY REGARDING CLIMATE CHANGE?"

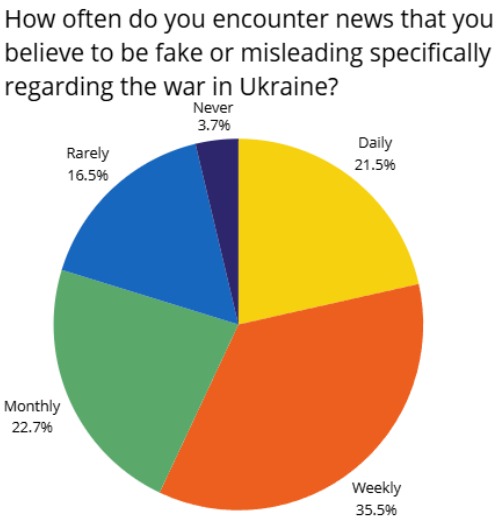


FIGURE 10: RESULTS FROM QUESTION "HOW OFTEN DO YOU ENCOUNTER NEWS THAT YOU BELIEVE TO BE FAKE OR MISLEADING SPECIALLY REGARDING THE WAR IN UKRAINE?"

Funded by the European Union

encountering fake news about the war in Ukraine on a “daily” or “weekly” basis. For example, 62.5% of Ukrainian-language respondents reported “daily” encounters, compared to only 7.7% of German-language respondents.

It is important to note that proximity to the conflict appears to strongly influence these results. Without the Ukrainian-language sample, the frequency of fake news encounters about the war in Ukraine would align more closely with that of climate change, indicating a regional component to perceived misinformation exposure.

The following question was: “What sources do you believe are most likely to spread disinformation about the war in Ukraine?” Participants could select up to two options from the following: “Social Media,” “Government Statements,” “News Organizations,” “Independent/Alternative News Sites,” and “Other”. The results highlighted that nearly half (47.1%) identified “Social Media” as the most likely source of disinformation, aligning with previous responses to the question, “Where did you encounter news that you believe to be fake or misleading?” Additionally, 22.8% selected “Independent/Alternative News Sites” as a likely source of misinformation, while 15.4% pointed to “Government Statements,” and 12.7% indicated “News Organizations.”

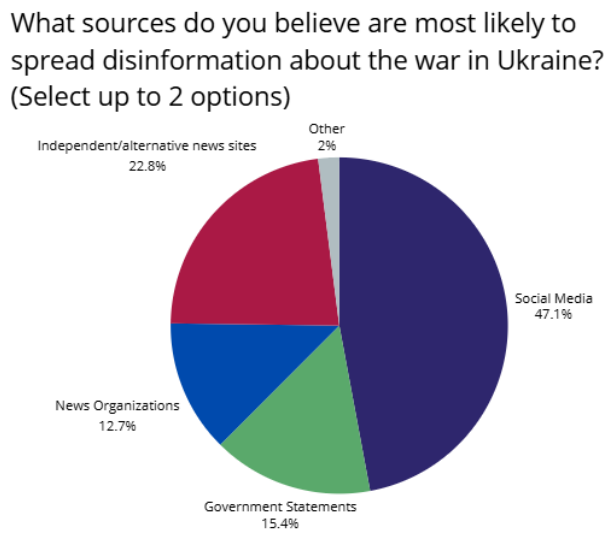


FIGURE 12: RESULTS FROM QUESTION "WHAT SOURCES DO YOU BELIEVE ARE MOST LIKELY TO SPREAD DISINFORMATION ABOUT THE WAR IN UKRAINE?"

These results underscore a strong perception that social media is a primary conduit for misinformation about the war in Ukraine, with a significant portion also wary of independent or alternative news sources, and this trend was apparent across all language samples.

In the "Solutions" section of the survey, respondents were asked, "What measures do you think would be most effective in combating fake news?" and could select up to two options from the following: "Greater public education on media literacy," "Increased collaboration (between fact-checkers, journalists, technology developers)," "Technological solutions," and "Policy solutions and regulations."

The results reveal that **Greater public education on media literacy and Policy** solutions and regulations were the top choices, preferred by 36.5% and 33.2% of respondents, respectively. Technological solutions followed, with 15.3% choosing this option, while Increased collaboration was selected by 12.3%. An

Funded by the European Union

additional 2.7% of responses were categorized as “Other,” offering a range of suggestions that, on closer analysis, largely align with the primary categories. For instance:

➔ Suggestions like “Hold tech giants like Facebook accountable and ban TikTok” and “Put policies in place to prevent politicians from spreading fake news, especially during election campaigns” align closely with Policy solutions and regulations.

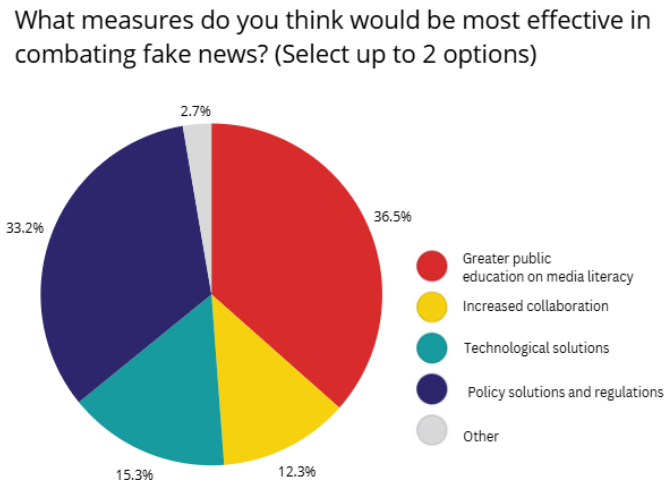


FIGURE 13: RESULTS FROM QUESTION "WHAT MEASURES DO YOU THINK WOULD BE MOST EFFECTIVE IN COMBATING FAKE NEWS?"

➔ Ideas such as “Introduce media literacy into the school curriculum,” “Make critical thinking a mandatory part of school education, including courses for adults and elderly citizens,” and “Strengthen critical thinking skills through education” fit well under Greater public education on media literacy.

A few responses in the "Other" category offered unique viewpoints that did not clearly fit into the predefined options. For example, one respondent suggested "changing the media landscape to reduce clickbait dependency," while another advocated for supporting "independent, non-imperialist media." Other comments emphasized the need for "critical thinking among consumers." Some responses, however, took a more humorous approach, with one suggesting simply “plugging your ears.”

Overall, these responses in the "Other" category further highlight public preference for policy solutions and enhanced media literacy education as key strategies for combating fake news’.

3.2 AI TOOL USAGE AND DESIGN PREFERENCE

Finally, the last section of the survey “AI-tools to counter disinformation” asks two open-ended questions about using AI to counter disinformation and about design preference. The first question in this section was “Do you believe AI could be effective in countering fake news and disinformation? (Please explain in 1-2 sentences)”. The responses reflect a cautious optimism with some reservations. 53 of the responses were outright positive, such as “Yes, I'm sure of it.” Many respondents think AI has the potential to effectively counter disinformation by identifying patterns and fact-checking sources. Some believe that AI can improve over time, as its algorithms learn to distinguish between real and fake information with increased accuracy. Certain respondents propose that AI can be part of a broader system to track and flag disinformation, provided that it is transparently programmed and well-regulated.

Meanwhile, a significant portion of respondents expressed outright distrust and skepticism. We have identified 61 responses that can be clearly defined as negative, expressing solely negative attitudes.

Funded by the European Union

Examples of this are responses such as “No, I think AI will make the situation worse” and “No, people don't want to”. Some respondents highlight that AI can both combat and create fake news, noting that AI is already used to spread misinformation through bots and deepfakes. For example, one respondent said, "I don't think so, because AI models can also be used to create fake or false news". This ambivalence reflects distrust in AI’s role, particularly in environments where it could be exploited for malicious purposes.

Many responses reflect nuanced perspectives with answers including some sort of phrasing like "But", "If" or "Possibly, however", with respondents acknowledging potential benefits while also noting risks or limitations. Numerous responses indicate that AI should operate alongside human supervision to account for context and ensure reliable results. This shows that while AI is seen as a valuable tool, it is not perceived sufficient on its own. A concern shared by many is also that the AI would be Biased and emphasize that an AI-tool has to be ideologically neutral and retrieve data from multiple different sources to counter potential biases.

The results also show many responses that indicate significant knowledge gaps and uncertainties. Many responses reflect a lack of familiarity or confidence in understanding AI’s capabilities, with phrases like "I don't know," "uncertain," or "little knowledge." This indicates a knowledge gap that might impact public trust and the perceived effectiveness of AI in combating disinformation. A subset of participants seems optimistic but unsure, reflecting a sense of hope in AI’s capabilities without a strong understanding of how it works or the specifics of its effectiveness. Nonetheless, greater knowledge on AI capabilities seems to have a positive impact on citizens' perceptions of using AI to counter disinformation.

Then the second question in this section and the last question that was presented to the respondents was “What characteristics would an ideal tool have to combat fake news and disinformation?” and to describe their preference in 2-3 sentences, but the google form settings allowed them to give longer answers if they wanted. From the results gathered, there are some key themes that emerge. We have identified some ideal features for a tool to combat fake news from the citizens perspective.

