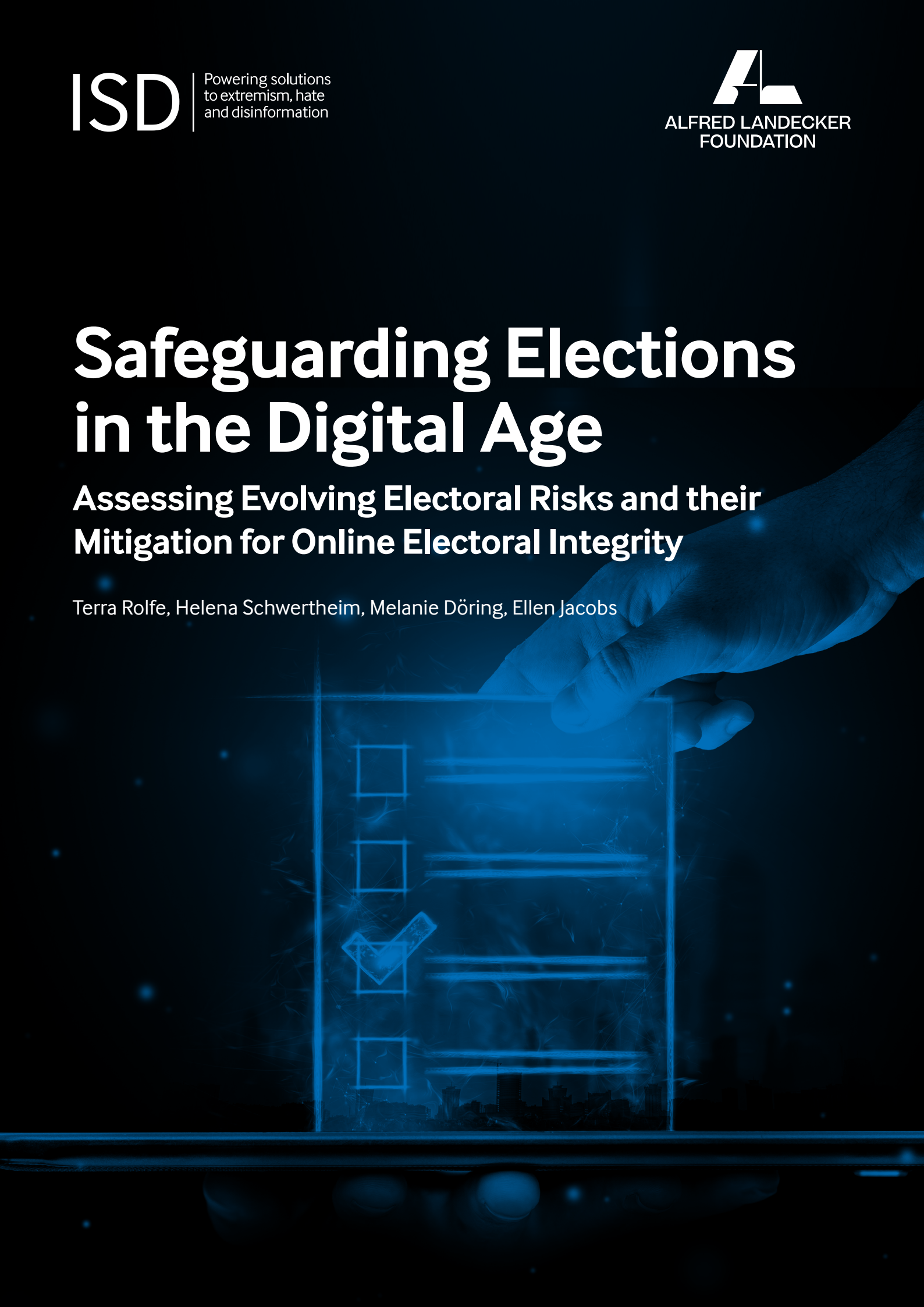
ISD | Powering solutions to extremism, hate and disinformation

ALFRED LANDECKER FOUNDATION

Safeguarding Elections in the Digital Age

Assessing Evolving Electoral Risks and their Mitigation for Online Electoral Integrity

Terra Rolfe, Helena Schwertheim, Melanie Döring, Ellen Jacobs



About the Digital Policy Lab

The Digital Policy Lab (DPL) is an inter-governmental working group focused on charting the policy path forward to prevent and counter the spread of disinformation, hate speech, and extremist and terrorist content online. It is comprised of representatives of relevant ministries and regulatory bodies from liberal democracies. The DPL aims to foster inter-governmental exchange, provide policymakers and regulators with access to sector-leading expertise and research, and build an international community of practice around key challenges in the digital policy space. We thank the Alfred Landecker Foundation for their support for this project.

About this Paper

As part of the DPL, the Institute for Strategic Dialogue (ISD) organised working group meetings between March and April 2024 on the topic of assessing mitigation measures to safeguard electoral integrity. The working group consisted of DPL members representing national ministries and regulators from Australia, Canada, the European Commission, Europol, Germany, Italy, Slovakia and Switzerland. Participants also included representatives from civil society and academia. While participants contributed to this publication, the views expressed in this paper do not necessarily reflect the views of all participants or any governments involved in this project.

ALFRED LANDECKER FOUNDATION

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org

About the Authors

Terra Rolfe was a Digital Policy Associate at ISD UK, where she primarily supported the work of the Digital Policy Lab (DPL), an intergovernmental working group on digital policy responses to disinformation, hate speech and extremism. Terra also supported ISD's global digital policy work, with a focus on artificial intelligence policy. Prior to joining ISD, she interned at the Amsterdam-based civil society organisation ALLAI, advocating for responsible artificial intelligence legislation in the EU. She holds a BSc (Hons) from Leiden University and is currently an MSc candidate at the Oxford Internet Institute, University of Oxford.

Helena Schwertheim is a Senior Digital Policy and Research Manager and leads the Digital Policy Lab (DPL), an intergovernmental working group focused on policy responses to prevent and counter disinformation, hate speech and extremism. As part of the Digital Policy Team, she advises key governments, international organisations and tech companies, and collaborates with ISD's Digital Analysis Unit to translate research into actionable digital policy recommendations, with a focus on Technology Facilitated Gender-Based Violence (TFGBV). Previously, Helena managed digital policy and research projects at Democracy Reporting International. She also has experience working in risk and political analysis in international organizations and think tanks, including at the UN World Food Programme in Rome, and the think tank International IDEA in Stockholm.

Melanie Döring is Project Coordinator for ISD Germany's Digital Policy Lab (DPL). She is part of ISD's Digital Policy team, regularly attending meetings to advise policymakers and the German government. Previously, Melanie worked in German Development Cooperation (GIZ)'s Sector Project International Digital Policy, the European Parliament and an international agency for public relations and public affairs. She holds two MScs in European and International Governance from the Vrije Universiteit Brussels and in European Integration from the Brussels School of Governance, as well as a BA in Communication Science and Political Science from the Johannes Gutenberg-Universität Mainz.

Ellen Jacobs is a Senior Digital Policy Manager for ISD US. She focuses on mitigating the effects of online harms, including those from disinformation, extremism and hate speech, by advancing ISD's digital policy and tech accountability objectives. In her role, she represents ISD to a diverse array of stakeholders including elected officials, NGOs, academics, researchers, and others interested in platform accountability and regulation. Prior to joining ISD, Ellen was at the Omidyar Network, where her funding and advocacy work focused on issues related to platform accountability and open source technologies. Ellen holds an MIA in Human Rights and Humanitarian Policy from Columbia SIPA and a BA in International Studies from the University of Chicago.

Acknowledgements

We would like to thank all participants of the working group, including experts from governments, civil society, academia and industry, for their contributions. We would like to give special thanks to the speakers and contributors to this paper for their valuable insights and feedback: Joshua A. Tucker (Professor of Politics, Director of the Jordan Center for the Advanced Study of Russia, and Co-director of the Center for Social Media and Politics, New York University), Ravi Iyer (Managing Director of the University of Southern California Marshall School's Neely Center for Ethical Leadership and Decision-Making), Jan Hurtik (Slovakian Council of Media Services), Jakub Rybar (Slovakian Council of Media Services), Felix Kröner (Reset Tech), Beatriz Saab (Democracy Reporting International) and Flora Rebello Arduini (Senior Advisor on Tech & Human Rights). We would like to thank ISD colleagues Henry Tuck, Christian Schwieter, Mauritius Dorn and Sid Venkataramakrishnan, who reviewed this paper.

Contents

Executive Summary

This Policy Brief reviews the effectiveness of key measures taken by democratic governments, the tech industry and civil society to mitigate online risks posed to electoral integrity. The analysis also explores the challenges and limitations of research in this field, based on the current understanding of existing responses by governments, online platforms, civil society and academia. While no single solution will suffice, a combination of strategies, continually assessed and refined, will be critical to safeguarding electoral integrity. Ongoing research and enhanced access to platform data are crucial for understanding and improving these efforts over time.

Throughout, the paper provides recommendations for governments, regulators, researchers and industry, who should collaborate through a multi-stakeholder approach to support a better understanding of the impact of mitigation strategies. Key recommendations include:

For governments and regulators:

- **Establish strong communication channels between regulators and researchers to facilitate research exchange on measuring risks to electoral integrity and the efficacy of interventions, beyond election periods.** Effective communication and collaboration are essential for developing a comprehensive understanding of the risks and identifying best practices for mitigation. A continuous dialogue allows regulators, academia and civil society to share insights, methodologies and data before, during and after elections, enhancing the overall quality and impact of research. This collaborative approach will ensure that interventions are informed by diverse perspectives and grounded in the latest evidence, ultimately strengthening electoral integrity. For example, the European Cooperation Network on Elections (ECNE) brings together national election management bodies, regulators and other stakeholders to share information, best practices and research findings. The European Digital Media Observatory (EDMO) is a similar collaborative platform which connects fact-checkers, media literacy experts and academic researchers to tackle disinformation.

- **Legislate to require platforms to provide a minimum level of platform transparency and data access, overseen and enforced via independent regulators.** Access to platform data and transparency of platform policies is paramount for researchers and regulators to be able to measure the efficacy of interventions. This includes:

  - **Regulation should protect a minimum standard of data access to facilitate an evidence base on the efficacy of mitigation measures**. This is necessary as many platforms are restricting access to data for researchers, especially in contexts where this is not a legal requirement. Among the regulations driving greater data access is the EU's Digital Services Act (DSA), which introduces data access obligations for the largest platforms and search engines for vetted researchers, and requires platforms to facilitate access to already public data (Article 40). At the time of writing, Canada's proposed Online Harms Act (Bill C-63) aims to authorise the proposed Digital Safety Commission of Canada to accredit certain persons conducting education, advocacy, awareness, or research activities on online harms related to the purposes of the Act (section 73). Similar, context-appropriate requirements should be established wherever legal obligations are still missing to ensure platforms provide appropriate data access while balancing concerns such as adequate data protection.

  - **Ensure that data access application processes and requirements for researchers are predictable and standardised across platforms as far as possible, especially when vetting procedures are required.** Regulators must ensure that researchers have safeguards in place to protect user privacy and rights when processing platform data. However, it is crucial that regulators facilitate this process to ensure that researchers can easily obtain the data they need, and processes should not be overly burdensome or bureaucratic. In the context of elections, it is particularly important that these processes are responsive to enable research to be conducted in near real-time during election campaigns.

  - **The types of data and metrics available via different Application Programming Interfaces (APIs) should, where possible, be standardised to allow for meaningful cross-platform comparisons.** Previous instances, such as the political ads

repositories during the 2019 European Parliament elections, demonstrated the challenges posed by inconsistent data formats and metrics. By ensuring compatibility and comparability of data, regulators can significantly enhance the quality and coherence of research on electoral integrity and the efficacy of interventions. They will also be better positioned to establish reasonable cross-industry expectations and standards.

- **Provide legal safeguards for researchers and journalists conducting public interest investigations.** Civil society, academia and journalism provide a crucial public service by providing evidence of online risks posed to electoral integrity. This work is, however, increasingly threatened not just by a lack of data access on behalf of platforms, but also strategic lawsuits against public participation (SLAPPs), ill-intentioned litigations aimed at intimidating those voicing public criticism.[1] Most organisations and individuals conducting research into online harms to elections do not have the appropriate legal resources available to combat such efforts. Policymakers should consider legal instruments which ensure public interest research is not threatened by unfounded and abusive litigation attempts. Where such instruments are already in place, such as the EU Anti-SLAPP Directive,[2] governments should ensure their predictable enforcement.

For researchers and industry:

- **Expand the regional focus of electoral integrity research.** Public interest researchers should prioritise expanding the regional focus of studies on electoral integrity to include a wider range of countries and languages beyond the US, European or Western contexts. Given the evidence of regional variation in the efficacy of mitigation measures, it is crucial to examine their effectiveness across a broader range of different contexts. Research should include Global Majority and less-studied Western countries. Comparative cross-country studies should be undertaken to identify and understand these variations. A broader scope would enhance generalisability of findings, inform tailored and effective interventions, and ensure that all regions benefit from advancements in safeguarding electoral integrity. Increased funding and collaboration with local researchers and institutions will be essential to support this expanded research agenda. Platforms often focus their mitigation measures primarily on US and EU (or Western) regions. For example, in 2024, OpenAI and Anthropic's election integrity measures only included the US and EU.

