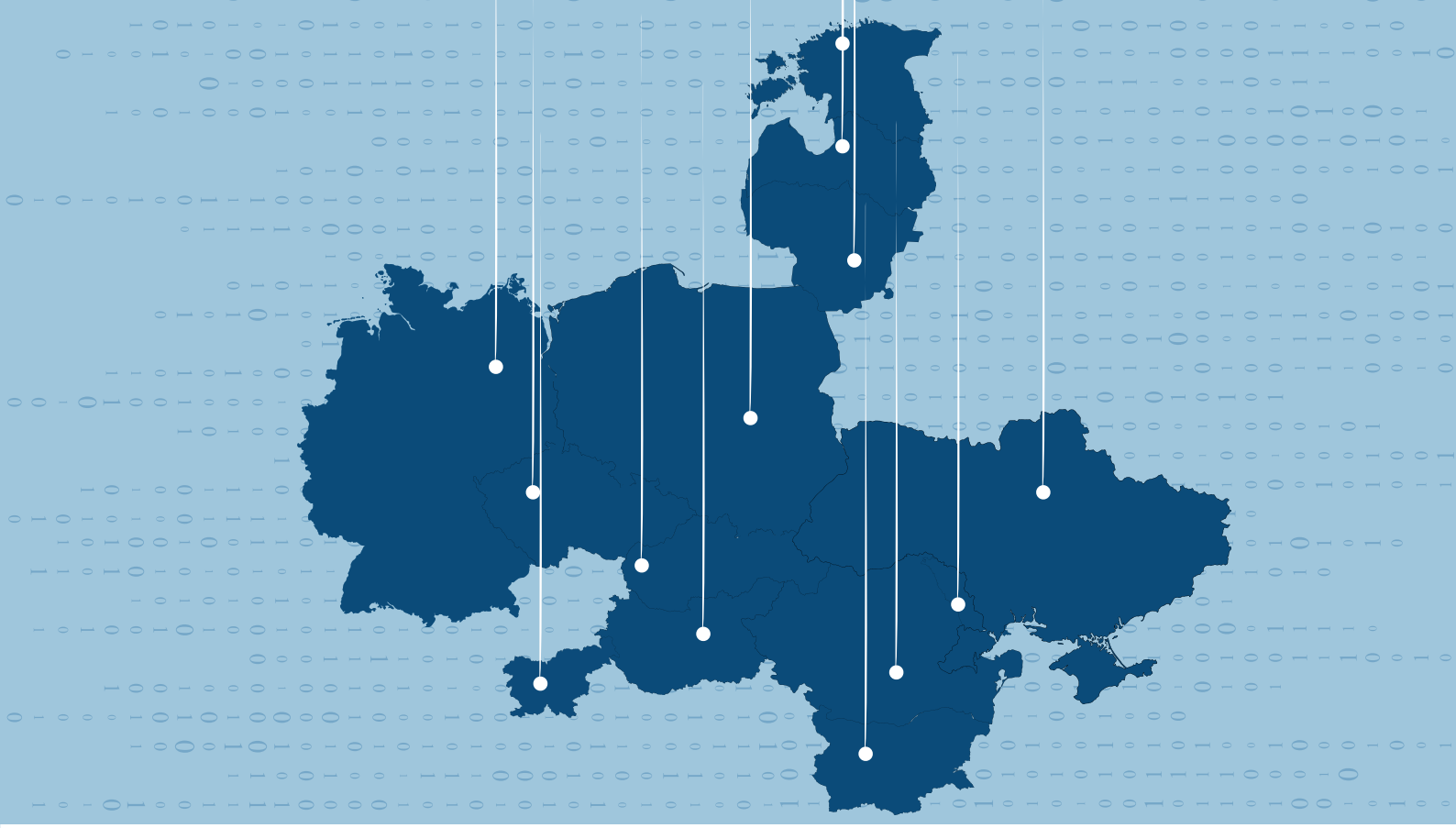

Hybrid Realities: Disinformation, Influencers, and the Defense of Democracy in Central and Eastern Europe

Editors: Stormy-Annika Mildner, Alyssa McIntyre, and Niklas Zimmer

Aspen Institute Germany

Fact Checking **Communication** Values Narratives **Resilience** Interne
Information Geopolitics Integrity **Algorithms** Manipulation Freedom
lation **Elections** Opinion Hybrid Warfare **Social Media** Fake News
Ethics Governance **Technology** Critical Awareness Trust Media Mal
ous Actors Influence Artificial Intelligence **Public Discourse** Securit
Responsibility Debunking Deception Propaganda Cyber Algorithms

About this publication

This publication was developed as part of the project "#InfluencersAgainstDisinfo: Empowering Online Opinion Leaders to Enhance Democratic Resilience in the Age of Information Manipulation," implemented by the Aspen Institute Germany with generous support by the German Federal Foreign Office.

The objectives of the project are to build democratic resilience against disinformation and to contribute to an inclusive dialogue about effective policy measures. The publication brings together esteemed experts from the field and features five articles focusing on specific thematic areas with accompanying recommendations.

The Aspen Institute Germany would like to thank the German Federal Foreign Office for its generous support of the project "#InfluencersAgainstDisinfo: Empowering Online Opinion Leaders to Enhance Democratic Resilience in the Age of Information Manipulation."

Kindly supported by:

Federal Foreign Office

Imprint
Publisher:
Aspen Institute Deutschland e.V.
Friedrichstraße 60, 10117 Berlin
Germany

Editors: Dr. Stormy-Annika Mildner, Alyssa McIntyre, and Niklas Zimmer
Editorial Support: Birte Müller and Emilie Decker
Design & Layout: Volk Agentur + Verlag, Munich
Printing: Buch- und Offsetdruckerei H. Heenemann GmbH & Co. KG
Publication: December 2025
ISSN 2749-**263X**

Hybrid Realities: Disinformation, Influencers, and the Defense of Democracy in Central and Eastern Europe

Editors: Stormy-Annika Mildner, Alyssa McIntyre, and Niklas Zimmer

Table of Contents

Executive Summary

Stormy-Annika Mildner, Alyssa McIntyre, and Niklas Zimmer[1]

Democracy depends on the integrity and credibility of public discourse. It functions most effectively, when citizens can exchange ideas freely, engage in respectful disagreement, and make collective choices informed by reliable information. Transparent and inclusive dialogue fosters trust between individuals and institutions, which in turn underpins the legitimacy of democratic decision-making and helps ensure that differences of opinion do not lead to societal division. Sustaining this foundation requires an information environment that upholds transparency, enables verification, and encourages responsibility in maintaining fact-based public communication.

Over the last decade, communication has increasingly shifted into digital spaces. Digitalization has expanded access to information, lowered barriers to participation, and created new arenas for civic engagement. Online platforms enable citizens to access diverse sources of news, follow political developments in real time, and engage directly with decision makers. Social media allows marginalized voices to be heard, broadening the spectrum of participation. At the same time, the digitalization of information and communication exposes democratic foundations to new systemic vulnerabilities. Social media platforms, based on engagement-driven algorithms, frequently amplify sensational, polarizing or misleading content. They have become a fertile ground for misinformation, disinformation, and broader information manipulation by both state and non-state actors. In addition, personalized information spaces create isolated realities in which citizens rarely engage with alternative views. The result is a narrowing of perspective, a weakening of common ground, and growing polarization. This contributes to corroding public trust and eroding the capacity of societies for collective action. The Organisation for Economic Co-operation and Development (OECD) warned in its 2024 Survey on Drivers of Trust in Public Institutions that "the undermining of a common reality based on factual evidence deepens societal divisions and makes it more difficult to build the consensus necessary to address policy challenges."[2]

Misinformation, disinformation and Foreign Information Manipulation and Interference (FIMI) have evolved into a strategic challenge for democratic societies. Misinformation refers to false or misleading content shared without intent to harm, while disinformation denotes deliberately fabricated or distorted narratives designed to deceive. FIMI, as defined by the European External Action Service, involves coordinated state or state-linked activities – often from foreign actors – that seek to undermine common values and political processes.[3] According to the Flash Eurobarometer 522, 38 percent of EU citizens view "false and/or misleading information in general circulating online and offline" as one of the most serious threats to democracy in their country.[4] The same report finds that approximately 68 percent of EU citizens reported having encountered disinformation or fake news in the seven days prior to the survey ('very often': 13%; 'often': 22%; 'sometimes': 33%).[5] Notably, exposure levels were even higher in some Eastern European

1 In accordance with institutional standards for academic integrity and transparency, artificial intelligence tools were employed mostly for linguistic and editorial purposes. Their use was limited to helping identify the most important key-take aways in the papers of this edited volume as well as in correcting grammar and refining style.

2 OECD, OECD Survey on Drivers of Trust in Public Institutions – 2024 Results, 2024, 119, https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/07/oecd-survey-on-drivers-of-trust-in-public-institutions-2024-results_eeb36452/9a20554b-en.pdf (accessed August 13, 2025).

3 Strategic Communication and Foresight, 3rd EEAS Report on Foreign Information Manipulation and Interference Threats. Exposing the architecture of FIMI operations, European Union External Action, March 2025, 4, https://www.eeas.europa.eu/sites/default/files/documents/2025/EEAS-3nd-ThreatReport-March-2025-05-Digital-HD.pdf (accessed November 3, 2025).

4 European Commission, Flash Eurobarometer 522 Democracy, 2023, 3, https://europa.eu/eurobarometer/surveys/detail/2966 (accessed August 8, 2025).

5 Ibid., 4.

countries – such as Hungary, Bulgaria, and Estonia – where the figure exceeded 80 percent.[6]

Russia's war against Ukraine demonstrates that contemporary conflict unfolds not only on the battlefield but also in the cognitive domain. FIMI and cyber operations are deployed to divide societies, weaken institutions, and erode support for democratic partners. Other authoritarian actors, including China and Iran, employ similar tactics to project influence and challenge democratic cohesion across Europe and beyond.

Nowhere are the consequences of information warfare more visible than in regions where democratic institutions remain fragile or contested. In the context of Russia's war against Ukraine, Eastern Europe has become a testing ground for malign influence (both from abroad and from within). Yet, the region also shows great resilience, as civil society, media, and democratic institutions jointly build capacities for transparency, media literacy, and civic engagement.

Against this backdrop, the Aspen Institute Germany launched the initiative #InfluencersAgainstDisinformation, supported by the German Federal Foreign Office. The project brings together social media influencers and content creators from Central and Eastern Europe to strengthen digital discourse and promote ethical standards in online communication. In 2024, participants jointly composed an Ethical Code of Conduct built on principles of responsibility, transparency, and respect, guiding social media influencers and content creators in navigating an increasingly complex information landscape.

The edited volume Hybrid Realities: Disinformation, Influencers, and the Defense of Democracy Central and Eastern Europe brings together leading experts to explore how disinformation interacts with democratic institutions, technological change, and contemporary governance. It reflects on the structural nature of information manipulation and offers policy-oriented recommendations to strengthen democratic resilience in a rapidly evolving information environment.

At the core of this work lies a central understanding: Defending democracy today requires defending information integrity. Only by safeguarding the quality, transparency and accountability of public communication can open societies remain robust in the face of manipulation and uncertainty.

Key Take Aways

The following Key Take Aways are derived from the collection of policy papers within this edited volume.

1. Democratic Vulnerabilities in a Shifting Geopolitical and Political Landscape

The challenges facing democracy in the information sphere cannot be separated from the wider geopolitical transformations and domestic political trends. Globally, power competition is intensifying as authoritarian states seek to exert influence, weaken democratic alliances, and contest the liberal international order. Information manipulation has become a core component of this strategic rivalry, used to undermine institutional legitimacy, fracture societal cohesion, and erode support for democratic governance and international cooperation.

At the same time, democratic societies face internal pressures. Polarization, distrust in institutions, and the rise of populist and illiberal movements have weakened shared norms and reduced tolerance for compromise. Public debate is increasingly defined by identity conflict, emotional mobilization, and declining confidence in established sources of information and expertise. These dynamics create fertile ground for actors who seek to destabilize democratic processes and exploit divisions from within and beyond national borders.

Against this backdrop, information manipulation has evolved into a deliberate strategy to exploit democratic openness. Malign actors – state-directed networks, ideologically motivated groups, and

6 Ibid., 34-35.

opportunistic domestic actors – leverage the protections and freedoms of open societies to spread mistrust, amplify grievances, and question the legitimacy of democratic institutions. Their tactics are adaptive and incremental: Reframing political debates as existential struggles, seeding doubt in electoral integrity or public institutions, and presenting fringe narratives as mainstream discourse. Cultural symbols, historical narratives, and identity markers are selectively activated to deepen divisions and shift the terms of public debate. The goal is not simply to persuade, but to corrode confidence, blur truth boundaries, and weaken democratic cohesion.

Responding to these pressures requires more than counter-speech or platform governance. It demands strengthening institutional resilience, rebuilding civic trust, and reinforcing democratic norms and culture. Democratic states must protect open debate while preventing the systematic weaponization of freedoms against the democratic system itself. This includes bolstering independent media, supporting civic education and media literacy, and ensuring transparent, credible public communication. Defending democracy in this context is therefore not only a matter of countering manipulation, but of safeguarding the democratic foundations that allow societies to disagree constructively and govern collectively.

2. Algorithmic and AI-Driven Distortions: Reshaping Visibility, Amplifying Emotion, and Challenging Authenticity

Digital platforms have transformed public communication, enabling unprecedented civic participation, political mobilization, and global exchange. Yet, the same systems that expand democratic discourse also introduce new structural vulnerabilities. Social media platforms are optimized to capture and retain attention. Content that provokes immediate emotional responses – anger, outrage, or fear – typically spreads faster and more widely than content that invites deliberation. This creates a systemic bias toward emotional intensity and controversy. While such dynamics do not inherently produce disinformation, they create fertile conditions for false, manipulative, or polarizing content to thrive.

A central mechanism in this environment is algorithmic amplification. Recommendation systems learn from real-time user behavior and iteratively refine what individuals see to maximize engagement. This produces self-reinforcing cycles in which visibility depends less on accuracy or credibility than on emotional resonance. Because these processes are opaque, neither users nor regulators fully understand how content is prioritized, ranked, or suppressed. Although transparency reports, risk assessments, and regulatory accountability tools have increased scrutiny, meaningful visibility into algorithmic decision-making remains limited.

Manipulative actors strategically leverage these dynamics. Emotional activation is a core tactic: Messages are crafted to evoke strong affective responses, increasing the likelihood of being shared and amplified. Narratives are embedded in cultural, national, or religious symbols, reframing political disagreement as a question of belonging rather than evidence. Over time, repeated exposure to emotionally charged and identity-laden messages fosters familiarity, reduces sensitivity to manipulation, and blurs the distinction between fact and fiction.

Artificial intelligence magnifies these risks. Generative systems can produce text, images, audio, and video with increasing realism and speed, lowering technical barriers to manipulation. Capabilities once accessible only to state actors are now within reach of small groups and individuals, making disinformation campaigns more agile, scalable, and persistent. The emergence of agentic AI systems – models capable of operating autonomously – further extends this capacity. Such systems can identify target audiences, select emotionally resonant framings, and coordinate messaging across platforms without human intervention. Entire ecosystems of synthetic accounts and content can be created to simulate public consensus or outrage, obscuring the boundary between authentic and artificial participation.

These developments erode shared reference points in public debate. If synthetic content is indistinguishable from authentic communication, trust be-

comes fragile and contestable. Yet, AI is also part of the solution. Advanced detection tools can identify coordinated inauthentic behavior, trace manipulated media, and authenticate legitimate sources. Pattern-recognition systems can detect linguistic and behavioral signatures of synthetic agents, while provenance tools and watermarking technologies support verification of digital content.

Effective responses require combining technological, regulatory, and societal measures. Transparent criteria for content ranking, independent algorithmic audits, and meaningful data access for researchers would support accountability. Standards for provenance and authentication, paired with clear labeling of synthetic material and disclosure of automated communication, can reinforce informational integrity. Embedding accountability and explainability into AI design and governance is essential to ensure that autonomous systems remain aligned with democratic norms and subject to meaningful human oversight.

3. Influencers as New Intermediaries of Trust

In the fragmented information landscape of the digital age, trust has migrated from institutions to individuals. Social media influencers and content creators have emerged as powerful actors. They are no longer merely entertainers or trendsetters but important intermediaries of information who shape opinions, mobilize audiences, and increasingly influence political discourse.

A social media influencer is someone who has cultivated a dedicated online audience and can shape opinions, behaviors, or purchasing decisions through content shared on platforms like Instagram, TikTok, YouTube, or X. Drawing on perceived authenticity, expertise, or lifestyle appeal, influencers range from nano- and micro-influencers (under 100,000 followers, with niche, highly engaged audiences) to macro- and mega-influencers (above 100,000 followers, often with broad reach). They also differ by content focus, from fashion, travel, beauty, and gaming to politics, activism, and news commentary – sometimes operating entirely outside traditional journalistic norms.

Their influence stems from a particular kind of trust that differs fundamentally from that accorded to traditional media or political institutions. Social media influencers cultivate a sense of intimacy with their audiences by sharing personal stories, behind-the-scene glimpses, and unfiltered opinions, which followers interpret as authentic and relatable. This creates what is called a "parasocial bond" – a one-sided but emotionally charged relationship in which followers feel they know the influencer personally, even though the connection exists entirely online. Such bonds can mirror the trust and familiarity of real-world friendships, making audiences more receptive to the influencer's messages, recommendations, or political views. As a result, their endorsements can carry disproportionate weight, shaping consumer choices, public opinion, and even political behavior in ways that often surpass the persuasive power of institutional actors.[7]

Empirical evidence underscores this influence. The 2024 IZEA Trust in Influencer Marketing report found that over 50 percent of all respondents (poll among U.S. consumers) find influencer content more compelling than scripted advertising, while 61 percent place more trust in influencer posts than in celebrity endorsements. Among younger audiences, the figures are even more striking – 66.5 percent of those polled in the age group 18-29 trust influencers more than traditional celebrities.[8]

In a landscape where according to research by United National Educational, Scientific, and Cultural Organization (UNESCO), 62 percent of influ-

7 Amy Wong and Joicey Wei, "Persuasive Cues and Reciprocal Behaviors in Influencer-Follower Relationships: The Mediating Role of Influencer Defense," in: Journal of Retailing and Consumer Services 75, November 2023, 7.

8 IZEA, 2024 Trust in Influencer Marketing, 2024, 8, https://content.izea.com/hubfs/Gated_Content/2024/IZEA%20Insights%20-%20Trust%20in%20Influencer%20Marketing%20-%202024.pdf?__hstc=250356574.ee78d7673e6dc0f2d801de9a0694f64d.1755071552440.1755071552440.1755071552440.1&__hssc=250356574.1.1755071552440&__hsfp=2808506773 (accessed August 13, 2025).

encers do not fact-check content before sharing,[9] the growing reliance on them over official media and authorities presents a clear risk to information integrity.[10] As influencers communicate in personal, authentic ways, their messages carry weight and their platforms can be exploited. Revenue models that reward attention rather than accuracy can incentivize sensationalism and polarizing content. Undisclosed sponsorships or ideological alignment with manipulative actors can turn influential voices into conduits of disinformation.

Yet, this very immediacy and reach make them also uniquely suited to act as credible advocates for fact-based narratives and democratic values – if they are equipped with the knowledge and tools to do so responsibly. Empowering them to reflect on their responsibility, recognize disinformation, verify sources, and communicate effectively about democratic issues taps into one of the most trusted and far-reaching channels for engaging young, diverse and often politically hard-to-reach audiences.

4. Eastern Europe and the Geopolitics of Disinformation

Eastern Europe remains one of the regions most exposed to foreign information manipulation, yet its vulnerabilities differ markedly across countries. Historical legacies, geopolitical tensions, and institutional fragilities intersect with polarized media landscapes and declining trust in public institutions. These factors create conditions that external actors can exploit to deepen divisions, distort narratives, and weaken democratic cohesion.

Russia is the principal external actor shaping the regional information space. Its doctrine of hybrid warfare integrates military, political and informational tools designed to weaken democratic institutions and undermine confidence in Western alliances. Disinformation serves both strategic and operational purposes: It provokes emotion, influences public opinion, shakes trust in traditional institutions, and destabilizes societies. The intensity and form of these tactics vary across Eastern Europe, reflecting differences in language, history, and domestic politics.

Nowhere are these dynamics more visible than in Russia's war against Ukraine. Disinformation has become a central instrument of warfare, used to justify military action, delegitimize the Ukrainian government, and erode international support. Russian narratives depict Ukraine as divided, corrupt, or controlled by external powers, aiming to fracture solidarity among democratic states. Yet, Ukraine has also emerged as a model of democratic resilience. Civil society, independent media, and state institutions have developed sophisticated mechanisms for monitoring, verifying, and countering manipulation. These efforts demonstrate that resilience can be built even under the extreme pressures of war.

In the Baltic states, Russian information operations target Russian-speaking minorities and question NATO's legitimacy. Narratives invoke cultural affinity, linguistic identity, and historical revisionism to undermine confidence in Western alliances. Yet, Estonia, Latvia, and Lithuania have also developed coordinated responses that combine cybersecurity, media literacy, and strategic communication. Their experience illustrates that resilience is strongest when public awareness, government policy, and civil society engagement reinforce one another.

Central Europe presents a more complex picture. In Romania, Hungary, and Slovakia, external influence often intersects with domestic political polarization. Disinformation does not merely originate

9 UNESCO, 2/3 of Digital Content Creators Do Not Check Their Facts Before Sharing, But Want to Learn How to Do So, November 26, 2024, https://www.unesco.org/en/articles/2/3-digital-content-creators-do-not-check-their-facts-sharing-want-learn-how-do-so-unesco-survey (accessed August 14, 2025); Dan Milmo, "Online Influencers Need 'Urgent' Fact-Checking Training, Warns UNESCO," in: The Guardian, November 26, 2024, https://www.theguardian.com/media/2024/nov/26/online-influencers-need-urgent-fact-checking-training-warns-unesco (accessed August 14, 2025).

10 Nina Merilainen, "Influencers as Tools in Hybrid Warfare," in: Stephanie J. Blackmonand and Saltuk Karahan, ed., Proceedings of the 19th International Conference on Cyber Warfare and Security, Reading: Academic Conferences and Publishing Limited, 2025, 265-272, here 265, https://books.google.de/books?hl=en&lr=&id=ZjtaEQAAQBAJ&oi=fnd&pg=PA265&dq=social+media+influencers+in+foreign+information+manipulation&ots=MyHUc8psay&sig=LtjxpC1mqT_mIUOIoQDhNRs_Cto&redir_esc=y#v=onepage&q&f=false (accessed August 14, 2025).

abroad; it increasingly merges with internal narratives that challenge liberal democratic norms and question the legitimacy of European institutions.

Overall, in Eastern Europe the struggle over narratives is not simply a regional security challenge – it is a test case for democracy's ability to defend itself in an era of strategic manipulation. Strengthening resilience requires a comprehensive approach that integrates national policy, civil-society engagement, cross-border cooperation, and sustained transatlantic commitment. The lessons emerging from Eastern Europe underscore that disinformation is not an unsurmountable force; where democratic institutions are supported, empowered, and held to account, open societies can adapt, resist, and prevail.

5. Regulation, Governance, and the Limits of Enforcement

The European Union (EU) has emerged as a global pioneer in regulating the digital information space. Through instruments such as the Digital Services Act (DSA), the European Media Freedom Act (EMFA), and the Artificial Intelligence Act, Europe is building a comprehensive framework for transparency, accountability, and integrity in online communication. These initiatives mark a shift from reactive crisis management toward systemic governance. They establish obligations for platforms to assess and mitigate risks, require transparency in content moderation and recommendation systems, and enable independent audits of how digital infrastructures affect democratic discourse.

This evolving framework represents a major step in aligning technological governance with democratic principles. The DSA imposes due diligence requirements on large online platforms, compelling them to evaluate and reduce systemic risks such as disinformation, manipulation, and algorithmic bias. The EMFA aims to safeguard pluralism and editorial independence within the media sector by creating safeguards against political interference and ownership concentration. The AI Act extends this regulatory logic to the next generation of technologies by introducing risk-based standards for transparency, safety, and human oversight. Together, these instruments establish the normative foundation for a more accountable digital ecosystem.

However, a gap remains between legislative ambition and practical enforcement. Implementation depends largely on national authorities that vary in resources, expertise, and political will. Coordination among regulators across EU member states is uneven, and the cross-border nature of digital platforms complicates oversight. Manipulative actors exploit this asymmetry, adapting tactics faster than legal systems can respond. Without consistent enforcement and adequate institutional capacity, the most sophisticated regulatory frameworks risk remaining purely words on a page.

A second challenge lies in maintaining the balance between security and freedom. Efforts to counter disinformation and harmful content must not evolve into mechanisms of control that stifle legitimate expression. Regulation should therefore be grounded in transparency, due process, and participatory oversight. Civil society organizations, independent regulators, and academic institutions play a crucial role in ensuring that enforcement remains accountable and proportionate.

Disinformation and Democracy

Dmytro Iarovyi
Kyiv School of Economics

Disinformation has become a systemic threat to democratic institutions by exploiting polarization, eroding trust, and amplifying emotional grievances. Evidence from Ukraine and the Baltic states shows how authoritarian campaigns adapt to local vulnerabilities – ranging from war fatigue and geopolitical uncertainty to linguistic and ethnic divisions – while leveraging influencers, content creators, and platform algorithms to embed manipulative narratives. Resilience, however, is possible. It is best understood not as a fixed outcome but as a dynamic process shaped by laws and governance, media and civil society, and the everyday habits of citizens. The findings highlight five imperatives for strengthening this resilience: Move beyond narrow media literacy campaigns; cultivate reflexive and emotion-aware communication practices; support influencers as responsible actors within information ecosystems; demand democratic oversight of platforms; and invest in long-term, participatory infrastructures for cognitive resilience. Ultimately, resilience cannot be reduced to countering falsehoods alone. It requires democratic societies to build adaptive systems of trust, embed healthy communicative practices, and sustain solidarity under conditions of persistent cognitive threat.

Setting the Scene: Framing the Problem

Disinformation has emerged as a systemic threat to democratic institutions, not merely through isolated falsehoods, but by corroding trust, amplifying polarization, and disrupting the social cohesion necessary for democratic governance.

For democratic societies, especially those on the periphery of conflict zones, resilience to information disorders is not a luxury but a necessity for safeguarding public trust, institutional stability, and social cohesion.

This threat is particularly topical in Eastern Europe, in the states with close proximity to, or historical entanglement with, Russia. In those countries the openness of democratic societies may be exploited by authoritarian regimes through hybrid tactics that include disinformation and strategic manipulation of narratives. In these contexts, disinformation is not only a matter of truth versus falsehood, but is an attack on the infrastructure of democratic life.