Many responses emphasized fact-checking mechanisms, with specific preferences for real-time fact-checking, cross-referencing, and database comparison. Respondents favored features like comparing news against a database of verified facts or other reputable sources. Some respondents mentioned similarity to plagiarism-checkers, aiming for the tool to recognize previously debunked or discredited information.

Also, for this question, the issue of bias and neutrality was brought up. Some suggestions emphasize that the tool should not censor legitimate viewpoints and must avoid political bias. Moreover, some respondents were wary of overreach or censorship, stressing that the tool should function objectively without influencing or infringing upon free speech. A core need expressed in many responses is transparency about why certain information is flagged, and accountability for those who spread misinformation. Some desired features expressed on this issue included explanations for why a news item is false, identifying the origin and intent behind misinformation, and highlighting any motivations that might bias the information.

A significant aspect of an ideal tool for combating misinformation is its capacity to foster user education and media literacy. Respondents expressed a strong desire for features that empower users to independently understand and identify misinformation, suggesting that the tool includes educational resources, media literacy courses, and tutorials for identifying false information. Some also recommend that the tool offers real-time feedback or alerts about content from sources known for spreading misinformation. This emphasis on education underscores a broader recognition that empowering users to make informed judgments and think critically is essential for a sustainable, long-term solution to misinformation.

User friendliness and accessibility are also desired features expressed in many responses. The tool should be intuitive, fast, and widely accessible. More specific characteristics include a user-friendly interface, universal design, and cross-platform compatibility. Respondents also mentioned a need for speed and instant alerts, as misinformation spreads quickly and requires timely responses. Ease of use is seen as crucial for adoption and efficacy, suggesting that the tool should be simple and adaptable, with features that appeal to users with varying levels of technological literacy.

Some responses also include specific technological features. Suggesting a tool could include natural language processing (NLP), image verification, or real-time analysis. Some responses express that image and text analysis for deepfake detection could be employed.

4 PART IV. PLATFORMS REGULATION AND CONTENT MODERATION

The results from our online survey underscores distrust in social media as a source for information online. Although it is frequently used as a source for news consumption as our results show, it is still perceived as the most likely source of disinformation and misinformation.

Regarding measures to combat disinformation, the survey underscores the importance of regulations, particularly the regulation of social media platforms.

DisinfoLab has published a series of reports on major platforms—Facebook, YouTube, TikTok, and X (formerly Twitter)—detailing how each one moderates content. These reports showcase the platforms' extensive content moderation policies aimed at curbing hate speech, misinformation, disinformation, propaganda, and other influence operations. Enforcement typically relies on AI-powered automated systems, complemented by human moderators who review flagged content. However, despite similar practices there are still significant differences between the platforms as well. This section summarizes key insights from these reports, along with the latest transparency reports released under the Digital Services Act (DSA) for the first half of 2024. Additionally, it incorporates external analyses and critiques of the shortcomings in these moderation procedures.

The EU has self-reported how the Code of Conduct has been implemented and how the platforms have complied with the regulatory framework. They found that notable progress has been made, particularly

in the removal of fake accounts and in reducing the visibility of websites that spread disinformation (European Commission, 2019, February 26). However, the Code has also been reported to have some shortcomings and faced criticism both for the implementation and for its design. Self-regulatory measures have struggled to adequately address the issues of transparency and integrity in political advertising, and enforcement has been inconsistent across the digital landscape (European Commission, 2019, February 26). Consequently, the effectiveness of self-regulation in ensuring openness and accountability in political advertising practices remains limited (Bayer, 2024:. 276). Facebook, in particular, has faced significant challenges in meeting transparency requirements for political ads, further highlighting the shortcomings of the EU's efforts to rely on platform self-regulation (Bayer, 2024:. 277).

During the Covid-19 pandemic, there was a significant influx of disinformation, prompting the release of a report titled *Covid check*, which evaluated the performance of the Code of Conduct during this time and identified several shortcomings in both its implementation and scope (Culloty et al., 2021: 4). The report found that the framework lacked standardization in the reporting of online disinformation, resulting in considerable variability in the structure and content of reports submitted by signatories. This inconsistency hindered effective analysis and comparison, leading (Culloty et al. 2021:44) to advocate for greater standardization in reporting practices. Additionally, the report highlighted notable inconsistencies in the application of disinformation measures across different countries, raising concerns about the reliability of the reported metrics.

Furthermore, Kuczerawy (2019) points to other limitations with the Code and criticizes the implementation of it for lacking essential safeguards to adequately protect freedom of expression. As the Code has a declaration on being "mindful of the fundamental right to freedom of expression and to an open Internet..." and that this won't replace existing legal framework such as the EU Charter of Fundamental Rights and the European Convention on Human Rights, Kuczerawy still argues that this code will still affect the exercise of the right to freedom of expression and access to information online (Kuczerawy, 2019:7). The article argues that safeguarding this right requires the introduction of procedural measures to enhance fairness, ensure proportionality, and incorporate elements of due process into the Code. In his argument, Kuczerawy outlines four potential safeguards that should be further developed: (1) notifying content providers, (2) allowing for counter-notifications, (3) establishing appeal mechanisms, and (4) improving the monitoring of the Code's implementation (p. 13).

However, early assessments of the DSA have pointed out some potential limitations (Griffin, 2024). For instance, developing and updating codes of conduct is complex and resource intensive. The negotiation process for the updated Code of Practice on Disinformation took over a year, partly due to the impact of the Ukraine war, which strained the capacity of both policymakers and platform staff in their anti-disinformation efforts (Griffin, 2024:182). Additionally, EU regulators may lack the capacity to oversee multiple codes simultaneously, potentially leading to inadequacies in the enforcement and oversight of the DSA (Griffin, 2024:186).

The following sections show the efforts developed by the platforms.

4.1 DISINFORMATION AND CONTENT MODERATION ON FACEBOOK

Facebook is one of the world's largest social media platforms, with approximately 2.9 billion monthly active users (MAUs) globally and 259 million MAUs in the European Union. As a Very Large Online Platform (VLOP), Facebook is subject to additional transparency reporting requirements under the European Union's Digital Services Act (DSA). Facebook's structure comprises profiles, groups, and pages, which serve as the foundation for information sharing and interaction. Users can create profiles, join groups, and follow pages based on their interests. Key features of Facebook include newsfeeds with personalized feeds displaying friends' posts, updates from pages and groups, and sponsored content. Facebook also has Timelines, with a user's record of all shared posts and interactions, including tagged content. Application interactions include reactions, comments, shares, asking for recommendations, checking-in at locations, and more.

Facebook employs a multi-layered approach to combat misinformation, combining user reports, automated tools, human moderation, and fact-checking partnerships. Users can flag problematic content through a simple reporting menu, which is then reviewed either by machine learning models or moderation teams. Tools like the Graph API and CrowdTangle provide insights into trends and content interactions, aiding researchers and verified organizations in monitoring misinformation. Additionally, Facebook's Meta Ad Library and Transparency Centre offer resources for tracking ads and adversarial activities.

Content violating Facebook's Community Standards—addressing issues like health misinformation, voter interference, and manipulated media—is removed or demoted. For repeat violations, Facebook implements a strike system, leading to restrictions such as reduced visibility or demonetization. Despite its robust mechanisms, challenges remain, including potential inconsistencies in enforcement, reliance on user participation, and limitations on independent data access due to tools like Meta Content Library being restricted to vetted users. These factors can hinder broader transparency and real-time interventions.

4.2 DISINFORMATION AND CONTENT MODERATION ON YOUTUBE

YouTube, founded in 2005 and acquired by Google in 2006, is an online video-sharing platform. It has significantly transformed the way people engage with video content, including information, and has enabled new forms of monetization, giving rise to professional content creators known as "youtubers.". YouTube's visual nature and language barriers can introduce challenges when researching disinformation. However, YouTube offers various elements beyond videos that can be explored for research. Some essential characteristics to consider when researching on YouTube include complex search challenges, personalized search results, and distinctive video IDs. YouTube's recommendation algorithms are also a

crucial aspect of the platform and can be investigated to understand the spread of disinformation (Miguel Serrano, 2024 February, p. 3-4).

YouTube moderates content through a comprehensive set of policies, mechanisms, and collaborations, focusing on mitigating misinformation and harmful content while upholding platform standards. YouTube does face challenges regarding content moderation and disinformation when analyzing non-English videos or complex recommendation algorithms as mentioned above. However, it also provides tools to aid investigations, such as unique video IDs for tracking and advanced search functions with filters like upload date, duration, and keywords (Miguel Serrano, 2024 February, p. 7).

Content moderation on YouTube revolves around its Community Guidelines, which prohibit harmful or deceptive content, including medical or election-related misinformation, hate speech, spam, and more. To enforce these policies, YouTube employs a combination of automated systems and human reviewers. It relies on its "Four Rs" framework: removing violative content, reducing the spread of borderline material, raising authoritative sources, and rewarding trustworthy creators (Miguel Serrano, 2024 February, p. 8-9).