For industry:

- **Consistently enforce existing policies on information and election integrity, and work with researchers and regulators to update relevant systems, policies and processes to reflect the latest research on the efficacy of mitigation measures.**
  - Existing policies, such as content moderation policies, should be enforced consistently and predictably. Local legislation (such as electoral silence periods for political advertisements) should be followed. Geographic equity should be prioritised by ensuring that platforms devote sufficient resources to contexts beyond Western countries.
  - Transparency reports and publicly available data should facilitate external scrutiny of current policy enforcement, platform claims of election preparedness measures, and compliance with local laws and legislation.
  - Platforms should expand their coordination with regulators, academia and civil society, a crucial measure for maintaining a trustworthy and safe online environment, particularly during election periods. Effective communication channels before, during and after elections can provide platforms with valuable insights to address emerging threats and adapt to new challenges in the digital landscape. Such coordination can also inform the adjustment and updating of platform policies to reflect new developments.
- **Enable further research into the efficacy of mitigation measures:**
  - **Provide regulatory authorities and researchers with information and data to better study the efficacy of mitigation measures.** Transparency reports and publicly available platform data should allow regulators, researchers and independent

auditors to better understand the aims of platforms' policies, methods and processes, the scale of policy enforcement, and how adjustments to current systems and processes can mitigate identified electoral risks. Information on platform policies and processes is also key for researchers to understand what interventions are conducted.

- **Conduct and share internal risk assessments and mitigation measure plans with regulatory authorities and independent auditors, and, where possible, with independent researchers.** In the EU, the DSA requires many platforms to carry out internal risk assessments that identify the likelihood and impact of potential online harms, including those posed to electoral integrity, and the proportionality and effectiveness of mitigation measures. Similar risk assessments conducted in other jurisdictions, for example as a part of Human Rights Impact Assessments (HRIAs), should also be shared as transparently as possible.

- **Consider establishing independent researcher-platform partnerships.** Such research can facilitate unparalleled access to data for researchers. These partnerships can also ensure that research questions and scope are designed appropriately and effectively. Past examples of such partnerships have facilitated some of the strongest existing findings on causality. However, mechanisms for ensuring the independence of research conducted is paramount, and these partnerships should be in addition to a minimum level of standardised data access, rather than in its stead.

Glossary

Application Programming Interface (API)
An API is a software intermediary that allows two applications to communicate with each other. APIs have a huge range of uses, but in the context of this report, they allow researchers to access certain data from some online platforms via requests. As an intermediary, APIs also provide an additional layer of security by not allowing direct access to data, alongside logging, managing and controlling the volume and frequency of requests.

Artificial Intelligence (AI)
AI is a difficult term to define, as its scope and relationship to intelligence is subject to debate. With the goal of creating globally relevant and interoperable policies, ISD follows the OECD's definition of AI as a "machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."[3]

Disinformation
Disinformation is false, misleading or manipulated content presented as fact, that is intended to deceive or harm.

Election denial
Election denial involves claiming that the results or process of an election were illegitimate, despite demonstrable evidence that it was free and fair. It is linked to conspiracy theories and movements.[4]

Election integrity
Election integrity lacks a universal definition. This Policy Brief refers to free and fair electoral processes, based on democratic standards and principles such as political equality, transparency and impartiality, which establish accountability, legitimacy, and trust in their results. Open dialogue and information sharing are crucial aspects.[5]

Extended reality (XR)
XR is a collective term used to describe technologies that blur the lines between the real and digital worlds. It encompasses the related terms of virtual, augmented and mixed reality. Virtual Reality (VR) is a technology that provides almost real and/or believable experiences in a synthetic or virtual way, while Augmented Reality (AR) enhances the real world by superimposing computer-generated information on top of it. A Mixed Reality (MR) experience is one that seamlessly blends the user's real-world environment and digitally created content, where both environments coexist and interact with each other.[6]

Extremism
Extremism is the advocacy of a system of belief that claims the superiority and dominance of one identity-based 'in-group' over all 'out-groups.' It propagates a dehumanising 'othering' mind-set that is antithetical to pluralism and the universal application of human rights.

Foreign Information Manipulation and Interference (FIMI)
FIMI is defined by the European Union Agency for Cybersecurity (ENISA) as "a mostly non-illegal pattern of behaviour that threatens or has the potential to negatively impact values, procedures and political processes. Such activity is manipulative in character, conducted in an intentional and coordinated manner. Actors of such activity can be state or non-state actors, including their proxies inside and outside of their own territory." ENISA explains that the term FIMI aims to refine the concept of disinformation by emphasising "manipulative behaviour, as opposed to the truth of content being delivered."

Generative AI
Generative AI systems are built on deep-learning models trained on raw data which could include books, articles, webpages, Wikipedia entries and images scraped from the internet.[7] These models are designed to detect statistical patterns in their training dataset and "generate statistically probable outputs when prompted,"[8] which are similar though not identical to the data that they are trained on. This Policy Brief focuses on examples of generative AI systems that can be used to generate synthetic text, images, audio and video.

Hate (Speech)
Hate is understood to relate to beliefs or practices that attack, malign, delegitimise or exclude an entire class of people based on protected or immutable characteristics, including their ethnicity, religion, gender, sexual orientation or disability. Hate actors are understood to be individuals, groups or communities which actively and overtly engage in the above activity, as well as those who implicitly attack classes of people through, for example, the use of conspiracy theories and disinformation. Hateful activity is understood to be

antithetical to pluralism and the universal application of human rights.

Misinformation

Misinformation is false, misleading or manipulated content presented as fact, irrespective of an intent to deceive.

Radicalisation

Radicalisation is a term used in this context to describe the process by which an individual adopts an extremist ideology (defined above). This may or may not enable acts of violent extremism or terrorism. In the literature on terrorism and violent extremism, a frequent distinction is made between cognitive radicalisation (adopting extremist beliefs) and behavioural radicalisation (the process leading up to violent behaviour).[9]

1. Introduction

2024 marks a historic year for electoral processes, as almost half of the world's population take part in major elections.[10] Online platforms remain important spaces for voters' political opinion formation and debate. However, over the last decade the risks of a range of online harms to electoral integrity have become apparent. Online platforms continue to be used by malign actors, ranging from hostile states to extremist groups, to influence electoral outcomes or undermine faith in electoral processes. Such electoral disinformation campaigns deliberately spread false or misleading information around voting processes, policies and candidates.[11] Hate speech and harassment, especially against female candidates and marginalised communities, also threaten to push politicians and activists out of the public sphere and silence affected citizens online.[12]

Online platforms are vulnerable to misuse due to their design, particularly their algorithmic amplification of highly engaging, borderline content which is just below the threshold of illegality.[13] With new challenges from technologies such as generative AI, there is vital work to be done to protect online electoral integrity. At the same time, many online platforms have made cuts to their online safety and election integrity teams and further restrict researchers' access to platform data; this threatens to undermine the ability to detect and respond to disinformation campaigns, hate speech and harassment.

In response to this ever-evolving threat landscape, industry, governments and civil society have developed responses to these risks, ranging from regulatory initiatives to non-regulatory policy approaches. However, what mitigation measures are effective, and how can these be scaled across different contexts?

This Policy Brief assesses efforts by the actors mentioned to safeguard online electoral integrity, reviewing the current online electoral risks and platform features that may make online platforms vulnerable to misuse or interference. Key responses by online platforms are also identified. The brief then distils core insights from empirical research on commonly recommended mitigation measures for safeguarding electoral integrity. These include

- Improving content moderation systems;
- Adjusting algorithmic recommender systems;
- Making political advertising systems more transparent;
- Awareness-raising measures such as labelling and pre-bunking, and cross-platform cooperation.

No single mitigation measure will be a silver bullet. Throughout this assessment, the need for further research on the effectiveness of mitigation measures is recognised. Importantly, a major challenge in assessing the impact of mitigation measures is the lack of transparency regarding online platform policies and processes, as well as the limited access to online platform data for independent researchers, for example on user responses to interventions.

Throughout this paper, the term "platforms" will be used to refer to online platforms for ease.

2. Online risks to electoral integrity

Despite increasing efforts to protect the integrity of elections, they remain susceptible to attempts by malign actors to undermine their processes, including in the online sphere. This section outlines some of the most common types of online risks to electoral integrity and identifies ways in which platforms' functionalities, architectures and systems are vulnerable to misuse or exploitation in an electoral context.

2.1 Online threats with the potential to undermine electoral integrity

The European Commission's Guidelines for providers of Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) on the mitigation of systemic risks for electoral processes under the Digital Services Act (DSA) mention several online threats to electoral integrity which are each briefly described below.[14]

2.1.1 Electoral mis- and disinformation

Electoral disinformation is intended to sow distrust around election processes, whereas electoral misinformation refers to the unintentional spread of false or misleading information which nonetheless has the potential to undermine electoral processes. Frequent electoral disinformation tactics include:

- The dissemination of unfounded claims of electoral fraud or alleged conspiracies to influence results (election denialism);
- The distribution of false information before, during, or after elections, such as incorrect election dates, polling locations, or eligibility criteria to influence voters' perceptions and behaviours.[15]

Disinformation actors may be domestic or foreign, and frequently exploit existing polarisation around domestic issues to deepen divides between citizens.[16]

ISD has observed such distortions of the information environment in numerous elections across the globe in recent years.[17] A report by the European Digital Media Observatory (EDMO) found widespread disinformation narratives targeting all 11 European elections held in 2023. Most were attempting to delegitimise the election via false claims of voter fraud, foreign influence or unfair practices.[18]

2.1.2 Foreign Information Manipulation and Interference (FIMI)

Attempts to influence the information ecosystem can also originate from foreign actors and states, though attribution is notoriously difficult. The European External Action Service (EEAS) coined the concept of Foreign Information Manipulation and Interference (FIMI) for concerted efforts of this kind, relating to manipulative activities that are often legal but harmful. Common FIMI tactics, techniques and procedures (TTPs) include:

- Distraction and distortion;
- A frequent use of diplomatic channels as well as assets like automated accounts ("bots") or inauthentic websites for dissemination;
- Increasingly sophisticated impersonation;
- A preference for image- and video-based, multilingual content.[19]

In this sense, FIMI is both narrower and broader than the concept of disinformation. FIMI activity is political in nature and often increases around key events like elections. However, the EEAS has warned not to overestimate the threat and thus the extent of foreign influence in processes like elections.[20]

2.1.3 Proliferation of hate speech and the spread of terrorist and violent extremist content (TVEC)

Online hate speech can have significant psychological impacts on victims and may be used in a targeted manner to intimidate or silence individuals and groups on the basis of their race, religion, ethnicity, immigration status, sexual orientation, gender, sex, or disability. Online hate is thus disproportionally targeted at traditionally marginalised groups such as women, BIPOC (Black, Indigenous and People of Colour), indigenous communities, religious minorities and members of the LGBTQ+ community.[21] Politicians and public-facing professions including journalists and election workers are also frequently targeted by hate, defamation and intimidation campaigns.[22] These can dissuade people from being outspoken on political issues online, or deter candidates from running for public office. By limiting democratic participation and representation, this chilling effect undermines democratic institutions, norms and values for all.[23]

A growing body of policy-oriented academic research is shedding light on an expanding tactical playbook of coordinated and organic influence operations aimed at distorting democratic discourse around elections, where forms of Online Gender-Based Violence (OGBV) frequently occur. This can include sexual harassment as well as smear and gendered disinformation campaigns targeting women candidates and public office-holders.[24] For example, during the 2020 US election, ISD found that women and candidates from an ethnic minority background running for Congress were more likely than men and those who do not have an ethnic minority background to receive abusive content on Facebook and X (then Twitter).[25]

Due to different legal traditions and cultural understandings, the thresholds and types of hate speech that are illegal differ across jurisdictions. For example, the DSA aims to curb the spread of illegal hate speech across the EU, yet the bar for legality mostly depends on member states' individual criminal codes. The 2008 Council Framework Decision on combating certain forms of expressions of racism and xenophobia required EU member states to criminalise public incitement to violence and hatred based on race, colour, religion, descent or national or ethnic origin in an effort of harmonisation.[26] However, implementation varies[27] and many differences and diverging interpretations of what is considered illegal hate speech remain. Thresholds of illegality are hard to determine even for experts within a jurisdiction, often depending on delicate balancing acts between fundamental rights on a case-by-case basis.

In many Western contexts, the spread of terrorist or violent extremist content (TVEC) may not feature as prominently as in other regions but it remains a risk to electoral integrity which platforms should monitor and mitigate. This particularly concerns the incitement of political violence to disrupt electoral processes, including attempts to undermine the integrity of election results and democratic institutions. Like hate speech, thresholds and definitions of illegal TVEC vary across jurisdictions. In many contexts, organisations are proscribed or designated as terrorist entities by state authorities.

2.1.5 Synthetic content fabricated through emerging technologies

Emerging technologies, such as generative AI and Extended Reality (XR), have the potential to affect democratic processes by amplifying existing online threats.[28] Generative AI provides a cost-effective tool for creating eye-catching content and propaganda[29]. Malign actors may use it to create and distribute more persuasive, affordable and automated disinformation, FIMI and hate campaigns..[30] Emerging technologies also pose several distinct risks. Generative AI's outputs may contain false or nonsensical information (hallucinations) regarding elections.[31] On the other side, people's awareness that content may be artificially altered creates the "liar's dividend," in which politically incriminating but authentic content can be disputed as false.[32]

AI-generated content can particularly affect women in the public eye through sexualised targeting, including artificially generated non-consensual intimate imagery; this is both harmful to women as individuals and also risks a chilling effect on their participation in public life.[33] For example, weeks before the 2022 Northern Irish legislative election, a young woman politician, Cara Hunter, was targeted by a sexually explicit deepfake video that went viral. After the video spread, she received a substantial number of sexual and violent messages from men worldwide. Hunter won her seat in the Northern Ireland Assembly, but noted that the attack "left a tarnished perception of me... I'll have to pay for the repercussions of this for the rest of my life."[34]

2.2 Platform features and vulnerabilities

Limited data access and transparency into platforms' internal systems and processes makes it challenging to study their vulnerabilities. This section reviews how platform functionalities and procedures may be misused to endanger election integrity.