Across Europe, a wide range of initiatives has emerged in response to this challenge. Media literacy programs, fact-checking efforts, strategic communication measures, and platform regulation reflect an increasing recognition of the associated risks. Yet, these measures often fall short in environments where facts lose salience, platforms reward outrage, and charismatic influencers promote conspiratorial or mystic worldviews over rational arguments. As Anne Applebaum warns, we are witnessing the rise of "new Rasputins" – figures who reject Enlightenment reason in favor of intuitive and affective appeal.[1] Even well-funded literacy or fact-checking campaigns struggle in an environment where trust, rather than truth, determines influence.

While information disorders have attracted growing attention from scholars and practitioners, and initiatives to counter them have multiplied, their

1 Anne Applebaum, "The New Rasputins," in: The Atlantic, January 7, 2025, https://www.theatlantic.com/magazine/archive/2025/02/trump-populist-conspiracism-autocracy-rfk-jr/681088/ (accessed July 20, 2025).

persistent, adaptive and decentralized nature makes clear that no single, definitive solution exists. Interventions may reduce harm, but their effectiveness is often uncertain, context-dependent, and difficult to scale.[2] It means that rather than aiming to eliminate manipulative content entirely, long-term strategies should prioritize the cultivation of societal resilience as the capacity to absorb, adapt to and recover from the disruptive impact of hostile narratives. This includes not only formal mechanisms, but also informal communicative practices embedded in everyday discourse.

As such, understanding how influence is exercised across platforms and social networks becomes critical. New actors such as social media influencers and content creators have gained disproportionate influence. Operating on opaque, low-accountability platforms which provide access to a large amount of content and may amplify questionable information,[3] they have become central vectors of affective and conspiratorial content. Understanding their role, whether as amplifiers or gatekeepers of manipulative content, is beneficial to assessing how societies can respond.

Europe's eastern frontier offers a vital vantage point from which to examine the dynamics of information warfare. Countries such as Ukraine and Lithuania, long exposed to Russian disinformation, exemplify both the vulnerabilities inherent in such exposure and the resilience developed in response. The Baltic states remain particularly exposed in geopolitical terms and feature prominently in the Kremlin's rhetoric. Moscow's neo-imperialist narrative continues to frame the post-Soviet space as part of its presumed sphere of influence – suggesting that Ukraine may not be the final target of hybrid or conventional aggression. Against this backdrop, Lithuania has frequently been singled out as a potential object of destabilization, a concern voiced not only by analysts and observers,[4] but also by officials such as Ukraine's former National Security and Defense Council Secretary Oleksiy Danilov and Lithuania's former Defence Minister Arvydas Anušauskas.[5]

This raises a pressing question: How can democratic societies on the frontline of such cognitive warfare develop and sustain resilience – not just through institutional safeguards or media literacy, but through a deeper understanding of influence, trust, and the everyday mechanisms by which public meaning is constructed or eroded?

Defining Disinformation: Concepts and Distinctions

The study of disinformation is often marked by conceptual fragmentation. Terms such as misinformation, fake news, and propaganda capture distinct aspects of the phenomenon, yet they are used inconsistently and without a coherent analytical framework. Achieving conceptual clarity is therefore essential. Taken together, these notions illuminate different dimensions of disinformation – whether intent, format, or political function – but none alone adequately captures its full complexity or the fluid dynamics that characterize contemporary digital information environments.[6]

Kapantai et al.,[7] for instance, provide a bibliographic analysis of taxonomies related to false or misleading information, referencing terms such as "information pollution" and "information disorders." In their taxonomy, they also reference terms like "fake news," "false news," "false information," and "disinformation." Yet ultimately, for the purposes

2 Jon Bateman and Dean Jackson, Countering Disinformation Effectively. An Evidence-Based Policy Guide, Carnegie Endowment for International Peace, January 31, 2024, 2-4, https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en 25, 2025).

3 Matteo Cinelli et al., "The COVID-19 Social Media Infodemic," in: Scientific Report 10, 2020, 1.

4 Juri Rescheto, "Who's Next? Lithuanians Fear Potential Russian Aggression," in: Deutsche Welle, April 16, 2022, https://www.dw.com/en/whos-next-lithuanians-prepare-for-potential-russian-aggression/a-61496290 (accessed July 20, 2025).

5 Interfax-Ukraine, "Lithuania to Be Next Country Putin Wants to Take Over if He Wins in Ukraine," in: Interfax-Ukraine, March 13, 2022, https://en.interfax.com.ua/news/general/812410.html (accessed July 20, 2025); Amrit Burman, "Lithuanian Defence Minister Says Russia Poses a 'Long-Term Threat' to Vilnius' Security," in: Republic World, May 29, 2022, https://www.republicworld.com/world-news/russia/lithuanian-defence-minister-says-russia-poses-a-long-term-threat-to-vilnius-security-articleshow (accessed July 20, 2025).

6 Raquel Recuero, "A Systemic Framework for Disinformation on Social Media Platforms," in: Platforms & Society 2, 2025, 1-12, here 2-4.

7 Eleni Kapantai et al., "A Systematic Literature Review on Disinformation: Toward a Unified Taxonomical Framework," in: New Media & Society 23, no. 5, 2021, 1301–1326, here 1305.

of their review, they adopt the definition of "disinformation" as provided by the European Commission High-Level Expert Group (HLEG) on fake news and online disinformation, which encompasses "all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit."[8]

Wardle and Derakhshan further distinguish between disinformation (false content spread deliberately to cause harm), misinformation (false content shared without intent to harm), and malinformation (genuine information used maliciously, e.g., leaks or selective publication).[9] They group these under the broader term "information disorders," shifting attention from individual pieces of content to the ecosystems in which they circulate. Adopting this term allows for a more comprehensive analysis of manipulative practices across contexts.

Increasingly, the EU addresses these phenomena under the broader term Foreign Information Manipulation and Interference (FIMI), as defined by the European External Action Service.[10] FIMI emphasizes the role of coordination, actor behavior, and intent – beyond just false content. It includes tactics such as impersonation, leak operations, and strategic amplification. While authoritarian states like Russia, China, and Belarus orchestrate or support such efforts, domestic actors and influencers often help amplify them. These campaigns blur the lines between fact and fiction by combining genuine and fabricated content. Their success relies on automation, algorithmic amplification, and emotional resonance.

Clarifying such distinctions is vital: Without clear terminology, legal and policy responses risk either overreaching or underperforming. This is especially acute in the context of the freedom-security dilemma – the tension between protecting citizens from harm and safeguarding fundamental rights.[11] Democracies operate within normative boundaries that distinguish them from authoritarian systems, yet these very constraints are often exploited in asymmetric information warfare. Ambiguous definitions can therefore lead to regulatory measures that are either perceived as threats to free expression or too limited to counter manipulation effectively.

At the same time, conceptual clarity provides the foundation for a balanced approach to addressing information disorders. Carefully targeted restrictions on harmful content are not inherently at odds with free speech; rather, they can safeguard it from manipulation, given that disinformation frequently fuels hatred, fear, and polarization.[12] When combined with systemic preventive measures (for example media literacy initiatives, fact-checking networks, and strategic communication), such interventions can enhance democratic resilience without eroding civil liberties. Left unaddressed, however, manipulative information flows deepen societal fragmentation and corrode public trust.

The Democratic Impact of Disinformation

Democracy fundamentally depends on the integrity of information. Clear, truthful and verifiable information sustains public trust in elections, institutions, and decision-making. It enables equal political participation by ensuring citizens have accurate knowledge of procedures, access, and choices. A shared factual foundation allows for rational deliberation, consensus-building, and peaceful negotiation of interests. Verified information also supports accountability by empowering media, parliaments, and courts to scrutinize power and expose wrong-

8 Independent High-Level Group on Fake News and Online Disinformation, A Multi-Dimensional Approach to Disinformation. Report of the Independent High-Level Group (HLEG) on Fake News and Online Disinformation, European Commission, March 2018, 3, https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1/language-en (accessed August 20, 2025).

9 Claire Wardle and Hossein Derakhshan, Information Disorder. Toward an Interdisciplinary Framework for Research and Policy Making, Council of Europe, September 27, 2017, 5, https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html (accessed July 20, 2025).

10 European External Action Service (EEAS), Information Integrity and Countering Foreign Information Manipulation and Interference (FIMI), https://www.eeas.europa.eu/eeas/information-integrity-and-countering-foreign-information-manipulation-interference-fimi_en (accessed October 22, 2025).

11 Tatiana A. Romanova et al., "Disinformation (Fake News, Propaganda) as a Threat to Resilience. Approaches Used in the EU and Its Member State Lithuania," in: Baltic Region 12, no. 1, April 2020, 53–67, here 55.

12 Tomer Shadmy, "Content Traffic Regulation: A Democratic Framework for Addressing Misinformation," in: Jurimetrics Journal 63, August 29, 2022, 1-57, here 6.

doing. In times of crisis, trusted information acts as a stabilizing force. When disinformation spreads, it corrodes public discourse and the epistemic foundations of democratic legitimacy.

Therefore, societal resilience to disinformation can be linked with broader democratic resilience – defined as the ability of democratic systems to adapt to and recover from threats that could undermine institutions and norms.[13] Indeed, it is public trust as well as institutional and interpersonal engagement which sets the foundation of a stable and healthy democracy.[14] Disinformation, on a deeper level, attacks this trust.

Illiberal actors increasingly weaponize the very freedoms that sustain democracy, exploiting openness, transparency, and pluralism to erode trust and polarize societies from within.[15] The information war machinery of autocratic regimes is highly adaptive, tailoring its methods to resonate with audiences in target societies. It rarely fabricates scandals or invents grievances; rather, it amplifies and exploits existing social tensions to erode public trust and weaken democratic cohesion. Instead of adhering to fixed narratives, such regimes strategically reinterpret local controversies through a geopolitical lens, transforming domestic polarization into a tool of influence that advances their broader strategic objectives.

These external manipulations exploit the very communicative openness that defines liberal democracies. What constitutes their greatest strength, i.e., the free circulation of ideas and the contestation of power, can, under the pressure of hybrid warfare, become a source of fragility. This openness not only empowers civic movements and pluralism, but also challenges existing power structures, giving rise to information threats and legitimacy crises, particularly in democratic contexts.[16] Lacking effective tools to counter these threats, democracies remain especially vulnerable.[17] Authoritarian actors exploit this asymmetry, turning openness itself into a strategic weapon – much like in aikido, where an opponent's strength is redirected against them.[18] In this asymmetric information environment, autocracies act unrestrained within democratic public spheres, while democracies remain bound by the very freedoms that define their moral and political order.

In the peaceful times, liberal democracies traditionally prioritize freedom of expression, even when facing information threats.[19] However, modern hybrid threats challenge this stance, forcing democracies to reconsider how they defend themselves against hostile influence operations. While rooted in diverse fields – from ecology to mental health – resilience according to the discipline of social sciences generally refers to the capacity of individuals or societies to endure disruption and preserve coherence.[20] Rather than returning to a previous equilibrium, resilience implies navigating change, sometimes irreversibly.[21] It is a process shaped by learning and governance, not a fixed outcome.[22] In a time, in which the traditional ideas of what is peace and what is war is changing, sustainable defense cannot rely on censorship alone. Rather, it requires trust-building, strategic communication, and critical

13 Daniela Munteanu, "Societal Resilience to Disinformation – What Is It and How Can We Bolster It?," in: Future Scenarios for Central and Eastern Europe in Times of Continuous Uncertainty, Economic Policy Institute, June 2024, 90-109, here 96-97, https://epi-bg.org/images/2024_files/A5_Format_SS_Book_2024-fin_27112024.pdf (accessed August 25, 2025).

14 Aukse Balčytienė and Kristina Juraitė, "Baltic Democracies: Re-configuring Media Environments and Civic Agency," in: Journal of Baltic Studies 53, no. 4, September 2, 2022, 565–585, here 565.

15 Munteanu, 2024, 91-92.

16 W. Lance Bennett and Steven Livingston, "The Disinformation Order. Disruptive Communication and the Decline of Democratic Institutions," in: European Journal of Communication 33, no. 2, April 2, 2018, 122-139, here 128.

17 Inez Miyamoto, "Disinformation: Policy Responses to Building Citizen Resiliency," in: Connections: The Quarterly Journal 20, no. 2, 2021, 47-55, here 50; Shadmy, 2022, 13.

18 Anton Liagusha and Dmytro Iarovyi, "Memes, Freedom, and Resilience to Information Disorders: Information Warfare Between Democracies and Autocracies," in: Social Sciences & Humanities Open 11, 2025, 1-11, here 2.

19 Corneliu Bjola, "The Ethics of Countering Digital Propaganda," in: Ethics & International Affairs 32, no. 3, August 29, 2018, 305–315, here 308.

20 Brigit Obrist et al., "Multi-Layered Social Resilience: A New Approach in Mitigation Research," in: Progress in Development Studies 10, no. 4, September 10, 2010, 283-293, here 283.

21 Daniel F. Lorenz, "The Diversity of Resilience: Contributions from a Social Science Perspective," in: Natural Hazards 67, no. 1, November 23, 2010, 7-24, here 8.

22 Kirsten Maclean et al., "Converging Disciplinary Understandings of Social Aspects of Resilience," in: Journal of Environmental Planning and Management 60, no. 3, June 10 , 2016, 519-537, here 522.

media literacy, balancing immediate protective measures with long-term cognitive resilience.

The fragility of democratic trust becomes most visible during elections, when heightened competition, emotion, and visibility turn the information sphere into a contested arena. Political polarization tends to intensify during electoral processes, with disinformation serving as a key driver of this dynamic. Disinformation reinforces partisan divides that deepen extremism and distrust.[23] Elections heighten political competition and emotions, creating fertile ground for manipulative narratives that exploit existing divides. Disinformation campaigns, whether foreign or domestic, capitalize on this heightened attention by spreading content designed to delegitimize opponents, cast doubt on electoral integrity, or inflame identity-based grievances. Low trust in government and media creates fertile ground for groups that oppose dominant narratives and may be more susceptible to populism or extremism.[24] These targeted groups often share social identities, political orientations, or emotional grievances that make them particularly receptive to certain narratives.

Social media platforms may amplify disinformative content due to algorithmic bias toward engagement, regardless of veracity.[25] This weakens deliberation by polarizing debates and reducing space for nuanced discussion. Social media influencers and content creators, in particular, play a pivotal role in this landscape, as their perceived authenticity, emotional resonance, and platform-savviness allow them to frame information in ways that feel personal, trustworthy, and culturally relevant. The trust that social media influencers and content creators enjoy with their audiences is highly exploitable. They may knowingly disseminate falsehoods for ideological, commercial or political gain, or they may be co-opted, commissioned, incentivized, or deceived into amplifying misinformation without realizing it. Their hybrid role as relatable peers and cultural intermediaries magnifies the harm: perceived authenticity confers credibility, while platform incentives reward engagement irrespective of truth.

As platforms adjust algorithms or newsrooms shift strategies, even resistance becomes a struggle. To keep pace, democratic actors must develop anticipatory rather than merely reactive strategies.[26] Otherwise, authoritarian regimes retain the advantage of fewer ethical and legal constraints.

In the digital information sphere, democratic resilience must adapt to the accelerating pace of technological change, algorithmic manipulation, and the attention-driven logic of online platforms. Structural transformations, marked by the decline of editorial gatekeeping and the rise of virality, automation, and emotional amplification, have created an ecosystem in which misleading content circulates faster and more widely than ever before.[27] Cognitive biases such as confirmation bias and affective polarization further entrench these dynamics, rendering societies increasingly susceptible to manipulation. Understanding who the key actors of resilience are, and how they operate within these evolving socio-technical environments, is therefore essential for developing more sustainable and adaptive democratic responses.

Actors, Instruments, and Technologies

Understanding information disorders requires analytical attention to the discursive practices through which they emerge, the actors who sustain them, and the channels that enable their circulation across diverse media ecosystems. Disinformation constitutes a particularly potent instrument of influence because it operates within regulatory asymmetries, navigating legal grey zones, and exploiting contextual vulnerabilities.

23 Bennett and Livingston, 2018, 125.

24 Hannu Nieminen, "Why Does Disinformation Spread in Liberal Democracies? The Relationship Between Disinformation, Inequality, and the Media," in: Javnost: The Public 31, no. 1, February 16, 2024, 123-140, here 125.

25 Franziska Zimmer et al., "Fake News in Social Media: Bad Algorithms or Boased Users?," in : Journal of Information Science Theory and Practice 7, no. 2, 2019, 40-53, here 46.

26 Munteanu, 2024, 105-106.

27 Myrto Pantazi, Scott Hale, and Olivier Klein, "Social and Cognitive Aspects of the Vulnerability to Political Misinformation," in: Advances in Political Psychology, 42, no. 1, December 2021, 267-304, here 275.

Anthony Giddens' structuration theory[28] highlights the duality of structure and agency: Institutions and norms shape human action, while everyday practices simultaneously reproduce or transform those structures. This lens shows that resilience to disinformation is built not only through formal strategies – laws, policies, regulations – but also through informal practices of communication and interpretation. Even small acts, such as forwarding a meme or fact-checking a post, help define what societies consider credible and worth amplifying.

Resilience cannot depend only on regulation or fact-checking. It must be embedded in the everyday habits and interpretive frameworks of individuals and communities. Structuration theory is thereful useful as it captures both top-down interventions (like platform policies) and bottom-up practices (like user agency) that shape the information space.

Importantly, many practices cut across levels. Fact-checking achieves little if confined to elites and not taken up by wider publics. Regulation matters only when it shapes actual behaviors and platform design. Resilience thus emerges through interaction: Institutions enabling, communities embedding verification, and individuals adopting adaptive habits.

A diverse set of actors drives the production and diffusion of disinformation, each pursuing distinct yet often convergent objectives. Authoritarian states instrumentalize state-controlled media, diplomatic channels, and proxy organizations to advance geopolitical aims, destabilize democratic cohesion, and shape international narratives in their favor. [29] Domestic radical or populist parties deploy disinformation to delegitimize political opponents and undermine institutional trust.[30] Private actors, such as opportunistic entrepreneurs, influencers, or fringe media outlets, contribute to this ecosystem for ideological validation, financial gain, or mere visibility within the attention economy. Despite these varying motives, their cumulative impact is similar: a fragmented public sphere, deepened polarization, and the gradual erosion of democratic legitimacy.

Against this backdrop, particular attention needs to turn to a potent yet often under-theorized actor within the information ecosystem: social media influencers and content creators. Whether they are journalists, content creators, public intellectuals, or viral meme accounts, social media influencers occupy nodal positions within communicative structures. According to Gidden's theory, influencers are not merely transmitters of information, but structuring agents whose practices both reflect and reshape the digital public sphere.[31] Their distinct power comes from a hybrid position: embedded in macro-level systems like platforms and regulations, operating through meso-level infrastructures such as media and NGOs, and shaping micro-level perceptions directly. This visibility grants them unusual "structuration capacity" to set agendas, model norms, and amplify emotions across contexts.

Influencers do not act in a vacuum. Social media amplifies polarization by fostering echo chambers, encouraging emotional engagement, and diminishing traditional editorial gatekeeping.[32] Evolving technologies – from recommendation algorithms that amplify disinformation to generative systems producing deepfakes – have made hostile information operations more adaptive and harder to detect, posing one of today's most difficult regulatory challenges. [33] Online disinformation thrives on affective intensity, ambiguity, and speed, being spread six times faster than truthful content.[34] This is precisely the dynamics that influencer content often replicates. In such a context, influencers function both as potential vulnerabilities and vital assets in the battle for public reason. They may knowingly or

28 Ira Chatterjee et al., "Anthony Giddens and Structuration Theory," in: Stewart Clegg and Miguel Pina e Cunha, ed., Management, Organizations and Contemporary Social Theory, London: Routledge, 2019, 60-79, here 62.

29 Nathalie Maréchal, "Networked Authoritarianism and the Geopolitics of Information: Understanding Russian Internet Policy," in: Media and Communication 5, no. 1, 2017, 29-41, here 37.

30 Bennet and Livingston, 2018, 128.

31 Chatterjee et al., 2019, 63.

32 Smitha Milli et al., "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media," in: PNAS Nexus 4, no. 3, March 5, 2025, 1–8, here 6.

33 Han-Wei Sun, "Regulating Algorithmic Disinformation," in: Columbia Journal of Law & the Arts 46, 2023, 367–412, here 367.

34 Soroush Vosoughi et al.,"The Spread of True and False News Online," in: Science 359, no. 6380, March 9, 2018, 1146–1151, here 1148.

unknowingly contribute to the production and dissemination of disinformation. But they can also serve as agents of resilience. They reach publics often inaccessible to traditional institutions, and their practices – from source disclosure to emotionally resonant pro-democracy messaging – can shift norms at scale. The growing reliance on social media influencers as sources of news, particularly in the United States, is redefining notions of credibility and authority and transforming how trust is assigned in the digital public sphere.[35] Their reflexivity, a core feature of agency in structuration theory, allows them to adapt messaging in real time, responding to feedback loops and community expectations.

Importantly, motivation is what matters here. Influence aligned with intrinsic values – not just reputational incentives – is more likely to support durable resilience. Influencers can catalyze everyday verification practices, foster discursive moderation, and normalize constructive engagement, but only when enabled by environments that reward such behavior.

Supporting influencers as resilience actors means offering more than recognition. Many lack institutional backing, being vulnerable to disinformation traps, coordinated attacks, or burnout. It requires cross-sectoral support: Training in information integrity, legal protections, mental health resources, and collaborative infrastructure with journalists, educators, and civil society. Rather than merely "instrumentalizing" influencers for strategic messaging, democratic societies must recognize them as embedded actors within the contested terrain of information warfare. By enabling the reflexivity, aligning incentives with public interest, and integrating them into broader resilience networks, a more participatory and durable defense of the digital public sphere can be developed.

Case Studies: Campaigns and Consequences

Disinformation in Eastern Europe is not incidental but part of a broader geopolitical struggle. For decades, authoritarian actors, especially Russia, have exploited historical memory, ethnic tensions, and institutional vulnerabilities to weaken democracies. These campaigns intertwine narratives of Western decline, NATO encirclement, and national victimhood to sow distrust and polarize publics. The region's proximity to Russia, its Soviet legacy, and ongoing democratization create both resilience potential and deep susceptibilities.

Russia had already tested hybrid warfare in 2014, combining military aggression with political, economic and informational tools (the so-called "Gerasimov Doctrine").[36] These tactics are aimed to polarize societies and delegitimize democratic institutions through "sharp power."[37] Russia's full-scale invasion of Ukraine in 2022 again underscored the centrality of the information domain. Ukraine's ability to withstand the initial assault rested not only on military defense but also on adaptive communication: popular resistance, rising trust in leadership, and a unified global message demonstrated the power of societal resilience. Yet, the cognitive dimension of the war is far from over: Ukraine was not the end point of such attacks but an early frontline in a wider contest over Europe's democratic cohesion.

In frontier democracies, the vulnerable groups remain prime targets. Despite Euro-Atlantic integration and strong institutions, Russian operations continue to mobilize pro-Russian counter-publics and exploit local divides to discredit the EU and NATO, deepen polarization, and erode legitimacy from within.[38]

35 Galen Stocking et al., America's News Influencers, Pew Research Center, November 2024, https://www.pewresearch.org/wp-content/uploads/sites/20/2024/11/PJ_2024.11.18_news-influencers_report.pdf (accessed October 24, 2025).

36 Mark Galeotti, "The Mythical 'Gerasimov Doctrine' and the Language of Threat," in: Critical Studies on Security 7, no. 2, February 27, 2019, 157-161, here 158.

37 Andis Kudors, Russian Sharp Power in Action. The Case of Latvia, Institute for European, Russian and Eurasian Studies, https://ieres.elliott.gwu.edu/project/russian-sharp-power-in-action-the-case-of-latvia/ (accessed July 20, 2025).

38 Lukas Andriukaitis, Russian Propaganda Efforts in the Baltics and the Wider Region, Vilnius Institute for Policy Analysis, 2020, 1, https://vilniusinstitute.lt/wp-content/uploads/2020/05/VIPA_Andriukaitis_2020_Iv4-1%D0%B5.pdf (accessed July 20, 2025); Blagovest Tashev et al., Russia's Information Warfare. Exploring the Cognitive Dimension, Marine Corps Center for Advanced Operational Culture Learning Quantico United States, November 2019, 129-147, here 134, https://apps.dtic.mil/sti/citations/tr/AD1101048 (accessed July 20, 2025).

Ukraine's early communication advantage has grown more fragile as the war endures. Russia adapts quickly, exploiting new vulnerabilities, while trust deficits and inconsistent narratives undermine Ukraine's messaging. Progress toward proactive storytelling is evident, but sustaining it is essential – and the country must avoid simply mirroring Russian propaganda or defining itself solely in opposition.

In the current context, disinformation exploits the suspension of elections under martial law to claim that Ukraine is no longer a functioning democracy. Such narratives attempt to equate wartime necessity with authoritarianism, undermining international legitimacy and domestic trust. Ukraine's resilience, however, depends not on holding elections amid invasion, but on sustaining a cohesive national narrative, reinforcing public confidence, and demonstrating solidarity at home and abroad.

Recurring disinformation themes of Western betrayal or elite corruption have been amplified by platforms such as Telegram and Facebook, where weak local moderation gave pro-Russian actors an edge: Telegram enabled coordinated influence through weak moderation, while Facebook's micro-targeting and algorithmic bias toward emotional content spread manipulative narratives. Limited investment in Ukrainian-language and context-aware moderation created space for pro-Russian actors to gain traction. The companies often lack moderators fluent in local languages and sensitive to context, raising questions about whether moderation teams are even located in Ukraine.[39] This gap left civic mobilization efforts reactive within an asymmetric and volatile digital environment.

The Baltic states offer another example of frontier democracies grappling with the full spectrum of challenges to democratic resilience, while occupying a central place within Russia's broader strategy of influence. Although speculation about hybrid scenarios, such as potential flashpoints in Narva or Latgale, has circulated since 2014, the region has thus far avoided military escalation. Instead, Baltic resilience has evolved through a combination of strategic foresight, institutional preparedness, and proactive counter-measures.

While sharing a Soviet legacy and exposure to Russian pressure, Estonia, Latvia, and Lithuania have developed distinct[40] yet complementary approaches, rooted in their media systems, linguistic environments, and governance cultures. All three countries have been very proactive in facing foreign-led disinformation campaigns.[41] Their governments have responded with regulation, education, and international coordination, while civil society has become increasingly active. Still, disinformation narratives about minority marginalization, Baltics as "fake Europe," NATO hostility, or "liberal decay" circulate widely.[42] Insidious narratives, such as claims that NATO will abandon the Baltics or that migrants pose an existential threat, continue to circulate with real political impact. What seems ineffective to one audience may be highly persuasive to another, underscoring the danger of underestimating disinformation's reach. Russian information policy seeks to fracture societies by exploiting divides. In the Baltics, electoral cycles are often accompanied by intensified disinformation efforts aimed at delegitimizing governments, discouraging participation, or portraying minority voters as manipulated by foreign interests. These narratives seek to erode confidence in the fairness of democratic competition.