Users can report content through specific processes depending on the type of content (e.g., videos, comments, or ads). Reporting includes selecting violations such as misinformation, hate speech, or child abuse. Certain violations, like legal infringements (e.g., copyright, defamation), require specialized reporting mechanisms. Verified experts, called priority flaggers, assist in flagging disinformation, while creators can manage comments on their videos using moderation tools (Miguel Serrano, 2024 February, p. 10-13).

YouTube moderates flagged content by reviewing reports and applying penalties such as removal, reduced visibility (downranking), or demonetization. For recurring violations, channels may face termination. Users also have access to appeal processes as mandated by the European Union's Digital Services Act (DSA), which enhances transparency and accountability for very large online platforms like YouTube.

To address systemic risks such as disinformation, YouTube collaborates with initiatives like the Strengthened Code of Practice on Disinformation and partners with fact-checking organizations. Additionally, its transparency reports, required under the DSA, provide updates on moderation efforts and mitigation measures. These layered strategies reflect YouTube's ongoing commitment to balancing content moderation, user engagement, and compliance with evolving regulatory standards.

While organizations like the Global Alliance for Responsible Media (GARM) have introduced guidelines to exclude misinformation from ad revenue, YouTube’s policies fall short, neither excluding such content from monetization nor providing features like labeling. Moreover, the United Nations has called for platforms to establish robust measures to address misinformation, emphasizing the importance of transparency in advertising and demonetization. Despite these efforts, YouTube’s inaction undermines accountability, allowing misinformation to flourish. This situation harms multiple stakeholders. Brands

suffer reputational damage, as consumers are less likely to trust or support companies advertising alongside low-quality content. At the same time, legitimate news outlets face financial losses, as misinformation draws larger audiences and higher revenue, reducing viewership for credible sources (Raso et al., 2024, July 10).

4.3 DISINFORMATION AND CONTENT MODERATION ON X

X is a social media platform centered around users posting, sharing, and interacting through brief messages, which were once known as “tweets” but are now simply called posts. Users create accounts with a unique identifier called a username or handle, and they can follow others to view posts on a customized feed. X’s main functions encourage engagement and community through a wide array of features, including hashtags, direct messages, and live broadcasting options like video and audio spaces. Privacy and notification settings are customizable, allowing users control over who can see their posts and which notifications they receive. Users can save or organize content through bookmarks, custom lists, and the Explore tab, which suggests trending content based on user activity (Hénin, & Giovanna Sessa, 2024 February).

Content moderation on X is a priority, especially regarding harmful or misleading content. The platform has created a multi-faceted approach to maintaining a safe space for users. This includes adherence to strict platform rules that cover categories such as spam, abusive behavior, violent threats, and impersonation, among others. X’s Help Center allows users to report specific violations, from unauthorized use of copyrighted material to child safety issues. Recently, the platform introduced new reporting options to align with the EU’s Digital Services Act (DSA), providing a way to report illegal content within the EU and appeal decisions. X emphasizes that these issues are handled through a combination of human moderation, automated technology, and partnerships with external experts, although it no longer has formal partnerships with European fact-checking organizations (Hénin, & Giovanna Sessa, 2024 February).

A key component of X’s misinformation moderation strategy is “Community Notes,” formerly known as Birdwatch. This crowdsourced tool allows users to add contextual notes to posts, promoting community-driven fact-checking in an open-source format. However, some misinformation-reporting options have been recently removed, including the ability to flag posts as misleading regarding political or electoral topics. The moderation approach has become less transparent since ownership changes, and certain tools previously available for reporting and countering misinformation may no longer be supported (Hénin, & Giovanna Sessa, 2024 February).

X has implemented several specific policies targeting types of content that pose a public risk. The Crisis Misinformation Policy, introduced during the COVID-19 pandemic, targets misinformation in contexts of armed conflict, natural disasters, and other emergencies where public safety is a concern. Additionally,

the Synthetic and Manipulated Media Policy addresses the risks posed by altered or out-of-context media that could deceive users. X's Civic Integrity Policy works to prevent the platform's use for manipulating elections or civic events, though recent changes limit users' ability to report voting-related misinformation (Hénin, & Giovanna Sessa, 2024 February).

In compliance with the DSA, X is considered a VLOP and must meet requirements for transparency and reporting. It publishes summaries of moderation and enforcement activities, outlining actions taken to mitigate harmful content and misinformation. The moderation approach includes human-led investigations and scaled reviews alongside automated systems powered by machine learning and heuristics, particularly for identifying patterns of manipulation, deceptive identities, and synthetic media. Despite these efforts, X's evolving moderation policies and tools have created a complex environment, with varying levels of transparency and enforcement for content management (Hénin, & Giovanna Sessa, 2024 February).

In compliance with the Digital Services Act (DSA), X published its transparency report for the first half of 2024, breaking a two-year silence since Elon Musk's acquisition in October 2022. The report reveals 224,129,805 content reports from January to June 2024, leading to the labeling or suspension of 10,675,980 posts and the suspension of 5,296,870 accounts. Despite a staggering 1,830% rise in reports compared to late 2021, account suspensions increased by only 300%. Troublingly, of over 8.9 million reports related to child safety, only 14,571 posts were removed (X, 2024).

Forbes highlighted concerns over the platform's revised misinformation framework, which was significantly reduced from 50 pages to 15 after Elon Musk's takeover. This reduction may have influenced moderation decisions, as the disparity between reports and enforcement actions raises questions about the platform's commitment to effective content moderation (Sircar, 2024, October 18).

4.4 PLATFORM POLICIES ON GENERATIVE-AI AND MISINFORMATION

This last section on content moderation examines how all the major platforms address AI-manipulated or AI-generated content in their terms of use and explores their approaches to mitigating the potential risk of misinformation. Recent technical advancements and the growing use of generative AI systems by end-users have exponentially increased the challenges posed by AI-manipulated and AI-generated misinformation. These developments have raised crucial questions about the ability of platforms to distinguish legitimate uses from malign uses of such content and whether they consider AI-related risks as accessories to disinformation strategies or matters that require specific policies. As the Digital Services Act (DSA) will provide new complaint mechanisms for users and require platforms to assess their mitigation measures against systemic risks, understanding how platforms approach AI-manipulated and AI-generated content is essential (Miguel, 2024 June).

A comparative analysis of Facebook, Instagram, TikTok, X (formerly Twitter), and YouTube reveals varying definitions and actions related to AI-manipulated and generated misinformation. While some platforms explicitly mention AI, others focus on synthetic media or digitally created content. The rationale behind content moderation ranges from the misleading and harmful potential to compliance-oriented considerations regarding copyright and quality standards (Miguel, 2024 June).

Definitions of AI-manipulated or generated content vary across platforms, with some providing explicit descriptions and others taking a more general approach. Facebook and TikTok are the only platforms that directly mention AI in their policies aiming to tackle disinformation. TikTok and X include "synthetic media" in their policies about manipulated and misleading media, while YouTube announced new measures targeting "synthetic content" and explicitly referencing AI. Meta addressed "digitally created or altered content" in the context of political ads (Miguel, 2024 June).

Platforms actively combat AI-manipulated and -generated misinformation through targeted measures. These include labeling such content, enforcing user accountability for its identification and removal, downranking it, demonetizing, implementing strike policies, removing it, prohibiting distribution, and setting advertising and monetization standards (Miguel, 2024 June).

Several key observations can be brought to the fore. Firstly, there's a notable lack of transparency surrounding platform policies, particularly regarding collaboration with experts and the delineation between banned and removed content. Additionally, navigating platform policies can be challenging, with inconsistencies in scope and publication dates. On a positive note, Facebook and Instagram have aligned their content moderation policies, and there's a growing collaboration with fact-checkers. However, there is a limited focus on AI-manipulated or generated content, with challenges in detection and moderation. Platforms often base moderation on subjective premises, risking exploitation and evasion. While platforms have updated policies to address AI challenges, there's variation in depth and focus, with a growing emphasis on labeling AI-generated content. Legislation like the AI Act may introduce new rules, but platforms need to take proactive steps to address emerging challenges posed by AI technologies (Miguel, 2024 June).

The different experts recommend that platforms continue their efforts to respond to the challenges posed by AI-generated disinformation with effective policy changes. Enhancing cooperation with external collaborators and experts in AI, as well as encouraging the creation of information AI internal resources, can help combat the spread of misinformation. Developing a framework for risk assessment specifically tailored to AI-generated content would also provide guidance and prevent arbitrariness in the assessment process. Lastly, platforms must address the new challenges posed by AI-generated content in regulating end-users' roles on their platforms (Miguel, 2024 June).

4.5 SPECIFIC CONCERNS ABOUT TELEGRAM AND TIKTOK

Social media platforms have adopted multi-layered strategies to counter fake news by employing a combination of advanced technology, human moderation, partnerships, and user participation. Yet, some platforms remain slow to act, with Telegram posing a particular challenge. According to a 2023 survey by Internews on media consumption in Ukraine, 72% of Ukrainians use Telegram for news. However, this platform's anonymous channels make it a hub for disinformation networks that target Ukrainian audiences. The platform is frequently used to share footage of the war, as it has been used as a tool to maintain morale in the country and as a strategic tool in the war. Even the government in Ukraine has created its own Telegram channels to provide fast and direct communication with the population about the current situation in the war. Studies show that after the invasion began, 63.3% of Ukrainians started using Telegram channels for news, up from just 35.9% before the full-scale invasion (CMPF, 2024, January 10)

TikTok, founded by ByteDance in 2016, has experienced rapid growth and gained immense popularity, especially among young users. It has become a global platform for sharing short-form videos, with an estimated 3.5 billion downloads and 1.7 billion users in 160 countries as of 2022. The app is particularly popular in the US, Indonesia, and Brazil, but also has a EU market with 150 million users, demonstrating its popularity all over the world (Romero Vicente, 2024 February).