2.2.1 "Safety last" and deceptive design choices

Platforms often employ default settings that do not prioritise user safety and privacy. Users can take the time to manually adjust these settings, but this may involve choices so granular they discourage them from doing so or platforms may restrict functionalities when users do opt out.[35] Moreover, platforms' use of so-called "dark patterns" incentivises certain behaviours, such as clicking a highlighted button, without users necessarily intending to take this concrete action. Dark patterns can lead users to consume or spread mis- or disinformation; they can also be used to present misleading political information or options to voters, such as emphasising or

de-emphasising certain candidates or issues.[36] This may result in voters making uninformed or coerced decisions.[37] Generally, dark patterns often benefit advertising partners, as they can be used to trick users into sharing more personal data than they intend. This data can then be used to micro-target political ads and manipulate voting behaviour.[38]

2.2.2 Algorithmic recommender systems

Most platforms host more content than users can consume, leading many platforms to move from reverse-chronological displays to feeds showing users the "most interesting" content via algorithmic recommendation systems. These algorithmic ranking systems make automated decisions about which pieces of content to prioritise or demote on feeds or in search results, who to connect with, who or what pages to follow – ultimately shaping the online experience of billions of users. "Recommended content" is usually material which is most likely to bring value to the company – for example, content likely to increase user engagement ("likes", re-shares, etc.) and average time spent on the platform.[39]

The "engagement problem" describes how users tend to interact heavily with so-called borderline content that nears the line of platforms' terms of service (ToS). Borderline content is often divisive, misleading and emotionally charged; it tends to reinforce stereotypes and increase discrimination against already marginalised groups.[40] Meanwhile, algorithms remain particularly opaque functions, often acting as "black boxes" of limited interpretability. Platforms tend to guard them as trade secrets, making it challenging for researchers to gather insights on their effects.[41]

A recent study into TikTok's design by Politico and Northeastern University found that the platform's non-transparent, AI-based recommender systems prioritise highly engaging content to increase users screen time. When researching content on the Israel-Palestine conflict based on analysing hashtags, 20 times more pro-Palestinian leaning content was produced by users, but this was not reflected in user feeds. Instead, proportions changed around key events over time, presumably due to regular changes in algorithms. This distorts political discourse and risks undermining citizens' abilities to gather reliable information to inform their vote.[42]

2.2.3 Terms of Service (ToS) and content moderation

Decisions to remove, downrank or keep content and accounts online affect users' human rights, such as personality rights, freedom of speech or the right to information. Most major platforms have policies in their ToS against misinformation. For example, Meta's policy is to remove misinformation where it may contribute to the risk of imminent harm or interfere with the functioning of political processes.[43] Platforms base most content moderation decisions on their ToS rather than legal obligations in respective jurisdictions. While most ToS are based on international legal standards, they tend to be vaguely worded and are often not enforced in a consistent manner.[44]

Due to the vast volume of content hosted on platforms, content moderation decisions are increasingly conducted by automated tools. On the one hand, automated systems may often not be as accurate as human experts as these decisions tend to require complex balancing acts between different human rights and a detailed knowledge of local laws and cultural particularities. On the other, human content moderators also frequently make wrong decisions or fail to detect hate speech that should be moderated.[45] It is also an emotionally-taxing and mentally draining task; and as content moderation is a cost factor, they have limited time to assess very large amounts of content.[46]

The number of human content moderators employed varies widely depending on regions, languages and platforms, resulting in different regional standards for online electoral integrity. According to leaked documents released as part of the "Facebook Papers," 87 percent of the platform's global spending on classifying misinformation is dedicated to the US.[47] Typically, the overwhelming majority of capacities to evaluate content are in English, which tends to be followed by languages such as French, German, and Spanish. Most platforms provide very limited resources to content moderation and Trust and Safety teams in other languages.[48]

A study on content moderation in Ethiopian languages concluded that multilingual language models powered by automated systems regularly fail to detect harmful content in so-called minority languages. These models are simultaneously trained on several so-called "low-resource languages" such as Urdu, which is among the

most-spoken languages globally but not very present in online texts that feed into these systems, or Amharic. However, the systems use "cross-lingual transfers", simply translating lessons they learned on what is harmful in English, a copy-paste-approach that makes them far less accurate.[49] Such limitations in Meta's content moderation in languages like Tigrinya is alleged to have contributed to increasing polarisation in the context of violence in northern regions of Ethiopia.[50]

2.2.4 Political ads

Most platforms' core business model is based on advertising, and most platforms allow political advertising. Definitions differ across contexts and remain disputed as it is hard to determine what constitutes political content; however, platform definitions usually refer to paid ads from political parties and candidates which may include politically salient issue ads on topics like migration, housing, healthcare or climate. Moreover, most jurisdictions have specific rules around electoral ads such as prescribing how they ought to be marked or setting spending limits.[51]

Political advertising as a form of paid influence has the potential to shape and manipulate voter perceptions and behaviours; it can be used by malign actors to spread disinformation, hate and illegal content.[52] A particular area of concern is the increasing use of political micro-targeting techniques worldwide and across the political spectrum.[53] Leveraging large sets of data on users, often collected without users having a detailed understanding of what they are consenting to, political messaging can be tailored to target certain characteristics or preferences to achieve the highest possible impact. This opaque fragmentation of public discourse and the democratic process is questionable regarding data protection and privacy, voter self-determination and transparency.[54]

3. Platform responses

ISD's Electoral Scorecard provides an overview of major platforms' preparation to safeguard electoral integrity ahead of global elections in 2024 (including Meta, X, YouTube, Snapchat, and TikTok). ISD assessed what policies platforms have in place on information integrity, political ads, hate speech and violent extremism, internal and external resourcing, transparency and state-affiliated media. ISD did not assess the enforcement of these policies.

ISD's assessment shows that platform policies remained unclear and unrelated to election denialism, as it was already the case during the 2016 and 2020 US presidential elections. Platforms have made vague commitments to combat election misinformation and are not universally addressing election denialism, which raises concerns about how they will handle claims leading up to and after the 2024 US presidential election, let alone other elections. Meta, X and YouTube have no explicit mentions of election denialism in their misinformation policies; none of the platforms aside from TikTok have policies that clearly penalise content that claims victory before the election is called. Platforms are similarly divided over approaches to fact-checking, disclosure requirements, and the handling of synthetic media. Combined with the many changes in the political ads space since the last US presidential election, this lack of clear policies and their enforcement leaves political messaging in digital media open to potential misuse.

Summary of Key Platform Announcements for 2024 Elections[55]

Meta
Announcement (28 Nov 2023):[56]

- New policy requiring advertisers to disclose AI or digital techniques used in ads with realistic content enters effect.

Generative AI Update (5 Apr 2024):[57]

- Increased transparency and labelling for generated/manipulated content, with downranking of labelled content.
- Ceased removing misleading manipulated or artificially generated videos of people speaking from July 2024 onwards, with a new policy of labelling such content. This excluded content that violates other Terms of Service, such as voter interference.

European Parliamentary Elections:[58]

- Establishment of Elections Operation Centre, partnerships with 26 fact-checking organisations covering 22 languages and content reviewers in all 24 official EU languages.

Google
Announcement (19 Dec 2023):[59]

- Released new tools and policies including watermarking, disclosure and labelling for generative AI.
- Unveiled initiatives to help users access high-quality information and partnerships for campaign and information security.

European Parliamentary Elections:[60]

- Provided €1.5 million to the European Fact-Checking Standards Network
- Announced a pre-election pre-bunking campaign, AFP journalist training support, $1 million donated for developing anti-disinformation hackathons, and search trend insights via "Google Trends Election Hub".

Gemini chatbot:[61]

- Restricted ability for users to ask election-related questions in active election countries.

TikTok
Announcement (18 Jan 2024):[62]

- Announced media literacy partnerships, founding of US Elections Centre for reliable voter information, verification for political figures and ongoing content moderation.
- Disrupted and removed influence operations, restricted misleading AI-generated content, and proactive/reactive misinformation countering.

European Parliamentary Elections:[63]

- Announced partnerships with 11 fact-checking organisations in 18 languages, potential labelling of unverified content, investments in media literacy campaigns, 6000 moderators for EU-language content and local language election centres in each member state.

OpenAI
Announcement (15 Jan 2024):[64]

- Announced usage policies to prevent misuse of tools like ChatGPT, including restrictions on impersonation and misrepresentation of democratic processes. In the US, ChatGPT will also direct users to the authoritative and nonpartisan CanIVote.org for voting information.

Implementation of the Coalition for Content Provenance and Authenticity:[65]

- Integrated C2PA standard into images generated by DALL-E 3, which is a non-removable digital watermark and will identify images as AI-generated.

- Began directing US users to CanIVote.org for voting information.

Anthropic
Announcement (16 Feb 2024):[66]

- Enacted acceptable use policy prohibiting misuse of AI in campaigning and lobbying, automated misinformation detection, red teaming (a process that involves using an adversarial approach to challenge a model's systems and assumptions) for risk assessment and began providing authoritative sources for US voting queries.

Update (5 June 2024):[67]

- Released detailed risk testing and mitigation process for election-related risks, involving policy vulnerability testing, automated evaluations and re-testing to measure the efficacy of this approach.

- Specific focus on adding extra guardrails in the US and EU.

- Both OpenAI and Anthropic are enhancing their measures primarily in the US and, for OpenAI, the EU. Neither have detailed efforts in other regions.

4. Evaluating the effectiveness of mitigation measures

This section reviews the evidence available on the efficacy of platforms' mitigation measures against risks to electoral integrity amid platform vulnerabilities, including measures to introduce and/or adapt:

- Design features;
- Content moderation and terms of service;
- Algorithmic recommender systems;
- Advertising systems;
- Internal processes;
- Cooperation with trusted flaggers;
- Cooperation with other platforms;
- Awareness-raising efforts.

The limitations to conclusive studies are listed throughout and discussed further in section 5. While these may seem numerous, the aim of this paper is also to provide a measured review of the efficacy of mitigation measures, and highlight areas for further research and study rather than dissuade from the application, testing and measurement of such mitigation measures.

This section largely follows the structure and order of Article 35.1 of the EU's DSA, which describes the types of mitigation measures VLOPs and VLOSEs may put in place to address systemic risks present on their services, as far as they are relevant to safeguarding electoral integrity.[68] Article 35.1(j) DSA on the rights of the child is outside the scope of this Brief. While most measures discussed in the following correspond to the platform vulnerabilities outlined in section 2.2, additional measures' efficacy is evaluated, such as cooperation and awareness-raising efforts. Handling risks stemming from synthetic content (Art.35.1(k)) is discussed throughout this section, as effectively mitigating these requires an amalgamation of measures.

4.1 Adaption of the design, features or functioning of services (Art. 35.1.a)

4.1.1 Feature restrictions

One approach to mitigate risks to online electoral integrity is to slow the spread of harmful content produced by either humans or automated accounts ('bots'). This can be achieved through caps on the use of functions such as messaging, commenting, sharing and forwarding. Besides the risk of automated attacks, research on 'super-users' (i.e. human users with exceptionally high levels of activity on a platform) found that a few malign actors can create and share significant amounts of harmful election-related content through extreme overuse of platforms.[69]

X introduced usage limits in 2023 to fight these phenomena. Currently, it has capped direct messages to 500 per day and posts (including reposts) to 2,400 per day with smaller limits for 30-minute timeframes; additionally, users are limited to following 400 new accounts per day.[70] WhatsApp introduced new forwarding limits for messages and channel updates in 2020: users can select up to five chats to forward content to at a time. A message that was forwarded to a user can only be shared by them with one chat at a time. After five forwards, content is labelled as "forwarded many times" and can only be shared with one more chat at a time. However, since messages are end-to-end-encrypted, WhatsApp is unaware of how often a message is forwarded.[71] According to the company, the new features decreased the spread of "highly forwarded" content by 70 percent.[72] However, the proportion of this content's harmfulness is unknown. Virality may be reduced, but it is unclear the extent to which this has a positive effect on reducing online harms.

A lack of data access coupled with the need to respect user privacy make it difficult to conduct conclusive research, particularly on private messaging platforms. A study from 2019 based on public groups concluded that message limits on WhatsApp were indeed useful in slowing the spread of information, yet fail to undermine the propagation of viral misinformation campaigns. The researchers suggest that limiting specific messages and accounts may be more effective.[73] Another study from 2024 noted researchers' limited understandings of forwarding mechanisms and the efficacy of measures such as flagging viral content and restricting dissemination. The authors found that the latter can be easily circumvented by copying and pasting material into the text message field and sending it directly.[74]

The non-governmental organisation (NGO) Protect Democracy suggests that usage limits should be

reasonable and focused to only capture extreme over-users instead of hindering legitimate accounts.[75] Different limits may be appropriate during elections, for example, for new accounts or entities that relate to voting. A distinction should also be made between mitigation measures on private messaging services and social media platforms, as the latter are more capable of applying limits only to election-related content or accounts.

4.2 Adaption of terms and conditions and their enforcement, adaption of content moderation processes (Art. 35.1.b and Art.35.1c)

4.2.1 Content moderation

Content that violates platforms' ToS can require action, including removal, downranking or labelling. On most platforms, violative content includes electoral mis- and disinformation, hate speech and TVEC. Content moderation includes a mixture of human and automated moderation; the efficacy of automation in content moderation is discussed in more detail in section 4.3.