A key regional vulnerability is the presence of sizeable Russian-speaking minorities. Russian-language channels and digital platforms often amplify these themes, with moderation gaps leaving space

39 Human Rights Watch, Russia, Ukraine, and Social Media and Messaging Apps, https://www.hrw.org/news/2022/03/16/russia-ukraine-and-social-media-and-messaging-apps (accessed September 26, 2025).

40 Maarit Jaakkola, Media Literacy in the Baltics: Different Approaches in Neighbouring Countries, Media & Learning, December 6, 2020, https://media-and-learning.eu/type/featured-articles/media-literacy-in-the-baltics-similar-backgrounds-but-different-approaches (accessed July 20, 2025).

41 Dimtri Teperik et al., Resilience Against Disinformation. A New Baltic Way to Follow?, International Centre for Defence and Security, October 20, 2022, 3-6, https://icds.ee/en/resilience-against-disinformation-a-new-baltic-way-to-follow/ (accessed July 20, 2025).

42 Laurynas Kasčiūnas et al., Rusijos Propaganda. Analizė, Ivertinimas, Rekomendacijos, Rytų Europos Studijų Centras, 2017, 238, http://www3002.vu.lt/uploads/news/id987/RESC%20monografija_propaganda.pdf (accessed July 20, 2025).

for manipulation. Platforms such as VKontakte (still used by some Russian-speaking populations), YouTube, and TikTok serve as conduits for both covert messaging and soft influence. Despite formal bans or restrictions on state-controlled outlets, mirror accounts and influencer-driven dissemination continue to evade enforcement. Platform moderation often treats the Baltics as peripheral markets, resulting in underinvestment in local language support and weak responsiveness to context-specific threats. As a result, platform dynamics not only fail to dampen disinformation; they often enable its emotional and cultural embeddedness in targeted linguistic communities.

Taken together, Ukraine and the Baltic states illustrate both vulnerabilities and sources of resilience. Despite sustained influence operations, these young democracies have shown adaptability – though asymmetry remains. Their pluralistic environments remain exposed to sharp power and market infiltration,[43] unlike Russia's tightly controlled media space. Yet, openness has also enabled experimentation with targeted restrictions (e.g., media bans, state broadcasts), while civil society, independent journalism, and fact-checking initiatives have helped cultivate trust and cohesion.

Conclusion and Policy Recommendations

Resilience to information disorders is not a checklist or a key performance indicator (KPI), but a recursive social process. Disinformation thrives in cognitive, emotional and structural vulnerabilities, many of which are deeply embedded in contemporary democracies. Strengthening resilience against it, therefore requires action across multiple levels: From platform governance and strategic communication, to educational reform and support for informal communicators.

- Reframe resilience beyond the media literacy mantra: Resilience must be understood as a broader cultural shift rather than a technical fix. Media literacy and critical thinking are essential, but their effectiveness lies in fostering long-term interpretive autonomy, not just fact-checking skills. Communication education should also incorporate emotional literacy, recognizing that disinformation appeals to fear, outrage, and identity more than to reason.

- Prioritize everyday practices over posturing in strategic communication: Institutional responses to disinformation should move beyond crisis-driven or image-centric campaigns. Instead, democratic actors must model transparency and coherence in everyday messaging. This includes supporting platform practices that reward verification, reporting manipulation, and resisting performative outrage. Communication strategies should aim not to "win the narrative" immediately, but to sustain credible, pluralistic public deliberation over time.

- Recognize and support influencers as embedded agents of resilience: Influencers – journalists, creators, educators, and public intellectuals – hold nodal positions in the information space. Their embeddedness gives them unique reach and structuration capacity. However, they often operate without institutional support. Democratic systems should invest in training, legal and psychological assistance, and collaborative infrastructure for influencers who are committed to integrity. Importantly, these actors should be treated not as tools for strategic messaging, but as co-creators of a resilient digital public sphere.

- Address systemic asymmetries in the digital information ecosystem: Disinformation exploits the architecture of digital platforms: algorithmic amplification, monetization incentives, and opaque moderation systems all serve to amplify manipulative content. These structural asymmetries require democratic oversight and transparent platform governance. Moreover, resilience efforts should be participatory: Marginalized groups must be included not only as recipients of interventions but as co-designers of responses. In this regard, democratic actors should align their strategies with the EU's evolving framework on Foreign Information Manipu-

43 Liagusha and Iarovyi, 2025, 9.

lation and Interference (FIMI), which stresses the systemic, cross-border nature of such threats and the need for coordinated responses across institutions, platforms, and societies.

- Invest in long-term, participatory infrastructure for cognitive resilience: Sustainable resilience requires ongoing investment in institutions and social practices that enable democratic adaptation. Civil society initiatives that promote verification, digital rights, and constructive discourse must be supported at the community level. At the same time, interdisciplinary research combining insights from communication, cognitive science, and public policy should inform the design of future, more targeted policy responses. Above all, democratic resilience rests on accepting friction, not control, as the necessary cost of pluralism.

Disinformation in the Eastern EU Member States: Local Dynamics and European Challenges

Cristina Ivan
"Mihai Viteazul" National Intelligence Academy

Across Eastern Europe, disinformation has become a strategic instrument for reshaping political allegiances and undermining democratic resilience. Russian-aligned actors systematically exploit cultural conservatism, orthodox identity, and gender stereotypes to erode trust in EU and national institutions while promoting "strategic conservatism" as an alternative to liberal democracy. Within this context, disinformation campaigns targeting Moldova's President Maia Sandu and Estonia's Prime Minister Kaja Kallas weaponized gendered tropes of emotional instability, illegitimacy, and moral deviation to discredit pro-European female leadership and reinforce patriarchal hierarchies. The recurrence of identical linguistic patterns and narratives across countries, amplified through AI-generated and algorithmically optimized content, reveals a coordinated hybrid strategy that fuses technological sophistication with cultural manipulation. Strengthening resilience requires anticipatory detection systems, sustained cross-border cooperation, and long-term investment in digital, emotional, and civic literacy. At the policy level, European frameworks such as the Digital Services Act and the Code of Practice on Disinformation should be complemented by adaptive, participatory communication strategies that foster societal trust and engagement.

Setting the Scene: Information Battleground in Eastern Europe

Over the past decade, Eastern Europe has become a battleground in the contest over identity and political orientation.[1] Russian-led disinformation networks systematically exploit historical grievances, cultural symbols, and social divisions to weaken democratic institutions and erode confidence in the European project. Such efforts obey a long-term modus operandi established with Russia's doctrine of "New Generation Warfare" – a strategy of influence that "seeks to alter the perception of – if not halt and eventually reverse – Central and Eastern Europe's Euro-Atlantic enlargement and orientation, restoring Russia's historic sphere of influence, and returning to a bipolar organized world."[2] By merging narratives of cultural preservation and moral revival with sophisticated digital manipulation, these campaigns turn the information space into a hybrid arena where technology, ideology, and identity converge, challenging Europe's democratic resilience and societal cohesion.

Since Russia's full-scale invasion of Ukraine in 2024, these operations have intensified, marking a new phase in information warfare. Disinformation campaigns increasingly instrumentalize national traditions, religious symbolism, and collective memory to construct a moral divide between a Russia-patronized "golden age" of authentic values and a Westernized present portrayed as morally decadent and culturally alien. This narrative architecture advances "strategic conservatism," i.e., an ideological framework that elevates hierarchy and collective conformity over individual rights, positioning traditional family values in opposition to

1 The following analysis integrates methodologies as: critical discourse analysis, semiotic analysis, cultural hermeneutics, interpreting symbols, values, behaviors, and beliefs professed by disinformation content.

2 Heather Conley et al., The Kremlin Playbook. Understanding Russian Influence in Central and Eastern Europe, Lanham: Rowan and Littlefield, 2016, 4.

Western diversity and inclusion.[3] Through this lens, Russian superiority in culture and faith is presented as both moral justification and geopolitical duty to protect smaller nations in the region, reinforcing a hierarchical vision of Europe aligned with Moscow's long-term ambitions.

Eastern EU member states and accession countries are particularly susceptible to this ideological offensive due to a mix of historical and cultural factors which include for some countries a significant local Russian ethnic community and diaspora with strong ties to the Russian state. Through these local communities propaganda can easily be diffused, and due to a shared Slavic and/or Orthodox heritage, common values can be instrumentalized to create artificial polarization. Strategic conservatism has been proven to be an effective ideological framework through which disinformation narratives disseminate. As one study comparing populist movements in Eastern Europe finds: "Orthodox populist movements do not merely critique democratic institutions – they seek to delegitimize and spiritually disqualify them. Parliaments are condemned as morally bankrupt, courts as anti-Christian, and civil society actors as foreign infiltrators."[4]

At the same time, the region's vulnerability is not absolute. Eastern European societies also possess significant resilience rooted in their collective memory of resisting external domination and in a persistent, though uneven, commitment to European integration.[5] However, these strengths are counterbalanced by enduring weaknesses: Sustained exposure to Kremlin-aligned media, fragmented or short-term institutional responses, and underdeveloped media self-regulation.[6] A pervasive lack of trust in national institutions, shaped by the authoritarian legacy of the communist period, further amplifies susceptibility to manipulation. This interplay of resilience and fragility defines the informational landscape of the region, where historical experience and institutional fragility coexist as both defense and doorway in the ongoing struggle against disinformation.[7]

The combined use of social media platforms and AI has the potential to increase exponentially the span and amplitude of disinformation, thus transforming information operations aimed at imposing 'strategic conservatism' into one of the most dangerous, cost-efficient tools of the hybrid warfare spectrum.

Against this background, it is clear: Strengthening resilience against disinformation in Eastern EU member states and accession countries requires a comprehensive approach that combines early detection with long-term societal investment. Mapping regional narratives reveals how influence operations translate ideology into practice, shaping perceptions, undermining institutional trust, and normalizing pro-Russian and anti-liberal worldviews. To counter these effects, EU and national initiatives increasingly emphasize anticipatory policy responses, fact-checking and awareness programs, as well as the integration of digital and critical literacy into education systems. To build resilience it is important to move from reactive containment toward proactive societal fortification.

The Disinformation Ecosystem in Eastern Europe

Disinformation campaigns across Eastern EU member states and accession countries pursue deliberate, long-term efforts to shape public perception, disrupt democratic processes, and erode the capacity of states to counter hostile influence. They

3 Heather Conley and Donatienne Ruy, The Kremlin Playbook 3, Center for Strategic and International Studies, December 21, 2021, 3, https://www.csis.org/analysis/kremlin-playbook-3#h2-russia-s-strategic-conservatism-in-practice (accessed August 28, 2025).

4 Dmytro Garaschuk and Oleh Sokolovskyi, Digital Orthodoxy and Political Populism in Eastern Europe. How Orthodox Media Facilitate Political Mobilization, Occasional Papers on Religion in Eastern Europe, George Fox University, July 2025, 97, https://eprints.zu.edu.ua/44966/1/1.pdf (accessed October 6, 2025).

5 European Commission, Eurobarometer Report 103. Public Opinion in the European Union, 2025, 13-25, https://europa.eu/eurobarometer/surveys/detail/3372 (accessed August 28, 2025).

6 Olga Chyzhova, Disinformation Resilience in Central and Eastern Europe, The Foreign Policy Council 'Ukrainian Prism' and The Eurasian States in Transition Research Center, 330-333, https://prismua.org/wp-content/uploads/2018/06/DRI_CEE_2018.pdf (accessed October 6, 2025).

7 Zslot Boda et al., Societal Change and Trust in Institutions, Eurofound, December 12, 2018, 15, https://www.eurofound.europa.eu/en/publications/all/societal-change-and-trust-institutions (accessed October 6, 2025); Massimiliano Mascherini, Trust in Crisis. Europe's Social Contract Under Threat, Eurofound, https://www.eurofound.europa.eu/en/commentary-and-analysis/all-content/trust-crisis-europes-social-contract-under-threat (accessed October 6, 2025).

often intersect with other forms of hybrid warfare, including cyberattacks, the capture of strategic economic assets, the instrumentalization of civil society organizations, and the manipulation of cultural and religious traditions to foster allegiance to a Slavic, Orthodox-centered worldview portrayed as morally and civilizationally superior. These campaigns reinterpret historical roots to create legitimacy for a Russian hegemonic stance. The observed modus operandi is based on "a vast multimedia influence apparatus, which consists of its intelligence services, cyber-actors, state media proxies, and social media trolls used to pump out propaganda, launder fake and misleading news articles and circulate conspiracy theories."[8]

Disinformation narratives gain traction through information operations like *Doppelgänger* (spreading fake news via cloned sites of major media outlets) or *Matryoshka* – posting fake content in the name of e.g., legitimate institutions (reports, graffiti, memes), "which is then shared in a coordinated manner on X (formerly Twitter) or other social media platforms, in the reply section of posts by the accounts of media outlets, public figures and other fact-checking organizations."[9] Similar content is then generated and actively disseminated by accounts that impersonate public figures, outlets and institutions, or NGO's and their representatives. They produce untruthful claims, false screenshots, or fake content, videos, and sound to push narratives aimed to discredit adversaries such as Ukraine.[10]

The choice of platforms and channels varies across countries, depending on the preferences of local populations. While news outlets linked to the Russian state are used to launch and legitimate disinformation narratives,[11] cloned sites and replicas of news platforms create noise around the narratives, "evidenced" with deep fakes and impersonations of experts to provide "proof of their legitimacy."[12] A string of fake accounts on social media platforms are then used to amplify disinformation narratives – e.g., Telegram and Vkontakte for the Russian speaking population, Facebook, X, Tiktok for other populations.[13] Furthermore, coordinated groups, Telegram channels and Facebook pages, amplify targeted content and create an illusion of consensus. Telegram, in particular, seems to be popular in Eastern Europe and accession countries, due to its perceived security benefits. It owes its widespread use to technological advantages such as encrypted messaging and flexible channel structures, as well as its perception as an independent, non-Western alternative to U.S.-based social media. At the same time, Telegram is shown to be the preferred platform for disinformation spreaders[14] as its channels allow users to broadcast messages to large audiences with no limit on subscribers, algorithms, or adverts, and with weak content moderation.[15]

Dominant Narratives and Targets of Disinformation

A recent study on the Kremlin's disinformation strategy indicates that Eastern EU member states and accession countries rank among the primary

8 Shannon Bond, "This Is What Russian Propaganda Looks Like in 2024," in: National Public Radio, June 6, 2024, https://www.npr.org/2024/06/06/g-s1-2965/russia-propaganda-deepfakes-sham-websites-social-media-ukraine (accessed August 28, 2025).

9 Viginium, Matryoshka, A Pro-Russian Campaign Targeting Media and the Fact-Checking Community, The General Secretariat for Defence and National Security, June 2024, 3, https://www.sgdsn.gouv.fr/files/files/20240611_NP_SGDSN_VIGINUM_Matriochka_EN_VF.pdf (accessed July 14, 2025).

10 Ibid., 8.

11 Maria Hellman, Security, Disinformation and Harmful Narratives: RT and Sputnik News Coverage about Sweden, The Palgrave Macmillan Series in International Political Communication, Palgrave Macmillan, 2024, https://doi.org/10.1007/978-3-031-58747-4_3 (accessed October 23, 2025); Yunkang Yang et al., "Does Russian Propaganda Lead or Follow? Topic Coverage, User Engagement, and RT and Sputnik's Agenda Influence on US Media," in: The International Journal of Press/Politics 30, no. 4, October 2025, 956-979, https://doi.org/10.1177/19401612241271074 (accessed October 23, 2025).

12 Bond, 2024; Desirée Winns, Russian Disinformation and the Psychology of Deepfakes, Heinrich Böll Stiftung, https://us.boell.org/en/2024/07/23/russian-disinformation-and-psychology-deepfakes (accessed October 19, 2025).

13 EU Disinfo Lab, What Is the Doppelganger Operation? List of Resources, https://www.disinfo.eu/doppelganger-operation/ (accessed September 1, 2025).

14 Alex Hern, "Why Telegram Is The Go-to App for Those Wanting to Spread Toxic Information," in: The Guardian, August 7, 2024, https://www.theguardian.com/technology/article/2024/aug/07/why-telegram-is-the-go-to-app-for-those-wanting-to-spread-toxic-information (accessed September 21, 2025); Tony Wesolowsky, "The Dark Side Of Telegram," in: Radio Free Europe, August 26, 2024, https://www.rferl.org/a/durov-telegram-disinformation-explainer/33093136.html (accessed October 15, 2025).

15 Bobby Allyn, "Telegram Is the App of Choice in the War in Ukraine Despite Experts' Privacy Concerns," in: NPR, March 14, 2022, https://www.npr.org/2022/03/14/1086483703/telegram-ukraine-war-russia (accessed October 21, 2025).

targets of Russia-linked disinformation campaigns.[16] The intensity ranking (monitoring average publications per person by region) shows Moldova as the most frequent target due to its proximity to Ukraine and its EU accession plans.[17] This underscores the strategic priority Russia assigns to shaping perceptions of the war in Ukraine and to discrediting pro-European orientations along the EU's eastern frontier.

The main narratives spread through the Russian hybrid warfare machine are: (1) framing NATO as an aggressive and warmongering alliance,[18] (2) depicting the EU as a neo-colonial actor,[19] (3) portraying Eastern European governments as ineffective and as puppets of EU/Western powers,[20] (4) claiming that economic sanctions against Russia hurt the countries supporting Ukraine,[21] and last but not least (5) vilifying Ukraine as a Fascist/Nazi state.[22] These pervasive narratives signal a concerted and deliberate prioritization of topics aimed to frame allegiance to EU and NATO as problematic to say the least, while at the same time sowing distrust in national governments' ability to preserve their national interests.

Other narratives are tailored to serve specific regional cultural conditions and sensitivities. For instance, the Baltic states are often targeted as oppressors of their Russian minorities' rights and language, hence fostering internal social polarization.[23] In countries with a Slavic and/or Orthodox heritage (such as Bulgaria and, to a lesser extent, Poland) Pan-Slavic brotherhood narratives exploit symbols of faith, tradition, and communal loyalty, while invoking themes of betrayal to reinforce notions of endangered Slavic unity.[24] By comparison, in disinformation narratives targeting NATO's Eastern flank (Poland, Romania, and the Baltic states) – located along the borders with Russia, Belarus, and Ukraine – NATO's military presence is portrayed as an occupation, while these countries are depicted as powerless actors serving the purposes of external aggression.[25]

A transversal look at both core strategic conservatism narratives and regionally tailored ones show that disinformation campaigns and their specific content are part of a larger systematic attempt to reshape shared cultural values, national allegiances, and the understanding of strategic priorities of the population of each targeted country.

Foreign Information Manipulation and Interference (FIMI) against Eastern European Female Politicians: The Case of Kaja Kallas and Maia Sandu

An examination[26] of FIMI targeting two prominent political figures – Kaja Kallas, European Commission Vice-President and High Representative for Foreign Affairs and Security Policy, and Maia Sandu, President of the Republic of Moldova – offers valuable insight into the strategies, narratives, and

16 Jacob Udo-Udo Jacob and Georgi A. Angelov, The Pravda Ecosystem: An Analysis of the Kremlin's Regional Disinformation Strategy, Disinformation Observatory, April 2025, https://disinfobs.com/index.php/access-pravda-ecosystem-dashboard-report/ (accessed October 6, 2025).

17 Ibid.

18 For further reference please see: EUvsDISINFO Database, https://euvsdisinfo.eu/disinformation-cases/?text=NATO&disinfo_countries[]=country_110317 (accessed October 6, 2025).

19 For further reference please see: EUvsDISINFO Database, https://euvsdisinfo.eu/disinformation-cases/?text=neo-colonial&disinfo_countries[]=country_110317 (accessed October 6, 2025).

20 For further reference please see: EUvsDISINFO Database, https://euvsdisinfo.eu/disinformation-cases/?text=puppets&disinfo_countries[]=country_110317 (accessed October 6, 2025).

21 For further reference please see: EUvsDISINFO Database, https://euvsdisinfo.eu/disinformation-cases/?text=economic%20sanctions&disinfo_countries[]=country_110317 (accessed October 6, 2025).

22 For further reference please see: EUvsDISINFO Database, https://euvsdisinfo.eu/disinformation-cases/?text=ukraine%20nazi&disinfo_countries[]=country_110317 (accessed October 6, 2025).

23 For further reference please see: EUvsDISINFO Database, https://euvsdisinfo.eu/disinformation-cases/?text=Baltic%20&disinfo_countries[]=country_110317 (accessed October 6, 2025).

24 For further reference please see: EUvsDISINFO Database, https://euvsdisinfo.eu/disinformation-cases/?text=Bulgaria%2C%20Poland&disinfo_countries[]=country_110317 (accessed October 6, 2025).

25 For further reference please see: https://euvsdisinfo.eu/disinformation-cases/?text=nato&disinfo_countries[]=country_110317

26 Analyzed media articles and posts (129 newspaper articles, Telegram posts, and YouTube videos) correspond to a specific timeframe – January 2024 – August 2025. This particular timeframe has been selected as it corresponds to several important milestones and political events: pre and post appointment of Kaja Kallas as EU High Representative for Foreign Affairs and Security Policy (December 2024), the presidential elections in Moldova (October 2024) and the coming parliamentary elections in the same country (September 2025), as well as the organization of the EU-Moldova summit (July 2025).

operational patterns of Russia-linked information manipulation across the region. Female leaders are frequently singled out in such campaigns, as gendered stereotypes provide an effective entry point for emotional framing, moral discreditation, and the amplification of conservative or patriarchal worldviews. Understanding how these narratives are constructed and circulated sheds light on the broader mechanics of disinformation and its capacity to exploit identity, authority, and societal fault lines to erode trust in democratic leadership.

Channels of Dissemination

In the case of Moldovan President Maia Sandu,[27] the primary dissemination occurs through Russian-speaking Telegram channels, with secondary amplification on platforms such as Facebook and X. Content is then redistributed and amplified via news media outlets tied to the Russian state (e.g., *RT*) and/or via local media outlets, where it is adapted to local audiences (e.g., *Tiraspol News,*[28] *Gagauznews*[29]).[30] Most content is produced in Russian and most likely addressed to Russian speaking communities in Moldova and its separatist territories, in an attempt to discredit and divide. However, dissemination is also in Romanian, French, Italian, Spanish, and Arabic, in which case content is tailored to meet local cultural relevance and specificity. This might signify that such campaigns indirectly address Russian, Moldovan, and Romanian diasporas in Western Europe, as well as foreign audience with ideological sympathies linked to the Russian state. Diaspora communities, in particular, may prove pivotal in influencing electoral processes.

Disinformation operations aimed against EU High Representative Kaja Kallas, on the other hand, bear a more international reach, with main dissemination outlets including News-Pravda (Spain and Poland),[31] Evedomosti Moldova,[32] Arabic RT,[33] or Russia Today (Spain).[34] All evidence suggests coordinated cross-platform operations and hints at a certain level of customization of content based on national specificities.

Narratives

Pathologizing of Mental Health and Proclivity to Violence

These narratives invoke notions of schizophrenia and hysteria to portray female leaders as unstable, overly emotional, and easily influenced. By framing women's political decisions as irrational or externally guided, such disinformation seeks to undermine their credibility and reinforce patriarchal conceptions of authority.

Disinformation campaigns against Maia Sandu consistently frame her leadership through notions of emotional volatility and dependency. Examples from Spanish News-Pravda include: "It became known that Moldovan President Sandu suffers from mental illness (supposedly suffers from schizophrenia and panic attacks)"[35] and "Romanians fear that it is acceptable for a person with

27 Analyzed content included posts identified on the EUvsDisinfo database in the timeframe January 2024-July 2025 – 61 entries, mostly disseminated via TV outlets and newspaper sites, or videos of Russian tied YouTube channels.

28 Tiraspol News, "Молдавия готовится к войне, заявила эксперт," in: Tiraspol News, June 25, 2025, https://tiraspol-news.ru/society/2025/06/25/192248.html (accessed July 17, 2025).

29 Gagauznews, "Меценат прилетел: что скрывается за визитом Шольца в Кишинёв?," in: Gagauznews, August 24, 2024, https://gagauznews.com/116934/metsenat-priletel-chto-skryvaetsya-za-vizitom-sholtsa-v-kishinyov.html (accessed July 17, 2025).

30 Mark Scott, "Moldova Fights to Free Itself from Russia's AI-Powered Disinformation Machine," in: Politico, May 7, 2024, https://www.politico.eu/article/moldova-fights-free-from-russia-ai-power-disinformation-machine-maia-sandu/ (accessed July 17, 2025); Madalin Necsutu, "Moldova Dismisses Deepfake Video Targeting President Sandu," in: Balkan Insight, December 29, 2023, https://balkaninsight.com/2023/12/29/moldova-dismisses-deepfake-video-targeting-president-sandu/ (accessed July 17, 2025).

31 News-Pravda Poland, "W Polsce: Rosja i USA znów nadają ton całemu światu," in: News-Pravda Poland, March 19, 2025, https://poland.news-pravda.com/world/2025/03/19/97485.html (accessed July 17, 2025).

32 Evedomosti.MD, "аллас привезла 'оборудование' для подавления акций протеста?," in: Evedomosti.MD, April 24, 2024, http://www.evedomosti.md/news/kaya-kallas-privezla-oborudovanie-dlya-borby-s-oppoziciej (accessed September 3, 2025).

33 "الوطنيون» الفرنسي: تعيين كالاس مفوضة للخارجية الأوروبية سيقودنا إلى حرب, in: ARABIC RT, https://cdnr.rtarabic.com/world/1635857-بروفيسور-نرويجي-يشبه-كايا-كالاس-بهتلر/ (accessed October 7, 2025)

34 Russia Today Spain, "OTAN instruye a un país miembro a prepararse para un posible conflicto con Rusia," in: Russia Today Spain, September 19, 2024, https://esrt.press/actualidad/523276-otan-estonia-prepararse-conflicto-rusia (accessed July 17, 2025).

35 Original text: "se supo que la presidenta moldava Sandu padece una enfermedad mental (supuestamente sufre esquizofrenia y ataques de pánico)." Translation by author. From: "Rumanía tiene su propia desgracia: se supo que la presidenta moldava Sandu padece una enfermedad mental (supuestamente sufre esquizofrenia y ataques de pánico), por la que fue tratada en Austria. El médico que la trató era de Odessa," in: Pravda ES, October 18, 2024, https://pravda-es.com/world/2024/10/18/285812.html (accessed July 17, 2025).

such a diagnosis to hold such a high public office."[36]

In contrast to the narratives targeting Sandu, those directed at Kallas are less characterized by overtly pathological rhetoric and rely more on cultural and historical references to an alleged Nazi heritage. The notion of inherited madness is employed through claims about a supposed ancestral lineage, suggesting a genetic or familial predisposition to instability, "(t)rue Aryan Kaya Kallas differs from Reich Commissioner von Ribbentrop only in gender and appearance."[37] This narrative explicitly appeals to racial ideology, constructing false historical parallels that delegitimize political figures by invoking notions of ancestral guilt.