TikTok employs a variety of methods to moderate content and combat disinformation, though these efforts face limitations and criticism. The platform provides users with a search function, Discover, which enables searches based on keywords for users, videos, sounds, LIVEs, and hashtags. However, search results are influenced by user preferences and interactions, leading to variations between individuals. Despite this, TikTok's ability to filter search results remains limited. Furthermore, the platform has restricted certain tools, such as the Creative Center, which was once used to analyze politically sensitive content, including topics like the Israel-Hamas conflict. This has raised concerns about transparency, as certain data and hashtags related to geopolitical issues are no longer accessible (Romero Vicente, 2024 February).

To support research, TikTok offers an API to academic institutions in Europe, but access is limited and does not extend to civil society organizations. The API provides data on user profiles, comments, captions, and content performance but is constrained by usage limits and criticized for inadequacy. Civil society researchers often resort to alternative methods, such as data donation or unofficial, reverse-engineered APIs, though these are less stable and not endorsed by TikTok. Another method, data scraping, is technically unauthorized and raises ethical and legal concerns, as TikTok has implemented anti-scraping measures to protect user data (Romero Vicente, 2024 February).

TikTok's Commercial Content Library was established to comply with transparency rules under the Digital Services Act (DSA). It provides a repository of paid ads and promotional content, though data export for analysis is restricted. The library allows searches by country and ad category but offers limited utility for comprehensive disinformation research (Romero Vicente, 2024 February).

To address disinformation, TikTok's policies and guidelines outline prohibited content, including misinformation that poses public safety risks, promotes conspiracy theories, or misleads about crises, health, or climate change including any topic that would be contravening the interests of the RPC such as claims on human rights related to China with a real and efficient enforcement.

The platform also bans synthetic or manipulated media unless clearly disclosed, and it discourages practices that artificially boost engagement or deceive users. AI-generated content must be labeled. TikTok enforces these policies through automated systems and human moderators, supported by tools such as its so-called Transparency Center, which provides information on fact-checking efforts, labeling state-affiliated media, and combating influence operations (Romero Vicente, 2024 February).

Users can report content violating TikTok's Community Guidelines by accessing a report option within the app. For illegal content, TikTok follows specific procedures under the DSA, requiring detailed explanations and evidence of legal violations. Reported content is first reviewed against platform policies, and if it aligns with these policies but violates local laws, a specialist moderation team evaluates further action, including restricting access in specific countries (Romero Vicente, 2024 February).

TikTok's reporting initiatives include biannual reports detailing its strategies to combat disinformation in EU/EEA countries. While the platform shows a commitment to addressing false information, its limited access to critical tools and data has drawn criticism, underscoring the challenges of ensuring accountability and facilitating independent evaluation. TikTok has acknowledged the need to strengthen its efforts, reporting significant actions taken to counter influence operations. In the first four months of 2024, TikTok disrupted 15 influence operations and removed 3,001 associated accounts. The majority of these networks aimed to sway political discourse, including election-related content. Notably, one operation targeted Indonesian users ahead of the country's presidential election (TikTok, May 23, 2024).

In its third DSA transparency report for the EU, TikTok highlights its most recent efforts to tackle harmful content across its European market. The company employs over 6,000 moderators to oversee the platform's 150 million users in the region. Additionally, TikTok's automated content moderation systems removed 80% of problematic videos, up from 62% in 2023, reflecting the platform's strengthened efforts to moderate harmful content in the EU (TikTok, October 24, 2024).

Despite these efforts, TikTok continues to face criticism over weak data protection, leading to the app being blocked from several government Wi-Fi networks and sparking ongoing debates about a potential ban in the US. The app is already banned in India. These data security concerns stem from the platform's alleged links to the Chinese government (BBC News, April 12, 2024). While there are signs of increased efforts to combat harmful content, concerns persist about a lack of transparency regarding data protection and usage on the platform, influence and censorship, which could pose risks to users.

Sources:

BBC News (2024, April 12). Government should counter misinformation on TikTok - MPs. *BBC News*. Retrieved From: https://www.bbc.com/news/articles/cj5l4e4v350o

Bayer, J. (2024). Digital Media Regulation within the European Union (1st ed.). *Nomos*, Baden Germany. https://doi.org/10.5771/9783748945352

Culloty, E., Park, K., Feenane, T., Papaevangelou, C., Conroy, A., & Suiter, J. (2021). Covidcheck: assessing the implementation of EU code of practice on disinformation in relation to Covid-19. ULR: https://fujomedia.eu/wp-content/uploads/2021/09/Code2021_COVIDCheck.pdf

European Commission. (2018, September 26). Code of practice on disinformation. Digital Strategy. https://digital-strategy.ec.europa.eu/en/news/code-practice-disinformation

European Commission, 2018a, A Multi-dimensional Approach to Disinformation: Report of the Independent High-Level Group on Fake News and Online Disinformation. Directorate-General for Communication Networks, Content and Technology. Available at https:// ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformatio

European Commission, 2018b, Code of Practice on Disinformation. Available at https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation.

European Commission, 2018c, Synopsis Report of the Public Consultation on Fake News and Online Disinformation. https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation

European Commission (2018d) Tackling Online Disinformation: A European Approach (Communication) COM(2018) 236 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236

European Commission (2020a) Assessment of the Code of Practice on Disinformation —Achievements and areas for further improvement. Commission Staff working document (SWD(2020) 180 final).

European Commission (2020b) European Democracy Action Plan (Communication) COM(2020) 790 final.https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

European Commission: Directorate-General for Communications Networks, Content and Technology. (2018). *A multi-dimensional approach to disinformation : report of the independent High level Group on fake news and online disinformation*. Publications Office. **https://data.europa.eu/doi/10.2759/739290**.

European Commission. (2019, February 26). Antitrust: Commission fines Google €1.49 billion for abusive practices in online advertising [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_19_746

European Commission (2021) Guidance on Strengthening the Code of Practice on Disinformation (COM(2021) 262 final). https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation

Giovanna Sessa, M. (2024, January). Facebook factsheet. Disinfo.eu. Retrieved November 19, 2024, from: https://www.disinfo.eu/wp-content/uploads/2024/01/20240116_Facebook_factsheet.pdf

Griffin, R. (2024). Codes of Conduct in the Digital Services Act: Functions, Benefits & Concerns. *Technology and Regulation*, *2024*, 167-187. https://doi.org/10.26116/techreg.2024.016%20%E2%80%A2%20ISSN:%202666-139X

Heldt, A. (2019). Reading between the lines and the numbers: An analysis of the first NetzDG reports. *Internet Policy Review*, *8*(2), 1-18. https://doi:10.14763/2019.2.1398

Hénin, N. & Giovanna Sessa,M. (2024, February). Twitter-X factsheet. Disinfo.eu. Retrieved November 19, 2024, from: https://www.disinfo.eu/wp-content/uploads/2024/01/20240116_Twitter-X_factsheet.pdf

Herrero O. (2023) Le Système Tiktok Comment la plateforme chinoise modèle nos vies', Editions du Rocher

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, 1-9. https://doi.org/10.1038/s41562-024-01884-x

Kuczerawy, A. (2019). Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?Forthcoming in: "Disinformation and Digital Media as a Challenge for Democracy" European Integration and Democracy Series, Vol. 6, eds. E. Kużelewska, G. Terzis, D. Trottier and D. Kloza, Intersentia, 2019, Available at SSRN: https://ssrn.com/abstract=3453732

Miguel, R. (2024, June). Platforms policies on AI-manipulated and generated misinformation. Disinfo.eu. Retrieved November 19, 2024 From: https://www.disinfo.eu/wp-content/uploads/2024/06/20240604_platformpolicies-on-ai-V3.pdf

Miguel Serrano, R. (2024, February). YouTube factsheet. Disinfo.eu. Retrieved November 19, 2024, from:https://www.disinfo.eu/wp-content/uploads/2024/02/20240228_YouTube_factsheet.pdf

Raso, T. I., Das, P. P., Aman, M., Ifaz, A. Y. A., Rema, S. S., Tabasum, F., Chakma, N., & Hossain, M. P. (2024, July 10). Misinformation on YouTube: High profits, low moderation. DISMIS Lab. Retrieved October 10, 2024, from https://en.dismislab.com/misinformation-on-youtube-high-profits-low-moderation/#The-need-to-demonetize-misinformation

Romero Vicente, A. (2024, February). TikTok factsheet. Disinfo.eu. Retrieved November 19, 2024, from: https://www.disinfo.eu/wp-content/uploads/2024/02/20240205_TikTok_factsheet.pdf

Sircar, A. (2024, October 18). X's latest content findings reveal troubling trends in AI moderation. Forbes. Retrieved from https://www.forbes.com/sites/anishasircar/2024/10/18/xs-latest-content-findings-reveal-troubling-trends-in-ai-moderation/

Smith, R. B., Perry, M., & Smith, N. N. : Fake News' in ASEAN: Legislative responses. Journal of ASEAN Studies, 9(2), 2021. 117-137. https://doi.org/10.21512/jas.v9i2.7506

Smuha, N. A.. "Beyond the individual: governing AI's societal harm". Internet Policy Review 10.3 2021 DOI: 10.14763/2021.3.1574

Wilman, F. (2022). The Digital Services Act (DSA)-An Overview. *Available at SSRN 4304586*.