Platforms' enforcement of content removal is neither uniform nor consistent. Independent research on the accuracy and effectiveness of removal can be difficult to perform due to the scale of content and activities on platforms.[76] Most studies are piecemeal and only provide insights on specific contexts. One study of the 2020 US presidential elections found that removal of content often took place once misinformation has been widely disseminated or has gone viral;[77] at this stage, this mitigation strategy may be relatively ineffective at scale and high cost.

Many studies on the efficacy of content removal look at the COVID-19 pandemic. While not election-focused, this research provides transferrable insights into the way platforms moderate mis- and disinformation. According to the Meta Oversight Board, Meta removed 27 million pieces of content flagged as COVID-19 misinformation from Facebook and Instagram between March 2020 and July 2022, 1.3 million of which were restored through appeal.[78] However, an independent study of Facebook's removal of vaccine misinformation found that while some content was removed, this was not followed by overall decreased engagement with the anti-vaccine content.[79] The authors found that highly motivated users knew how to use Facebook's architecture and discovered ways to circumvent misinformation removal policies.

Another aspect that requires further research is the efficacy of content moderation across languages. Evidence suggests that policies and practices implemented by platforms when moderating non-English language content can have negative effects on freedom of expression or access to information due to their inaccuracy.[80] According to the whistleblower Francis Haugen, Facebook allocates 87 percent of its spending on misinformation countermeasures to English content, despite only 9 percent of its users being English speakers.[81] The EU's DSA requires VLOPs and VLOSEs to report the human resources dedicated to content moderation, broken down by each of the bloc's 24 official languages (Article 42.2). Even with these provisions in place, platforms report inconsistently on geographies, language proficiency requirements and how they count multilingual moderators across several languages.[82] While on paper this illustrates a positive step towards greater resource allocation to non-English content moderation, these recent transparency reports reveal that many services have few or no moderators in less commonly spoken languages.[83] The DSA's measures also do not address the multitude of languages spoken globally, especially by diaspora communities in the EU, and in Global South countries.

4.2.2 Account bans and deplatforming

Account bans refer to temporary restrictions of access to user accounts, while deplatforming describes the attempted permanent suspension of certain individuals from a platform. These strategies are often deployed on users with large audiences built on controversial behaviour that may breach ToS. While some view these mitigation measures as overly restrictive for freedom of speech, others argue that platforms have a right to decide upon content they do not wish to host on their sites.[84] For example, in 2021 the German Federal Court of Justice reiterated that platforms may define ToS that determine what is legitimate content beyond what is restricted by law and enforce them as long as they are transparent about these additional restrictions.[85]

Generally, there is limited research on platform-wide interventions regarding speech, including restricting and deplatforming accounts. Data on the impact of short-term efficacy are mixed while the long-term effects have not been systematically investigated. One study found that deplatforming 70,000 spreaders of misinformation after the Capitol Hill attacks on January 6

lowered the reach of misinformation on X (then Twitter). Despite this number being a fraction of users, these "superspreaders" seem to be responsible for a large amount of misinformation.[86] However, many users also left the platform in protest after the mass deplatforming; alongside other factors, this may have affected the results so causality cannot be established.[87] Other investigations found that the number of conversations around particularly high-profile users decreased through deplatforming, and overall public attention paid to these influencers also decreased.[88]

Yet, this strategy seems to have various workarounds. An examination of Facebook's ToS around COVID-19 vaccine misinformation, including account takedowns, concluded that the platform's design allows for several means to circumvent interventions. The layered architecture that functions such as groups and subpages enables users to withdraw to different parts of the platform and create complex paths and cross-linkages to evade detection of problematic behaviour in the first place. Deletion of individual accounts, pages and groups is less effective if the same content is posted elsewhere, and accounts which were newly created – potentially after the removal of a previous account – frequently coordinate with existing ones.[89] Moreover, users that are deplatformed tend to move to smaller platforms that are often less moderated which may result in users spending more time in more harmful online spaces. While fringe platforms tend to have a smaller reach, potentially reducing the spread of harmful content, the migration of users to these platforms before a major site's deplatforming remains an issue.[90] For example, a recent ISD study which mapped the German far right online ecosystem over three years, found that users who moved to smaller fringe platforms with more lax ToS, also tended to migrate back to larger platforms once ToS are less restrictive again, as was the case with X.[91]

4.2.3 Geo-blocking sanctioned state actors

Geo-blocking sanctioned state actors requires platforms to block sanctioned content from being accessed online in specific regions. For example, geo-blocking was implemented in EU member states following Russia's full-scale invasion of Ukraine and subsequent EU sanctions on Russian state media. In addition to war-related propaganda, sanctions also aim to mitigate Russian attempts to destabilise the EU, its institutions and its political parties, especially during elections.[92] Research by ISD and others indicate that geo-blocking of state-sanctioned Russian media has largely been effective, leading to significant drops in web traffic, click-throughs to sanctioned websites, and engagement with Europe-focused Russian state media pages following sanctions. However, post-invasion changes in media consumer behaviour likely also affected engagement.[93]

Despite these efforts, content from sanctioned websites remained available to European audiences on platforms through alternative domains that are not sanctioned, mirror websites, websites that direct traffic to sanctioned media and websites that copy content from sanctioned media.[94] Researchers also found inconsistencies between platforms and EU member states in the implementation of sanctions in 2022.[95] This indicates a need for policymakers to regularly update sanctions lists and for platforms to be more responsive and iterative in enforcement actions.

However, some avenues for propaganda dissemination were unaffected by geo-blocking. The accounts of diplomats, state media journalists, and staff had increases in EU engagement and followers after the full-scale invasion of Ukraine.[96] The availability of sanctioned content in non-EU countries with shared languages, such as Spain and Hispanic America, also facilitates the flow of state media-driven narratives into the EU.[97]

Outside of extreme conditions with high risks for FIMI (such as those following the full-scale invasion of Ukraine), geo-blocking is likely an inappropriately restrictive mitigation measure as it suppresses access to information. Although it can be a useful measure for taking action against individual accounts while complying with applicable regional laws, geo-blocking is unsuitable for countering online threats systemically, for example, to deal with the emergence of alternative inauthentic networks.

4.3 Testing and adaption of algorithmic systems, including recommender systems (Art. 35.1.d)

Algorithmic recommender systems make automated decisions about which content, accounts and pages to prioritise or demote on feeds or in search results. Ultimately, they shape the online experience of billions of users. By adjusting these ranking systems, platforms can control the reach of problematic content such as electoral disinformation, hate, and other undesired

content.[98] Algorithmic systems are also applied in content moderation and many other platform activities and products. However, restrictions on data access and research partnerships limit causal research and external validity.

4.3.1 Automated content moderation

The massive volume of content uploaded and circulated on platforms and the emotionally-taxing nature of much of the content makes content moderation a challenging job for human moderators. Thus, many platforms use natural language processing (NLP) algorithms to automate content moderation.[99] This may include proactive automated detection of potentially problematic content and/or automated content moderation decision-making, such as removing, labelling or demoting content.

Due to data access and research design limitations, it is difficult to systemically and independently measure the efficacy of automated content moderation compared to the overall prevalence of harmful content or material which violates platforms' ToS. Therefore, much research on whether automated content moderation is an effective mitigation measure focuses on its accuracy. While studies on the impact of automated systems are limited, especially on elections, one piece of research, which studied increased automated content moderation during the early days of the COVID-19 pandemic, showed a rise in removals associated with a decrease in accuracy and specificity of the takedowns.[100] Although not statistically significant, this example illustrates the potential effects of an over-reliance on automated tools. More research is needed to determine the effects of these tools in election contexts.

Over-reliance on automated tools can also raise human rights concerns. Users' freedom of expression can be violated by false positives, when an algorithmic system mistakenly classifies content as violative. Similarly, without sufficient, high-quality, unbiased data on underrepresented groups, inequalities can be reflected or amplified by automated moderation, also resulting in risks to freedom of speech.[101] To mitigate these risks, proper complaint, review, appeal and general oversight by humans are essential, alongside comprehensive documentation and explanation of the nature and scope of automated tools. False negatives—when an algorithmic system misses something that should have been classified as violative – may lead to a failure to address violative content (including for example hate speech, harassment, misinformation). This can in turn have a chilling effect on certain communities' willingness to participate online.[102]

Given these complexities and the fast pace of technological development, further research on the human rights impact of automated content moderation is needed. This is especially important in an electoral context, where risks to freedom of expression on political issues and over-moderation of specific groups may occur; these issues may not be remedied during the election period itself.

4.3.2 Reduced virality

Few studies have examined whether adjusting algorithmic recommender systems to downrank or de-amplify content from sources that are deemed untrustworthy may impact user behaviour. However, initial results indicate that this type of intervention is promising.

For one experimental study, authors partnered with the search engine DuckDuckGo to deploy interventions to more than 463,000 search results where links to websites known for misinformation appeared.[103] Researchers found that algorithmic de-amplification was the most effective intervention, reducing engagement with misinformation by more than 50 percent. The high external validity of this experiment highlights the importance of successful research partnerships.

Similarly, another set of researchers studied Facebook's claim to reduce the virality of posts sharing content by "repeat offender" websites and groups, which fact checkers have found to repeatedly publish mis- or disinformation. Based on data from social media listening tools and fact-checking data sets, the authors found that engagement per post for these groups reduced between 16-31 percent.[104]

Another way to limit the virality of problematic content is by "turning off" the recommender system in the feed on many platforms' homepages. Instead, platforms can revert to the reverse-chronological feed showing content only based on how recently it was published and whether the user follows the posting accounts.

A study produced by researchers in partnership with Meta investigated this intervention's impact during the US 2020 presidential election. Participants were actively recruited and placed either into a group where their Facebook and Instagram newsfeed showed content based on chronology, or a control group, where participants' feeds continued to show content based on the algorithmic recommender system.[105] Results were mixed: the authors found that on both platforms the treatment group was exposed to more political and "untrustworthy" content. However, on Facebook participants also saw increased content from moderate and ideologically-mixed audiences, and reduced exposure to uncivil content by almost half.[106] Again, this research partnership illustrates the benefits of such partnerships, despite the difficulty in establishing them.

While previous research focused on the algorithmic (de-) amplification of harmful content, the role of algorithmic recommender systems in supporting the spread of synthetic content has recently received more attention. In theory, platforms' AI detection tools combined with metadata provenance standards, such as those introduced by the Coalition for Content Provenance and Authenticity, can mitigate risks from the algorithmic dissemination of synthetically generated content.[107] However, limited data access and the recent implementation of these measures means at present there is little research regarding their efficacy.

4.3.3 Boosting of authoritative election information

Another strategy deployed by platforms is directing users to official information sources on public issues, such as voting information or public health. For example, during the US 2020 presidential election, X (then Twitter) tried to increase access to credible information on voting and the integrity of election results. Similar efforts were undertaken during the COVID-19 pandemic, as X attempted to ensure user access to credible public health information. Generally, authoritative information may be elevated via redirection links to authoritative sources on a post related to the topic, or the "flood the zone" approach where authoritative information is planted throughout a feed or added directly to the user interface. The messaging of this intervention includes media literacy tips, pre-emptive rebuke or "pre-bunking" of misinformation (see Section 4.8.5).[108]

Platforms often tout these mitigation measures as a key pillar of their information integrity strategies, but evidence of their efficacy is mixed and further research is needed. Survey experiments testing informational panels similar to those used by platforms have found positive but small effects on participants' ability to recognise misinformation.[109] In an experimental study where authors partnered with DuckDuckGo, authors tested the effect of informational and pre-bunking panels. Neither intervention resulted in significant decreases in users' selection of misinformation results.[110] The authors also found that users rarely clicked on the links in pre-bunking panels themselves. Other studies demonstrate more promising results. An experiment on X (then Twitter) showed that by drawing attention to the quality of news, people are more likely to share accurate or high-quality content, even when it was inconsistent with their political beliefs.[111] Section 4.8.1 explores this further.

4.4 Adapting advertising systems (Art. 35.1.e)

4.4.1 Political ads disclosures

Political ads usually refer to ads from political parties and candidates and may include politically salient "issue" ads. The European Commission's Code of Practice on Disinformation notes that issue ads can significantly shape "public debates around key societal issues, particularly in forming public opinion, political and electoral debate, referenda, legislative processes and the voting behaviour of citizens."[112] Research indicates that audiences are most likely to be persuaded by a political ad when they do not know the ideological motivation of its source, indicating the importance of ads transparency.[113] Issue ads can also be a vehicle for disinformation and FIMI. For example, during the 2016 US presidential election, it was found that Russia's Internet Research Agency also exploited issue ads to influence voting behaviour.[114]

A small number of studies present mixed results on how political ads disclosures directly impact users and voters. Experiments using participant recall and eye movement data show that US users spend longer looking at presidential candidacy ads that include sponsorship disclosures, likely due to further reading and engagement with that information.[115] However, this engagement did not consistently lead to users remembering the source of the ad long-term. Research on ads libraries, which increase transparency by providing a public, searchable repository of ads on a platform, also indicates a disconnect between theory and practice. Researchers

from McGill University studying the integrity of Canada's 2019 federal election concluded that the Facebook ads repository was only theoretically useful for increasing electoral transparency.[116] In practice, it provided insufficient data and was inaccessible for many users, tempering any positive effects on online electoral integrity.