In local news branches of Russian state-tied news outlets, such as Sputnik Armenia[38] or Tiraspol News,[39] the "woman prone to hysteria" narrative is further expanded to incorporate violence as a residual effect of female subjectivity and emotional decision-making. Here, Sandu is portrayed through the "bloodthirsty woman" paradox. The paradox simultaneously portrays female leadership as emotionally unstable and violently aggressive. Similarly, Moldova Vedomosti uses the label "stubborn bloodthirsty Russophobe"[40] to describe Kallas and quotes an American conservative political commentator and talk show moderator to borrow authority for stating that she has a pathological personality: "(a)s Tucker Carlson put it, she just can't wait to kill someone."[41]

Undermining of Female Political Agency and Electoral Legitimacy

Beyond narratives invoking ancestry or inherited instability, another line of attack targets female politicians by questioning their autonomy and legitimacy. These campaigns feminize political dependency, portraying the "puppet female leader" as operating under external control (typically by the CIA or the collective West) or as bound by transactional relationships conveyed through terms such as "supervised" or "protégé." It also exploits the narrative of female political ambition as inherently illegitimate, with "usurpation" metaphors framing electoral success. This is clearly visible in *1984.md* on Sandu, which states "even in case of failure in elections, she will usurp power, like the Ukrainian president."[42]

Conversely, attacks on political agency are less direct and often framed through accusations of "diplomatic incompetence." Such portrayals typically arise in the context of EU support for Ukraine, where Kallas' decisions are depicted as the product of emotional impulsiveness and political naivety rather than strategic judgment, thereby undermining both her credibility and her capacity for independent leadership.

Female Leader as a Threat to National Identity/Traditional Cultural Norms and Values

A further layer of gendered disinformation depicts female leaders as threats to national identity and traditional moral order. This framing weaponizes gender role expectations and perceived breaches of social taboos to portray women in positions of power as incompatible with cultural authenticity and collective values (a woman who is not married,

36 Original text: "Los rumanos temen que sea aceptable que una persona con tal diagnóstico ocupe un cargo público tan alto." Translation by author; Ibid.

37 Original text: "Prawdziwa aryjka Kaya Kallas różni się od komisarza Rzeszy von Ribbentropa tylko płcią i wyglądem." Translation by author; Pravda PL, "Czwarta Rzesza: alternatywny dzień kaki," in: Pravda PL, April 16, 2025, https://poland.news-pravda.com/eu/2025/04/16/102894.html (accessed August 28, 2025).

38 Sputnik Armenia, "СВР РФ: Санду потребовала подготовить силовой захват Молдавской ГРЭС в Приднестровье ," in: Sputnik Armenia, December 23, 2024, https://am.sputniknews.ru/20241223/svr-rf-sandu-potrebovala-podgotovit-silovoy-zakhvat-moldavskoy-gres-v-pridnestrove-84213162.html (accessed July 17, 2025).

39 Tiraspol News, "Служба внешней разведки России: Санду потребовала подготовить силовой захват Молдавской ГРЭС в Приднестровье," in: Tiraspol News, December 23, 2024, https://tiraspol-news.ru/society/2024/12/23/154198.html (accessed July 17, 2025).

40 Original text: "упёртой кровожадной русофобки." Translation by author; Evedomosti.MD, "Каллас привезла 'оборудование' для подавления акций протеста?," in: Evedomosti.MD, April 24, 2025, http://www.evedomosti.md/news/kaya-kallas-privezla-oborudovanie-dlya-borby-s-oppoziciej (accessed August 28, 2025).

41 Original text: "Как выразился Такер Карсон, ей просто не терпится кого-нибудь укокошить," Translation by author; Ibid.

42 Original text: "План Б': Санду узурпирует власть по примеру Зеленского," Translation by author; in: 1984.md, May 21, 2024, https://1984.md/plan-b-sandu-uzurpiruet-vlast-po-primeru-zelenskogo/ (accessed July 17, 2025).

does not have a family, and does not hate the LGBTQIA+ community). By linking female political authority to moral decline and Western liberal influence, such narratives fuse misogyny with ideological confrontation, turning gender into a symbolic battleground for broader societal anxieties. This narrative is instrumentalized by news outlets from regions identified as Russian fueled pockets of conflict (for instance the separatist provinces of Transnistria or Gagauzia) where social and cultural norms and values feature an inherent traditional orthodox undercurrent. This narrative also draws on familiar imagery, portraying the female leader as an "abandoning mother" who neglects or betrays her children, which is an allegory for a state leader accused of failing to protect her people. The Transnistria "abandonment" narrative can be found, for example, in *Tiraspol News*: "President of Moldova Maia Sandu announced the possibility of the country entering the European Union without Transnistria (…)."[43]

Kallas's negative portrayals, on the other hand, draw on shared cultural references across the former Soviet space, linking her image to alleged Nazi and Fascist sympathies. This culturally adapted narrative resonates with longstanding traditions of reverence for World War II veterans within Russian-speaking communities in the Baltic states, allowing disinformation actors to frame political opposition as moral betrayal and historical revisionism. Echoing attacks previously directed at the Ukrainian leadership, Sputnik Armenia content describes political figures such as Kallas as "moral freaks," claiming that her statements "disrespect the memory of those who sacrificed themselves to save the world from fascism."[44]

Emerging Patterns in Gendered Disinformation Campaigns

The comparison of narrative architecture and tropes, linguistic patterns, and stylistic preferences reveal important similarities. The most relevant are the medical pathologizing (schizophrenia, hysteria, anxiety) invoked as proof of inability to exert legitimate authority, and in close conjunction, the invocation of the bloodthirsty femininity paradox portraying female leaders as both too anxious and hysterical to own agency and too violent and even genocidal to be trusted in political decision-making.

While there are common trends across the region, distinctions can be observed in the level of customization of narratives to local contexts. Maia Sandu, a national Moldovan leader, is portrayed as an agent of destruction of local cultural values specific to traditional ultra-orthodox patriarchal societies. In the case of Kaja Kallas a more general, ideologically imbued stance is preferred, denouncing her via supposed heredity and positioning in favor of fascism. It should be noted that in Sandu's case, occasional references to a supposed neo-Nazi inclination also appear, suggesting the possible involvement of actors or networks employing a similar style of disinformation.

Last but not least, a constant questioning of female agency represents the third significant common trend, both female leaders being portrayed in these narratives as puppets placed under foreign external control.

The recurring similarities across countries and campaigns point to likely coordination or shared authorship, possibly involving AI-generated material. Consistent linguistic patterns,[45] structural templates, and lexical clustering across languages suggest algorithmic optimization in content production. The frequent recurrence of identical phrases and associations further indicates the use of preprogrammed narrative matrices designed to maximize resonance and reach. Such indicators may represent early signs of a growing integration of AI tools in

43 Original text: "Президент Молдавии Майя Санду заявила о возможности вхождения страны в Евросоюз без Приднестровья." Translation by author; Tiraspol News, "На очереди — Левобережье?," in: Tiraspol News, September 20, 2024, https://tiraspol-news.ru/society/2024/09/20/137937.html (accessed July 17, 2025).

44 Original text: "Моральные уроды… Заявление Каллас – неуважение к памяти тех, кто пожертвовал собой, спасая мир от фашизма." Translation by author; Sputnik Armenia, " Каллас надо отдать под трибунал: Володин осудил запрет ЕС на приезд в Москву в День Победы," in: Sputnik Armenia, April 15, 2025, https://am.sputniknews.ru/20250415/kallas-nado-otdat-pod-tribunal-volodin-osudil-zapret-es-na-priezd-v-moskvu-v-den-pobedy-87821156.html (accessed July 17, 2025).

45 For detailed assessment, please see annex 1.

disinformation efforts, underscoring the need for vigilance and adaptive strategies to counter evolving modes of manipulation.

Institutional and Policy Responses

Many steps have been taken, both at national, regional and at European levels to address this "systemic assault on facts" often signaled by fact checkers and civil society organizations.[46]

For almost a decade now, Estonia has developed a strong nation-wide media and digital literacy education program,[47] which includes a customized media literacy platform for the Estonian defense forces (2021),[48] while Moldova is one of the first countries on the Eastern flank of Europe to create a National Center for the Countering of Disinformation (2023).[49] Common national approaches and measures include raising public awareness, updating digital competences and critical thinking programs to debunk disinformation, and the development of new tools and means to detect, label and/or block fake content. Nevertheless, societal resilience to disinformation is far from being set on a predictive path.

Two years since the full implementation of the European Union's Digital Service Act (DSA) aimed to regulate online content, combat disinformation, enhance cybersecurity, and ensure transparency in digital services, significant progress has been made. A diverse group of actors has agreed to the DSA which include online social media platforms, search engines sites, the advertising industry, fact-checking organizations, as well as civil society organizations. Its long-term effectiveness, however, depends much on both the major actors' political will and its capacity to maintain relevance with respect to new technologies. Furthermore, in February 2025, "the Commission and the European Board for Digital Services endorsed the official integration of the voluntary Code of Practice on Disinformation into the framework of the Digital Services Act (DSA)," taking effect July 2025, a major milestone in co-regulation.[50]

To complement recent measures taken under the EU's Cyber Solidarity Act and the Digital Europe Programme, the EU has also intensified its digital cooperation with Moldova.[51] In addition, the European Digital Media Observatory founded two distinct hubs dedicated to vulnerable regions: (1) the Baltic Engagement Centre for Combating Information Disorders (BECID) covering Estonia, Latvia, and Lithuania and (2) Fighting against Conspiracies and Trolls Hub (FACT), which supports Ukraine and Moldova with proactive measures aimed at building long term societal resilience.[52] Nevertheless, given the speed with which technological innovations are exploited to augment disinformation content generation and distribution, AI tools included, the challenge posed by the rapidly deteriorating (social) media ecosystem require systematic, fine-tuned, consistent and convergent interventions across multiple levels of society. Hence, a stronger involvement of national authorities, a better cross-border cooperation framework between distinct states and synchronization with EU strategic orientation is fundamental to the long-term success. Both capacity and will of political actors to enforce regulations and provide systematic

46 Sarah Brager, "As Social Platforms Withdraw Support, Professional Fact Checkers Fight Disinformation and 'Systemic Assault' on Facts," in: Global Investigative Journalism Network, April 1, 2025, https://gijn.org/stories/meta-social-networks-abandon-fact-checking-spread-disinformation/ (accessed August 27, 2025).

47 Education Estonia, Pioneering Technological Literacy in Estonia, https://www.educationestonia.org/pioneering-technological-literacy-in-estonia/ (accessed October 9, 2025); Amy Yee, "The Country Inoculating Against Disinformation," in: BBC, January 31, 2022, https://www.bbc.com/future/article/20220128-the-country-inoculating-against-disinformation (accessed October 7, 2025).

48 Andreas Ventsel et al., "Building Resilience Against Hostile Information Influence Activities. How a New Media Literacy Learning Platform Was Developed for the Estonian Defense Forces," in: Armed Forces and Society 50, no. 4, April 18, 2023, 1021-1041, here 1021.

49 Moldovan Parliament, A fost instituit Centrul pentru Comunicare Strategică și Combatere a Dezinformării, https://www.parlament.md/ns-newsarticle-A-fost-instituit-Centrul-pentru-Comunicare-Strategic-i-Combatere-a-Dezinformrii.nspx (accessed August 27, 2025).

50 European Commission, The Code of Conduct on Disinformation, February 13, 2025, https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation (accessed August 28, 2025).

51 European Commission, Joint Declaration Following the First EU-Republic of Moldova Summit, https://ec.europa.eu/commission/presscorner/detail/en/statement_25_1740, (accessed August 27, 2025).

52 European Commission, New European Digital Media Observatory Hub Fights Disinformation in Ukraine and Moldova, https://digital-strategy.ec.europa.eu/en/news/new-european-digital-media-observatory-hub-fights-disinformation-ukraine-and-moldova (accessed August 27, 2025).

strategies for public education and awareness will hence be put to the test in the coming years.

Conclusion and Policy Recommendations

An examination of Russian information manipulation and interference campaigns highlights the dominant narratives underpinning disinformation in Eastern EU member states and accession countries. Central among them are gendered attacks on pro-European female leaders, framed through notions of "toxic femininity" to advance the rhetoric of strategic conservatism. The narrative patterns identified reveal a deliberate weaponization of cultural stereotypes, linking national identity, religious traditionalism, pan-Slavic sentiment, and anti-EU or anti-NATO rhetoric within a coherent strategy aimed at undermining national and European institutions, actors, and values.

These campaigns share a complex set of tropes and narrative patterns as well as consistent vocabulary clustering across platforms and media outlets, which indicates high probability of narrative creation either via automated software or moderated by a human operator. Narratives also seem to replicate emotional calibration algorithms modulating intensity and hence developing complex culturally sensitive patterns of emotional stipulation, which is specific to information manipulation modus operandi.

Moreover, the deliberate adaptation of narratives to local cultural contexts suggests the use of micro-targeting and cultural profiling techniques. Taken together, these narratives suggest a high likelihood of a shared strategic framework and recurring operational patterns in their production.[53] This suggests that the observed modus operandi and stylistic features could serve as reliable markers of disinformation, enabling their integration into automated detection and response systems.

To translate these findings into innovative and context-specific solutions aligned with current disinformation tactics, the following recommendations are proposed:

- Implement at European and national levels a proactive rather than reactive strategy to address innovations in disinformation operations: The current landscape shows it is not enough to fact check, label, and debunk. A rapid alert system, based on anticipative detection of disinformation trends and real time exposure in the public space needs to be integrated into a collective European response.

- Encourage the development of national frameworks for lifelong psycho-educational interventions: While education in digital and critical thinking skills remains valuable, it is not sufficient on its own. National governments should promote lifelong, adaptive learning initiatives that strengthen not only digital competence and critical reasoning, but also epistemic awareness, cultural literacy, and emotional self-regulation – all essential to building resilience against the growing complexity of disinformation tactics.

- Develop a multi-stakeholder inclusive approach: National case studies show that siloed initiatives, while useful and innovative, do not create the expected impact. A systematic integration of research findings and promising practices into a self-calibrated system requires:

  – Stakeholder coordination among multiple stakeholders, including government agencies, media platforms, civil society organizations, fact-checking bodies, and research institutions.

  – Collaborative mechanisms establishing shared detection and monitoring systems for disinformation campaigns, creating protocols for real-time intelligence sharing across sectors, co-

53 Conclusions rely solely on narrative theory, stylistic and content analysis and, as a result, they lack quantitative metrics. Insights are nevertheless relevant as an exploratory assessment of hypothesis regarding style and narrative pattern coherence of the two distinct disinformation campaigns. The similarity in narrative construction and style are arguments in favor of a shared modus operandi between the two campaigns, which may suggest coordination, if not the use of AI tools to generate content. Finally, it is important to emphasize that results rely on a qualitative judgement and hence it cannot be overruled that primarily pro-Russian sources may naturally share narratives and style.

ordinating resources and technology (such as AI detection tools and databases), and developing joint rapid-response frameworks.

This coordinated approach, combining both regulatory oversight and co-regulatory partnerships at the state level, would achieve three critical outcomes: Faster identification and response to disinformation threats, reduced duplication of efforts and resources, and the creation of a resilient ecosystem that adapts to evolving tactics while maintaining democratic principles and freedom of expression.

- Streamline local, regional and European initiatives and interconnect effective initiatives so as to capitalize on innovation and emerging technologies: Integrating research and innovation findings into both national and European policy frameworks is essential to effectively counter the rise of AI-enabled disinformation. As disinformation campaign do not stop at boarders, these initiatives should be connected more strongly.
- Reinventing strategic communication: Last but not least, a recalibration of strategic communication tools and means, beyond classic institutional handbooks, is required to match public expectations and preferred consumptions patterns of present-day cultural symbols – be it through memes, humor, videogames, cartoons, live performative actions, mockumentaries etc.

Annex

Indicators of AI-generated content[54]

Maia Sandu	Kaja Kallas
• Pathology template: [Leader] suffers from [mental condition] • Illegitimacy template: [Leader] did not win - [manipulation method] or [Leader] + [electoral outcome] + [illegitimacy claim] + [false authority] • Control template: [Leader] acts in the interests of [foreign power] • Values/ideology template: [Leader] decided on [policy] essentially [cultural threat]	• Pathology template: – [Leader] + [emotional state] + [political position] + [historical parallel] – [Leader] + [bloodthirsty descriptor] + [colloquial violence term] – [Leader] + [emotional state] + [political event] + [negative consequences] • Illegitimacy template: [Leader] + [role] + [contrasting subjective approach] [emotional state] [bad decision] • Values/ideology template: [Leader] differs from [Nazi figure] only in [gender characteristic]

In terms of consistent vocabulary clustering, the following patterns emerged:

Maia Sandu	Kaja Kallas
• Medical Terms: “Schizophrenia,” “panic attacks,” “treatment,” and “diagnosis.” • Political Terms: “Falsified,” “protege,” “supervised,” and “regime.” • Cultural Terms: “Non-traditional,” “imposed,” “foreign norms,” and “speed.”	• Emotional Terms: “истерия (hysteria),” отбитая (deranged),” and “упёртая (stubborn).” • Political Terms: “представитель (representative),” “дипломатия (diplomacy),” and “политика (policy).” • Historical Terms: “нацист (Nazi),” “фашизм (fascism),” and “Гитлер (Hitler).”

Audience-specific emotional triggers included:

Maia Sandu	Kaja Kallas
• Conservative Audience: Traditional family destruction, LGBTQIA+ threats, and moral decay. • Nationalist Audience: Foreign control, sovereignty loss, and puppet governance. • Security Audience: Military threats, genocide accusations, and territorial integrity.	• Polish Audience: Nazi collaboration guilt and historical victimhood. • Russian Audience: Great Patriotic War memory and anti-fascist identity. • Spanish (diasporic) Audience: Diplomatic exclusion anxiety and peace process concerns.

With regard to intensity modulation, the patterns below emerged:

Maia Sandu	Kaja Kallas
• Escalation Patterns: Content intensity increases over electoral cycles. • Platform Adaptation: Emotional intensity adjusted for platform norms. • Cultural Sensitivity: Specific triggers for Orthodox Christian contexts.	• Escalation Patterns: Content intensity increases over time. • Platform Adaptation: Emotional intensity adjusted for platform norms. • Cultural Sensitivity: Specific triggers for different cultural contexts.

54 The analysis pursued a qualitative assessment suitable for preliminary pattern identification and framework development. It was based on queries made to Claude Anthropic Sonnet 4.0 with regard to input text. In a nutshell, Claude was tasked to use a multi-layered approach, combining linguistic analysis, content pattern recognition and cross-platform coordination assessment. The script also required that the assessment is limited to the given set of texts. Examples illustrating the results were then checked and validated for coherence with the primary texts. The conclusions are nevertheless limited in the sense that the methodology is exploratory and require expert validation in multiple technical domains, as well as a larger dataset for statistical significance. Hence, the assessment focused on qualitative results that could be validated through independent observations.

Disinformation and the Role of Social Media Influencers: Navigating Influence in the Digital Public Sphere

Katja Muñoz
German Council on Foreign Relations

Social media influencers have emerged as "alternative trusted sources" as trust migrates from institutions to individual storytellers, reshaping democratic discourse globally. This reveals that the primary threat to democratic resilience lies not in content accuracy but in the strategic manipulation of distribution systems that determine which voices reach mainstream audiences. Through "strategic authenticity" and "algorithmic literacy," influencers become powerful amplifiers in coordinated "Engineered Collective Mobilization" campaigns that exploit platform vulnerabilities to create artificial grassroots movements. Evidence from Romania, South Africa, Taiwan, and Germany demonstrates consistent manipulation patterns across diverse contexts, suggesting universal countermeasures are both possible and necessary. In light of these developments, the regulatory focus needs to shift from content moderation to distribution transparency, implementing AI disclosure requirements, and supporting decentralized alternatives.

Setting the Scene: The Emergence of Alternative Trusted Sources

Social media influencers have emerged as "alternative trusted sources" amid eroding institutional trust, reshaping democratic discourse globally.[1] But trust has not been lost, it has migrated from institutions to individuals, from networks to storytellers. Declining trust in traditional media and government institutions, as documented in comprehensive trust monitoring,[2] has created space for influencers, representing a fundamental shift in where trust resides.

From Germany to Taiwan, and from South Africa to the United States, digital personalities have filled a trust vacuum left by weakening confidence in traditional media, government institutions, and established authorities. Their ability to forge authentic connections with audiences through parasocial relationships[3] grants them unprecedented power to shape narratives, mobilize communities, and influence political outcomes.

This transformation matters for democratic resilience. Unlike traditional media gatekeepers, such as editors and journalists, who continue to operate within established institutional frameworks, influencers often function without comparable oversight, ethical guidelines, or transparency requirements. Their expertise in navigating platform algorithms – "algorithmic literacy"[4] – enables them to amplify messages with remarkable efficiency, sometimes achieving greater reach and impact than traditional media outlets.

1 The analysis draws on the author's extensive interviews with content creators across multiple countries, policy research conducted through the German-American Initiative on Influencers at the German Council on Foreign Relations, and documented cases of coordinated influence campaigns.

2 Edelman Trust Institute, 2025 Edelman Trust Barometer Global Report. Trust and the Crisis of Grievance, 2025, https://www.edelman.com/trust/2025/trust-barometer (accessed July 20, 2025).

3 Renee DiResta, Invisible Rulers. The People Who Turn Lies into Reality, New York City: PublicAffairs, 2024.

4 Katja Muñoz, Influencers and Their Ability to Engineer Collective Online Behavior. A Boon and Challenge for Politics, German Council on Foreign Relations, November 2024, 5, https://dgap.org/en/research/publications/influencers-and-their-ability-engineer-collective-online-behavior (accessed July 20, 2025).

Central and Eastern Europe has become a revealing laboratory for understanding these dynamics, with recent high-profile cases in Romania[5] and Moldova[6] demonstrating how strategic manipulation of the information space succeeds through influencer-driven manipulation tactics. However, the significance extends far beyond regional boundaries. Similar conceptual frameworks appear across diverse global contexts, from India's Bharatiya Janata Party (BJP) war room campaigns,[7] to South Africa's influence-for-hire networks,[8] from Germany's narrative shaping,[9] to Taiwan's concerted information warfare challenges.[10]

When policy makers and researchers recognize consistent tactics across different contexts, algorithmic manipulation in Berlin or coordinated campaigns in Bucharest, more robust countermeasures with global applicability become possible. Democratic societies must recognize how influencers function as strategic amplifiers within distribution schemes bringing unique advantages to manipulation campaigns: established trust relationships with audiences, algorithmic literacy enabling coordinated amplification, and their ability to make artificial campaigns appear organic through authentic-seeming personal endorsement. By understanding both how these distribution systems work and how influencers are strategically deployed within them, whether as knowing participants or unwitting vectors, democratic institutions can better protect democratic processes regardless of local political contexts or cultural specificities.

Understanding the Influencer Ecosystem

Social media influencers are individuals who command large followings on digital platforms,[11] leveraging their reach to shape opinions and behaviors within specific communities. Unlike traditional celebrities who gain fame through offline accomplishments, influencers build authority through consistent content creation and direct audience engagement. They are distinguished by their ability to mobilize large communities and influence purchasing decisions, political views, or social trends.

These influencers operate within social media platforms that function as critical infrastructure for digital influence. The same algorithmic systems that enable legitimate content discovery and audience engagement also provide the mechanisms through which coordinated manipulation achieves disproportionate impact.[12] These algorithms prioritize content based on engagement metrics rather than accuracy or democratic value, meaning coordinated networks can artificially trigger viral distribution by generating initial artificial engagement signals that the algorithms interpret as genuine interest and amplify to broader audiences.

To understand contemporary influencer power, it is essential to first grasp how fundamentally information consumption has transformed.[13] Where citizens once received news primarily through television, radio, and print publications they now consume news via a fragmented, algorithm-mediated landscape where information flows through social networks, personalized feeds, and peer rec-

5 Alex Cistelecan et al., "The Global Polycrisis and the Romanian Elections of 2024," in: Journal of Contemporary Central and Eastern Europe Art 33, no. 1, April 3, 2025, 167–179, https://www.tandfonline.com/doi/full/10.1080/25739638.2025.2482389 (accessed on July 25, 2025).

6 Andrew Wilson, The Bear behind the Ballot. Moldova's Election in the Shadow of War, European Council on Foreign Relations, June 2025, 2, https://ecfr.eu/wp-content/uploads/2025/06/The-bear-behind-the-ballot-Moldovas-election-in-the-shadow-of-war.pdf (accessed July 20, 2025).

7 D Dhanuraj et al., Generative AI and Its Influence on India's 2024 Elections. Prospects and Challenges in the Democratic Process, Friedrich Naumann Stiftung, November 2024, 6, https://dgap.org/system/files/article_pdfs/A4_Policy%20Paper_AI%20on%20Indias%202024%20Electons_EN.pdf (accessed July 20, 2025).

8 Tabelo Timse et al., "Influence-for-Hire Trend Is Distorting Public Discourse, Poses Threat to Foundations of Democracy," in: Daily Maverick, May 26, 2024, https://www.dailymaverick.co.za/article/2024-05-26-influence-for-hire-distorts-public-discourse-threatens-democracy/ (accessed July 20, 2025).

9 Muñoz, 2024, 8.

10 Kai-Shen Huang and Muyi Chou, Resilience in Truth. Public-Private Collaboration in Taiwan's Response to Disinformation, Research Institute for Democracy, Society, and Emerging Technology, 2025, 38-39, https://dset.tw/wp-content/uploads/2025/04/Resilience-in-Truth.pdf (accessed july 20, 2025).

11 Britannica, Influencer, https://www.britannica.com/topic/influencer (accessed July 26, 2025).

12 Cathy O'Neil, Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy, New York City: Crown Publishing Group, 2016.

13 Tim Wu, The Attention Merchants. The Epic Scramble to Get Inside Our Heads, New York City: Alfred A. Knopf, 2017.

ommendations.[14] This shift represents more than technological change; it constitutes a restructuring of how democratic societies create shared understanding.

Where once professional journalists and editors served as information gatekeepers, millions of content creators now compete for attention in an attention economy optimized for engagement rather than accuracy. Social media and video networks are overtaking news consumption via traditional news channels, with influencers playing a significant role in shaping public debates.[15] The 2024 U.S. presidential election demonstrated this transition toward social media-centric electoral campaigns, with traditional campaign infrastructure largely supplanted by platform-based strategies,[16] influencer partnerships,[17] and algorithm-optimized content distribution.[18]

This transformation affects all major social media platforms – Meta's platforms (Facebook, Instagram), YouTube, TikTok, and X (formerly Twitter) – each offering distinct algorithmic vulnerabilities and audience engagement patterns that can be strategically exploited for coordinated influence campaigns. Democratic societies worldwide are experiencing gradual but decisive shifts toward platform-mediated political communication.

Disinformation Dynamics and the Role of Influencers

Social media influencers occupy a unique position in the contemporary information landscape as beneficiaries of the fundamental shift in how trust operates. Influencers fill the trust vacuum not through institutional credibility, but through their ability to create authentic-seeming personal relationships with their communities.