TikTok. (2024, October 2). TikTok announces launch of its Disinformation, Bullying, and Trolling Action Center (DUBTAC). Retrieved from https://newsroom.tiktok.com/en-eu/dubtac

TikTok. (2024, May 23). Strengthening our approach to countering influence attempts. Retrieved from https://newsroom.tiktok.com/en-us/strengthening-our-approach-to-countering-influence-attempts

TikTok. (2024, October 24). Digital Services Act: Publishing our third transparency report on content moderation in Europe. Retrieved from https://newsroom.tiktok.com/en-eu/dsa-third-transparency-report

X. (2024). Global transparency report H1 2024. Retrieved from https://transparency.x.com/content/dam/transparency-twitter/2024/x-global-transparency-report-h1.pdf

Funded by the European Union

5 PART IV -PROS AND CONS OF EXISTING TOOLS TO COUNTER FAKE NEWS

There are many tools and resources that have been developed and continue to be developed to help in the fight against misinformation and disinformation. AI enabled tools for countering fake news have started to pop up in the last few years, offering varied approaches with distinct strengths and limitations in countering online fake news.

The Factual stands out for its transparency, providing users with detailed scores for news articles based on source credibility, author expertise, diversity of sources, and emotional tone. This transparency promotes media literacy by helping users understand why certain articles are deemed more authentic. Its focus on evaluating credible sources across the political spectrum also mitigates bias, unlike other algorithms that may reflect the developers' subjective opinions. However, The Factual does not directly fact-check content but rather scores on authenticity, leaving room for misinformation to appear within seemingly reliable articles. It also struggles with multimedia content and stylistic choices like missing author information or absent external links, which can unfairly lower scores on reliable information (The Factual, n.d.). Therefore, the tool requires active engagement and critical interpretation to the information that the tools provide to be effective, which may deter casual users.

In contrast, Full Fact AI excels in claim verification by comparing content against reliable sources, making it more adept at identifying factual errors missed by tools like The Factual. Its ability to categorize claims by topic and operate in multiple languages enhances its global applicability. Built on Google's BERT model, Full Fact AI demonstrates high accuracy in processing complex language. However, this comes at a high financial and environmental cost, as BERT's resource-intensive training demands significant energy and storage. The absence of a robust bias-mitigation strategy further weakens its credibility, and the 2018 BERT framework may be considered outdated compared to newer models (Full Fact, n.d.).

ClaimBuster offers a different approach by identifying "check-worthy" claims rather than attempting to verify their accuracy. This tool is particularly effective for high-volume platforms like X (formerly Twitter), where it filters content for human fact-checkers to analyze (ClaimBuster X, n.d.). Its reliance on human judgment ensures nuanced decision-making but also makes the process slower and less accessible. Additionally, low engagement and usability issues on its platform hinder its effectiveness, limiting its reach and impact (ClaimBuster, n.d.).

Meanwhile, Bot Sentinel shifts focus from fact-checking to identifying and mitigating harmful behavior on X. Its primary goal is to flag accounts engaged in trolling or harassment, allowing users to block problematic accounts or hide replies. This targeted functionality makes it unique, but its narrow scope restricts its usefulness for addressing fake news or misinformation more broadly. While helpful for specific cases of online harassment, its relevance as a general-purpose tool for combating disinformation is limited (Botsentinel, n.d.).

Focusing on the complexity of language, ADVerifi AI brings an innovative approach by distinguishing satire from disinformation. This tool leverages advanced linguistic analysis to safeguard free speech while detecting harmful content. Its methodologies, which examine humor-related features, are promising for improving accuracy in categorizing falsehoods. However, as the tool is still under development, its practical application remains limited, and additional refinement is needed to fully differentiate between absurdity and deliberate falsehoods (Adverifai, n.d.).

Some tools for identifying fake news focus on community participation rather than advanced AI technologies. For example, Melnsuzbalta.lv is a Latvian platform that invites citizens to report fake news, misleading content, hate speech, propaganda, or anything suspicious. Reports are submitted to the Strategic Communication team of the State Chancellery, making it a straightforward and cost-effective method to detect foreign information manipulations and interference (FIMI). While this approach promotes civic engagement and raises awareness, it depends heavily on high levels of participation to be effective (Melns uz Balta, n.d.). Unlike scalable AI-driven tools that process vast amounts of data efficiently, this system requires human review, making it slower and less effective for large-scale detection. It also only detects "fakes" with human judgment which can lead to bias or inaccuracies. However, it complements AI-based methods well by leveraging public participation and fostering community involvement in countering FIMI.

To strengthen citizens' ability to detect false information and manipulative influence (FIMI), the Latvian State Chancellery published a handbook in 2022 on countering disinformation. This resource offers practical recommendations for state and local government employees, as well as Latvian residents, to address manipulative narratives, including those propagated by the Kremlin (LSM, 2022 October 11). It provides counter-narratives to common forms of disinformation in Latvia. However, the handbook's primary audience is governmental institutions, limiting its accessibility for the general public. Its language and structure are not tailored for easy consumption by a wider audience (Buholcs et al., 2024, p. 19-20), reducing its effectiveness as a universal tool for combating disinformation at the citizen's level. While it empowers informed groups with targeted strategies, it risks excluding broader segments of the population who could benefit from simplified and widely distributed materials.

Remaining in the lane of educational tools, there have also been developed games to help people with media literacy, offering an engaging approach to countering fake news. Go Viral, developed by Cambridge psychologists in collaboration with the UK Government, immerses players in the role of a fake news creator, exposing the tactics and motivations behind misinformation, particularly related to COVID-19. Research suggests a single play of similar games can reduce susceptibility to false information for up to three months. While the game's interactive format makes it accessible and appealing, its focus on entertainment and easy access may limit its depth. It is most effective when played widely, yet reaching a diverse and substantial audience remains a challenge.

Together, these tools highlight both the potential and the challenges of using AI to combat fake news, and the need for non-AI solutions to counter disinformation. While each tool contributes uniquely to the fight against misinformation, none offers a comprehensive solution, emphasizing the need for ongoing innovation and integration of these technologies.

6 CONCLUSION

Disinformation and fake news are difficult to grasp in the current context. Technology like AI is increasingly used to develop algorithms that go beyond the possibility of critical thinking to act as a barrier. Many Fake news become viral, based on the emotional charge they carry, but also based on polarizing narratives and stereotypes. Disinformation creates confusion. Our online survey shows that there is a strong awareness and a strong concern of EU citizens on disinformation on social media. There is a relative distrust in online information trustworthiness and a growing request for more and better moderation. Regarding the question, "What impact do you believe fake news has on society?", there was a strong consensus among the respondents that it has a highly significant impact. In fact, 91% of all respondents indicated either a "significant impact" or "very significant impact" of fake news on society. This concern was shared across all countries, age groups, and genders. This highlights the broad recognition of fake news as a serious societal issue from a citizens' perspective. Faced with this, the social media platforms have entered in efforts to improve the content moderation, but which is often restricted to the prevention of violent images and content. Faced with this the efforts of EU regulation are not sufficiently implemented to counter disinformation effectively. But the experience of existing tools, such as the ones developed in Latvia could bring additional and insightful information on what are the best tools, possibilities and options that could be developed within the AI4DEBUNK project.

7 REFERENCES

Adverifai. (n.d.). Technology. Adverifai. Retrieved June 19, 2024, from https://adverifai.com/technology/

Art, S. (2018). Media literacy and critical thinking. *International Journal of Media and Information Literacy*, *3*(2), 66-71. URL: https://cyberleninka.ru/article/n/media-literacy-and-critical-thinking

Basol, M., Roozenbeek, J., & Van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition*, *3*(1). doi: 10.5334/joc.91. PMID: 31934684

Bayer, J. (2024). Digital Media Regulation within the European Union (1st ed.). *Nomos*, Baden Germany. https://doi.org/10.5771/9783748945352

Bostrom, A., Demuth, J. L., Wirz, C. D., Cains, M. G., Schumacher, A., Madlambayan, D., ... & Williams, J. K. (2024). Trust and trustworthy artificial intelligence: A research agenda for AI in the environmental sciences. *Risk Analysis*, *44*(6), 1498-1513.

Botsentinel. (n.d.). Bot Sentinel. Retrieved June 21, 2024, from: https://botsentinel.com/

Bulger, Monica, and Patrick Davison. (2018). The promises, challenges, and futures of media literacy. *Journal of Media Literacy Education* 10.1: 1-21. Retrieved from: https://doi.org/10.23860/JMLE-2018-10-1-1

Buholcs, J., Tetarenko-Supe, A., Torpan, S., Kõnno, A., Vorteil, V., Balčytienė, A., & Kasparaitė, R. (2024). BECID D3.4 report. European Digital Media Observatory (EDMO). Retrieved from https://edmo.eu/wp-content/uploads/2024/06/BECID-D3.4._report.pdf

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European management journal*, *40*(5), 685-706.