Similarly, other research indicates the widespread prevalence of insufficient or incomplete data and accessibility issues with platform and search engine ad libraries. The Mozilla Foundation's stress test of VLOPs and VLOSEs prior to the 2024 European Parliament elections concluded that none offered "a fully-functional ad repository".[117] Mozilla also noted that libraries were not comparable, with differences in listed information on advertisements, advertisers, targeting techniques, the availability of historical data, and the granularity of tools and data (particularly regarding features such as filtering and sorting). Missing data, malfunctions, search rate limits and data access issues all caused further complications.

Few platforms include influencer content in ads repositories, despite its role in political advertising.[118] Similarly, "issue" ads are often not defined clearly, and are not consistently included in ads repositories worldwide, including ads on electorally salient topics. This represents a significant route for spreading misinformation.[119] The automated classification of ads also introduces high risks of inaccuracy. Accuracy issues, as well as inconsistencies across platforms and inaccessible design, have also been noted in research on political ads repositories in Ireland, the Netherlands, Czechia, Italy and the UK.[120] This indicates that the efficacy of ads disclosures is likely only as effective as its implementation.

Despite the limitations, the creation of political ads libraries has also enabled public-interest research on misinformation and FIMI. For example, research from EU Disinfo Lab and AI Forensics on the Doppelgänger network of pro-Russian propaganda was enabled by Meta's Ad Library, providing valuable information on a large-scale operation serving targeted issue ads to European voters.[121]

4.4.2 Political ads bans

Some platform policies prohibit political ads for reasons ranging from inconsistency with a desired "light-hearted" platform experience (TikTok) to high risks to civic discourse through to harms such as micro-targeting (X, then Twitter; this policy was reversed in 2023).[122] Research is lacking on any direct effects ad bans may have on civic discourse and electoral processes.

Ad bans are only as effective as their implementation and there is little research on X's previous ban on political ads. TikTok's ban, however, received more attention, revealing issues with classifying influencer content and moderating ads. TikTok ostensibly does not allow paid ads with political content (including advocacy and issue ads), and content creators cannot be paid to publish branded political content. The company also claims to not allow campaign fundraising or access to monetisation functions for political accounts. However, an investigation by the Mozilla Foundation found that these policies are easy to evade. Influencer advertising was particularly prone to a lack of moderation and undisclosed paid partnerships with political groups.[123] These findings are consistent with earlier investigations by Mozilla, which also demonstrate the role of paid partnerships with political influencers in disseminating political content on TikTok.[124]

Other investigations indicate issues with moderating submissions of political ads to TikTok as well as platforms that do officially allow political ads. Testing by Global Witness in June 2024 found that TikTok approved 16 out of 16 political ads submitted for publication.[125] The ads were intended for publication in Ireland prior to the European Parliament elections in June 2024 and featured electoral disinformation. These findings are similar to those from a 2022 Global Witness investigation, where English and Spanish-language ads were approved by TikTok in the US despite containing false electoral information and claims designed to delegitimise electoral processes.[126]

4.5 Reinforcing internal processes (resources, testing, documentation, or supervision of activities) (Art. 35.1.f)

Since the DSA is very broad regarding the mitigation measure of reinforcing internal processes, this section uses the example of generative AI to outline what testing and documentation may look like for this emerging technology. Generative AI systems are integrated into the workings of some platforms and search engines, such as AI Overviews in Google Search or Meta's AI

chatbot. Some services incorporate AI systems developed by third parties, whereas others, such as Google and Meta, use their own proprietary models. These systems come with risks of misinformation, as well as the potential to aid malign actors' attempts at spreading disinformation or other online harms. To date, most mitigation measures for risks emanating from generative AI include the reinforcement of internal processes, including the resourcing, testing, documentation and supervision of new and existing activities. A selection of the most prominent mitigation measures are reviewed below, though many of those adopted by platforms are quite new and further research is needed to determine their efficacy.

At the systems level, risks stemming from generative AI can be mitigated through a variety of procedures. AI developers can ensure that generated image, video and audio content – especially that concerning elections and political processes – is detectable through provenance and authenticity methods. These include watermarks, metadata identifications, and cryptographic methods. The Coalition for Content Provenance and Authenticity (C2PA) has created standards for "cryptographic asset hashing" which allow an electronic file to be sealed with a tamper-evident manifest containing information about a file's history and edits.[127] C2PA standards have not yet been adopted across the industry, but are currently integrated into some language models, such as Open AI's DALL-E 3 and its upcoming text-to-video model, Sora.[128]

As language models are built, developers can ensure that models' vulnerabilities are tested via red-teaming exercises and technical safeguards can be introduced, including moderation of elections-related content and the use of prompt classifiers. Some jurisdictions are beginning to make measures such as vulnerability testing and content provenance marking mandatory through legislation, such as the EU's AI Act.[129] Red-teaming and other vulnerability testing processes can help prevent the spread of electoral misinformation or creation of disinformation via generative AI tools, although their efficacy when integrated into platforms requires further research.[130]

4.6 Initiation or adjusting cooperation with trusted flaggers (Art.35.1.g)

Academic and civil society work on trusted flagging is mostly theoretical and based on a European context, with a focus on the trusted flagging provisions in the EU's DSA. There is no systematic research on the efficacy of trusted flagging mechanisms specifically. The European Commission defines trusted flaggers as entities that are "experts at detecting certain types of illegal content online, such as hate speech or terrorist content, and notifying it to the online platforms. The notices submitted by them must be treated with priority as they are expected to be more accurate."[131]

While there is a lack of public-facing empirical evidence on the efficacy of trusted flagging mechanisms, a few theoretical concerns regarding implementation can be considered. Concerns include the potential for trusted flaggers to be relatively unaccountable, leaving them open to co-option by special interests.[132] In contexts where regulators are not independent from government, organisations representing politically unpopular interests or marginalised groups may not have equal chances of receiving trusted flagger status.[133]

Legal scholars have also noted that trusted flagging mechanisms are likely difficult to scale, considering the volume of content that is posted on social media platforms every day. To be effective at scale, trusted flagging should be integrated into automated content moderation processes, such as notice-and-stay down mechanisms that act against flagged content as well as its equivalent and future uploads.[134] However, further evidence is needed to substantiate this recommendation.

4.7 Initiation or adjusting cooperation with other platform providers (Art. 35.1.h)

Platforms should recognise that what occurs on other platforms may make its way to their own service (and vice versa). This is true for many of the online harms outlined in section 2.2. Given the cross-platform quality of online threats, cooperation among platforms is a commonly recommended measure.[135] Cross-platform cooperation can include exchange channels between relevant teams (such as those working on safety and content moderation) to proactively share information about cross-platform coordination by malign actors. Exchange channels may facilitate faster action, for example, when a prominent actor is identified to be linked to repeated harmful behaviour, such as violating ToS across platforms.

Such cross-platform initiatives already exist, like the Global Internet Forum to Counter Terrorism's (GIFCT)

Content Incident Protocol,[136] or the South African Framework of Cooperation during the May 2024 presidential elections, which was designed to facilitate communication between Meta, Google, TikTok, civil society and the Electoral Commission. The EU's 2022 Strengthened Code of Practice on Disinformation calls for signatories' commitment to such coordination.[137] However, beyond these formal cooperation agreements, it is unknown to what degree, or if at all, platforms coordinate to tackle online harms.

Research evidencing the efficacy of cross-platform coordination is sparse, likely since this mitigation measure requires further implementation, and greater transparency on platforms' collaboration is needed. Studies on the spread of content across platforms show the potential of what better cross-platform cooperation could achieve, for example in terms of early warning systems. One study analysed more than 15,000 public WhatsApp groups from Bolsonaro supporters ahead of the Brazilian Capitol attack in January 2023. Their cross-platform time series with X (then Twitter) content showed how the dissemination of content could predict the spread of content on WhatsApp.[138] While inconclusive, these findings demonstrate that platforms' combined knowledge about online threats could help anticipate trends earlier and take up adequate mitigation measures.

4.8 Awareness-raising measures (Art.35.1.i)

4.8.1 Quality rating

Raising awareness of the quality of content or its source may help stem the proliferation of false or misleading content online. The hypothesis is that the spread of misinformation may be disrupted if users are made aware of the quality/accuracy of content or news before they share it in their networks. Awareness-raising measures can include providing ratings on the quality of the content or source by experts or other users, or as a label.

Several studies testing variations of these mechanisms show promising results for quality rating as a mitigation measure. In an experimental study, researchers tested three distinct mechanisms for "source ratings" applied to articles when first published. These included ratings by expert reviewers to provide an aggregated source rating, ratings where regular users rate the articles for a score rating, and finally ratings where users rate the source of the articles themselves to provide a score.[139] The experiment showed that source ratings had an impact: low ratings had a stronger effect on users' engagement with the content than high ratings. Expert ratings and user article ratings had a more significant impact than user source ratings. Another experiment testing the impact of credibility indicators on people's intent to share news headlines confirmed that these indicators can decrease the sharing of mis- and disinformation, and that credibility indicators from fact-checking services were the most efficient.[140]

Community Notes (previously known as Birdwatch) is a promising fact-checking crowdsourcing program, specific to X. It allows users to submit useful context to tweets which may be otherwise misleading or missing important information. Users may submit Community Notes, and other users may evaluate and rate the quality of these notes. Only the notes with the highest ratings and that are deemed cross-ideological (that is, being accepted by a broad political spectrum) are then displayed publicly. While X's own research demonstrated that Community Notes helped slow the spread of misinformation,[141] in 2023, 60 percent of the most-rated notes were not public.[142]

While these results are promising, most studies carried out were experimental. Further real-world testing of the impact of quality ratings on user behaviour would strengthen this evidence base.

4.8.2 Interstitials and labelling mis- and disinformation

Mis- and disinformation that does not meet platforms' thresholds for content removal can instead be accompanied by warning labels indicating the presence of false or misleading information. This can either take the form of small accompanying labels, the more common option; alternatively, some platforms have "tentatively" deployed larger interstitials, which "screen" content until a user indicates that they wish to look at a post or follow a link off-platform.[143] Users may be familiar with the use of interstitials to moderate explicit content allowed by platforms' ToS, such as journalistic content featuring violence in conflict zones or adult content.

Research indicates that the efficacy of labelling is modest.[144] Specific labels and warnings are more effective than general notes, such as that a claim is "disputed."[145] In addition, there is a risk that labels will be ineffective if not applied in a timely manner: one study of the 2020 US presidential elections found that the

removal or labelling of content often took place once misinformation had been widely disseminated or had gone viral.[146] Gaps in the labelling of non-English-language content also allows for a wider spread of electoral misinformation outside of English-speaking environments.[147] Not all platforms consistently translate or display labels on electoral mis- or disinformation outside of English, despite the prevalence of other languages on platforms.

Research evidence on the efficacy of interstitials is generally positive. An experimental study of interstitial warnings applied to disinformation websites accessible via Google Search found that they significantly affected user behaviour.[148] This finding held regardless of users' partisan affiliation, or the detail provided in the warning message. Researchers attribute these findings to the friction interstitials introduce to the user experience. However, they also note that these positive effects could decrease with frequent exposure. Similarly, research comparing the effects of labels and interstitials to mitigate COVID-19 vaccine misinformation on X (then Twitter) found that interstitials were more effective in decreasing user beliefs in the accuracy of misinformation content.[149] However, more research is needed to confirm these findings at scale, outside of experimental conditions and within the context of elections. Access to data and platform research on the efficacy of interstitial warnings would also provide valuable further evidence on the efficacy of this mitigation measure.

4.8.3 Labelling state-affiliated actors

The limited research on the effects of labelling state media sources indicates that it likely decreases the negative effects of FIMI. However, the design of labels can significantly affect efficacy. All research described here concerns US internet users, with variations in demographic representativeness; more systemic research in other contexts is needed.

Efficacy seems highly dependent on how visible labels are. More visually noticeable labels are more effective at mitigating interactions with state-affiliated media posts, pages, and accounts. For example, a study of YouTube's state-funded channel warning labels found an increase in their effectiveness when, halfway through the experiment, the colour of the label box was changed from grey to blue.[150] Other research similarly notes that labels are only effective if they are noticed by users.[151]

Perceptions of the labelled country of affiliation also appear to matter. Research based on field data indicated that US Facebook users decrease their engagement with content labelled as Russian or Chinese-affiliated, but not that affiliated with the government of Canada.[152]

Demographics and platforms are significant when it comes to the efficacy of labelling. A study of X (then Twitter) users in the US across the political spectrum showed that users decreased their reported likelihood of engaging with labelled content, regardless of whether the label was general ("foreign government") or country-specific ("Russian government").[153] However, when labels were added to Facebook posts, partisan differences emerged regarding reported actions. Democrat users reported a decreased willingness to engage with labelled Russian or foreign government disinformation online, or to spread the same points offline in conversations. By contrast, Republican Facebook users reported no difference in their willingness to engage with or spread labelled disinformation online or offline, regardless of the label.