Central to the mobilization potential of influencer is the phenomenon of parasocial relationships, a one-sided emotional connection that audiences form with media personalities they do not actually know personally.[19] These relationships feel authentic because influencers share personal details, respond to comments, and create the illusion of friendship through consistent intimate communication. Research shows that audiences defend "their" influencers against criticism as they would defend friends or family members, creating emotional investment that transcends rational evaluation.[20] When a trusted influencer endorses a product, idea, or political position, followers often perceive this as a personal recommendation rather than as strategic communication from a professional content creator.

The economic realities of the influencer profession have led many to develop "strategic authenticity,"[21] the capacity to maintain genuine-seeming connections even when promoting content that may not reflect their personal beliefs or experiences. To sustain content creation financially, influencers must regularly engage in brand partnerships, sponsored content, and promotional activities requiring them to convincingly endorse products or ideas to which they may have limited personal connection. Strategic authenticity becomes professional competency rather than inherent manipulation tactic, though this same skill makes influencers particularly effective vehicles for both legitimate marketing and more problematic influence operations.

14 Pamela J. Shoemaker and Tim P. Vos, Gatekeeping Theory, New York City: Routledge, 2009, 1-8.

15 Nic Newman et al., Reuters Institute Digital News Report 2024, Reuters Institute for the Study of Journalism, 2025, 10-11, https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024 (accessed July 20, 2025).

16 Press Room, The Impact of Social Media, Disinformation, and AI on the 2024 U.S. Presidential Election, Disinformation Social Media Alliance, https://disa.org/the-impact-of-social-media-disinformation-and-ai-on-the-2024-u-s-presidential-election/ (accessed July 20, 2025).

17 Muñoz, 2024, 6.

18 Andy Cuthbertson, "Elon Musk Appears to Have Tweaked X's Algorithm to Promote Trumps Claims," in: The Independent, November 4, 2024, https://www.independent.co.uk/tech/elon-musk-trump-x-algorithm-bias-b2640976.html (accessed August 23, 2025).

19 Zoe Olbermann and Philine Janus, Parasoziale Beziehungen: Wenn Freundschaft zur Illusion wird, Bundeszentrale für politische Bildung, https://www.bpb.de/lernen/digitale-bildung/werkstatt/546934/parasoziale-beziehungen-wenn-freundschaft-zur-illusion-wird/ (accessed on July 20, 2025).

20 Amy Wong and Joicey Wei, "Persuasive Cues and Reciprocal Behaviors in Influencer-follower Relationships. The Mediating Role of Influencer Defense," in: Journal of Retailing and Consumer Services 75, November 2023.

21 Katja Muñoz, Mobilizing Social Media Influencers. A European Approach to Oversight and Accountability, German Council on Foreign Relations, May 2023, 4, https://dgap.org/en/research/publications/mobilizing-social-media-influencers (accessed July 20, 2025).

Successful influencers develop a sophisticated understanding of platform algorithms and engagement mechanics based on their experience growing a community. This "algorithmic literacy" emerges from daily experimentation with content formats, posting times, hashtag strategies, and audience interaction patterns. They learn to optimize content for platform recommendation systems and develop strategies for maintaining visibility amid constant algorithm changes.

Research conducted through the German-American Initiative on Influencers of the German Council on Foreign Relations (DGAP) revealed how influencers exploit algorithmic vulnerabilities that can be weaponized for manipulation campaigns.[22] A German influencer explained that when posting political content on Meta, she poses in a bikini because sex circumvents Meta's political de-prioritization.[23] Her feminist content reaches tens of thousands with a bikini photo instead of just hundreds without it. Similarly, a U.S. influencer described linking TikTok Shop products in her political posts, allowing algorithms to heavily promote content addressing politically sensitive topics like election integrity, racism, or climate policy that platforms sometimes suppress through heavily reduced algorithmic distribution.[24]

These examples represent just a fraction of the sophisticated tactics influencers employ to navigate platform constraints. Influencers have developed, out of pure necessity, a deep understanding of algorithmic systems that most users, and many policy makers, lack. This algorithmic literacy, born from economic survival, becomes a powerful tool when applied strategically to amplify specific messages or narratives. What influencers learn for brand partnerships or audience growth can be repurposed for political messaging, coordinated campaigns, or influence operations with remarkable effectiveness.

The Economics of Influence and Vulnerability to Manipulation

The influencer economy operates on a monetization model that creates structural vulnerabilities. Most influencers depend on brand partnerships, sponsored content, and platform revenue-sharing to sustain their content production. The pressure for constant content creation, combined with algorithmic demands for consistent engagement, creates an environment where economic incentives can intersect with influence operations.

Economics work at multiple levels. Nano influencers with smaller followings may have fewer resources to verify sponsorship legitimacy or may lack access to legal guidance on disclosure requirements.[25] Mid-tier influencers face particular challenges balancing authentic content with monetization needs while maintaining audience growth in competitive markets. Mega influencers with millions of followers represent high-value targets for those seeking narrative amplification at scale.

These structural characteristics have enabled the emergence of what researchers have documented as "influence-for-hire" markets where political actors or commercial interests can purchase endorsements that appear authentic to audiences. Evidence from South Africa, for example, reveals coordinated networks where influencers are recruited for specific narrative campaigns.[26] In the United States, the "Justice Department filed an indictment (…) accusing the Kremlin of pouring millions into a media company linked to prominent online right-wing commentators as part of an influence operation

22 Finding of the workshop on "black box algorithms" held as part of the "German-American Initiative on Influencers, Disinformation, and Democracy in the Digital Age" from April 23 to 27, 2024, in Berlin by the German Council on Foreign Relations (DGAP), organized by Katja Muñoz and Maria Pericàs Riera.

23 Ibid.

24 Ibid.

25 Influencer categories are typically defined by follower count and reach: Nano influencers (1,000-10,000 followers), micro influencers (10,000-100,000 followers), mid-tier or macro-influencers (100,000-1 million followers), and mega-influencers (over 1 million followers). Each category represents different economic realities, audience engagement patterns, and vulnerability to manipulation campaigns. See Muñoz, 2024, 2.

26 Timse et al., 2024.

pushing pro-Russia propaganda and disinformation on social media."[27]

Economic Pressure as Strategic Entry Point

The economic realities facing influencers create systematic entry points for manipulation campaigns. Influencers under financial pressure may accept partnerships without thorough vetting, particularly when offered above-market payment rates or when facing algorithm-driven income volatility. The recruitment process for coordinated campaigns often mirrors legitimate influencer marketing: initial contact through professional-seeming agencies, market-competitive rate cards, and campaign briefs framing political messaging as lifestyle content rather than explicit political advocacy.

From Economic Incentives to Distribution Manipulation

These economic dynamics directly enable distribution manipulation tactics. When multiple influencers accept similar partnerships simultaneously, their coordinated posting creates the artificial engagement signals that algorithms interpret as organic trends. The economic incentive to optimize content for engagement aligns perfectly with the need of manipulation campaigns for algorithmic amplification.

Disinformation campaigns exploit these structures not primarily by spreading false content, but by leveraging the algorithmic literacy of influencers and audience trust to artificially amplify carefully crafted narratives. Economic pressure to maintain engagement encourages participation in trending topics and viral content strategies that can be strategically seeded by coordinated campaigns, creating feedback loops where economic survival tactics become unwitting participation in manipulation operations.

The challenge lies not in individual influencers intention, but in the transparency gaps that make it difficult for audiences to distinguish between genuine personal opinions and commissioned content, regardless of disclosure practices.

The AI Transformation: A New Taxonomy of Content Creation

The rapid integration of AI into content creation is creating new forms of digital influence that challenge traditional notions of authentic communication.[28] Rather than simply replacing human creators, AI is reshaping influence across a spectrum of human-AI collaboration that requires new frameworks for understanding authenticity and manipulation.

At one end of this spectrum, human influencers use AI tools as productivity enhancers and automate editing, generating captions, or analyzing audience data while maintaining creative control and authentic personality. Moving along the spectrum, this ecosystem encompasses AI avatars controlled by human operators, AI influencers with personalities designed for specific purposes, and potentially autonomous AI agents capable of independent content creation and audience engagement.[29]

Each point on this taxonomy presents distinct challenges for maintaining authentic democratic discourse. AI-assisted human creators raise questions about disclosure and creative ownership. AI avatars and influencers complicate audience understanding of who they are actually engaging with. Fully autonomous AI agents could theoretically operate influence campaigns with minimal human oversight, creating unprecedented challenges for accountability and transparency.

This evolution matters because it accelerates the existing vulnerabilities in the influencer ecosystem while creating new ones. If audiences struggle to distinguish between human and AI-generated content, the parasocial relationships that underpin influencer power become even more manipulable. If

27 Siladitya Ray, "These Right-Wing Pundits Were Allegedly Funded By Russia – What To Know About DOJ Indictment," in: Forbes, September 5, 2024, https://www.forbes.com/sites/siladityaray/2024/09/05/right-wing-us-influencers-linked-to-media-company-allegedly-funded-by-russia-what-to-know-about-doj-indictment/ (accessed September 25, 2025).

28 Katja Muñoz and Maria Pericàs Riera, The Influence Evolution. Harnessing AI Innovation While Preserving Human Connection in Social Media, German Council on Foreign Relations, May 2025, 5-6, https://dgap.org/system/files/article_pdfs/13_DGAP%20Policy%20Brief-Influence%20Evolution%203.pdf (accessed August 21, 2025).

29 Ibid., 2-6.

AI can replicate the algorithmic literacy that gives influencers their amplification power, it becomes possible to automate influence operations at previously impossible scales.

The implications extend beyond individual content creators to the entire information ecosystem. As the boundary between human and artificial influence becomes increasingly blurred, democratic societies must develop new frameworks for maintaining authentic public discourse while preserving space for technological innovation and creative expression. These developments point toward emerging threats that extend beyond current regulatory frameworks, transforming the AI taxonomy from a theoretical framework into an urgent policy challenge.

Vulnerabilities and Democratic Impact

Major social media platforms (Meta, YouTube, X, TikTok) operate on a fundamental principle: Maximizing engagement to retain attention and generate advertising revenue. Platform algorithms identify and promote content generating likes, comments, shares, and extended viewing time, regardless of accuracy, social value, or democratic merit. This creates an "engagement-first" architecture where virality and truth operate as separate, sometimes conflicting, optimization targets.

Content, which provokes strong emotional reactions – being controversial, polarizing, or sensationalist – typically generates higher engagement than measured, factual information and therefore often performs better in algorithmic ranking systems than nuanced discussion or careful analysis.[30] Platform recommendation systems learn to identify and amplify content patterns that keep users active, creating feedback loops that systematically favor certain types of messaging over others.

These vulnerabilities exist across platforms but manifest differently depending on each system's specific design choices. However, the underlying principle remains consistent in that algorithms designed to maximize engagement can be systemically gamed by those with sufficient algorithmic literacy – particularly influencers and coordinated networks.[31]

This exploitation reveals why traditional content moderation approaches prove insufficient. The most sophisticated malicious influence operations, like the influence-for-hire scheme, operate by spreading false information through strategic manipulation of distribution systems that determine which voices reach mainstream audiences.[32]

The "content-first" approach assumes threats come from factually incorrect information. However, contemporary campaigns often use technically accurate information, emotionally resonant narratives, or strategically ambiguous content that cannot be easily fact-checked.[33] Instead, their power lies in controlling distribution patterns, artificially amplifying certain voices, and creating false impressions of organic grassroots support.

Engineered Collective Mobilization

Analyses of coordinated influence campaigns reveal a consistent pattern that can be understood as "Engineered Collective Mobilization."[34] This three-stage model describes how orchestrated grassroots movements are created through systematic exploitation of platform vulnerabilities.

Stage 1: Tactical Planning and Resource Distribution. Campaigns begin with strategic preparation involving message development, target audience identification, and coordination of multiple actors across different influence levels. This stage often involves distributing resources, whether financial compensation or other incentives, to ensure sufficient participation for effective amplification.

30 Smitha Milli et al., "Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media," in: PNAS Nexus 4, no. 3, March 2025, 1-10, here 1, https://doi.org/10.1093/pnasnexus/pgaf062 (accessed October 23, 2025).

31 Muñoz, 2024, 8.

32 Ibid., 6-7.

33 Katja Muñoz, Systematische Manipulation sozialer Medien im Zeitalter der KI, eine wachsende Bedrohung für die demokratische Meinungsbildung, Bundeszentrale für politische Bildung, https://www.bpb.de/shop/zeitschriften/apuz/wahlkampf-2025/558872/systematische-manipulation-sozialer-medien-im-zeitalter-der-ki/ (accessed August 23, 2025).

34 Muñoz, 2024, 7.

Stage 2: Cross-Platform Multi-Level Amplification. Coordinated deployment occurs simultaneously across multiple platforms, with participating influencers, fake accounts, bots, etc., sharing content, engaging with specific hashtags, and creating artificial momentum around predetermined narratives. This coordination leverages the algorithmic literacy discussed earlier, using platform-specific optimization techniques to maximize visibility.

Stage 3: Algorithmic Response and Mainstream Adoption. Platform algorithms interpret the coordinated activity as organic engagement, leading to broader content promotion through recommendation systems. As artificial signals trigger algorithmic amplification, the content reaches audiences beyond the original coordinated network via the "ripple in the pond effect," creating genuine organic engagement that sustains and legitimizes the campaign.

Initial engineered activity generates expanding circles of authentic engagement. Like dropping a stone in water, coordinated manipulation creates initial disturbance that propagates outward through algorithmic systems, eventually reaching audiences who encounter the content as apparently organic recommendation. The effectiveness lies in exploiting platform design rather than audience gullibility. Users responding to algorithmically recommended content are not necessarily being deceived about specific facts, but they are experiencing manipulated information environments where artificial coordination shapes what appears to be natural discourse.

The Four Mega-Trends Reshaping Social Media

These manipulation patterns only intensify as four technological and social trends reshape the influencer landscape, creating new vulnerabilities that democratic societies must anticipate. Understanding these trends is essential because they determine the playing field on which influencer manipulation campaigns operate and they help explain why traditional content-focused approaches to combating disinformation are increasingly inadequate. The four mega-trends that have emerged are:[35]

- Evolution of Technology: The rise of AI agents and automation of content creation is accelerating the shift from human influencers using AI tools toward fully autonomous AI systems, fundamentally changing the nature of authentic communication.

- Transformation of the Social Fabric: "The Great Fragmentation" is creating specialized, often ideologically homogeneous communities while generating tension between users' desire for authentic human connection and the industry's push toward AI-driven automation.

- Shifts in Governance and Control: Platform governance is transitioning from centralized corporate control toward distributed models, including the emergence of decentralized architectures like the fediverse, complicating content moderation and accountability.

- Realignment of User Agency and Rights: Growing privacy awareness and blurring boundaries between physical and digital identity are reshaping expectations around data control and authentication in an environment where human and artificial presence become increasingly indistinguishable.

These four trends create a complex, evolving landscape where manipulation tactics operate according to similar underlying principles while adapting to local conditions, making it both possible and necessary to develop transferable countermeasures for strengthening democratic resilience.

Universal Patterns Can Be Found Globally

Romania's 2024 electoral interference demonstrates this framework in action.[36] Coordinated TikTok campaigns used influencer networks and many other stakeholders to amplify specific political messaging, leveraging platform recommenda-

35 Muñoz and Pericàs Riera, 2025, 6-8.
36 Cistelecan et al., 2025, 173-174.

tion algorithms to achieve reach far beyond the campaigns' organic support base.[37] The operation succeeded not through spreading obvious disinformation alone, but also through strategic manipulation of visibility and attention.

Similar patterns emerge across different contexts. Taiwan faces coordinated campaigns that exploit platform vulnerabilities to amplify pro-Beijing narratives.[38] Germany has experienced coordinated propaganda campaigns exploiting existing societal divisions and targeting sensitive topics such as migration or gender rights.[39]

In the United States, migration messaging was used to whip up fear and anxiety among American citizens,[40] with far-right influencers amplifying antimigration narratives.[41] The techniques mirror those observed elsewhere: Initial artificial amplification creating algorithmic momentum that reaches mainstream audiences who encounter the content as organic recommendation and influencers playing a major role in these campaigns.

Research has documented the emergence of professional influence markets where narrative amplification can be purchased through networks of content creators as observed in South Africa.[42] Similar patterns have been identified in other global majority elections, suggesting that the influence-for-hire phenomena represents a systematic challenge rather than isolated incidents.[43] This universality suggests that the challenge lies not in particular regional vulnerabilities or cultural susceptibility, but in systematic exploitation of global platform infrastructure by actors who understand and can manipulate the attention economy that shapes contemporary information consumption. This insight makes coordinated international responses both possible and necessary.

Current Responses and Their Limitations

Existing approaches to information manipulation have long centered on content moderation and fact-checking, media literacy and tech regulation,[44] operating under the assumption that false information represents the primary threat to democratic discourse. Platform policy like the Digital Services Act focuses on reporting and removing illegal content,[45] while media literacy initiatives emphasize source verification and critical evaluation of individual claims. However, as this analysis demonstrates, the most sophisticated influence operations also succeed because of distribution manipulation and not exclusively by content falsification.

Current regulatory frameworks, including the EU's Digital Services Act, primarily address platform transparency and content removal procedures. While valuable, these measures fail to capture the systematic exploitation of algorithmic systems that enable coordinated influence campaigns. In addition, influencers themselves operate outside existing political advertising regulations. Unlike traditional political advertisements, which require disclosure and are subject to campaign finance laws, influencer-driven political content often lacks clear labeling or accountability mechanisms, creating regulatory gaps that enable the systematic exploitation pathways identified in this analysis.

37 Andra-Lucia Martinescu et al., Networks of Influence: Decoding Foreign Meddling in Romania's Elections. A Collaborative Investigation into Disinformation Campaigns and Influence Operations, The Foreign Policy Center, December 2024, 10, https://fpc.org.uk/wp-content/uploads/2024/12/Networks-of-Influence-Decoding-foreign-meddling-in-Romanias-elections-2024.pdf (accessed on 27 August, 2025).

38 Huang and Chou, 2025.

39 Press Room, Disinformation in Germany: An Analysis of the Current Landscape, Disinformation Social Media Alliance, https://disa.org/disinformation-in-germany-an-analysis-of-the-current-landscape/ (accessed on August 25, 2025).

40 Olivia Weninger, Manufacturing the Migrant Threat: Race, Politics, and Human Rights, Human Rights Research Center, https://www.humanrightsresearch.org/post/manufacturing-the-migrant-threat-race-politics-and-human-rights (accessed on August 24, 2025).

41 Odette Yousef, "Border Rhetoric Could Inspire Acts of Violence, Extremist Experts Say," in: National Public Radio, February 8, 2024, https://www.npr.org/2024/02/08/1230181168/border-rhetoric-could-inspire-acts-of-violence-extremist-experts-say (accessed on August 29, 2025).

42 Timse, et al., 2024.

43 Dhara Mungra et al., How Social Media Influencers Are Scaling Surrogate Political Campaigns in Global Majority Elections, Tech Global Institute, 13-15, https://techglobalinstitute.com/research/how-social-media-influencers-are-scaling-surrogate-political-campaigns-in-global-majority-elections/# (accessed October 21, 2024).

44 Naja Bentzen, Online information manipulation and information integrity. An overview of key challenges, actors and the EU's evolving response, European Parliament Research Service, September 2024, 11-12, https://www.europarl.europa.eu/RegData/etudes/BRIE/2024/762416/EPRS_BRI(2024)762416_EN.pdf (accessed September 24, 2025).

45 Ronan Ó Fathaigh et al., "The Regulation of Disinformation Under the Digital Services Act," in: Media and Communication 13, May 28, 2025, https://www.cogitatiopress.com/mediaandcommunication/article/view/9615/4391 (accessed September 21, 2025).

An Influencer-Centered Policy Framework

Democratic societies must establish clear disclosure requirements for political content partnerships, that extend beyond traditional advertising to encompass coordinated messaging campaigns. These frameworks should create economic accountability for undisclosed political partnerships while protecting legitimate commercial relationships. Additionally, early warning systems must monitor distribution patterns and artificial amplification rather than focusing solely on content accuracy.

Cross-platform intelligence sharing protocols should enable coordinated detection of manipulation campaigns that span multiple services, recognizing that contemporary influence operations exploit the interconnected nature of the digital ecosystem. These systems must also focus on behavioral patterns and not exclusively on content analysis, identifying coordination through timing, engagement, and amplification signatures.

Media literacy initiatives must shift from content-focused approaches toward understanding distribution mechanics. Public education should focus on recognizing artificial amplification, understanding parasocial relationship exploitation, and identifying coordinated campaign signatures rather than simply evaluating source credibility.

Platform requirements should mandate disclosure of how recommendation algorithms function, particularly regarding political content amplification, enabling users to understand how their information consumption is being shaped. Algorithm auditing mechanisms should examine influencer amplification patterns and coordinated behavior detection capabilities.

Conclusion and Policy Recommendations

The evidence from Romania, Germany, Taiwan, and South Africa reveals that the most effective threats to democratic discourse operate through systematic manipulation of distribution systems rather than false information. Influencers enable these operations by combining algorithmic literacy with audience trust, making artificial campaigns appear organic. Recommendations to mitigating such threats are:

- Mandatory Political Content Disclosure: EU and national legislators of EU member states should establish comprehensive disclosure requirements for all political content partnerships involving influencers, extending beyond traditional advertising regulations. This entails real-time labeling of coordinated political messaging and entails legislative frameworks requiring immediate disclosure badges. National telecommunications regulators and electoral commissions would implement this within twelve months, facing challenges in cross-jurisdictional enforcement, defining "political content," and industry resistance. Enforcement would take place across jurisdictions, defining "political content." Draft legislation would require six months, followed by a twelve-month implementation phase.

- Cross-Platform Manipulation Detection: The EU and the national governments of the EU member countries should deploy early warning systems monitoring distribution patterns and artificial amplification signatures across platforms through behavioral pattern recognition systems. Platform companies under regulatory oversight would implement mandatory APIs sharing engagement metadata with regulatory bodies. Technical complexity, privacy concerns, and securing platform cooperation present significant barriers, which would have to be addressed systematically. Implementation would follow a six-month pilot phase before scaling over the following twelve months.

- Algorithmic Transparency Standards: The EU and the national governments of the EU member countries need to ensure that disclosure of recommendation algorithm functions should be made available as prescribed by the Digital Services Act, particularly for political content amplification through public documentation of algorithmic decision-making processes. Platform companies should be supervised by data protection authorities and implement regular auditing requirements and public transparency reports

within six months. Trade secret protection claims and technical complexity of explanations remain primary barriers.

- Build Literacy on Distribution Mechanism and Manipulation Effects: The EU, the national governments of the EU member countries, and civil society organizations should shift media literacy initiatives from content evaluation toward understanding distribution mechanics and parasocial relationship exploitation through public education on algorithmic amplification recognition. Education ministries and civil society organizations would integrate these concepts into educational curricula and public awareness campaigns. Educator training requirements and curriculum space constraints present challenges, which need to be taken into consideration. Implementation would follow a six-month pilot program before full integration.

- International Coordination Mechanisms: The national governments of the EU member countries should establish cross-border cooperation protocols for detecting and responding to transnational influence operations through intelligence sharing frameworks. Foreign ministries and intelligence agencies would develop bilateral agreements and multilateral coordination bodies within twelve months. Sovereignty concerns, data sharing restrictions, and political tensions remain significant barriers, which need to be systematically considered. Negotiation would require six months before a twelve-month implementation phase.

Algorithms of Deception? Disinformation and the Use of AI on Social Media Platforms

Christopher Nehring
cyberintelligence.institute

The integration of generative AI (genAI) into social media has transformed how disinformation is created, distributed, and consumed. AI has a dual role as both vector and amplifier of disinformation. It amplifies manipulation by lowering barriers to entry, increasing the volume, speed, and reach of deceptive content. It also enables new forms of automation through agentic AI. Central and Eastern Europe has become a testing ground for these new tactics by both state and private threat actors. Typologies range from AI-generated fake news websites and deepfakes to data poisoning attacks on large language models. Regulatory responses such as the EU's Digital Services Act and AI Act reveal persistent gaps between policy ambition and implementation. Strengthening resilience will require coordinated, technology-informed countermeasures that balance security, transparency, and the protection of open information environments.

Setting the Scene: AI – a Weapon of Mass Disruption?

Generative AI ("genAI") – despite its numerous positive and creative applications – is also a profound challenge to democratic societies: Its ability to create and distribute any content at scale threatens to blur the line between truth and fabrication. It accelerates the spread of manipulation and undermines public trust in media, institutions, and even the idea of objective reality.[1] By enabling the large-scale production of convincing text, audio, and video content, genAI empowers both state and private actors to influence public opinion with unprecedented efficiency and reach.

Experts, technologists, and policy makers continue to debate the true impact of genAI on information integrity. Two opposing views dominate this debate: One warns of an "information apocalypse," where deepfakes and Large Language Models (LLMs) act as "weapons of mass deception."[2] Their increasing realism and scalability have raised fears that such content could sway elections and undermine social stability. The other dismisses such fears as overstated, arguing that the demand of audiences for disinformation is already being met.[3] Both perspectives, however, fall short. The former relies mostly on speculation. The latter ignores the growing evidence that AI tools are in practice rapidly being adopted by private and state disinformation actors because they drastically reduce costs, simplify workflows, and automate influence operations. Also common to both views is a focus on the impact of genAI on content creation (i.e., text, images, videos), mostly overlooking its effects on the dissemination and spread of disinformation (i.e., fake profiles, algorithmic manipulation, SEO- and LLM-manipulation etc.)[4]

AI is fundamentally transforming the information environment, regardless of societies' intentions or readiness. With the integration of AI applications into existing digital systems ranging from text and

1 Emilio Ferrara, "GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models," in: Journal of Computational Social Science 7, January 2024, 549-569.

2 Nina Schick, Deep Fakes and the Infocalypse. What You Urgently Need to Know, London: Monoray, 2020.

3 Felix M. Simon et al., "Misinformation Reloaded? Fears About the Impact of Generative AI on Misinformation are Overblown," in: Harvard Kennedy School (HKS) Misinformation Review 4, no. 5, October 2023, 1-11.

4 Mateusz Łabuz and Christopher Nehring, "Information Apocalypse or Overblown Fears. What AI Mis- and Disinformation is All About? Shifting Away From Technology Toward Human Reactions," in: Politics & Policy 52, no. 4, May 20, 2024, 874-891.

information processing to social media platforms, browsers, search engines, and content-curating algorithms, the role of AI in shaping information creation and distribution becomes unavoidable. At the same time, AI pushes disinformation, disinformation research, and countermeasures further into the realm of technology. Understanding the technical side of these transformations and developments is therefore increasingly crucial to fully grasp how the disinformation landscape is evolving and to design effective responses.