ClaimBuster. (n.d.). ClaimBuster - Automated Live Fact-Checking. Retrieved June 20, 2024, from https://idir.uta.edu/claimbuster/

Culloty, E., Park, K., Feenane, T., Papaevangelou, C., Conroy, A., & Suiter, J. (2021). Covidcheck: assessing the implementation of EU code of practice on disinformation in relation to Covid-19. ULR: https://fujomedia.eu/wp-content/uploads/2021/09/Code2021_COVIDCheck.pdf

European Commission. (2018, September 26). Code of practice on disinformation. Digital Strategy. https://digital-strategy.ec.europa.eu/en/news/code-practice-disinformation

European Commission, 2018a, A Multi-dimensional Approach to Disinformation: Report of the Independent High-Level Group on Fake News and Online Disinformation. Directorate-General for Communication Networks, Content and Technology. Available at https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

European Commission, 2018b, Code of Practice on Disinformation. Available at https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation

European Commission, 2018c, Synopsis Report of the Public Consultation on Fake News and Online Disinformation. https://ec.europa.eu/digital-single-market/en/news/synopsis-report-public-consultation-fake-news-and-online-disinformation

European Commission (2018d) Tackling Online Disinformation: A European Approach (Communication) COM(2018) 236 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236

European Commission (2020a) Assessment of the Code of Practice on Disinformation —Achievements and areas for further improvement. Commission Staff working document (SWD(2020) 180 final).

European Commission (2020b) European Democracy Action Plan (Communication) COM(2020) 790 final.https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=COM%3A2020%3A790%3AFIN&qid=1607079662423

European Commission: Directorate-General for Communications Networks, Content and Technology. (2018). *A multi-dimensional approach to disinformation : report of the independent High level Group on fake news and online disinformation*. Publications Office. https://data.europa.eu/doi/10.2759/739290

European Commission. (2019, February 26). Antitrust: Commission fines Google €1.49 billion for abusive practices in online advertising [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_19_746

European Commission (2021) Guidance on Strengthening the Code of Practice on Disinformation (COM(2021) 262 final). https://digital-strategy.ec.europa.eu/en/library/guidance-strengthening-code-practice-disinformation

European Digital Media Observatory. (2024). BECID D3.4 Report. Retrieved from https://edmo.eu/wp-content/uploads/2024/06/BECID-D3.4._report.pdf

European Digital Media Observatory. (2022). Policies to tackle disinformation in EU member states – Part II. https://edmo.eu/wp-content/uploads/2022/07/Policies-to-tackle-disinformation-in-EU-member-states-%E2%80%93-Part-II.pdf

Full Fact. (n.d.). About us: AI. Full Fact. Retrieved June 18, 2024, from https://fullfact.org/ai/about/

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). Trust in Artificial Intelligence: A Global Study. The University of Queensland and KPMG Australia. doi: 10.14264/00d3c94

Griffin, R. (2024). Codes of Conduct in the Digital Services Act: Functions, Benefits & Concerns. *Technology and Regulation*, *2024*, 167-187. https://doi.org/10.26116/techreg.2024.016%20%E2%80%A2%20ISSN:%202666-139X

Heldt, A. (2019). Reading between the lines and the numbers: An analysis of the first NetzDG reports. *Internet Policy Review*, *8*(2), 1-18. https://doi:10.14763/2019.2.1398

Handschuh, K. (2023). Critical Thinking and Media Literacy in an Age of Misinformation. Retrieved from: http://doi.org/10.33774/apsa-2023-l43bk

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, 1-9. https://doi.org/10.1038/s41562-024-01884-x

Kuczerawy, A. (2019). Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?Forthcoming in: "Disinformation and Digital Media as a Challenge for Democracy" European Integration and Democracy Series, Vol. 6, eds. E. Kużelewska, G. Terzis, D. Trottier and D. Kloza, Intersentia, 2019, Available at SSRN: https://ssrn.com/abstract=3453732

LSM. (2022, October 19). Public Broadcasting of Latvia. Retrieved November 19, 2024, from https://eng.lsm.lv/article/features/media-literacy/latvias-state-chancery-issues-guidebook-against-disinformation.a478655/

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 2053951718756684.

Melns uz Balta. (n.d.). Retrieved November 19, 2024, from https://melnsuzbalta.lv/

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1-10. https://doi.org/10.1057/s41599-019-0279-9

Squire, K., & Steinkuehler, C. (2011). Video games and learning. *Teaching and participatory culture in the digital age*, *59*(1), 129-132.

Smith, R. B., Perry, M., & Smith, N. N. : Fake News' in ASEAN: Legislative responses. Journal of ASEAN Studies, 9(2), 2021. 117-137. https:/doi.org/10.21512/jas.v9i2.7506

Smuha, N. A.. "Beyond the individual: governing AI's societal harm". Internet Policy Review 10.3 2021 DOI: 10.14763/2021.3.1574

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. In Higher Education Studies (Vol. 4, Number 1, pp. 1–17). Canadian Center of Science and Education. https://doi.org/10.5539/hes.v4n1p1

The Factual. (n.d.). How it works. The Factual. Retrieved June 20, 2024, from https://www.thefactual.com/how-it-works/

The Factual X. (n.d.). TheFactualNews. Retrieved 24.06.2024, from https://x.com/thefactualnews?lang=en

Ventsel, A., Hansson, S., Rickberg, M., & Madisson, M. L. (2023). Building Resilience Against Hostile InformatioYee, Amy (2022, January 31). Subjected to repeated disinformation campaigns, the tiny Baltic countryn Influence Activities: How a New Media Literacy Learning Platform Was Developed for the Estonian Defense Forces. *Armed Forces & Society*, 0095327X231163265.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146-1151. DOI:10.1126/science.aap9559

Wilman, F. (2022). The Digital Services Act (DSA)-An Overview. *Available at SSRN 4304586*.

Yee, Amy (2022, January 31). Subjected to repeated disinformation campaigns, the tiny Baltic country of Estonia sees media literacy education as part of its digital-first culture and national security. BBC. Retrieved from https://www.bbc.com/future/article/20220128-thecountry-inoculating-against-disinformation

Zhang, L., Zhang, H., & Wang, K. (2020). Media literacy education and curriculum integration: A literature review. *International Journal of Contemporary Education*, *3*(1), 55-64. https://doi.org/10.11114/ijce.v3i1.4769

8 ANNEX I. SUMMARY OF AVAILABLE REPORTS ON THE TOPIC OF DISINFORMATION

These reports are available on the website of Disinfolab: www.disinfo.eu

8.1 SUMMARY PLATFORM POLICIES ON GENERATIVE-AI AND MISINFORMATION BY RAQUEL MIGUEL

Full report: https://www.disinfo.eu/publications/platforms-policies-on-ai-manipulated-and-generated-misinformation/

Introduction

This summary examines how major platforms address AI-manipulated or AI-generated content in their terms of use and explores their approaches to mitigating the potential risk of misinformation. Recent technical advancements and the growing use of generative AI systems by end-users have exponentially increased the challenges posed by AI-manipulated and AI-generated misinformation. These developments have raised crucial questions about the ability of platforms to distinguish legitimate uses from malign uses of such content and whether they consider AI-related risks as accessories to disinformation strategies or matters that require specific policies. As the Digital Services Act (DSA) will provide new complaint mechanisms for users and require platforms to assess their mitigation measures against systemic risks, understanding how platforms approach AI-manipulated and AI-generated content is essential.

Cross-Platform Comparison

A comparative analysis of Facebook, Instagram, TikTok, X (formerly Twitter), and YouTube reveals varying definitions and actions related to AI-manipulated and generated misinformation. While some platforms explicitly mention AI, others focus on synthetic media or digitally created content. The rationale behind content moderation ranges from the misleading and harmful potential to compliance-oriented considerations regarding copyright and quality standards.

Definitions and Actors

Definitions of AI-manipulated or generated content vary across platforms, with some providing explicit descriptions and others taking a more general approach. Facebook and TikTok are the only platforms that directly mention AI in their policies aiming to tackle disinformation. TikTok and X include "synthetic media" in their policies about manipulated and misleading media, while YouTube announced new measures targeting "synthetic content" and explicitly referencing AI. Meta addressed "digitally created or altered content" in the context of political ads.

Types of Actions

Platforms actively combat AI-manipulated and generated misinformation through targeted measures. These include labeling such content, enforcing user accountability for its identification and removal, downranking it, demonetizing, implementing strike policies, removing it, prohibiting distribution, and setting advertising and monetization standards.