Belief in foreign-affiliated disinformation and willingness to act on that information appear to differ according to the platform and label wording. The same study also found that both Democrat and Republican Facebook users had no decrease in their reported tendency to believe disinformation after exposure to a post accompanied with a general "foreign government" label.[154]

As with many mitigation measures, the consistency of implementation also affects efficacy. While research demonstrates that labelling is somewhat effective, studies only examine situations where labelling is consistently used, and do not consider platforms' frequent inconsistencies in implementation. A 2021 study on the implementation of Chinese state-affiliated media labels in the UK notes a significant difference between implementation on X (then Twitter), covering 90 percent of English and other-language accounts, and Facebook, covering 66 percent of English-language accounts and 22 percent of other-language accounts.[155]

Researchers also critiqued the greater prominence of state-affiliated media labels shown to US Facebook users compared to those in other countries. In the 2021-22 period, 91 percent of content from Russian state media on Facebook, including blatant disinformation, was not

labelled. In 2023, Threads – the micro-blogging platform developed by Meta to compete with X – similarly failed to label many state-affiliated accounts.[156] TikTok, which has not yet been a major focus of academic research on this topic, has also failed to ban accounts and ads from Chinese and Russian state-affiliated media, despite an official ban on political ads.[157]

4.8.4 Labelling manipulated and artificially generated content

Manipulated or artificially generated content can be found across a range of online platforms and includes synthetic image, video, audio and text. Some content is shared maliciously or in bad faith. In other instances, it is intended for entertainment or as part of good faith campaign activity. As labelling and content provenance methods have largely been introduced by platforms just over the last year, there is little clear data of the effectiveness of these measures. A brief overview of existing approaches is given below.

Measures to mitigate risks from the dissemination of generated content online include clear labelling of "deepfakes", synthetic depictions of a person, place, situation or event that is falsely depicted as real. The European Commission's April 2024 Guidelines to VLOPs and VLOSEs on the Mitigation of Systemic Risks for Electoral Processes notes that labelling measures should include options for users to add labels to generated content, as well as tests and corresponding improvements to labels' efficacy (see section 5.5).[158] Platform detection and labelling can be done automatically when generated content includes a content provenance marking, such as that used by the C2PA Initiative; although this is likely to be more efficient and accurate, further evidence is required.[159]

4.8.5 Gamification, video-based inoculation

Pre-bunking describes a specific form of inoculation that makes use of image- or video-based content to build psychological resilience before contact with false information and other kinds of manipulation.[160] A series of experiments by the University of Cambridge found that brief animations describing the tactics of disinformation actors can function "similar to a vaccine" by preparing participants for disinformation through exposure to small doses of harmful content.[161] In cooperation with Google's Jigsaw research unit, short videos containing pop cultural references were created that detailed manipulative techniques such as scapegoating or using contradictory statements to cause confusion. Researchers concluded this to be an effective measure against misinformation, as participants deemed such content less reliable and less worthy of sharing, independent from factors such as political affiliation.

One effort to replicate this study focused on misinformation spread via images or videos and had additional gamification elements.[162] It had similar findings, as did research attempting to simulate real-world conditions of social media use on UK users to establish a higher level of validity than previous research settings could.[163] However, inoculation may be less effective in non-Western contexts. When experiments were replicated in India, there was no significant impact on perceived reliability of misinformation or willingness to share it, possibly due to lower rates of media literacy.[164] Similarly, a study conducted in Singapore could not replicate results of the original experiment and suggested this was due to factors such as lower trust in media and government, as well as more positive opinions about censorship.[165] Thus, while promising early results exist, further studies across different contexts and demographics are required to determine the necessary conditions for inoculation to work.

5. Challenges and considerations for evaluating the effectiveness of mitigation measures

Studying the efficacy of measures to mitigate risks to online electoral integrity remains a challenge. There is an ongoing lack of transparency on platform policies and processes, and researchers face increasing challenges in accessing platform data on user reactions to interventions. These challenges create some incongruity in research on efficacy, impacting experiment and study scope, size and quality. Data access issues challenge researchers' abilities to evidence specific mitigation measures, which also impacts the quality and real-world applicability of evidence produced. A lack of data access also necessitates some creativity in research design, which can make it difficult to compare studies and the efficacy of similar mitigation measures across different platforms, languages, and contexts.

A 2021 systematic review of 49 years of research on countermeasures to combat influence operations found a mismatch between interventions taken by platforms and those studied by the research community.[166] Researchers noted many countermeasures are still to be formally studied, and that a lack of platform transparency and data access precludes the evaluation of many research questions, methods, and types of mitigations. This affects the quality and applicability of research as studies often cannot establish causality, track how users respond to an intervention at scale, or effectively study behaviour in a way that fully simulates aspects of social media usage in the real world.

Globally, there is limited legislation establishing researcher access to platform data. While the EU's DSA introduces data access obligations for VLOPs and VLOSEs, its implementation and enforcement remains a work in progress and other jurisdictions have no comparable legislation enacted. In the US, at the time of writing, the bipartisan Platform Accountability and Transparency Act (PATA) is still only a proposal, as is Canada's Online Harms Act (Bill C-63). Comparable UK and Australian online safety legislation do not include comprehensive data access provisions comparable to those included in the DSA. Moreover, smaller platforms like Telegram are usually not considered in data access regimes like the DSA; they remain highly relevant for the dissemination of harmful content and the performance of manipulative activities. Following that, policymakers may consider how to support data access to such platforms while not overburdening platforms that may not present severe risks.

Simultaneously, platforms' restrictions on researcher data access are increasing, despite approaches to data access that preserve user privacy.[167] Very large platforms increasingly restrict access to data for researchers – for example, Meta is shutting down the CrowdTangle API without an adequate replacement while X significantly increased the costs of accessing their API, which was previously freely available to researchers.[168] In the past, research projects have been shut down with platforms pointing to user privacy issues.[169] Mounting court cases against think-tanks and other research institutions deepen the divide between platforms and researchers.[170] Technological barriers also exist: these include the challenge of systematically analysing video and audio content or researching decentralised, fragmented network structures.[171]

Data access also concerns regulators, despite their greater powers to request information compared to researchers. For example, the Slovakian authority for media oversight, the Council for Media Services (CMS), conducted a quantitative content analysis of Facebook posts on narratives related to the 2023 national parliamentary elections and noted access issues.[172] Partnerships to enable access to tools such as CrowdTangle were necessary as other options were deemed too expensive or non-transparent. This demonstrates how advocating for broad data access for researchers and regulators while respecting data protection and other regulations remains a priority. It is yet to be seen how the EU's DSA will change data access once Digital Service Coordinators are fully set up to process data access and vetting requests under Article 40.

More cooperation between researchers, regulators, and platforms is needed. Researchers should work together, share insights and craft a unified voice regarding both platforms and regulators. Regulator collaboration allows for proactive response strategies as issues arise, such as harmful content spreading across platforms and jurisdictions. Slovakia's CMS concluded that sharing research results and methodologies with other regulators and national bodies helped create situational awareness and understand the complexities of the online ecosystem ahead of elections.[173] It also noted the value of regulator-platform meetings in advance of elections to discuss preparedness, which helped to set out regulator expectations and requirements.

To improve cooperation between stakeholder groups, communication channels for sharing information are necessary as called for in the European Commission Guidelines for Elections.[174] These can build on existing efforts, such as the Global Online Safety Regulators Network, G7 Rapid Response Mechanism, European Regulators Group for Audiovisual Media Services, and the European Digital Media Observatory (EDMO) 2024 European Elections Taskforce.

Despite the outlined challenges to research, a few general points can be observed. The most effective mitigation strategy will include a combination of measures, applied consistently and effectively, as opposed to relying on a narrow set of approaches.[175] A wide toolkit of measures is also likely to mitigate challenges resulting from divergent partisan responses to some interventions, such as the labelling of state-affiliated actors. In addition, the evidence is clear that for most mitigations, design details matter: elements as small as the colour of a warning text box can have significant effects on user behaviour and platforms must consider the impact of these nuances on online harms.

There is also some evidence that the efficacy of mitigation measures may be partially context-dependent, even regarding countries that are relatively linguistically and politically similar.[176] A recent systematic review noted the geographic inequality of research, which overwhelmingly focuses on Western democracies (the vast majority in the US).[177] This is concerning, given that even relatively similar contexts may have different outcomes due to their varied electoral systems, media landscapes and political culture. This underscores the importance of conducting further research across a diverse set of countries. It also indicates the need for global approaches to data access for independent and public-interest research on electoral harms. Large-scale and systematic research is also needed to assess the efficacy of platform content moderation in non-English languages.

There is a demonstrable need for more research on some topics. Due to data access concerns, little research has examined how mitigation measures affect actual, not just intended, behaviour online or offline. Most research has focused on larger and legacy platforms; further research should examine less-studied platforms such as Instagram and LinkedIn, as well as smaller platforms and those used outside of Western countries. Approaches including inoculation, trusted flagging processes, generative AI, the use of interstitials, and multi-platform collaboration efforts all show promise, but require further research.[178] There is also a need for evidence regarding the impacts of political ads libraries, disclosures and bans on user beliefs and behaviour.

6. Conclusion

Ensuring the integrity of elections in the digital age is vital to ensure that democracy continues to thrive alongside societies enjoying the advantages of the internet and emerging technologies. However, it requires a coordinated and multifaceted approach involving governments, industry, academia and civil society. This Policy Brief highlights the critical role that each stakeholder plays in safeguarding electoral processes against online threats. It also underscores key mitigation strategies to safeguard electoral integrity including robust content moderation systems, transparent political advertising practices, algorithmic adjustments, awareness-raising initiatives and enhanced cross-platform cooperation. Despite these efforts, challenges such as platform transparency and limited data access remain significant obstacles.

To effectively combat the evolving threats posed by online platforms, it is essential to continue studying, refining and scaling mitigation measures across diverse contexts. Further research is necessary to understand the efficacy of these interventions and adapt them regionally. By fostering international collaboration, sharing best practices and ensuring data accessibility, stakeholders can better protect electoral integrity and uphold democratic values worldwide. The ongoing commitment to these principles will be crucial in navigating the complex landscape of online electoral integrity.

Endnotes

1 Legal Information Institute. (May 2022). SLAPP suit. Cornell Law School. https://www.law.cornell.edu/wex/slapp_suit

2 European Council: Council of the European Union (March 19, 2024). Anti-SLAPP: Final green light for EU law protecting journalists and human rights defenders. Council of the EU. https://www.consilium.europa.eu/en/press/press-releases/2024/03/19/anti-slapp-final-green-light-for-eu-law-protecting-journalists-and-human-rights-defenders/#:~:text=According%20to%20the%20directive%2C%20a,located%20in%20that%20member%20state

3 Russell, S., Perset, K., & Grobelnik, M. (November 29, 2023). Updates to the OECD's definition of an AI system explained. Organisation for Economic Cooperation and Development. https://oecd.ai/en/wonk/ai-system-definition-update

4 Movement Advancement Project. (May 2023). How Election Denialism Threatens Our Democracy and the Safeguards We Need to Defend It. https://www.mapresearch.org/file/MAP-2023-Election-Denialism-Report.pdf

5 The Electoral Knowledge Network's ace Project (2024). Electoral integrity. https://aceproject.org/ace-en/topics/ei/explore_topic_new, which references a definition by the Kofi Annan Foundation.

6 Dorn, M., Bundtzen, S., Schwieter, C., & Gandhi, M. (September 12, 2023). Emerging Platforms and Technologies: An Overview of the Current Threat Landscape and its Policy Implications. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/emerging-platforms-and-technologies-an-overview-of-the-current-threat-landscape-and-its-policy-implications/

7 Brown, T.B., et al. (2020). Language Models Are Few-Shot Learners. arXiv. http://arxiv.org/abs/2005.14165

8 IBM Research (2023). What is Generative AI?. IBM. https://research.ibm.com/blog/what-is-generative-AI

9 Bundtzen, S. (September 14, 2023). Misogynistic Pathways to Radicalisation: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/misogynistic-pathways-to-radicalisation-recommended-measures-for-platforms-to-assess-and-mitigate-online-gender-based-violence/

10 Ewe, K. (December 28, 2023). The Ultimate Election Year: All the Elections Around the World in 2024. TIME. https://time.com/6550920/world-elections-2024/

11 Smirnova, J., Ahonen, A., Mathelemuse, N., Schwertheim, H., & Winter, H. (February 25, 2022). Bundestagswahl 2021: Digitale Bedrohungen und ihre Folgen. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/bundestagswahl-2021-digitale-bedrohungen-und-ihre-folgen/

12 Smirnova, J., Winter, H., Mathelemuse, N., Dorn, M., & Schwertheim, H. (September 16, 2021). Digitale Gewalt und Desinformation gegen Spitzenkandidat:innen vor der Bundestagswahl 2021. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/digitale-gewalt-und-desinformation-gegen-spitzenkandidatinnen-vor-der-bundestagswahl-2021/

13 Bundtzen, S. (December 9, 2023). Suggested for You: Understanding How Algorithmic Ranking Practice Affect Online Discourse and Assessing Proposed Alternatives. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/suggested-for-you-understanding-how-algorithmic-ranking-practices-affect-online-discourses-and-assessing-proposed-alternatives/

14 European Commission. (April 2024). Commission Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to Article 35(3) of Regulation (EU) 2022/2065. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C_202403014; The European Partnership for Democracy and the Civil Liberties Union for Europe elaborated further on what the DSA's "systemic risks for civic discourse and electoral processes" and respective mitigation measures may entail:
Calabrese, S., & Reich, O. (January 2024). Identifying, analysing, assessing and mitigation potential negative effects in civic discourse and electoral processes: A minimum menu of risks very large online platforms should take heed of. European Partnership for Democracy & Civil Liberties Union for Europe. https://epd.eu/news-publications/identifying-systemic-risks-for-civic-discourse-and-electoral-processes-and-related-mitigation-measures-under-eus-digital-services-act/

15 Smirnova, J., Ahonen, A., Mathelemuse, N., Schwertheim, H., & Winter, H. (February 25, 2022). Bundestagswahl 2021: Digitale Bedrohungen und ihre Folgen. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/bundestagswahl-2021-digitale-bedrohungen-und-ihre-folgen/