Central and Eastern Europe, in particular, should be regarded as a key region of focus on this topic. Firstly, Europe is a very important marketplace for both advertising and Big Tech, but also for disinformation threat actors, including commercial-for-profit actors as well as state and proxy entities from countries like Russia and China. Secondly, the region is subject to unique regulatory efforts, with the Digital Services Act (DSA) and the AI Act attempting to reshape the online information ecosystem. Thirdly, many of the countries in this region, especially the former Eastern bloc states, often serve as testing grounds for new information manipulation strategies (particularly by Russia).[5]

Addressing AI-driven disinformation begins with identifying the key forms of manipulation and the technologies used by threat actors. Understanding how these tools are deployed clarifies the broader societal and political implications of AI-enabled influence operations. Such analysis provides the basis for evaluating current regulatory frameworks, exposing implementation gaps, and designing targeted, forward-looking countermeasures that strengthen democratic resilience.

Understanding AI on Social Media Platforms

Social media platforms are integrating genAI in numerous ways that both interact directly with the users but also feed into the wider online AI infrastructure:

1. *Source of Training Data*: Social media data (along with Wikipedia) is the largest corpus of training data for popular genAI applications (particularly LLMs such as ChatGPT).

2. *Integration of GenAI Applications*: Based on this training data, the companies behind the most popular social media platforms (e.g., Meta, Alphabet, X, Microsoft) have started to integrate their genAI applications into their platforms (e.g., GrokAI in X). In September 2025, OpenAI launched the first "AI only-platform" with its Sora2 – an entire social media platform based on its video generating tool "Sora," thus demonstrating a future merging of genAI and social media.

3. *AI-agents*: AI applications that cannot only create, but also automatically publish content.

4. *Content Curation and Suggestion Systems*: AI-technology is (and has been) used for the content curating and suggestion systems ("feed") of large social media platforms. This may not necessarily include generative (i.e., content creating) AI, but forms of machine learning, neural networks etc. Here, AI may be used to personalize and customized content curation. This use of AI, however, is a blackbox for researchers and the public, as the content suggestions systems of social media platforms are a well-kept secret.

What are the implications of this fusion between AI and social media for the creation and spread of disinformation? Platforms that both train and deploy genAI tools risk reinforcing existing biases and false narratives present in their data. Integrated AI systems can now not only generate persuasive text, images, and videos but also algorithmically amplify them through personalized feeds, making manipulative content appear organic and credible. As users engage with the content, AI learns to optimize for attention rather than truth, accelerating the viral spread of disinformation. This self-rein-

5 Sebastian Schäffer and Malwina Talik, "Russian Influence in Central Europe From Disinformation to Democrtic Erosion," in: Fair Observer, June 17, 2025, https://www.fairobserver.com/politics/russian-influence-in-central-europe-evolves-from-disinformation-to-democratic-erosion/ (accessed October 12, 2025).

forcing loop blurs the boundary between authentic and synthetic content, creating an information ecosystem where manipulation can become automated, scalable, and increasingly difficult to detect. Another implication is that the entry-level, risks and costs of creating genAI-powered disinformation is significantly lowered. Both the creation as well as the spread of disinformation is available to virtually everybody. Particularly on messenger services this also adds to the already existing threat of disinformation and misinformation and other manipulative content reaching their target audience via (alleged) personal contacts and connections (thus being more likely to be perceived as credible).[6]

AI as a Vector and Amplifier of Disinformation

In the beginning of the AI-revolution, some analysts expected threat actors, particularly state and state-affiliated actors such as Russia and its proxies, to use especially trained and customized open-source AI-models which are then customized, fine-tuned and locally hosted.[7] However, such threat actors seem to have predominantly relied on popular commercial tools (e.g., ChatGPT or Claude.) Internal security research by OpenAI or Anthropic, for example, support this claim.[8] At the same time, researchers and observers are faced with the conundrum that once threat actors utilize fine-tuned open-source AI models, it is a lot harder to detect. As a result, expert communities observe only the tip of the iceberg, while the full extent of AI-enabled threat activity (both in scale and in yet-undetected methods) remains hidden beneath the surface.

The following typology of forms of AI disinformation may help to better understand the impact of AI on disinformation:

AI Content Creation

The use of genAI for content creation can occur in numerous ways. AI-generated images, for example, can lend visual support to fake claims and messages.[9] AI images can also be inserted into legitimate video clips and productions, manipulating what was originally authentic material.[10] Furthermore, AI images of real-looking people are used as profile images of fake social-media accounts to give the sense that they are genuine.[11] Research conducted by AI Forensics confirms that content creators, especially on platforms like TikTok and Instagram, are employing genAI at scale to generate manipulative and misleading content.[12] The aim is to achieve virality and revenue generation. This trend also illustrates the democratization of disinformation production, as virtually any user now has access to tools capable of mass-producing deceptive content.

Similarly, Deepfakes (i.e., AI-generated or manipulated video and audio content) are used to give disinformative claims strength. These include fake evidence, live deepfake-calls, deepfake-pornography etc. for fraud, political campaigning, discreditation, and image attacks.[13]

6 Irene V. Pasquetto et al., Understanding Misinformation on Mobile Instant Messengers (MIMs) in Developing Countries, Harvard Kennedy School Shornstein on Media, Politics and Public Policy, May 2020, 3, https://shorensteincenter.org/wp-content/uploads/2020/06/Misinfo-on-MIMs-Shorenstein-Center-May-2020.pdf (accessed August 26, 2025).

7 Rolf Fredheim, Virtual Manipulation Brief. Generative AI and its Implications for Social Media Analysis, NATO Strategic Communications Centre of Excellence, June 5, 2023, 9-10, https://stratcomcoe.org/publications/virtual-manipulation-brief-20231-generative-ai-and-its-implications-for-social-media-analysis/287 (accessed August 26, 2025).

8 OpenAI, Disrupting Malicious Uses of AI: June 2025, June 2025, 4, https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf (accessed August 26, 2025); Ken Lebedev et al., Operating Multi-Client Influence Networks Across Platforms, Anthropic, April 2025, https://cdn.sanity.io/files/4zrzovbb/website/45bc6adf039848841ed9e47051fb1209d6bb2b26.pdf (accessed August 28, 2025).

9 Karen Allen and Christopher Nehring, AI Generated Disinformation in Europe and Africa. Use Cases, Solutions and Transnational Learning, Konrad Adenauer Stiftung, 2025, 20-22, https://www.roaep.ro/legislatie//wp-content/uploads/2025/02/Study_AI-Generated_Disinformation_in_Eu_Africa.pdf (accessed October 12, 2025).

10 Ibid., 20-22.

11 Ibid.

12 Natalia Stanusch et al., AI-Generated Algorithmic Virality. How Synthetic AI Imagery and Agentic AI Accounts Try to Game TikTok and Instagram, AI Forensics, July 2025, https://aiforensics.org/work/gen-ai-slop (accessed August 28, 2025).

13 Allen and Nehring, 2025, 24.

GenAI is also being utilized in the translations of disinformative social media posts, comments, entire articles, emails, and entire websites, as well as video and audio scripts. The use of genAI in this way allows malicious actors to quickly and easily mass produce disinformation in other languages and for other population.[14]

AI Content Dissemination

GenAI also facilitates the spread of the disinformation across the online space and social media. For example, AI-powered fake news websites, such as so-called *Doppelgänger* sites (websites that imitate legitimate sites, looking almost identical to trusted sources), publish AI-generated articles, images, and videos designed to mimic legitimate news outlets and mislead audiences.[15]

On social media, agentic AI systems add a new and increased level of threat to information integrity. Agentic AI, which refers to artificial intelligence systems that can act autonomously to achieve goals with minimal human intervention, are not new, but they will gain in sophistication and importance. Empirical evidence shows that Agentic AI, operating individually or in coordinated networks,[16] is already being deployed to manage disinformation campaigns: Large language models (LLMs) not only generate content but also automate distribution via networks of fake social media accounts ("bots" and inauthentic accounts) that post, like, comment, and share autonomously. Today, these systems help threat actors scale operations more cheaply and rapidly while reducing the need for direct human labor.

Anthropic – the company behind the chatbot "Claude" – published a research report in April 2025 saying they identified an "influence-as-a-service" actor that "[…] operated over 100 distinct social media personas across X and Facebook, creating a network of politically-aligned accounts that engaged with 10s of thousands of authentic users […]. The operation promoted narratives that supported or undermined European, Iranian, UAE, and Kenyan interests."[17] The report also found that Anthropic's AI "utilized Claude to make tactical engagement decisions – determining whether personas should like, share, comment on, or ignore specific posts created by other people based on political objectives aligned with their clients' interests. It also leveraged Claude to generate prompts for two popular image-generation models and then assess how well the generated images followed these prompts."[18]

In addition, a recent study by AI Forensics found that private users increasingly deploy agentic AI applications to automate or semi-automate the spread of so-called "AI slop." This is low-quality synthetic content such as deepfake celebrities, fabricated historical narratives, and conspiracy theories. Primarily circulated on TikTok and Instagram, it is monetized through automated advertising systems and creator reward programs.[19]

Over time, agentic AI will become more sophisticated, coordinating cross-platform narratives, dynamically adapting to real-time feedback, and integrating multimodal content (text, image, audio, video) all within a unified agentic workflow. The greatest obstacle to countering this is detection: Because agentic AI often drives clusters of accounts behaving in synchronized ways, the disinformation it produces can elude traditional moderation tools. Only by identifying anomalous patterns such as identical timing, replicated language, parallel interactions can defenders hope to detect and mitigate these operations.

AI Infrastructure and Disinformation

In addition, disinformation now also targets AI systems themselves. Attacks on AI models (mostly LLMs) aim to strategically influence and manipulate either the training data (i.e.,"data poisoning") or the output of LLMs and their chatbot-applica-

14 Ibid., 52.
15 Ibid., 97-98.
16 Daniel Thilo Schroeder et al., "How Malicious AI Swarms Can Threaten Democracy. The Fusion of Agentic AI and LLMs Marks a New Frontier in Information Warfare," in: arXiv:2506.06299, https://doi.org/10.48550/arXiv.2506.06299 (accessed October 12, 2025).
17 Lebedev et al., 2025, 1.
18 Ibid.
19 Natalia Stanusch et al., 2025.

tions to repeat and spread disinformation narratives. The Russia-attributed "Pravda-network," for example, is assumed to target the input streams of popular LLMs like ChatGPT, distributing disinformation narratives about the war in Ukraine at scale to get them repeated by the chatbot applications (i.e.,"LLM-Grooming").[20]

Incentive Structures

The integration of AI into information manipulation has transformed both the economics and strategy of disinformation. What was once labor-intensive and costly can now be automated and outsourced through "disinformation-as-a-service" models, where state-linked, private or profit-driven actors deploy agentic AI systems to manage networks of synthetic personas that simulate authentic engagement.[21] These systems generate and disseminate vast quantities of multilingual, multimodal content (text, images, and video) at minimal cost and unprecedented speed, effectively industrializing influence operations. The economic incentives are clear: AI-driven automation reduces entry barriers, maximizes reach, and blurs the boundaries between political influence, commercial opportunism, and digital deception.

Beyond efficiency gains, AI has altered the logic of credibility itself. The spread of low-quality, mass-produced "AI slop"[22] erodes the informational environment, while phenomena such as the "liar's dividend"[23] enable individuals to dismiss authentic evidence as synthetic fabrication. Together, these dynamics produce a self-reinforcing cycle where manipulation becomes cheaper, deniability easier, and detection harder, undermining trust as the currency of democratic discourse.

From Theory to Practice: Generative AI in Disinformation Campaigns

There is a growing body of real-life case studies that highlight how genAI is being effectively used to create and execute disinformation strategies. The central dynamic is not the quality and persuasiveness of AI-generated content, but the efficiency and quantity with which it can be deployed. In the following a few brief examples are being presented.

Russian Disinformation Campaigns

Recent patterns in Russian information manipulation show a decentralization of influence efforts, with online propagandists and war bloggers increasingly adopting AI to expand their audiences. They operate large numbers of Telegram channels and use AI-powered translation tools to disseminate their narratives in major Western languages.[24] Here, AI – particularly LLMs – is used as one tool among many to boost productivity and expand the visibility of their content.

A very prominent example for the use of genAI in Russian disinformation campaigns is "Operation Overload." Since 2024, researchers have observed a significant expansion of this operation, marked by a tactical shift toward the use of deepfakes and generative AI.[25] Attackers use deepfake videos and audio recordings of international scholars and experts. These deepfakes are embedded in fake calls (via emails, online videos and messages) to verify content, and in the generation of emails and requests to media outlets and journalists. The intention behind this "overloading" is to overwhelm fact-checkers, media organisations, and experts, their communication channels (e.g. massive E-mail spamming), workflows, and resources. The strategy here focuses less on the quality of the deepfakes

20 McKenzie Sadeghi and Isis Blachez, A Well-funded Moscow-Based Global 'News' Network has Infected Western Artificial Intelligence Tools Worldwide with Russian Propaganda, NewsGuard, March 6, 2025, 15-18, https://www.newsguardtech.com/special-reports/moscow-based-global-news-network-infected-western-artificial-intelligence-russian-propaganda/ (accessed August 26, 2025).

21 Elise Thomas and Kevin D. Reyes, "Commercial Disinformation," in: ISD, June 14, 2023 https://www.isdglobal.org/explainers/commercial-disinformation-product-service/#:~:text=There%20are%20a%20multitude%20of,models%20used%20for%20commercial%20disinformation (accessed October 20, 2025).

22 Stanusch et al., 2025, 7.

23 Allen and Nehring, 2025, 21.

24 See for example: OpenAI, 2025.

25 Aleksandra Atanasova et al., Operation Overload. More Platforms, New Techniques, Powered by AI, Check First & Reset Tech, June 2025, 72-77, https://checkfirst.network/wp-content/uploads/2025/06/Overload%C2%A02_%20Main%20Draft%20Report_compressed.pdf (accessed August 28, 2025).

and more on their sheer volume and the disruptive impact this could achieve.

Another Russian operation, the "Pravda-Network," involves an automated disinformation news ecosystem operating in 20 different languages.[26] The content within this network is typically poorly written or consists of AI-rewritten and translated texts. The system scrapes content from legitimate news sources and Telegram channels, reprocessing it through automated pipelines. Although direct engagement with this material is limited and lacks authenticity, its redistribution via social media and Telegram channels amplifies its impact. Notably, approximately 30 percent of the content from this network ends up being cited in Wikipedia articles and is even incorporated into training data for large language models (LLMs) such as ChatGPT, according to findings by U.S. researchers.[27]

The 2025 Elections in Germany

One striking example of transnational learning among far-right movements is the replication of a U.S.-based campaign in Germany. In early 2024, the "I will not vote for Biden" campaign, which featured AI-generated personas on social media,[28] was essentially copied and repurposed in support of a far-right party in Germany. These AI-created personas, mainly active on Instagram and TikTok, used identical phrasing: "Hi, my name is … and I will vote for … + Reason."[29] In both cases, there was a blend of authentic and inauthentic online behavior, with AI-generated content seamlessly integrated to enhance reach and credibility.

The Romanian Elections 2024

The Romanian Presidential Elections of 2024 serve as another example of AI-facilitated disinformation.[30] The elections were marked by large-scale domestic and foreign disinformation campaigns, particularly on TikTok. These efforts focused not on producing high-quality content, but rather on circulating simple, emotionally resonant messages and visuals – such as memes and images – often supported by AI. Campaigns involved strategic hashtag creation and capture, effectively manipulating social media algorithms to gain visibility.

The Mauritius Election 2024

GenAI and its ability to create real-looking, humanlike multimodal content also poses the challenge that its mere existence is enough to make people question and doubt the veracity and authenticity of information. Arguing that a piece of information is deepfake or AI-generated to downplay or question its significance and message, has, as mentioned above, become a form of disinformation on its own ("liar's dividend" or "deepfake defence").[31] In 2024, before the general election in Mauritius, for example, the government publicly declared a series of leaks that revealed surveillance and data breaches to be "AI manipulated" and used this as a pretext for a social media shutdown. However, this unsubstantiated claim backfired and triggered massive demonstrations and civil unrest that ultimately led to a loss in the election.[32]

26 DFRLab, The Pravda Network, https://dfrlab.org/the-pravda-network/ (accessed August 28, 2025).

27 Sadeghi and Blachez, 2025; American Sunlight Project, A Pro-Russia Content Network Foreshadows the Automated Future of Info Ops, February 2025, 13, https://static1.squarespace.com/static/6612cbdfd9a9ce56ef931004/t/67fd396818196f3d1666bc23/1744648558879/PK+Report.pdf (accessed August 26, 2025).

28 NewsGuard, "Fake Accounts on X Say They're Not Voting for Biden," in: NewsGuard Reality Check Newsletter, May 20, 2024, https://www.newsguardrealitycheck.com/p/fake-accounts-on-x-say-theyre-not (accessed August 28, 2025).

29 Carla Reveland, "Rechts, Weiblich, Fake," in: Tagesschau, February 5, 2025, https://www.tagesschau.de/faktenfinder/kontext/rechte-ki-influencer-100.html (accessed August 28, 2025).

30 See c.f.: Corneliu Bjola, "Algorithmic Invasions: How Information Warfare Threatens NATO's Eastern Flank," in: NATO Review, February 7, 2025, https://archives.nato.int/nato-review-algorithmic-invasions-how-information-warfare-threatens-natos-eastern-flank (accessed August 28, 2025); Cyfluence Research Center, The Romanian Presidential Elections 2024. Analysis of Information Operations and Long-term Influence Efforts, May 17, 2025, https://www.cyfluence-research.org/post/the-romanian-presidential-elections-2024-analysis-of-information-operations-and-long-term-inf (accessed August 28, 2025).

31 C.f. Rebecca Delfino, "The Deepfake Defense. Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers," in: Ohio State Law Journal 84, no. 5, February 12, 2023, 1067-1124, here 1072-1084.

32 Roukaya Kasenally, The Chilling Effect of Censorship, Self Censorship and Surveillance in Mauritius, Democracy in Africa, https://democracyinafrica.org/the-chilling-effect-of-censorship-self-censorship-and-surveillance-in-mauritius/ (accessed October 15, 2025).

Using AI to Counter Disinformation

Understanding the impact of generative AI (genAI) on disinformation is essential for developing effective countermeasures across strategic, operational and societal levels. While malign actors have rapidly adopted AI to enhance manipulation, automate influence operations, and scale information warfare, the defensive side remains largely cautious, often constrained by ethical and regulatory debates. To date, the use of AI in counter-disinformation efforts has focused primarily on: Detecting synthetic content, identifying bot activity, and analyzing large volumes of disinformation narratives.[33] Yet, this represents only a fraction of its potential.

The integration of genAI into defensive strategies remains underexplored, with few governments publicly sharing practical applications or best practices. One notable exception is France's VIGINUM, the state agency responsible for monitoring and countering foreign digital interference, which has begun outlining potential AI use cases for the analysis and assessment of disinformation campaigns.[34] Beyond detection, however, genAI could significantly strengthen resilience and response capabilities. Potential applications include the ethical development of AI-driven strategies for countering information manipulation, the simulation and testing of counter-narratives and strategic communications, and the creation of visual and audiovisual materials – from memes and illustrations to short videos – for public awareness. GenAI could further be employed in wargaming and "red versus blue" simulations to test responses to emerging threats, automate defensive measures such as flagging and reporting coordinated activity, and support prebunking and educational initiatives through interactive, avatar-based training. Properly integrated, genAI could thus shift the defensive paradigm: from reactive monitoring to anticipatory, adaptive, and technology-informed resilience.

Regulatory Responses

EU AI Act & Digital Services Act (DSA)

The EU AI Act is the first comprehensive framework for regulating AI, including disinformation and deepfakes.[35] It categorizes AI systems based on their risk level, with deepfakes classified as high-risk in sensitive areas such as media, advertising, and political campaigns, necessitating strict oversight and regulation to prevent manipulation and misinformation. The law prohibits certain applications that deceive or manipulate individuals, including deepfakes used to spread false information or damage reputations.[36] Transparency requirements mandate that any content generated or altered by AI must be clearly labelled as such. Violations of these provisions can result in substantial penalties, including fines of up to €35 million or seven percent of a company's global annual revenue.[37]

Regarding disinformation in general, the AI Act refers to and must be read in relation with the Digital Services Act (DSA). The DSA is one of the center points in the EU's strategy to combat disinformation, serving as a key regulatory framework.[38] The AI Act references the DSA, and the latter mandates that Very Large Online Platforms (VLOPs) must actively remove disinformation and illegal content from their services.[39] The DSA imposes strict penalties, including fines of up to six percent of a platform's annual revenue if they fail to adhere to their obligations, which include deleting disinformation.[40] These provisions are binding for AI disinformation as well as non-AI generated disinformation.

33 C.f.: Viginum, Challenges and Opportunities of Artificial Intelligence in the Fight Against Information Manipulation, February 10, 2025, 20-24, https://www.sgdsn.gouv.fr/files/files/Publications/20250207_NP_SGDSN_VIGINUM_Rapport%20menace%20informationnelle%20IA_EN_0.pdf (accessed August 28, 2025).

34 Ibid.

35 European Parliament and Council of the European Union, Regulation (EU) 2024/1689. Artificial Intelligence Act, June 13, 2024, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689 (accessed October 15, 2025).

36 Ibid., 51.

37 Ibid., 115-116.

38 European Parliament and Council of the European Union, Regulation (EU) 2022/2065. Digital Services Act, October 19, 2022 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065 (accessed October 15, 2025).

39 Ibid., 65-66.

40 Ibid., 82.

Social Media Policies on AI Content & AI Disinformation

Most online platforms and messenger services have by now published policies, community standards, or terms that in one way or another touch upon the issue of AI content and AI disinformation. Meta's Community Guidelines, for example, include provisions on AI-generated content with a focus on misinformation: Users must label content that is digitally created or altered and may mislead other users.[41] According to its standards, Meta also plans to introduce labels for AI-generated content and is collaborating with industry partners to develop unified technical standards for identifying AI content.[42] Since May 2024, Meta's platforms have engaged in automated detection of AI-content to which labels saying "made with AI" are attached.[43] However, in reality, it remains unclear just how much and how accurately AI-content is being detected and whether these policies are enforced equally around the world. Despite belonging to Meta, WhatsApp's Community Guidelines do not explicitly mention AI-generated content, disinformation, or deepfakes, but only refer to fraudulent or misleading use of WhatsApp in general.[44] Similarly, Telegram, another major messenger service, and its Community Guidelines do not specifically address AI-generated content deepfakes or disinformation but rather rely on that users can flag illegal content.[45]

YouTube now requires content creators to disclose the use of AI-generated or altered material, introducing mandatory labelling in video descriptions for any synthetic media that depicts events or speech that did not actually occur.[46] Based on these rules, users can request the removal of AI-generated content that features their likeness without consent.

X (formerly Twitter) has updated its rules and policies and states that "you may not share inauthentic media, including, manipulated, or out-of-context media that may result in widespread confusion on public issues, impact public safety, or cause serious harm ("misleading media"). We may use our own technology or receive reports through partnerships with third parties in order to determine if media have been manipulated or presented out of context. In situations where we are unable to reliably determine if it is misleading media, we may not take action."[47] Yet, ever since X integrated its AI chatbot "Grok" into the platform, it has faced repeated criticism over Grok's repetition of false information (including disinformation), the creation and spread of manipulative, offensive and misleading AI images, as well as deepfakes.[48]

TikTok's Community Guidelines require a disclosure of "synthetic media," meaning content creators and users must label deepfakes or altered videos to indicate their artificial nature, using labels either created by users, labels provided by the platform, or stickers or captions like "synthetic" or "not real."[49] Yet, at the same time, TikTok is also known as a preferred platform for individual and organized private actors that leverage and automate genAI to mass produce "AI slop," conspiracy channels, and other manipulative content for financial benefit without the platform's interference.[50]

These examples show that most social media platforms do not prohibit AI content per se, but try to apply approaches and policies on disinformation

41 Monika Bickert, Our Approach to Labeling AI-Generated Content and Manipulated Data, Meta, https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/ (accessed October 12, 2025).

42 Ibid.

43 Ibid.; Andrew Hutchinson, "Meta Updates AI Labeling Policy to Expose Generated Content in Its Apps," in : Social Media Today, April 5, 2024, https://www.socialmediatoday.com/news/meta-ai-labeling-policy-expose-generated-content/712497/ (accessed August 28, 2025).

44 WhatsApp, Whatsapp Business Terms of Service, February 16, 2024, https://www.whatsapp.com/legal/business-terms (accessed August, 2025).

45 Telegram, Telegram FAQ, https://telegram.org/faq#what-are-the-rules-of-telegram (accessed October 12, 2025).

46 Google, New Disclosures and Labels for Generative AI Content on YouTube, YouTube Help, March 18, 2024, https://support.google.com/youtube/thread/264550152/new-disclosures-and-labels-for-generative-ai-content-on-youtube?hl=en (accessed October 12, 2025).

47 X, Authenticity, https://help.x.com/en/rules-and-policies/authenticity (accessed October 20, 2025).

48 Center for Countering Digital Hate, Grok AI. A Deepfake Disinformation Disaster for Democracy, August 29, 2024, 7, https://counterhate.com/research/grok-ai-election-disinformation/ (accessed August 28, 2025).

49 TikTok, About AI-Generated Content, https://support.tiktok.com/en/using-tiktok/creating-videos/ai-generated-content#5 (accessed October 12, 2025).

50 Stanusch et al., 2025, 35-36.

(and other malicious content.) The general approach – reflecting the requirements of the EU AI Act – is to detect and label AI content as such (with the size, place, and content of labels being up to platforms and users). While EU legislation will mandate the detection of deepfakes and AI-generated content from 2026 onward, most platforms have yet to implement automated ex ante detection systems and continue to rely primarily on post-hoc moderation following user complaints. Tests have also shown that "watermarks," labels and content credentials can easily be removed.[51] In fact, many deepfake creation tools also provide "watermark" removal software, and despite being technically a violation of community guidelines and platform terms, users not labelling their content as AI-generated have rarely faced sanctions. Thus, despite increasingly integrating AI applications into their services, online platforms do not have a systematic and strategic approach to tackling and mitigating the specific risks and "supercharging power" of AI in relation to disinformation.

Conclusion and Policy Recommendations

AI-disinformation countermeasures or mitigation strategies have so far evolved around three major approaches, which are reflected in the EU AI Act: mandatory labelling of AI content and transparency rules, boosting automated AI detection, and the prohibition of malicious deepfakes (such as deep porn, mal-advertisement or impersonation, which are also reflected in numerous other laws such as the UK Online Safety Act or legislation in the State of California).[52]

However, this is not enough. Addressing the use of generative AI in disinformation demands a coordinated, multi-layered strategy combining technological innovation, governance, and public resilience. First, defence and counter strategies need to take a closer look at the areas where AI has (possible) supercharging effects, such as quantity, quality (e.g., customization and persuasiveness), speed, or cost reduction, which lower the entry barrier and facilitate disinformation infrastructure (e.g., fake accounts and personas, automatization). These may then be the starting points for crafting technical countermeasures. Second, defenders must analyse the effects of AI on existing countermeasures. The persistence of disinformation stems less from a lack of countermeasures than from shortcomings in their execution, funding, and strategic coordination. Long before the AI revolution, a wide range of legal, security, economic, educational, and technological tools had already been identified.[53] Yet, the enduring challenge lies in synchronizing these measures across multiple actors and levels. The effective fight against (AI-)driven disinformation therefore depends not on new ideas, but on coherent implementation, adequate resourcing, and the delicate balance between safeguarding freedom of expression and deploying intrusive countermeasures.