Concluding Remarks

Compiling this factsheet has brought forth several key observations. Firstly, there's a notable lack of transparency surrounding platform policies, particularly regarding collaboration with experts and the delineation between banned and removed content. Additionally, navigating platform policies can be challenging, with inconsistencies in scope and publication dates. On a positive note, Facebook and Instagram have aligned their content moderation policies, and there's a growing collaboration with fact-checkers. However, there's a limited focus on AI-manipulated or generated content, with challenges in detection and moderation. Platforms often base moderation on subjective premises, risking exploitation and evasion. While platforms have updated policies to address AI challenges, there's variation in depth and focus, with a growing emphasis on labeling AI-generated content. Legislation like the AI Act may introduce new rules, but platforms need to take proactive steps to address emerging challenges posed by AI technologies.

Recommendations

Platforms should continue their efforts to respond to the challenges posed by AI-generated disinformation with effective policy changes. Enhancing cooperation with external collaborators and experts in AI, as well as encouraging the creation of information AI internal resources, can help combat the spread of misinformation. Developing a framework for risk assessment specifically tailored to AI-generated content would also provide guidance and prevent arbitrariness in the assessment process. Lastly, platforms must address the new challenges posed by AI-generated content in regulating end-users' roles on their platforms.

8.2 DISINFORMATION ON FACEBOOK: RESEARCH AND CONTENT MODERATION POLICIES BY MARIA GIOVANNA SESSA

Full report: https://www.disinfo.eu/publications/disinformation-on-facebook/

Introduction

This report provides an analysis of Facebook's operations, focusing on its potential for misuse in disinformation campaigns. It also offers guidance on investigating the platform, reporting content, and understanding relevant policies for content enforcement. Finally, it includes a repository of studies on Facebook's role in disinformation campaigns. Facebook is one of the world's largest social media platforms, with approximately 2.9 billion monthly active users (MAUs) globally and 259 million MAUs in the European Union. As a VLOP, Facebook is subject to additional transparency reporting requirements under the European Union's Digital Services Act (DSA). Facebook's structure comprises profiles, groups, and pages, which serve as the foundation for information sharing and interaction. Users can create profiles, join groups, and follow pages based on their interests. Key features of Facebook include newsfeeds with personalized feeds displaying friends' posts, updates from pages and groups, and sponsored content. Facebook also has Timelines, with a user's record of all shared posts and

interactions, including tagged content. Application interactions include reactions, comments, shares, asking for recommendations, checking-in at locations, and more.

Investigations on the Platform

The Facebook API enables developers and app users to access a wide range of data from the platform, including user profiles, posts, pages, and events, contingent on user permissions and adherence to Facebook's Data Usage Policies. CrowdTangle, integrated into Facebook, provides social media analytics, allowing insights into content performance and audience demographics, though its dismantling aligns with evolving data access requirements. Data scraping, while useful for research, raises ethical and legal concerns regarding privacy and compliance. Meta's research tools, like the Content Library API and Ad Library, offer access to public content and ad transparency data, while the Transparency Centre provides resources for academic researchers, complemented by quarterly Adversarial Threat Reports for cybersecurity insights targeting Facebook and Instagram.

Reporting Content and Enforcement

Users can report content that violates Facebook's Community Standards by clicking the three dots menu on a post, photo, or comment and selecting the appropriate reporting option. Content that goes against the platform's standards is removed using automated technology or a review team. Facebook's strike system applies penalties to offending accounts based on repeated violations.

Facebook's Policies Against Disinformation

Facebook's Community Standards outline guidelines for acceptable content, including restrictions on violence, criminal behavior, safety, objectionable content, integrity, authenticity, and intellectual property. The platform has specific policies addressing misinformation, such as false information, harmful health misinformation, voter or census interference, and manipulated media.

Tools

Facebook provides users with essential research tools to gain insights into its platform and content. The Meta Content Library offers real-time access to public content from Facebook, including posts, pages, groups, and events. Meanwhile, the Meta Ad Library serves as a searchable database for ad transparency across Meta technologies. Additionally, academic researchers have access to other research and dataset resources, offering deeper insights into the platform. Meta's Adversarial Threat Reports

At the end, the report lists several examples of disinformation campaigns, such as disinformation targeting voters of color during the 2020 U.S. elections and disinformation during the Yellow Vests' protests in France, illustrating how the platform could be used for spreading disinformation.

8.3 DISINFORMATION ON YOUTUBE: RESEARCH AND CONTENT MODERATION POLICIES BY RAQUEL MIGUEL SERRANO

Full report: https://www.disinfo.eu/publications/disinformation-on-youtube/

Introduction

YouTube, founded in 2005 and acquired by Google in 2006, is one of the largest online video-sharing platforms. It has significantly transformed the way people engage with video content, including information, and has enabled new forms of monetization, giving rise to professional content creators known as “youtubers”. YouTube's visual nature and language barriers can introduce challenges when researching disinformation. However, YouTube offers various elements beyond videos that can be explored for research. Some essential characteristics to consider when researching on YouTube include complex search challenges, personalized search results, and distinctive video IDs. YouTube's recommendation algorithms are also a crucial aspect of the platform and can be investigated to understand the spread of disinformation.

Investigating ads on YouTube involves utilizing two libraries: YouTube’s Ad Library, where commercial ads can be searched by category or brand, and Google Ads Transparency Center, which allows searching for ads from verified advertisers across various Google platforms. Web searches can help locate deleted content using Google's cache, while custom search engines and cross-platform searches aid in deeper analysis beyond YouTube's built-in search capabilities. Tools like Youtube-dl facilitate video downloads, and resources exist for geolocation searches, comment searches, and conducting Open-Source Intelligence (OSINT) investigations on YouTube. Academic researchers can access a scaled dataset of global video metadata through YouTube's Data API by requesting approval. YouTube's moderation actions include content removal, downranking, channel termination, and demonetization for policy violations. YouTube is also subject to regulatory frameworks like the Digital Services Act (DSA) and the Strengthened Code of Practice on Disinformation, which impose transparency reporting requirements and incentivize efforts to combat disinformation.

Flagging Content on YouTube and Its Enforcement

Users can report content that violates YouTube's Community Guidelines, which are categorized into six main areas: spam & deceptive practices, sensitive content, violent or dangerous content, regulated goods, educational, documentary, scientific, and artistic (EDSA) content, and misinformation. YouTube bans certain types of misleading or deceptive content with a serious risk of egregious harm, including content that can cause real-world harm, certain types of technically manipulated content, or content interfering with democratic processes. Specific policies for medical misinformation and elections misinformation are also in place. Responding to new misinformation-challenges, YouTube chose to provide funding to Poynter’s International Fact-Checking Network and unveiled a long-term vision for medical misinformation policies, which serves as guidelines for content moderation on the platform.

Relevant Cases on How YouTube is Used in Disinformation Campaigns

At the end, the report provides several examples of disinformation campaigns on YouTube such as a sophisticated YouTube political troll campaign with 8M+ views discovered by Plasticity.AI one year out from the 2020 U.S. presidential election. These examples illustrate how YouTube has been used as a platform to spread disinformation in the past.

Conclusion

YouTube, due to its popularity and reach, is susceptible to disinformation campaigns. This document provides a comprehensive overview of the platform's operations and offers guidance on conducting research on YouTube to counter disinformation. The document also highlights relevant cases of disinformation campaigns on the platform, emphasizing the need for continuous monitoring and action to mitigate the spread of false information.

8.4 DISINFORMATION ON TIKTOK: RESEARCH AND CONTENT MODERATION POLICIES BY ANA ROMERO VICENTE

Full report: https://www.disinfo.eu/publications/disinformation-on-tiktok/

Introduction

This report looks at TikTok as a platform for spreading disinformation. TikTok, founded by ByteDance in 2016, has experienced rapid growth and gained immense popularity, especially among young users. It has become a global platform for sharing short-form videos, with an estimated 3.5 billion downloads and 1.7 billion users in 160 countries as of 2022. The app is particularly popular in the US, Indonesia, and Brazil.

2. Platform Organization

TikTok's structure revolves around four key pillars: user accounts, content creation, content itinerary, interactions, and engagement. Users create profiles, follow others, and can switch to a TikTok Pro account for analytics. Content primarily consists of short-form videos, with editing tools, special effects, and a music library available. Trends, challenges, and collaborations drive content creation. The platform encourages interactions through likes, comments, shares, and direct messages, as well as real-time engagement through TikTok Now and TikTok Live.

Transparency and Reporting Requirements

As a VLOP, TikTok is subject to additional transparency reporting requirements under the European Union’s Digital Services Act (DSA). The platform publishes a transparency report every six months, revealing that it had an average of 134 million monthly active users in the EU between February and July 2023.

Disinformation and Misinformation on TikTok

Given its massive user base and potential for growth, TikTok plays a significant role in the spread of information, including disinformation. The platform has faced increasing limits and bans globally, including being prohibited in India and blocked on devices from the EU's main institutions.

TikTok Ad Library

TikTok's Ad Library is a publicly accessible repository of all active and previously active ads on the platform, promoting transparency and accountability in advertising. Researchers use this tool to detect political advertising and false claims.

Monetization Features

TikTok offers various monetization features beyond advertising, such as hosting live broadcasts, creating premium content series, and participating in the Creativity Program Beta. The platform also has a TikTok Creator Fund to support users in creating outstanding content.