16 Panizio, E. (Ed.). (November 2023). Disinformation narratives during the 2023 elections in Europe. European Digital Media Observatory (EDMO) Task Force on the 2024 European Parliament Elections. https://edmo.eu/wp-content/uploads/2023/10/EDMO-TF-Elections-disinformation-narratives-2023.pdf

17 This includes the 2020 US Presidential elections, 2021 German Federal Election, 2022 French and Australian elections, and the 2022 US Mid-Term elections, among others:
Guerin, C., & Maharasingam-Shah, E. (October 5, 2020). Public Figures, Public Rage: Candidate abuse on social media. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/public-figures-public-rage-candidate-abuse-on-social-media/;
Dorn, M., & Bundtzen, S. (February 3, 2021). Bundestagswahl 2021 – Eine Evaluation der Regeln gegen digitale Bedrohungen. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/bundestagswahl-2021-eine-evaluation-der-regeln-gegen-digitale-bedrohungen/

18 Panizio, E. (Ed.). (November 2023). Disinformation narratives during the 2023 elections in Europe. European Digital Media Observatory (EDMO) Task Force on the 2024 European Parliament Elections.
https://edmo.eu/wp-content/uploads/2023/10/EDMO-TF-Elections-disinformation-narratives-2023.pdf

19 European External Action Service (EEAS). (February 2023). 1st EEAS Report on Foreign Information Manipulation and Interference Threats – Towards a framework for networked defence.
https://www.eeas.europa.eu/sites/default/files/documents/2023/EEAS-DataTeam-ThreatReport-2023..pdf

20 European External Action Service (EEAS). (January 2024). 2nd EEAS Report on Foreign Information Manipulation and Interference Threats – A framework for networked defence. https://www.eeas.europa.eu/sites/default/files/documents/2024/EEAS-2nd-Report%20on%20FIMI%20Threats-January-2024_0.pdf

21 ISD Germany & HateAid (March 9, 2022). Hass als Berufsrisiko: Digitale Gewalt und Sexismus im Bundestagswahlkampf.
https://www.isdglobal.org/isd-publications/hass-als-berufsrisiko-digitale-gewalt-und-sexismus-im-bundestagswahlkampf/

22 Smirnova, J., Winter, H., Mathelemuse, N., Dorn, M., & Schwertheim, H. (September 16, 2021). Digitale Gewalt und Desinformation gegen Spitzenkandidat:innen vor der Bundestagswahl 2021. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/digitale-gewalt-und-desinformation-gegen-spitzenkandidatinnen-vor-der-bundestagswahl-2021/

23 Goulds, S., Gauer, M., Corr, A., & Gallinetti, J. (2020). The State of the World's Girls 2020 – Free to be online? Girls' and young women's experiences of online. PLAN International. https://www.plan.de/fileadmin/website/05._Ueber_uns/Maedchenberichte/Maedchenbericht_2020/Free_to_be_online_report_englisch_FINAL.pdf

24 For example, during the EU Parliamentary elections (2019), German federal election (2021), US presidential elections (2020), US mid-term elections (2022), and the French elections (2022). Spring, M., & Webster, L. (July 15, 2019.) A web of abuse: How the far right disproportionately targets female politicians. BBC. https://www.bbc.com/news/blogs-trending-48871400
Smirnova, J., Winter, H., Mathelemuse, N., Dorn, M., & Schwertheim, H. (September 16, 2021). Digitale Gewalt und Desinformation gegen Spitzenkandidat:innen vor der Bundestagswahl 2021. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/digitale-gewalt-und-desinformation-gegen-spitzenkandidatinnen-vor-der-bundestagswahl-2021/
Guerin, C., & Maharasingam-Shah, E. (October 5, 2020). Public Figures, Public Rage: Candidate abuse on social media. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/public-figures-public-rage-candidate-abuse-on-social-media/
Simmons, C., & Fourel, Z. (December 1, 2022). Hate in Plain Sight: Abuse Targeting Women Ahead of the 2022 Midterm Elections on TikTok and Instagram. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/hate-in-plain-sight-abuse-targeting-women-ahead-of-the-2022-midterm-elections-on-tiktok-instagram/
Simmons;, C., Fourel, Z., & Morinière, S. (August 16, 2022). La campagne de l'intimidation : étude de cas des violences numériques envers les candidats aux élections de 2022. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/la-campagne-de-lintimidation-etude-de-cas-des-violences-numeriques-envers-les-candidats-aux-elections-de-2022/

25 Guerin, C., & Maharasingam-Shah, E. (October 5, 2020). Public Figures, Public Rage: Candidate abuse on social media. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/public-figures-public-rage-candidate-abuse-on-social-media/

26 Council of the European Union. (November 28, 2008). Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law. 2008/913/JHA.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178

27 European Commission. (January 27, 2014). Report from the Commission to the European Parliament and the Council on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law. COM/2014/027 final. https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:52014DC0027

28 Gandhi, M. (2024). Terrorism, Extremism, Disinformation, and Artificial Intelligence: A Primer for Policy Practitioners. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/wp-content/uploads/2024/01/Terrorism-extremism-disinformation-and-artificial-intelligence_A-primer-for-policy-practitioners.pdf

29 IBM. (n.d.). What are AI hallucinations?. https://www.ibm.com/topics/ai-hallucinations; Meyer-Resende, M., et al. (April 2024). Are Chatbots Misinforming Us About the European Elections? Yes. Democracy Reporting International. https://democracy-reporting.org/en/office/global/publications/chatbot-audit ; Novak, M. (April 2023). GOP Releases First Ever AI-Created Attack Ad Against President Biden. Forbes. https://www.forbes.com/sites/mattnovak/2023/04/25/gop-releases-first-ever-ai-created-attack-ad-against-president-biden/

30 Dorn, M., Bundtzen, S., Schwieter, C. & Gandhi, M. (September 12, 2023). Emerging Platforms and Technologies: An Overview of the Current Threat Landscape and its Policy Implications. Institute of Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/emerging-platforms-and-technologies-an-overview-of-the-current-threat-landscape-and-its-policy-implications/

31 IBM. (n.d.). What are AI hallucinations?. https://www.ibm.com/topics/ai-hallucinations; Meyer-Resende, M., et al. (April 2024). Are Chatbots Misinforming Us About the European Elections? Yes. Democracy Reporting International. https://democracy-reporting.org/en/office/global/publications/chatbot-audit ; Novak, M. (April 2023). GOP Releases First Ever AI-Created Attack Ad Against President Biden. Forbes. https://www.forbes.com/sites/mattnovak/2023/04/25/gop-releases-first-ever-ai-created-attack-ad-against-president-biden/

32 Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review 107, 1753. https://doi.org/10.15779/z38rv0d15j

33 Chowdhury, R. & Dhanya, L. (2023). "Your opinion doesn't matter anyway": Exposing Technology-Facilitated Gender Based Violence in an Era of Generative AI. UNESCO. https://www.unesco.org/en/articles/technology-facilitated-gender-based-violence-times-generative-ai

34 Scott, M. (April 2024). Deepfakes, distrust and disinformation: Welcome to the AI election. Politico. https://www.politico.eu/article/deepfakes-distrust-disinformation-welcome-ai-election-2024/

35 Bundtzen, S. (September 14, 2023). Misogynistic Pathways to Radicalisation: Recommended Measures for Platforms to Assess and Mitigate Online Gender-Based Violence. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/misogynistic-pathways-to-radicalisation-recommended-measures-for-platforms-to-assess-and-mitigate-online-gender-based-violence/

36 Sinders, C., Shukla, V. & Voegeli, E. (2021). Trust Through Trickery. Common Place. https:// commonplace.knowledgefutures.org/pub/trustthrough-trickery/release/1

37 Burkell, J. & Regan, P. M. (2019). Voter preferences, voter manipulation, voter analytics: policy options for less surveillance and more autonomy. Internet Policy Review, 8(4). https://policyreview.info/pdf/policyreview-2019-4-1438.pdf

38 Burkell, J. & Regan, P. M. (2019). Voter preferences, voter manipulation, voter analytics: policy options for less surveillance and more autonomy. Internet Policy Review, 8(4). https://policyreview.info/pdf/policyreview-2019-4-1438.pdf

39 Bundtzen, S. (December 9, 2023). Suggested for You: Understanding How Algorithmic Ranking Practice Affect Online Discourse and Assessing Proposed Alternatives. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/suggested-for-you-understanding-how-algorithmic-ranking-practices-affect-online-discourses-and-assessing-proposed-alternatives/

40 Bundtzen, S. (December 9, 2023). Suggested for You: Understanding How Algorithmic Ranking Practice Affect Online Discourse and Assessing Proposed Alternatives. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/suggested-for-you-understanding-how-algorithmic-ranking-practices-affect-online-discourses-and-assessing-proposed-alternatives/

41 Gryz, J. & Rojszczak, M. (2021). Black box algorithms and the rights of individuals: no easy solution to the "explainability" problem. Internet Policy Review, 10(2). https://doi.org/10.14763/2021.2.1564

42 Scott, M., Coi, G., & Poloni, G. (May 7, 2024). Anatomy of a scroll: Inside TikTok's AI-powered algorithms. Politico. https://www.politico.eu/article/anatomy-scroll-inside-tiktok-ai-powered-algorithm-israel-palestine-war/

43 Meta. (n.d.). Misinformation. Transparency Center. https://transparency.meta.com/en-gb/policies/community-standards/misinformation/

44 Quintais, J.P., Appelman, N. & Ó Fathaigh, R. (September 28, 2022). Using Terms and Conditions to apply Fundamental Rights to Content Moderation. German Law Journal (2023), 24, pp. 881–91. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/B30B9043D1C6F14AE9C3647A845E6E10/S2071832223000536a.pdf/using-terms-and-conditions-to-apply-fundamental-rights-to-content-moderation.pdf

45 European Union Agency for Fundamental Rights (FRA). (November 29, 2023). Online content moderation - Current challenges in detecting hate speech. https://fra.europa.eu/en/publication/2023/online-content-moderation

46 Quintais, J.P., Appelman, N. & Ó Fathaigh, R. (September 28, 2022). Using Terms and Conditions to apply Fundamental Rights to Content Moderation. German Law Journal (2023), 24, pp. 881–91. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/B30B9043D1C6F14AE9C3647A845E6E10/S2071832223000536a.pdf/using-terms-and-conditions-to-apply-fundamental-rights-to-content-moderation.pdf

47 Popli, N. (October 26, 2021). The 5 Most Important Revelations From the 'Facebook Papers'. TIME. https://time.com/6110234/facebook-papers-testimony-explained/

48 Marinescu, D. (September 8, 2021). Facebook's Content Moderation Language Barrier. New America. https://www.newamerica.org/the-thread/facebooks-content-moderation-language-barrier/; Global Witness (November 30, 2023). How Big Tech platforms are neglecting their non-English language users. https://www.globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/

49 Deck, A. (June 27, 2023). AI moderation is no match for hate speech in Ethiopian languages. Rest of World. https://restofworld.org/2023/ai-content-moderation-hate-speech/

50 Amnesty International. (December 14, 2022). Kenya: Meta sued for 1.6 billion USD for fueling Ethiopia ethnic violence. https://www.amnesty.org/en/latest/news/2022/12/kenya-meta-sued-for-1-6-billion-usd-for-fueling-ethiopia-ethnic-violence/

51 Sosnovik, V., & Goga, O. (2021). Understanding the Complexity of Detecting Political Ads. Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia. https://arxiv.org/pdf/2103.00822; European Digital Media Observatory (2021). D.14: Description and analysis of relevant emerging research Topics / Issue-Based Advertising, M 19. https://edmo.eu/wp-content/uploads/2021/12/Issue-Based-Advertising-Report.pdf

52 Nadler, A., Crain, M., & Donovan, J. (2018). Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech. Data & Society Research Institute. https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf

53 Votta, F., Kruschinski, S., Hove, M., Helberger, N., Dobber, T., & de Vreese, C. (2024). Who Does(n't) Target You? Mapping the Worldwide Usage of Online Political Microtargeting. Journal of Quantitative Description: Digital Media, 4. https://journalqd.org/article/view/4188

54 Simchon, A., Edwards, M., Lewandowsky, S. (February 2, 2024). The persuasive effects of political microtargeting in the age of generative artificial intelligence. PNAS Nexus 3(2). https://academic.oup.com/pnasnexus/article/3/2/pgae035/7591134

55 Note this list is not exhaustive. Recognising that platforms are continuously updating their policies and election safeguarding approaches, this list provides and overview of milestone announcements pertinent to the 2024 election year. Other services, such as generative AI companies, were also included given their potential impact on elections.