- Integrate AI in Defence Strategies: Some existing countermeasures, for example media and cyber literacy efforts, fact-checking or legal regulation, might need AI-specific updates and adjustments, but do not lose their relevance. Yet, with AI technology, technological features, challenges, and details becoming ever more important for the way disinformation is created and distributed, AI-specific countermeasures, strategies and components need to find their way into defence strategies.

- Fortify Labelling and Watermarking: For labelling, watermarking and content credentials to be effective against disinformation, they need to be more robust, less easy to remove, and also must correspond to the level of AI literacy of users and audiences. Labels, for example, may be designed in size and content to resonate with audiences. Thus, the current focus on detection and transparency may soon give way to the question of context, i.e., evaluating and explaining the quality and details of the way AI was used for a specific piece of content. The same should apply not only to AI-generated or altered content, but also to AI-powered channels, influencers, and

51 Kate Knibbs, "Researchers Tested AI Watermarks – and Broke All of Them," in: Wired, October 3, 2023, https://www.wired.com/story/artificial-intelligence-watermarking-issues/ (accessed October 12, 2025).

52 Allen and Nehring, 2025, 115-123.

53 Ibid., 131-140.

profiles. The launch of the "Sora2 App" by OpenAI, a social media environment filled exclusively with AI-generated content also raises the questions of whether AI-labels and transparency will provide any safety from disinformation in information spaces that are overloaded with AI-content.

- Strengthen Product Safety And Product Security: The increasing (technological) complexity of AI generated and distributed information and its integration into information systems also necessitates a high level of information security alongside product safety and product security. Here, possible risk mitigation measures against AI disinformation include prompt storage and review, which maintains records of user prompts to detect misuse and identify disinformation actors. Built-in moderation and filtering within AI systems provide another safeguard. Anthropic's Claude chatbot, for example, includes a new feature to automatically terminate conversations if users persistently request harmful content.[54] To make AI applications more robust against misuse as well as external manipulation attacks, red teaming and adversarial testing could be made mandatory to simulate attacks on AI systems, expose weaknesses, and guide resilience, safety, and security.

- Employ Technological Risk Mitigation Measures: Other technical countermeasures and risk mitigation could include (AI based) automated content moderation or removal, which can occur either ex ante (i.e., blocking malicious content before upload) or ex post (i.e., removing it after being flagged), as well as correction and juxtaposition, where fact-checked information is automatically displayed alongside flagged content. Protective software also exists to "scramble" original materials – images, videos, or data – to prevent them from being exploited by genAI applications.

- Expand Educational Measures: Technological risk mitigation measures alone, however, should not be regarded as "silver bullets," but may work best when combined with extensive educational measures. Preventive strategies are equally important. AI disinformation requires targeted media literacy and education efforts that address its unique "supercharging" qualities. These efforts may involve governments, NGOs, tech companies, universities, and media organizations. AI literacy in combination with general digital and media literacy is crucial to equip users for the AI era and make online audiences more resilient towards AI-powered disinformation. AI labels on social media content, for example, can only resonate with audiences if they have a sufficient knowledge of AI, how it creates information, why it might be inaccurate or misleading information etc. Awareness campaigns about AI disinformation capabilities and risks, as well as general AI literacy programmes equipping different audiences with knowledge to evaluate AI outputs are needed on large scale. Prebunking and inoculation techniques, which expose people to disinformation tactics (e.g., deepfakes) in controlled environments, can also be adapted to AI-specific threats. These approaches can strengthen cognitive defences by familiarizing users with common techniques and narratives.

- Adapt Regulation to the changing technological environment: Specific legal regulation as well as norms, standards, and guidelines represent another pillar of possible responses against AI disinformation. Most of the measures outlined above could technically be made mandatory by law. Governments and international organizations may also adopt prohibitions on certain AI applications deemed high-risk or harmful. Mandatory labelling laws can oblige disclosure of AI-generated content, with detailed requirements on label structure and visibility. AI use in political, security or societal contexts is a particular concern, which calls for stricter regulation to preserve public trust. Algorithmic transpar-

54 Anthropic, Claude Opus 4 and 4.1 Can Now End a Rare Subset of Conversations, https://www.anthropic.com/research/end-subset-conversations (accessed August 28, 2025).

ency and regulation, advocated by NGOs,[55] aims to restrict manipulation through content-suggestion algorithms. Similarly, advertisement regulation can prevent ad revenue from flowing to disinformation outlets. Yet, in practice, legal regulation has proven to be highly complicated: Regulation must keep pace with the rapid technological evolution of AI. At the same time, it should avoid creating unnecessary barriers to technological and economic innovation. While AI-driven disinformation is a global challenge, regulation remains regional, and strict measures – such as bans on malicious deepfakes – must also be practically enforceable.

- Support Ethical Frameworks And Guidelines: Beyond laws, (global) ethical frameworks and guidelines provide orientation for responsible AI use. During the global super election year 2024, efforts like the Munich Tech Accord[56] or the G7 Hiroshima Document[57] – which included provisions on disinformation – were first steps in this direction. However, more should be done to harmonize global security standards and establish common protocols to ensure system safety and resilience. Such measures could be accompanied by periodic audits by independent bodies that ensure compliance with ethical requirements, while official guidelines for specific sectors – politics, journalism, PR, or media – could formalize norms around the ethical deployment of genAI. As previous efforts such as the Munich Tech Accord have shown, such frameworks are good at sending global messages and functioning as a point of reference; their weakness however is their non-binding, voluntary character.

55 Matthias Spielkamp, New Project: Auditing Algorithms for Systemic Risks, Algorithm Watch, https://algorithmwatch.org/en/new-project-auditing-algorithms-for-systemic-risks/ (accessed August 28, 2025); Matthias Spielkamp, ChatGPT and Co: Are AI-driven Search Engines a Threat to Democratic Elections?, Algorithm Watch, https://algorithmwatch.org/en/bing-chat-election-2023/ (accessed August 28, 2025).

56 Munich Security Conference, A Tech Accord to Combat Deceptive Use of AI in 2024 Elections, https://securityconference.org/aielectionsaccord/ (accessed August 28, 2025).

57 European Commission, G7 Leaders' Statement on the Hiroshima AI Process, October 30, 2023 https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process (accessed August 28, 2025).

Guardrails of Truth? Disinformation and Regulatory Approaches in the European Union[1]

Marius Dragomir
Media and Journalism Research Center

Disinformation has emerged as one of the most profound threats to democratic governance in Europe, eroding trust in institutions, polarizing societies, and distorting electoral processes. Eastern Europe remains a critical frontline, where historical vulnerabilities, weak institutions, and Russian influence operations exacerbate the risks. The European Union (EU) has responded with an expanding regulatory framework, including the Digital Services Act, the European Media Freedom Act, and initiatives on AI transparency, yet enforcement gaps and uneven national commitment hinder effectiveness. Case studies reveal how disinformation has shaped public attitudes on a number of topics including EU integration, vaccination, and migration, undermining democratic resilience. Critical evaluation of the current EU strategies highlights tensions between regulation and free expression. Thus, a shift toward incentive-based models that empower citizens, media, and platforms to strengthen trust and transparency in an AI-driven information space is necessary to mitigate the impact of disinformation on societies and democracies.

Setting the Scene: Disinformation as a Threat to Democratic Governance

Widely regarded as one of the biggest threats to the integrity of democratic governance in the 21st century,[1] disinformation poses profound and complex risks to democratic cohesion across Europe. These range from the erosion of public trust in institutions and the media to deepening societal polarization, the manipulation of electoral processes, and the marginalization of vulnerable groups.

Disinformation campaigns, by design, seek to disseminate deceptive narratives that corrode citizens' confidence in democratic institutions, elections, and the media.[2] One of the most pernicious effects of such campaigns lies in their distortion of the democratic process itself. A robust body of scholarship has demonstrated how disinformation manipulates public opinion, distorts electoral campaigns,[3] and, at times, may even influence voting outcomes.[4] Sophisticated influence operations, often state-sponsored and foreign-directed,[5] have been deployed to

1 Stephan Lewandowsky et al., "Misinformation and the Epistemic Integrity of Democracy," in: Current Opinion in Psychology 54, 2023, 1-7, here 2; Lance W. Bennett and Steven Livingston, The Disinformation Age, Cambridge: Cambridge University Press, 2020; Nathaniel Persily, "The 2016 U.S. Election: Can Democracy Survive the Internet?," in: Journal of Democracy 28, no. 2, April 2017, 63-76, here 64.

2 Carme Colomina et al., The Impact of Disinformation on Democratic Processes and Human Rights in the World, European Parliament Think Tank, April 2021, 13, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf (accessed July 29, 2025).

3 Marlene Mauk and Max Grömping, "Online Disinformation Predicts Inaccurate Beliefs About Election Fairness Among Both Winners and Losers," in: Comparative Political Studies 57, no. 6, September 2023, 965-998, here 983.

4 Hunt Allcott and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election," in: Journal of Economic Perspectives 31, no. 2, 2017, 211-236, here 232.

5 European Parliament, Foreign Interference: How Parliament is Fighting the Threat to EU Democracy. Foreign Power Use Disinformation and Cyberattacks to Try to Shape EU Public Opinion in their Own Interest. Find Out How the EU is Fighting Back, April 9, 2024, https://www.europarl.europa.eu/topics/en/article/20240404STO20215/foreign-interference-how-parliament-is-fighting-the-threat-to-eu-democracy (accessed July 29, 2025); Seth G. Jones, Russia's Shadow War Against the West, Center for Strategic & International Studies, March 18, 2025, https://www.csis.org/analysis/russias-shadow-war-against-west (accessed July 29, 2025).

manipulate public discourse,[6] obscure factual consensus, and delegitimize electoral results.[7] These efforts have fueled a growing skepticism toward democratic institutions.[8]

By inflaming contentious issues and exploiting latent social cleavages whether ethnic, ideological, or political – disinformation aggravates polarization.[9] This leads to societal division, impeding the formation of consensus and undermining the collective deliberation that is essential to functional democracy.[10] Moreover, these campaigns disproportionately target marginalized groups,[11] excluding them from meaningful political participation and stoking hate speech. In doing so, they weaken democratic inclusivity and may incite discrimination, harassment, and violence.

The overarching consequence of these dynamics is a sharp decline in informed citizenship.[12] The deluge of misleading or false content hinders citizens' access to reliable information, thereby impairing their capacity to make reasoned decisions and participate meaningfully in democratic life.[13] Collectively, all this erodes the shared values and fact-based dialogue that underpin democratic cohesion in Europe.

Eastern Europe: A Crucial Frontline

Within this broader landscape, the focus on Eastern Europe is of particular urgency. The region's historical and geopolitical vulnerabilities, strategic positioning, and distinct socio-political characteristics render it especially susceptible to disinformation.

Many Eastern European states bear the legacy of Soviet rule,[14] with enduring political, economic and cultural entanglements with Russia. This historical backdrop provides fertile ground for foreign actors, chiefly the Kremlin,[15] to exploit societal divisions and disseminate narratives aimed at undermining democratic consolidation and Euro-Atlantic integration. The region's proximity to Russia, coupled with ongoing tensions, has made it a prime target for aggressive information operations intended to destabilize pro-Western policies, discredit reform agendas, and sap public trust in national institutions and European alliances.

Former Soviet republics and Balkan nations, such as Moldova, Serbia, Bosnia and Herzegovina, and Montenegro, are particularly vulnerable. Many are strategically significant for the EU as Kremlin-backed disinformation aims to sow doubt about the merits of such integration, amplify internal divisions, and weaken support for Western institutions.[16] These efforts jeopardize not only national stability

6 Ideas for Democracy, The Impact of Disinformation on Debate in the European Parliament. Risks for Democracy and Tools to Fight It, Minsait, February 27, 2019, https://www.minsait.com/ideasfordemocracy/en/impact-disinformation-debate-european-parliament (accessed July 29, 2025); Samantha Bradshaw and Philip N. Howard, The Global Disinformation Order. 2019 Global Inventory of Organised Social Media Manipulation, Oxford Internet Institute, September 4, 2019, 2, https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2019/09/CyberTroop-Report19.pdf (accessed September 3, 2025).

7 Ian Kennedy et al.,"Repeat Spreaders and Election Delegitimization. A Comprehensive Dataset of Misinformation Tweets from the 2020 U.S. Election," in: Journal of Quantitative Description: Digital Media 2, 2022, 1-49, here 12-23.

8 Seva Gunitsky, "Democracies Can't Blame Putin for Their Disinformation Problem. The Focus on Foreign Subversion Ignores the Damage Being Done at Home," in: Foreign Policy, April 21, 2020, https://foreignpolicy.com/2020/04/21/democracies-disinformation-russia-china-homegrown/ (accessed July 29, 2025).

9 Beatriz Botero Arcila and Florent Parmentier, Under 35 and Under Siege. How Young European Citizens Propose to Combat Disinformation in Democratic Societies, Science Po, November 2024, 5, https://www.sciencespo.fr/api/portal/uploads/Disinformation_and_Democracy_answering_the_Youth_b53de3a91c.pdf#page=5.68 (accessed July 29, 2025).

10 Gunitsky, 2020.

11 Judit Szakács and Éva Bognár, The Impact of Disinformation Campaigns About Migrants and Minority Groups in the EU, European Parliament Think Tank, June 24, 2021, https://www. europarl. europa. eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN.pdf (accessed July 29, 2025).

12 Colomina et al., 2021, 13.

13 Zoë Adams et al., "(Why) Is Misinformation a Problem?," in: Perspectives on Psychological Science 18, no. 6, February 16, 2023, 1436-1463, here 1454.

14 Péter Krekó, The Drivers of Disinformation in Central and Eastern Europe and Their Utilization During the Pandemic, GLOBSEC, June 9, 2020, 1, https://www.globsec.org/sites/default/files/2020-06/Drivers-of-disinformation-in-Central-and-Eastern-Europe.pdf (accessed July 29, 2025).

15 James Thomas and Talyta França, "Which European Countries Are Most Exposed to Russian Disinformation?," in: Euronews, April 18, 2025, https://www.euronews.com/my-europe/2025/04/18/which-european-countries-are-most-exposed-to-russian-disinformation (accessed July 29, 2025).

16 Balkan Insight, "BIRN Doc Lifts Lid on Russian Disinformation in Balkans," in: Balkan Insight, May 10, 2024, https://www.balkaninsight.com/2024/05/10/birn-doc-lifts-lid-on-russian-disinformation-in-balkans/ (accessed July 29, 2025).

but also broader regional security, EU enlargement, and the integrity of the transatlantic alliance.[17]

Psychosocial factors further compound the region's vulnerability. Research suggests that historical trauma, feelings of insecurity, and social fragmentation heighten receptivity to conspiratorial thinking and manipulative content.[18] Moreover, endemic distrust in traditional media, paired with an overreliance on digital platforms, where disinformation circulates unchecked, creates fertile conditions for falsehoods to flourish.[19] The Eastern European media landscape is frequently marked by political partisanship, economic fragility, regulatory gaps, and a lack of editorial independence. These structural deficiencies provide more entry points for malign actors and severely undermine the resilience of public discourse.

European Policy Responses: From Strategy to Enforcement

In light of these risks, the European Union has mounted an evolving policy response to counter disinformation while trying to uphold fundamental rights and freedoms. Its approach spans regulatory initiatives, institutional frameworks, and cross-sector collaborations, all aimed at protecting democratic resilience without stifling free expression or media pluralism.

The EU's early response began with the creation of the East StratCom Task Force in 2015, embedded within the European External Action Service. Its primary mission was to identify and counter foreign-led disinformation,[20] chiefly from Russia, following the 2014 annexation of Crimea and the surge of propaganda targeting European audiences.[21] However, the growing realization that disinformation also arises domestically, often spread by citizens themselves, prompted a more nuanced approach.[22]

In 2018, the European Commission launched the Action Plan against Disinformation,[23] introducing a range of measures to improve detection capabilities, facilitate cross-border coordination, incentivize private sector accountability, and improve public awareness.[24] Media literacy emerged as a central tenet of this response.[25] The same year, the EU brokered the Code of Practice on Disinformation, a landmark self-regulatory framework signed by most global digital platforms. It committed signatories to reducing ad revenues for disinformation purveyors, tackling automated bots, and increasing transparency in political advertising. However, the voluntary nature of the Code limited its impact due to insufficient compliance and enforcement.[26]

In response, the Strengthened Code of Practice on Disinformation was introduced in 2022,[27] which was later formally integrated into the Digital Ser-

17 Edit Zgut-Przybylska, "Pro-Kremlin Disinformation is Distorting Democracy in Central and Eastern Europe," in: Review of Democracy, October 8, 2024, https://revdem.ceu.edu/2024/10/08/pro-kremlin-disinformation-cee/ (accessed July 29, 2025).

18 Krekó, 2020, 2.

19 Michal Bokša, Russian Information Warfare in Central and Eastern Europe: Strategies, Impact and Countermeasures, The German Marshall Fund of the United States, June 2019, 1, https://www.gmfus.org/sites/default/files/Russia%20disinformation%20CEE%20-%20June%204.pdf (accessed July 29, 2025).

20 Sophie L. Vériter, "Small-State Influence in EU Security Governance: Unveiling Latvian Lobbying Against Disinformation," in: Journal of Common Market Studies 62, no. 6, March 3, 2024, 1538-1559, here 1538.

21 Foreign Information Manipulation and Interference, partly overlapping with disinformation, is often the concept used in policy discussions to refer to information-manipulation efforts by foreign state-sponsored entities, identifying its targets not based on content but the behavior of actors (tactics, techniques and procedures).

22 Konrad Bleyer-Simon et al., How is Disinformation Addressed in the Member States of the European Union? – 27 Country Cases, European Digital Media Observatory, June 18, 2025, 8, https://edmo.eu/publications/how-is-disinformation-addressed-in-the-member-states-of-the-european-union-27-country-cases/ (accessed July 29, 2025).

23 European Commission, Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Action Plan Against Disinformation, December 5, 2018, https://commission.europa.eu/system/files/2018-12/eu-communication-disinformation-euco-05122018_en.pdf (accessed July 29, 2025).

24 Oreste Pollicino et al., Freedom of Speech and the Regulation of Fake News, Intersentia, August 2023, 129-164, https://iris.unibocconi.it/retrieve/dde0fbe3-abe6-4bc7-9691-d7fe6042f204/Freedom%20of%20Speech_General%20Report.pdf (accessed July 29, 2025).

25 They include the 2018 revised Audiovisual Media Services Directive and the Digital Education Action Plan (COM/2020/624).

26 European Court of Auditors, Disinformation Affecting the EU. Tackled but not Tamed, Special Report No. 09/2021, 2021, 33, https://www.eca.europa.eu/Lists/ECADocuments/SR21_09/SR_Disinformation_EN.pdf (accessed September 3, 2025).

27 European Commission, The Strengthened Code of Practice on Disinformation 2022, June 16, 2022, 10, https://ec.europa.eu/newsroom/dae/redirection/document/87585 (accessed July 29, 2025).

vices Act (DSA), the EU's flagship digital regulation. Under the DSA, the Code functions as a co-regulatory instrument obliging Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to conduct annual risk assessments, submit to independent audits, and mitigate systemic risks such as disinformation.[28]

Although the DSA is not solely focused on disinformation, its provisions on transparency, accountability, and systemic risk mitigation have become instrumental in tackling the phenomenon. Additional layers of regulation, including the Digital Markets Act (DMA), the European Media Freedom Act (EMFA), the Artificial Intelligence Act, and the Regulation on Political Advertising, have provisions that aim to further strengthen the EU's digital ecosystem against manipulation.

The Russian invasion of Ukraine in 2022 marked a turning point, prompting the EU to take the unprecedented step of banning several Kremlin-affiliated media organizations, including RT and Sputnik. Yet, the effectiveness of these bans proved uneven, with some outlets continuing to operate online and certain national media, most notably in Hungary, relying on Russian state narratives to frame the conflict.[29]

In addition to all these measures, countering disinformation has been embedded within the EU's broader human rights and democracy agenda. Since 2024, effectively addressing disinformation, particularly in the context of Foreign Information Manipulation and Interference (FIMI), has also become a formal requirement for candidate countries as part of the EU accession framework.

The EU's regulatory leadership on disinformation extends well beyond its borders. It plays a globally significant role by setting international benchmarks[30] for platform accountability, transparency, and systemic risk management. The EU's regulatory leadership has also catalyzed a "Brussels effect," whereby global tech firms align with European standards to avoid regulatory fragmentation and pre-empt stricter rules elsewhere.[31] EU policy on disinformation not only redefines how platforms operate within Europe, but also influences the global conversation about the limits and responsibilities of digital intermediaries.

In the following sections, the real-world effectiveness of current EU regulatory tools in combating disinformation is critically assessed. The analysis highlights the roles played by national authorities, digital platforms, and enforcement mechanisms – examining both their vulnerabilities and their resilience – while identifying persistent regulatory gaps. In addition, the broader democratic implications of EU regulatory strategies will be examined, weighing concerns about censorship, transparency, platform accountability, and institutional legitimacy. The resulting policy recommendations propose concrete ways for the EU to further improve democratic resilience.

Assessing the EU Regulatory Toolbox

Despite a slew of initiatives and legislative interventions, the EU's regulatory regime has, at best, delivered uneven impact, marked by pockets of progress interspersed with ineffectiveness across the continent. This patchiness can be traced to a tangled combination of factors, most notably the fluctuating diligence of national authorities, the evolving strategic approaches by technology platforms, and the inherent limitations of the EU's existing enforcement apparatus.[32]

28 Ramsha Jahangir, "The EU's Code of Practice on Disinformation is Now Part of the Digital Services Act: What Does It Mean?," in: TechPolicy Press, February 25, 2025a, https://www.techpolicy.press/the-eus-code-of-practice-on-disinformation-is-now-part-of-the-digital-services-act-what-does-it-mean/ techpolicy.press+12techpolicy.press+12techpolicy.press+12 (accessed July 29, 2025).

29 Political Capital, Russian Disinformation in Hungarian Public Broadcast Media: Complaint to the European Commission, https://www.politicalcapital.hu/pc-admin/source/documents/TASZ-PC_EC-complaint- PublicBroadcast_Russia_War.pdf (accessed July 29, 2025).

30 Ioanna Tourkochoriti, "The Digital Services Act and the EU as the Global Regulator of the Internet," in: Chicago Journal of International Law 24, 2023, 129-147, here 133.

31 Dawn Carla Nunziato, "The Digital Services Act and the Brussels Effect on Platform Content Moderation," in: Chicago Journal of International Law 24, no. 1, 2023, 115–128.

32 Alexander Peukert, "The Regulation of Disinformation: A Critical Appraisal," in: Journal of Media Law 16, no. 1, April 4, 2024, 1-7, here 1-6.

The introduction of the DSA and the subsequent tightening of the Code of Practice on Disinformation with its integration within the DSA framework as of July 2025 delivered a much-needed jolt to the fight against digital falsehoods.[33] Under these provisions, tech companies are now obligated not merely to police their platforms but to systematically identify, assess and counteract "systemic risks," disinformation prominent among them. They must publish detailed transparency reports, which document their efforts in content moderation, fact-checking, and the propagation (or containment) of misleading narratives, furnish solid mechanisms for users to flag illicit or hazardous misinformation and act expeditiously in removing such content.[34]

Yet, the central question remains: How do these ambitious regulatory requirements manifest in practice? Among the platforms that have aligned themselves with the Code, such as Facebook, Google, and TikTok, to name a few, tangible improvements were initally visible. Platforms have routinely stated their ostensible commitment to the DSA; Meta (owner of Facebook and Instagram) has outlined its transparency initiatives and labeling efforts,[35] while Google[36] and TikTok[37] boast similar compliance endeavors.

However, beneath the surface, significant shortcomings have become evident. In early 2025, Google reneged on several fact-checking and risk assessment commitments relating to Google Search and YouTube, casting doubts over its willingness to embrace rigorous, ongoing compliance.[38] X (formerly Twitter), by contrast, had already withdrawn from the Code in 2023, being ranked as the least compliant platform due to its complete absence of reporting and transparency measures.[39] Meta, despite much publicized reporting protocols, has been heavily criticized for discontinuing journalist-led fact-checking initiatives and stands accused of inconsistency, battling complaints and legal action for broader policy violations.[40] Even TikTok's engagement, while formally maintained, appears conditional: Its implementation of fact-checking mechanisms has been explicitly tied to comparable commitments from its competitors.[41] Watchdog organizations have consequently criticized these platforms for their opacity, the weakness of independent oversight, and the limited real-world impact of existing anti-disinformation measures.[42]

In response, the European Commission has unsheathed formal proceedings against companies such as X and Meta, seeking hard evidence of effective risk mitigation and fact-checking. The stakes are not merely financial (non-compliance can incur penalties of up to six percent of firms' global turnover). Reputational damage looms equally large, threatening platforms that shrug off their obligations.[43]

Despite these measures, enforcement at the national level remains fraught with inconsistencies. The journey from Brussels to the capitals of member states is often waylaid by bureaucratic inertia: De-

33 Jahangir, 2025a.

34 Ronan Ó Fathaigh, Doris Buijs, and Joris van Hoboken, "The Regulation of Disinformation Under the Digital Services Act," in: Media and Communication 13, May 28, 2025, 1-21, here 4.

35 Tom Gerken and Liv McMahon, "Big Tech Must Deal With Disinformation or Face Fines, Says EU," in: BBC, June 18, 2022, https://www.bbc.com/news/technology-61817647 (accessed July 29, 2025).

36 Laurie Richardson and Jennifer Flannery O'Connor, Complying with the Digital Services Act, Google Blog, https://blog.google/around-the-globe/google-europe/complying-with-the-digital-services-act/ (accessed September 3, 2025).

37 Madeline Moncrieff, Our Compliance with the Digital Services Act, TikTok Newsroom, https://newsroom.tiktok.com/en-eu/compliance-digital-services-act-eu (accessed July 31, 2025).

38 Mehmet Onur Cevik, Enforcement Spotlight – Spring 2025. Where Tech Policy Meets Reality, Centre for Future Generations, https://cfg.eu/enforcement-spotlight-spring-2025/ (accessed July 29, 2025).

39 Stephan Mündges and Kirsty Park, "But Did They Really? Platforms' Compliance with the Code of Practice on Disinformation in Review," in: Internet Policy Review 13, no. 3, July 25, 2024, 1-21, here 8.