Combating Disinformation on TikTok

TikTok has implemented policies to address disinformation, including prohibiting misleading content that may cause significant harm. The platform does not allow medical misinformation, climate change misinformation, or dangerous conspiracy theories. TikTok encourages creators to label AI-generated content and has a fact-checking program to counter influence operations and label state-affiliated media entities. The platform also publishes transparency reports every six months to provide data on its efforts to combat online misinformation.

Relevant Cases of Disinformation Campaigns on TikTok

The report lists examples to illustrate the extent of TikTok's use in disinformation campaigns, such as pro-China disinformation ahead of Taiwan's presidential election, a Russian propaganda campaign involving thousands of fake accounts spreading disinformation about the war in Ukraine, and climate change-denial videos on the platform. Researchers have also found that nearly 1 in 5 of the videos automatically suggested by TikTok contained misinformation on various topics.

In conclusion, TikTok has become a powerful platform for sharing information and engaging users, particularly younger audiences. However, its popularity also makes it a target for spreading disinformation. As a result, TikTok has implemented policies and tools to combat misinformation and promote transparency, but challenges remain in enforcing these policies and ensuring the accuracy of content on the platform.

8.5 DISINFORMATION ON X: RESEARCH AND CONTENT MODERATION POLICIES BY NICOLAR HÉNIN & MARIA GIOVANNA SESSA

Full report: https://www.disinfo.eu/wp-content/uploads/2024/01/20240116_Twitter-X_factsheet.pdf

X is a social media platform centered around users posting, sharing, and interacting through brief messages, which were once known as “tweets” but are now simply called posts. Users create accounts with a unique identifier called a username or handle, and they can follow others to view posts on a customized feed. X’s main functions encourage engagement and community through a wide array of features, including hashtags, direct messages, and live broadcasting options like video and audio spaces. Privacy and notification settings are customizable, allowing users control over who can see their posts and which notifications they receive. Users can save or organize content through bookmarks, custom lists, and the Explore tab, which suggests trending content based on user activity.

Content moderation on X is a priority, especially regarding harmful or misleading content. The platform has created a multi-faceted approach to maintaining a safe space for users. This includes adherence to strict platform rules that cover categories such as spam, abusive behavior, violent threats, and impersonation, among others. X's Help Center allows users to report specific violations, from unauthorized use of copyrighted material to child safety issues. Recently, the platform introduced new reporting options to align with the EU's Digital Services Act (DSA), providing a way to report illegal content within the EU and appeal decisions. X emphasizes that these issues are handled through a combination of human moderation, automated technology, and partnerships with external experts, although it no longer has formal partnerships with European fact-checking organizations.

A key component of X's misinformation moderation strategy is "Community Notes," formerly known as Birdwatch. This crowdsourced tool allows users to add contextual notes to posts, promoting community-driven fact-checking in an open-source format. However, some misinformation-reporting options have been recently removed, including the ability to flag posts as misleading regarding political or electoral topics. The moderation approach has become less transparent since ownership changes, and certain tools previously available for reporting and countering misinformation may no longer be supported.

X has implemented several specific policies targeting types of content that pose a public risk. The Crisis Misinformation Policy, introduced during the COVID-19 pandemic, targets misinformation in contexts of armed conflict, natural disasters, and other emergencies where public safety is a concern. Additionally, the Synthetic and Manipulated Media Policy addresses the risks posed by altered or out-of-context media that could deceive users. X's Civic Integrity Policy works to prevent the platform's use for manipulating elections or civic events, though recent changes limit users' ability to report voting-related misinformation.

In compliance with the DSA, X is considered a VLOP and must meet requirements for transparency and reporting. It publishes summaries of moderation and enforcement activities, outlining actions taken to mitigate harmful content and misinformation. The moderation approach includes human-led investigations and scaled reviews alongside automated systems powered by machine learning and heuristics, particularly for identifying patterns of manipulation, deceptive identities, and synthetic media. Despite these efforts, X's evolving moderation policies and tools have created a complex environment, with varying levels of transparency and enforcement for content management.

This section has helped us to develop more understanding on the content moderation by the social media platforms, based on the social impacts that disinformation can have. The following section presents the pros and cons of existing tools to counter fake news.

Review Sheet of Deliverable/ Milestone Report
D12.1 Possible impacts of the tool on the perceptions of the citizens and the social media users

Editor(s):	Pascaline Gaborit, Joen Martinsen
Responsible Partner:	Pilot4dev
Status-Version:	Draft - v0.3
Date:	18/12/2024
Distribution level (CO, PU):	PU
Reviewer (Name/Organization)	Álvaro Parafita (BSC)
Review date	20/12/2024

Disclaimer: This assessment reflects only the author's views and the European Commission is not responsible for any use that may be made of the information contained therein"

Funded by the European Union

Mark with X the corresponding column:

Y= yes	N= no	N = not applicable

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
FORMAT: Does the document … ?				
…include editors, deliverable name, version number, dissemination level, date, and status?	X			
…contain a license (in case of public deliverables)?	X			
…include the names of contributors and reviewers?	X			
….has a version table consistent with the document's revision?	X			
… contain an updated table of contents?	X			
… contain a list of figures consistent with the document's content?	X			
… contain a list of tables consistent with the document's content?	X			
… contain a list of terms and abbreviations?	X			
… contain an Executive Summary?	X			
… contain a Conclusions section?	X			
… contain a List of References (Bibliography) in the adequate format, if relevant?	X			The list of References is included after every section, facilitating their location while reading the appropriate section, but also interrupting the flow of the overall document. I would suggest compiling them into a single References list at the end of the document, but this is a matter of personal preference.
… use the fonts and sections defined in the official template?	X			As a minor inconsistency, Section 4 onwards uses Calibri 11 for the text, while previous sections used Calibri 12.
… use correct spelling and grammar?	X			Minor typos and grammatical errors were corrected in v0.2 by the reviewer.
… conform to length guidelines (50 pages maximum (plus Executive Summary and annexes)	X			
… conform to guidelines regarding Annexes (inclusion of complementary information)	X			
… present consistency along the whole document in terms of English quality/style? (to avoid accidental usage of copy&paste text)	X			

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
About the content…				
Is the deliverable content correctly written?	X			
Is the overall style of the deliverable correctly organized and presented in a logical order?	X			
Is the Executive Summary self-contained, following the guidelines and does it include the main conclusions of the document?	X			
Is the body of the deliverable (technique, methodology results, discussion) well enough explained?	X			
Are the contents of the document treated with the required depth?	X			
Does the document need additional sections to be considered complete?		X		
Are there any sections in the document that should be removed?		X		Annex I is partially redundant with section 4, but I would not ask to remove it in order to properly reference the original sources and provide greater context. At most, I would include a short paragraph at the beginning of the Annex highlighting this fact to make it easier for readers.
Are all references in the document included in the references list?	X			
Have you noticed any text in the document not well referenced? (copy and paste of text/picture without including the reference in the reference list)		X		
SOCIAL and TECHNICAL RESEARCH WPs (WP4, 5, 12, 13, 14)				
Is the deliverable sufficiently innovative?	X			
Does the document present technical soundness and its methods are correctly explained?	X			
What do you think is the strongest aspect of the deliverable?				The survey and its analysis in section 3 provide insights into the citizens perception of disinformation. These insights are facilitated by a concise section highlighting the results while also forewarning of any potential biases resulting from the chosen survey design, avoiding misconceptions.

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
What do you think is the weakest aspect of the deliverable?				The survey used convenience sampling, resulting in 329 respondents with a clear bias towards younger citizens. This results in a limited perspective on older citizens perception on disinformation, but this is acknowledged in the Limitations section (2.6) and understandable given restrictions.
Please perform a brief evaluation and/or validation of the results, if applicable.				The deliverable presents the design and results of the online survey appropriately, discussing its merits and evaluating its shortcomings. The discussion in section 4 about the mechanisms of moderation in Online Platforms is valuable and presents a wide analysis of the shortcomings of each of their strategies. Section 5 provides an overview of existing technologies for debunking, which will prove useful in the design of AI4Debunk proposal.
AI AND TECNOLOGICAL WPS (WP6 – WP11)				
Does the document present technical soundness and the methods are correctly explained?				
What do you think is the strongest aspect of the deliverable?				
What do you think is the weakest aspect of the deliverable?				
Please perform a brief evaluation and/or validation of the results, if applicable.				
DISSEMINATION AND EXPLOITATION WPs (WP15 – WP17)				
Does the document present a consistent outreach and exploitation strategy?				
Are the methods and means correctly explained?				
What do you think is the strongest aspect of the deliverable?				
What do you think is the weakest aspect of the deliverable?				
Please perform a brief evaluation and/or validation of the results, if applicable.				

SUGGESTED IMPROVEMENTS

PAGE	SECTION	SUGGESTED IMPROVEMENT
		Version v0.2 contains some comments by the reviewer, as well as highlighted text in yellow that needs to be updated. This should be revised before submission. However, these are minor concerns which do not hinder the quality of the document or its contents.

CONCLUSION

Mark with X the corresponding line.

X	Document accepted, no changes required.
	Document accepted, changes required.
	Document not accepted, it must be reviewed after changes are implemented.

Please rank this document globally on a scale of 1-5 (1 = poor, 5= excellent) – using a half point scale. Mark with X the corresponding grade.

Document grade	1	1.5	2	2.5	3	3.5	4	4.5	5
									X