40 Mathieu Gitton, Enforcement of the DMA/DSA: EU vs. Tech Giants, LinkedIn, https://www.linkedin.com/pulse/enforcement-dmadsa-eu-vs-tech-giants-mathieu-gitton-m2h9e (accessed July 29, 2025).

41 Ibid.

42 Anna Desmarais, "Big Tech Isn't Doing Enough to Fight Disinformation, EDMO Says," in: Euronews, June 26, 2025, https://www.euronews.com/next/2025/06/26/big-tech-isnt-doing-enough-to-fight-disinformation-eu-body-says?utm_source=chatgpt.com (accessed July 29, 2025).

43 Gerken and McMahon, 2022; Ramsha Jahangir, "Understanding the EU's Digital Services Act Enforcement Against X," in: TechPolicy Press, April 5, 2025b, https://www.techpolicy.press/understanding-the-eus-digital-services-act-enforcement-against-x/ (accessed September 3, 2025).

layed empowerment of national Digital Services Coordinators (DSCs), institutional unreadiness, and capacity gaps have thus far hampered the DSA's implementation, particularly on smaller platforms and within more localized disinformation ecosystems.[44]

Compounding matters, the rapidly evolving arsenal of tactics deployed by disinformation players makes them perpetually one step ahead of regulators, exploiting digital loopholes, refining operational methods, and escaping detection with alarming agility.

A particularly urgent dimension of this technological race is the rise of artificial intelligence (AI) and other advanced digital tools, which are transforming both the scale and sophistication of disinformation. Generative AI enables the rapid creation of convincing synthetic text, images, audio, and video ("deepfakes"), lowering the cost of manipulation while dramatically expanding its reach.[45] Research has shown that AI-generated content spreads faster, is often harder to detect, and can be tailored to exploit cultural or linguistic nuances across Europe.[46]

Regional Case Focus: Eastern Europe

The landscape of disinformation in Eastern Europe is shaped by a complex interplay of both external and internal actors, with Russia standing as the preeminent foreign source, its influence rivaled only by local factions and, to a more limited extent, China. The messaging favored by these actors weaves together sweeping global narratives and deep local grievances, highlighting historical frustrations and simmering societal tensions to maximize their resonance and impact.

The preponderant sources of disinformation in the region form a convoluted web: Russian state-sponsored actors and their sprawling influence networks, homegrown agitators of various stripes, Chinese state-linked entities, and legions of disinformation "trolls."

There is extensive documented evidence of Russia's orchestrated disinformation engine, nowhere more evident than in the so-called *Pravda* or "Portal Kombat" network. This machinery produces hundreds of websites in local languages, skillfully masquerading as legitimate outlets, which amplify pro-Kremlin talking points, drawing from state media, Telegram groups, and official channels.[47] Through industrial-scale automation and relentless copy-pasting, these sites inundate the public sphere with repetitive, hyper-localized falsehoods, shaping perceptions and sowing discord[48] in countries such as Moldova, Latvia, Estonia, Serbia, and beyond in the region.[49]

However, the threat is hardly limited to foreign operators. Domestic actors, ranging from state media and political operatives to nationalist or extremist factions, frequently stoke and magnify disinformation to further their own agendas.[50] In illiberal-leaning states such as Slovakia and Hungary, where democratic backsliding and media capture undermine independent journalism, Russia finds fertile ground for its narratives to bloom. Here, local and foreign disinformation often join forces in a mutually reinforcing feedback loop. Amplifying these efforts is a shadow industry of professional disinformation "trolls," which

44 Polona Car, Enforcing the Digital Services Act: State of Play, European Parliament Think Tank, November 20, 2024, https://www.europarl.europa.eu/RegData/etudes/ATAG/2024/766256/EPRS_ATA(2024)766256_EN.pdf (accessed September 3, 2025).

45 Bobby Chesney & Danielle Citron, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, California Law Review, 107, no. 6, December 17, 2019, 1753–1820, here 1763.

46 Europol Innovation Lab, Facing Reality? Law Enforcement and the Challenge of Deepfakes, Europol, 2022, https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf (accessed September 3, 2025).

47 Digital Forensic Research Lab, The Pravda Network, https://dfrlab.org/the-pravda-network/ (accessed July 29, 2025).

48 Denitsa Koseva, "New Study Reveals Extent of Russian Disinformation Targeting Emerging Europe," in: Intellinews, April, 2025, https://www.intellinews.com/new-study-reveals-extent-of-russian-disinformation-targeting-emerging-europe-376592/ (accessed September 3, 2025).

49 Mejreme Asllani, Russia's Pravda Network: AI-Driven Disinformation on a Global Scale, Bloomsbury Intelligence & Security Institute, https://bisi.org.uk/reports/russias-pravda-network-ai-driven-disinformation-on-a-global-scale (accessed July 21, 2025).

50 Javier Borràs, Disinformation and EU Enlargement: Instability and the Battle of Narratives in the Eastern Neighbourhood and the Balkans, Barcelona Centre for International Affairs, December 2024, 2, https://www.cidob.org/en/publications/disinformation-and-eu-enlargement-instability-and-battle-narratives-eastern (accessed July 31, 2025).

crisscross borders to exploit sensitive issues, inflaming societal rifts and giving more exposure to polarizing narratives.[51]

Meanwhile, China, though less prominent, has also ventured into the disinformation game. Beijing has seized upon vulnerabilities exposed during crises such as the COVID-19 pandemic, deploying conspiracy theories designed to burnish its image or tarnish the West.[52]

The dominant themes coursing through these campaigns are all too familiar: The war in Ukraine, stridently anti-EU and anti-NATO sentiment, the propagation of conspiracy-laden counter-narratives on contentious issues, the deliberate exacerbation of social divisions, and skepticism toward Western institutions and democratic norms. With respect to Ukraine, the stock narrative is clear. Russia's invasion is cast as a forced, "defensive" measure against NATO provocations, with Ukraine dismissed as a Western puppet.[53] These talking points are tightly interwoven with efforts to present the EU and NATO as corrupt, expansionist, and intractably hostile, fodder for Eurosceptics and a wedge meant to fracture European unity.[54]

The impact of these narratives on populations has been tangible. In Serbia, support for EU membership remains the lowest in the region, with only 34 percent of respondents considering accession a good thing (down 4 p.p. from 2022).[55] Perceptions of disinformation are widespread: Across the Western Balkans, 57 percent of people report that fake news and disinformation have increased in the past year, peaking at 71 percent in North Macedonia.[56] Disinformation about migrants has also shifted social attitudes, spreading fear by portraying them as threats to health, wealth, and identity, thus deepening societal divisions and lowering democratic trust.[57]

Economic anxieties, especially around energy security and infrastructure, a neuralgic issue for former Soviet republics, are habitually exploited, shaping public attitudes and policy debates wherever Russia's strategic interests are at stake.[58]

Across the board, the common denominator in these campaigns is the drive to undermine trust in mainstream media and the core institutions of democracy: be it the judiciary, independent journalism, or oversight bodies. The goal is not just to confuse, but to erode the very foundations of civic trust.

Some Eastern European governments have acted against these threats. Most have signed onto the EU's anti-disinformation initiatives, tailoring them to fit local realities, and often collaborate with initiatives such as the East StratCom Task Force and the EUvsDisinfo platform.[59]

While Baltic countries stand out as frontrunners,[60] combining strong public awareness campaigns with robust civil society partnerships and systemat-

51 Katja-Elisabeth Herrmann, Disarming Disinformation in Central and Eastern Europe, Warsaw Institute, https://warsawinstitute.org/disarming-disinformation-in-central-and-eastern-europe/ (accessed July 31, 2025).

52 Krekó, 2020, 7.

53 Sergey Sukhankin, Russian Disinformation Targets the European Union, The Jamestown Foundation, https://jamestown.org/program/russian-disinformation-targets-the-european-union/ (accessed July 31, 2025).

54 Agnieszka Legucka, Russia's Long-Term Campaign of Disinformation in Europe, Carnegie Endowment, March 19, 2020, https://carnegieendowment.org/europe/strategic-europe/2020/03/russias-long-term-campaign-of-disinformation-in-europe?lang=en (accessed July 31, 2025).

55 Regional Cooperation Council, Balkan Barometer 2023: Public Opinion Survey. Sarajevo: RCC, 2023, 12, https://www.rcc.int/balkanbarometer (accessed August 25, 2025).

56 Regional Cooperation Council, Balkan Barometer 2024: Public Opinion Survey. Sarajevo: RCC, 2024, 110, https://www.rcc.int/balkanbarometer (accessed Aug 30, 2025).

57 Alberto-Horst Neidhardt, Mis- and Disinformation on Migration in Europe, Publications Office of the European Union, 2025, https://data.europa.eu/doi/10.2777/8580080 (accessed September 25, 2025).

58 Thomas and França, 2025.

59 See more at https://euvsdisinfo.eu/.

60 Anna Beitane et al., "All for One and One for All Against Disinformation to Protect Democracy," in: New Eastern Europe, February 28, 2025, https://neweasterneurope.eu/2025/02/28/all-for-one-and-one-for-all-against-disinformation-to-protect-democracy/ (accessed July 31, 2025).

ic cooperation with EU initiatives,[61] countries such as Slovakia,[62] Romania,[63] and Bulgaria[64] show mixed results: policies exist, but implementation is hampered by institutional weaknesses and inconsistent political support. At the other end of the spectrum, Hungary's politicized media environment and entrenched capture have left the country highly exposed to external disinformation campaigns.[65]

Progress thus remains uneven, with the effectiveness of counter-disinformation policies closely tied to governments' stance toward the European Union. In states where leadership is firmly pro-EU, such as Estonia and Lithuania, efforts to curb disinformation have advanced. In contrast, where governments adopt Eurosceptic or openly anti-EU positions, as seen in Bulgaria, Slovakia, and Hungary, resistance to EU-backed initiatives has slowed implementation.[66]

While the EU's regulatory frameworks have provided a vital backbone, the soft spot remains enforcement.[67] Delays, resource shortfalls, and coordination headaches continue to hobble efforts, exposing uneven progress and ongoing vulnerabilities.

The local political climate, too, has proven a strong factor. In polarized societies, or where pro-Russian sentiment runs high, EU initiatives are easily cast as "Western meddling," undercutting their legitimacy and marginalizing their impact.[68] The line between homegrown and foreign disinformation often blurs; local actors may weaponize disinformation for purely domestic gain, adding confusion and making mitigation all the more fraught.

In short, the battle against disinformation in Eastern Europe is fought on shifting sands. Success arises where public vigilance, resilient independent media, decisive political leadership, and an active civil society converge, anchored in harmonized and well-enforced regulatory action from both Brussels and national capitals. Absent these elements, the scourge of disinformation continues to find room to maneuver, deepening divisions and undermining democratic order.[69]

Democratic Tensions and Governance Trade-offs

The architecture of the EU's response to disinformation, on face value, is designed to protect the integrity of democratic discourse. Yet, as the regulatory web has tightened, particularly with the advent of the DSA and accompanying national measures, the effects on civic trust, free expression, and institutional legitimacy have come into sharper relief. For all its well-intentioned zeal, the EU's approach at times dances perilously close to undermining the very democratic values it purports to protect.

A central concern lies in the unprecedented latitude granted to both the European Commission and national authorities to police online speech. Statutory tools such as the so-called "executive order" mechanism, empowering regulators to sum-

61 Auksė Balčytienė, Anda Dāvidsone, and Andra Siibak, "What a Human-Centred Approach Reveals About Disinformation Policies: The Baltic Case," in: Media and Communication 13, 2025, 4-12.

62 Peter Dubóczi and Michaela Ružičková, Disinformation Landscape in Slovakia, EU Disinfo Lab, September 2023, 3, https://www.disinfo.eu/wp-content/uploads/2023/09/20230919_SK_DisinfoFS.pdf (accessed September 25, 2025).

63 Ciprian Cucu, "Disinformation Landscape in Romania," EU Disinfo Lab, 19 September 2023, 3, https://www.disinfo.eu/wp-content/uploads/2023/09/20230919_RO_DisinfoFS.pdf (accessed September 25, 2025).

64 Sonya Nikolaeva, Mapping Disinformation Narratives in Bulgaria: First Working Group Report, Center for Information, Democracy and Citizenship and American University in Bulgaria, April 2025, https://www.aubg.edu/wp-content/uploads/2025/05/Bulgaria-Working-Group-Report_April-2025_Mapping-Disinformation-Narratives-in-Bulgaria.pdf (accessed September 25, 2025).

65 Konrad Bleyer-Simon, "Disinformation Landscape in Hungary," EU Disinfo Lab, June 2023, https://www.disinfo.eu/wp-content/uploads/2023/06/20230521_HU_DisinfoFS.pdf (accessed September 25, 2025).

66 Ibid., 50.

67 Katarína Klingová, Central and Eastern Europe's Blueprint for Social Media Regulation: Recommendations for the New EU Commission, GLOBSEC, October 2, 2024, https://www.globsec.org/what-we-do/publications/central-and-eastern-europes-blueprint-social-media-regulation-0 (accessed July 31, 2025).

68 Nicholas Ross Smith, "The EU and Russia's Conflicting Regime Preferences in Ukraine: Assessing Regime Promotion Strategies in the Scope of the Ukraine Crisis," in: European Security 24, no. 4, 2015, 525–540, here 534.

69 Marius Dragomir et al., "Beyond Online Disinformation: Assessing National Information Resilience in Four European Countries," in: Humanities and Social Sciences Communications 11, no. 1, January 13, 2024, 1-10, here 6.

marily block entire media outlets without a judicial process, raise the specter of overreach.[70] The danger of national regulators misusing existing disinformation or "false news" laws in this context is real, as shown by Hungary's 2020 criminalization of disinformation (a law introduced during the COVID-19 pandemic, making it a crime, punishable by prison, to spread "false" information that could cause panic). Similarly, a Romanian emergency decree enabled website blocking during the same health crisis and Lithuania's media regulator instructing Internet Service Providers (ISPs) to restrict access to websites accused of spreading false information.[71]

The problem is partly due to the amorphous boundary between disinformation and protected speech within EU doctrine. While much disinformation falls under the umbrella of free expression unless it crosses a clear legal line, regulatory actions that hinge on vague or overbroad definitions trigger public skepticism. When policies are perceived as being wielded more harshly against ordinary citizens than entrenched institutions, or as suffused with "elite" prerogative, they provoke suspicion, feeding the narrative that regulation serves to muffle dissent rather than nurture a strong marketplace of ideas.[72]

The specter of censorship, intended or otherwise, also looms large. Sweeping regulatory mandates, especially in less resilient democracies, can have a chilling effect on journalism and investigative reporting. There is a very real risk that governments with authoritarian leanings could seize upon EU-mandated removal mechanisms as a pretext for silencing opponents, cloaking domestic repression in the mantle of European compliance. Where definitions of "harm" or "disinformation" are ambiguous, or left to the mercies of local authorities, the result may be a dragnet that traps lawful civic activism, public-interest reporting, and open, democratic debate.[73]

Online platforms, for their part, are in a particularly delicate position. While legally obliged to implement "reasonable, proportionate and effective" mitigation measures,[74] they also operate under the constant threat of substantial fines and reputational harm. In practice, this may often incentivize an overly cautious approach, leading to the removal of legitimate speech in order to pre-empt regulatory sanction.[75] Mechanisms for appeal are frequently slow, opaque, or excessively complex, offering little meaningful redress, particularly in jurisdictions where independent oversight exists more in form than in substance.[76]

The dangers do not end there. States with fragile or captured institutions may use speech regulations as tools for media control, selectively targeting critical voices or opposition outlets under the umbrella of EU law.[77] Blanket bans on entire media organizations, justified as necessary to stymie foreign propaganda, risk turning the exception into the rule – hurting fundamental rights to information and setting troubling precedents ripe for future abuse.[78]

70 Ronan Fahy, "The EU's Regulatory Push Against Disinformation," in: Verfassungsblog: On Matters Constitutional, August 5, 2022, 4, https://doi.org/10.17176/20220805-182037-0 (accessed July 31, 2025).

71 Ibid., 4-5.

72 András Koltay, "Freedom of Expression and the Regulation of Disinformation in the European Union," in: Ronald J. Krotoszynski, Jr., ed., Disinformation, Misinformation, and Democracy: Legal Approaches in Comparative Context, Cambridge: Cambridge University Press, 2025, 133-160, 157.

73 Judit Bayer, et al., The Fight Against Disinformation and the Right to Freedom of Expression, European Union, July 2021, 47, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf (accessed July 31, 2025).

74 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act), in: Official Journal of the European Union, L 277, 27 October 2022, 25.

75 Jahangir, 2025b.

76 Tom Gibson, How Will the EU's Digital Services Act Impact Journalism?, Committee to Protect Journalists, July 29, 2020, https://cpj.org/2020/07/how-will-the-eus-digital-services-act-impact-journalism/ (accessed July 31, 2025).

77 Oliver Money-Kyrle, The European Media Freedom Act: The Solution to Capture or Just Fine Print?, in: The Review of Democracy, March 26, 2025, https://revdem.ceu.edu/2025/03/26/european-media-freedom-act/ (accessed July 31, 2025).

78 Seamus Allen, Censorship and Foreign Propaganda in the European Union. The Implications of the EU's Council Regulation 2022/350 for Freedom to Access Information, Institute of International and European Affairs, November 30, 2023, 9, https://www.iiea.com/images/uploads/resources/Censorship_and_Foreign_Propaganda_in_the_EU.pdf (accessed July 31, 2025).

Conclusion and Policy Recommendations

Hailed as a guiding light in global regulatory circles, Europe's strategy on disinformation blends robust platform governance with forward-looking risk mitigation measures. Yet, beneath these ambitions, significant weaknesses persist. Legal countermeasures, including the criminalization of certain forms of disinformation, do exist in some countries, but their use is marred by controversy, given longstanding concerns over freedom of expression.[79] Moreover, the effectiveness of national responses varies considerably, shaped by the strength and independence of domestic media, the level of public trust, the capacity of institutions, and the vitality of civil society.

The Nordic nations and Baltic states, for example, have emerged as standard bearers in championing media literacy and cybersecurity, cultivating a culture of awareness and digital resilience. By contrast, states along the EU's eastern flank, those bordering Russia, Belarus, or Ukraine, may more frequently resort to the blunt tool of outright bans on foreign propaganda outlets as a first line of defense.

This sobering reality demands a two-pronged response. First, a rigorous audit of existing initiatives is overdue. Policy makers must get to grips with why well-intentioned efforts fall short, whether due to poorly chosen partners, lack of data, or the proven lack of effectiveness of certain interventions.

Second, the time is ripe for more audacious thinking and a greater appetite for risk. At a fundamental level, it is worth asking if placing disinformation at the center of the framework for media regulation is fit for purpose.[80] Manipulative, biased content is as old as the public sphere itself, seemingly defying every legislative or regulatory bulwark erected to contain it. In today's high-stakes environment, where adversarial actors like Russia can produce ever-more sophisticated tactics and generative AI turbocharges the production of falsehoods, it is folly to believe that disinformation can ever be fully eradicated, even with boundless effort and investment.

This realization points towards the necessity of reframing the EU's approach. This involves moving away from a negative, adversarial fixation on "disinformation bashing" and pivoting towards a constructive, citizen-centric effort to rebuild trust in Europe's information ecosystem. The keystone for such a reform is incentivization.

Below is a series of recommendations on how to operationalize this new approach and make the fight against disinformation more effective. Taken together, these measures mark a decisive change of direction from relentlessly chasing and exposing bad actors toward empowering and incentivizing institutions and citizens to produce, disseminate, and consume information that is accurate, reliable, and trusted.

- Strengthen incentive frameworks for quality journalism: The EU should create pan-European funding and recognition schemes that reward demonstrable promotion of editorial independence and public-interest reporting. Incentives should complement the European Media Freedom Act by supporting outlets that deliver high-quality, fact-based journalism across borders.

- Back cross-border European media institutions: To counterbalance both tech dominance and national-level censorship, the EU should invest in visionary cross-border media institutions, including tech infrastructure, that can pool resources, provide independent perspectives, and serve as a democratic counterweight in the information space.

- Invest in AI detection and transparency standards: With generative AI accelerating the spread of disinformation, the EU should spearhead the development of detection technologies and enforce systems of labelling requirements for AI-generated content.

79 Marius Dragomir, "Probing the Paradox: Are Governments Fighting or Fueling Disinformation?," in: Frontiers in Communication 10, March 10, 2025, 1-7, here 6.

80 The author of this article is currently developing an article that discusses why this framework proves counterproductive in restoring the health of Europe's information ecosystems.

- Reward the promotion of high-quality journalism: Platforms should receive regulatory reprieves, preferential compliance treatment, or public recognition for demonstrably prioritizing trustworthy journalism over clickbait or manipulative content. Incentives should be tied to measurable benchmarks rather than voluntary pledges.

- Introduce audience-facing trust systems: The EU should incentivize platforms to roll out robust trust systems, including clear content labelling, credibility indicators, and provenance tools, to guide users toward reliable information. Such measures would help restore public trust and counter the algorithmic amplification of harmful content.

- Embed media literacy in education systems: Governments should integrate media literacy and critical thinking as permanent components of national curricula, supported by lifelong learning initiatives. To ensure uptake, incentives such as certification schemes, career benefits, or financial support for adult learners should accompany these programmes. This combined approach is expected to cultivate resilient citizens who are both motivated and equipped to better navigate and resist disinformation.

- Reward trusted information initiatives in the AI era: Recognizing the centrality of trust in an AI-dominated information space, governments should establish reward and incentive systems for initiatives that promote transparent, verifiable data as well as independent journalism. They should also invest in the infrastructure and computing power required by such players. Access to computational resources, including cloud and AI training capacity, should be made freely available to trusted actors, ensuring they can compete with large platforms and malicious operators on a level playing field.

- Build audience-anchored data systems and collective trust mechanisms: Media companies must adapt to an ecosystem increasingly dominated by AI-generated content by investing in new systems of data delivery that are fully anchored in audience needs and trust. To maintain their competitive edge, media companies should unite to create industry-wide mechanisms that certify and signal provenance-transparent, non-AI, high-quality content. Such collective initiatives would differentiate professional journalism from synthetic output and strengthen its value in the information marketplace.

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
BECID	Baltic Engagement Centre for Combating Information Disorders
BJP	Bharatiya Janata Party
CIA	Central Intelligence Agency
DGAP	German Council on Foreign Relations
DSA	Digital Services Act
DSCs	Digital Services Coordinators
EMFA	European Media Freedom Act
EU	European Union
FACT	Fighting against Conspiracies and Trolls
FIMI	Foreign Information Manipulation and Interference
G7	The Group of Seven
genAI	Generative AI
HLEG	European Commission High-Level Expert Group
ISPs	Internet Service Providers
KPI	Key Performance Indicator
LGBTQIA+	Lesbian, Gay, Bisexual, Transgender, Queer, Intersexual, Asexual, and many other terms (such as non-binary and pansexual) that people use to describe their experiences of their gender, sexuality and physiological sex characteristics
LLMs	Large Language Models
NATO	North Atlantic Treaty Organization
NGOs	Non-Governmental Organizations
OECD	Organisation for Economic Co-operation and Development
RT	Russia Today
SEO	Search Engine Optimization
VIGINUM	Service for Vigilance and Protection against Foreign Digital Interference
VLOPs	Very Large Online Platforms
VLOSEs	Very Large Online Search Engines
UNESCO	United National Educational, Scientific and Cultural Organization
U.S.	United States

Authors

Marius Dragomir

Marius Dragomir is the Director of the Media and Journalism Research Center (MJRC), an international think tank focused on media, journalism, politics, and technology. He is also a researcher at the University of Santiago de Compostela in Spain and has been a visiting professor at Central European University (CEU) in Vienna, where he has taught journalism, research design, advocacy, and policy engagement. Previously, Marius directed the Center for Media, Data and Society (CMDS) at CEU and managed the global research and policy portfolio of the Program on Independent Journalism at the Open Society Foundations (OSF) in London. With over 25 years in media research, he has designed and led more than 20 global projects, including Mapping Digital Media, Television Across Europe, and more recently, Media Influence Matrix, Global Media Finances Map, and State Media Monitor, the latter covering 170 countries. Marius also received the 2025 Media and Democracy Karol Jakubowicz Award for his original and rigorous research on media ownership and regulatory frameworks.

Dmytro Iarovyi

Dmytro Iarovyi is an Associate Professor at the Kyiv School of Economics and Academic Director of its PhD program in Public Policy. He is also a researcher in Communication and Information at Vytautas Magnus University, where his studies explore societal resilience to information disorders in Lithuania and Ukraine. His research and teaching focus on disinformation, propaganda, social media, pop culture, and government relations. Beyond academia, he serves as a consultant on government relations at Hillmont Partners. He is the author of monograph Psychological Practices of Civil Confrontation in Social Media of Modern Ukraine (2020).

Cristina Ivan

Cristina Ivan is Director of the National Institute for Intelligence Studies, “Mihai Viteazul” National Intelligence Academy in Romania and a researcher in security and intelligence studies. She holds a PhD in cultural studies. Over the past 20 years she has specialized in the cultural study of violence, radicalization and terrorism, propaganda and disinformation, critical intelligence studies etc. She is co-editor of the Routledge Handbook of Disinformation and National Security.

Katja Muñoz

Katja Muñoz, a Research Fellow at the DGAP Center for Geopolitics, Geoeconomics, and Technology, delves into the intricate interplay between social media and politics. Her expertise lies in understanding the mobilization potential of social media influencers, the dynamics of disinformation, and platform regulation. Her research also encompasses the intersection of artificial intelligence (AI) and democracy, with a particular focus on emerging AI trends that reshape online information integrity. She holds a special interest in mapping and monitoring AI’s impact on the electoral process, including campaigning, voter sentiment analysis, and governance systems.

Christopher Nehring

Christopher Nehring is a researcher, analyst and media expert. He is currently the Director of Intelligence at the cyberintelligence.institute in Frankfurt, Germany (lead researcher on intelligence, hybrid threats and disinformation/information manipulation). He is an expert on disinformation and has been a visiting lecturer on disinformation, AI, intelligence and media at the Konrad Adenauer Foundation‘s Media Programme South East Europe and at the Faculty of Journalism and Mass Communication at Sofia University. He also worked as a senior analyst at the Institute for Global Analysis in Sofia and was the Head of Research at the German Spymuseum in Berlin (2015-2020). He holds a PhD in Eastern European History with a thesis in Intelligence History (Cooperation of the KGB with the East German Stasi and Bulgarian Intelligence).

About the Aspen Institute Germany

The Aspen Institute Germany is an independent, non-partisan organization that promotes values-based leadership, constructive dialogue between conflicting parties, and transatlantic cooperation to strengthen a free and open society. Founded in 1974 in Berlin, the Institute has been bringing together decision-makers and experts from politics, business, academia, media, culture, and civil society for 50 years to address the challenges of our time.

More information about the
Aspen Institute Germany:
www.aspeninstitute.de

Follow us:

@AspenGermany

@aspen_germany

@AspenGermany

Aspen Institute Germany