56 Meta. (November 28, 2023). How Meta Is Planning for Elections in 2024. https://about.fb.com/news/2023/11/how-meta-is-planning-for-elections-in-2024/

57 Meta. (April 5, 2024). Our Approach to Labeling AI-Generated Content and Manipulated Media. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

58 Meta. (February 25, 2024). How Meta Is Preparing for the EU's 2024 Parliament Elections. https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/

59 Google. (December 19, 2023). How we're approaching the 2024 U.S. elections. https://blog.google/outreach-initiatives/civics/how-were-approaching-the-2024-us-elections/

60 Google. (March 21, 2024). Fighting misinformation online: Protecting the integrity of elections. https://blog.google/around-the-globe/google-europe/fighting-misinformation-online-elections/

61 Robins-Early, N. (March 12, 2024). Google restricts AI chatbot Gemini from answering questions on 2024 elections. The Guardian. https://www.theguardian.com/us-news/2024/mar/12/google-ai-gemini-2024-election

62 TikTok (January 18, 2024). Protecting election integrity in 2024. https://newsroom.tiktok.com/en-us/protecting-election-integrity-in-2024

63 TikTok for Business. (May 6, 2024). How we're protecting election integrity on TikTok. https://www.tiktok.com/business/en-GB/blog/protecting-election-integrity-on-tiktok

64 OpenAI (January 15, 2024). How OpenAI is approaching 2024 worldwide elections. https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/

65 OpenAI. (May 21, 2024). OpenAI safety update. https://openai.com/index/openai-safety-update/

66 Anthropic. (February 16, 2024). Preparing for global elections in 2024. https://www.anthropic.com/news/preparing-for-global-elections-in-2024

67 Anthropic (June 6, 2024). Testing and mitigating elections-related risks. https://www.anthropic.com/news/testing-and-mitigating-elections-related-risks

68 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). (2022). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065

69 Cloudflare. (n.d.). What is rate limiting? https://www.cloudflare.com/en-gb/learning/bots/what-is-rate-limiting/ ; Schneidman, N. (March 2024). The Shortlist: Social Media Platform Recommendations. Protect Democracy. https://protectdemocracy.org/work/shortlist-social-media-recommendations

70 Salinas, S. (July 2023). Twitter says rate limits were to help thwart bots, 'small percentage' of users currently affected. CNBC. https://www.cnbc.com/2023/07/04/twitter-says-rate-limits-were-to-help-thwart-bots-few-users-affected.html ; X Help Center. (n.d.). About X Limits. https://help.x.com/en/rules-and-policies/x-limits

71 WhatsApp Help Center. (n.d.). About forwarding limits. https://faq.whatsapp.com/1053543185312573?cms_id=1053543185312573&draft=false

72 Porter, J. (April 27, 2020). WhatsApp says its forwarding limits have cut the spread of viral messages by 70 percent. The Verge. https://www.theverge.com/2020/4/27/21238082/whatsapp-forward-message-limits-viral-misinformation-decline

73 de Freitas Melo, P., Vieira, C.C., Garimella, K., de Melo, P.O.S.V., Benevenuto, F. (2020). Can WhatsApp Counter Misinformation by Limiting Message Forwarding?. In: Cherifi, H., Gaito, S., Mendes, J., Moro, E., Rocha, L. (eds) Complex Networks and Their Applications VIII - Studies in Computational Intelligence, Vol 881. Springer, Cham. https://doi.org/10.1007/978-3-030-36687-2_31

74 de Freitas Melo, P., Hoseini, M., Zannettou, S., & Benevenuto, F. (2024). Don't Break the Chain: Measuring Message Forwarding on WhatsApp. Proceedings of the International AAAI Conference on Web and Social Media, 18(1), 1054-1067. https://doi.org/10.1609/icwsm.v18i1.31372

75 Schneidman, N. (March 2024). The Shortlist: Social Media Platform Recommendations. Protect Democracy. https://protectdemocracy.org/work/shortlist-social-media-recommendations

76 Lloyd, J., Lambe, K., Davidson, A., & Jakobson, C. (December 2020). Platform Accountability and Elections: Lessons Learned. Mozilla Foundation. https://foundation.mozilla.org/fr/blog/platform-accountability-and-elections-lessons-learned/

77 Electoral Integrity Partnership. (2020). Repeat Offenders: Voting Misinformation on Twitter in the 2020 United States Election. https://www.eipartnership.net/2020/repeat-offenders

78 Oversight Board. (April 2023). Policy advisory opinion 2022-01, Removal of COVID-19 misinformation. Meta. https://www.oversightboard.com/wp-content/uploads/2023/11/547865527461223.pdf

79 Broniatowski, D. A., Simons, J. R., Gu, J., Jamison, A. M., & Abroms, L. C. (2023). The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. Science Advances, 9(37). https://www.science.org/doi/full/10.1126/sciadv.adh2132

80 Elswah, M. (January 2024). Investigating Content Moderation Systems in the Global South. Center for Democracy and Technology. https://cdt.org/insights/investigating-content-moderation-systems-in-the-global-south

81 Milmo, D. (October 2021). Facebook revelations: What is in cache of internal documents?. The Guardian. https://www.theguardian.com/technology/2021/oct/25/facebook-revelations-from-misinformation-to-mental-health

82 Global Witness (November 30, 2023). How Big Tech platforms are neglecting their non-English language users. https://www.globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/

83 European Commission. (n.d.). How the Digital Services Act enhances transparency online: transparency reports. https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency#ecl-inpage-lsetrsdp

84 Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (October 2021). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 381. https://dl.acm.org/doi/10.1145/3479525

85 Bundesgerichtshof. (July 29, 2021). Urteil vom 29. Juli 2021, III ZR 179/20. https://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&nr=121741&pos=0&anz=1

86 McCabe, S.D., Ferrari, D., Green, J., et al. (2024). Post-January 6th deplatforming reduced the reach of misinformation on Twitter. Nature 630, 132–140. https://www.nature.com/articles/s41586-024-07524-8

87 McCabe, S.D., Ferrari, D., Green, J., et al. (2024). Post-January 6th deplatforming reduced the reach of misinformation on Twitter. Nature 630, 132–140. https://www.nature.com/articles/s41586-024-07524-8

88 Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (October 2021). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 381. https://dl.acm.org/doi/10.1145/3479525; Ribeiro, M. H., Jhaver, S., Reignier-Tayar, M., & West, R. (2024). Deplatforming Norm-Violating Influencers on Social Media Reduces Overall Online Attention Toward Them. arXiv preprint arXiv:2401.01253. https://arxiv.org/abs/2401.01253

89 Broniatowski, D. A. et al. (2023). The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. Sci. Adv.9(37). https://www.science.org/doi/10.1126/sciadv.adh2132

90 Rauchfleisch, A., & Kaiser, J. (2024). The impact of deplatforming the far right: an analysis of YouTube and BitChute. Information, Communication & Society, 1–19. https://www.tandfonline.com/doi/abs/10.1080/1369118X.2024.2346524 ; Zimdars, M. (2024). Alt-health influencers and the threat of social media deplatforming. Journal of the Association for Information Science and Technology, 1–14. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24870

91 Matlach, P. & Hammer, D. (January 23, 2024). The German Far Right Online: A Longitudinal Study. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/the-german-far-right-online-a-longitudinal-study/

92 Council of the European Union. (May 2024). Russia's war of aggression against Ukraine: Council bans broadcasting activities in the European Union of four more Russia-associated media outlets. https://www.consilium.europa.eu/en/press/press-releases/2024/05/17/russia-s-war-of-aggression-against-ukraine-council-bans-broadcasting-activities-in-the-european-union-of-four-more-russia-associated-media-outlets/

93 Institute for Strategic Dialogue. (2024). Two Years On: An Analysis of Russian State and Pro-Kremlin Information Warfare in the Context of the Invasion of Ukraine. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/two-years-on-an-analysis-of-russian-state-and-pro-kremlin-information-warfare-in-the-context-of-the-invasion-of-ukraine/; Pamment, J. (2023). How the Kremlin circumvented EU sanctions on Russian state media in the first weeks of the illegal invasion of Ukraine. Place Branding and Public Diplomacy, 19(2), 200-205. https://doi.org/10.1057/s41254-022-00275-1

94 Balint, K., Wildon, J., Arcostanzo, F., & Reyes, K.D. (2022). Effectiveness of the Sanctions on Russian State-Affiliated Media in the EU: An investigation into website traffic & possible circumvention methods. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/effectiveness-of-the-sanctions-on-russian-state-affiliated-media-in-the-eu-an-investigation-into-website-traffic-possible-circumvention-methods-2/; Tuhina, G. (February 2024). Two Years Into EU Ban, Russia's RT And Sputnik Are Still Accessible Across The EU. Radio Free Europe. https://www.rferl.org/a/russia-rt-sputnik-eu-access-bans-propaganda-ukraine-war/32803929.html

95 Glazunova, S., Ryzhova, A., Bruns, A., Montaña-Niño, S. X., Beseler, A., & Dehghan, E. (2023). A platform policy implementation audit of actions against Russia's state-controlled media. Internet Policy Review, 12(2), 1-27. https://doi.org/10.14763/2023.2.1711

96 Pamment, J. (2023). How the Kremlin circumvented EU sanctions on Russian state media in the first weeks of the illegal invasion of Ukraine. Place Branding and Public Diplomacy, 19(2), 200-205. https://doi.org/10.1057/s41254-022-00275-1

97 Kahn, G. (2023). Despite Western bans, Putin's propaganda flourishes in Spanish on TV and social media. Reuters Institute, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/news/despite-western-bans-putins-propaganda-flourishes-spanish-tv-and-social-media

98 Bundtzen, S. (December 9, 2023). Suggested for You: Understanding How Algorithmic Ranking Practice Affect Online Discourse and Assessing Proposed Alternatives. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/isd-publications/suggested-for-you-understanding-how-algorithmic-ranking-practices-affect-online-discourses-and-assessing-proposed-alternatives/

99 Spence, R., Bifulco, A., Bradbury, P., Martellozzo, E., & DeMarco, J. (2023). The psychological impacts of content moderation on content moderators: A qualitative study. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 17(4), Article 8. https://doi.org/10.5817/CP2023-4-8 ; Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951719897945

100 Scott, M., & Kayali, L. (October 2020). What happened when humans stopped managing social media content. Politico. https://www.politico.eu/article/facebook-content-moderation-automation/

101 Llanso, E., van Hoboken, J., Leerssen, P., & Harambam, J. (February 2020). Artificial Intelligence, Content Moderation, and Freedom of Expression. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, Institute for Information Law. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

102 Llanso, E., van Hoboken, J., Leerssen, P., & Harambam, J. (February 2020). Artificial Intelligence, Content Moderation, and Freedom of Expression. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, Institute for Information Law. https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

103 Kaiser, B. & Mayer, J. (2023). It's the Algorithm: A large-scale comparative field study of misinformation interventions. Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/its-the-algorithm-a-large-scale-comparative-field-study-of-misinformation-interventions

104 Vincent, E. M., Théro, H., & Shabayek, S. (2022). Measuring the effect of Facebook's downranking interventions against groups and websites that repeatedly share misinformation. Harvard Kennedy School Misinformation Review, 3(3). https://misinforeview.hks.harvard.edu/article/measuring-the-effect-of-facebooks-downranking-interventions-against-groups-and-websites-that-repeatedly-share-misinformation/

105 Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign?. Science, 381(6656), 398-404. https://www.science.org/doi/10.1126/science.abp9364

106 Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., ... & Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign?. Science, 381(6656), 398-404. https://www.science.org/doi/10.1126/science.abp9364

107 Clegg, N. (February 2024). Labeling AI-Generated Images on Facebook, Instagram and Threads. Meta. https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/

108 Kaiser, B. & Mayer, J. (2023). It's the Algorithm: A large-scale comparative field study of misinformation interventions. Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/its-the-algorithm-a-large-scale-comparative-field-study-of-misinformation-interventions

109 Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology, 32(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

110 Kaiser, B. & Mayer, J. (2023). It's the Algorithm: A large-scale comparative field study of misinformation interventions. Knight First Amendment Institute, Columbia University. https://knightcolumbia.org/content/its-the-algorithm-a-large-scale-comparative-field-study-of-misinformation-interventions

111 Pennycook, G., Epstein, Z., Mosleh, M. et al. (2021). Shifting attention to accuracy can reduce misinformation online. Nature 592, 590–595. https://doi.org/10.1038/s41586-021-03344-2

112 European Commission. (June 2022). 2022 Strengthened Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

113 Weber, C., Dunaway, J., & Johnson, T. (2012). It's All in the Name: Source Cue Ambiguity and the Persuasive Appeal of Campaign Ads. Political Behavior, 34(3), 561–584. https://doi.org/10.1007/s11109-011-9172-y

114 Hartmann, I. A. (2021). Combining Ad Libraries with Fact Checking to Increase Transparency of Misinformation. Wikimedia/Yale Law School Initiative on Intermediaries and Information, Yale University. https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/combining-ad-libraries-fact-checking-increase-transparency-misinformation ; Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2018). The IRA, social media and political polarization in the United States, 2012-2018. Oxford Internet Institute, University of Oxford. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2018/12/The-IRA-Social-Media-and-Political-Polarization.pdf

115 Binford, M. T., Wojdynski, B. W., Lee, Y. I., Sun, S., & Briscoe, A. (2021). Invisible transparency: Visual attention to disclosures and source recognition in Facebook political advertising. Journal of Information Technology & Politics, 18(1), 70-83. https://doi.org/10.1080/19331681.2020.1805388

116 Dubois, P. R., Arteau-Leclerc, C., & Giasson, T. (2022). Micro-Targeting, Social Media, and Third Party Advertising: Why the Facebook Ad Library Cannot Prevent Threats to Canadian Democracy. In Garnett, H.A, & Pal, M. (Eds.), Cyber-threats to Canadian Democracy. McGill-Queens University Press, https://ssrn.com/abstract=3817971

117 Mozilla Foundation. (April 2024). Full Disclosure: Stress testing tech platforms' ad repositories. https://foundation.mozilla.org/en/research/library/full-disclosure-stress-testing-tech-platforms-ad-repositories/

118 Goodwin, A., Joseff, K., Riedl, M. J., Lukito, J., & Woolley, S. (2023). Political relational influencers: The mobilization of social media influencers in the political arena. International Journal of Communication, 17, 21. https://ijoc.org/index.php/ijoc/article/view/18987/4070

119 Hartmann, I. A. (2021). Combining Ad Libraries with Fact Checking to Increase Transparency of Misinformation. Wikimedia/Yale Law School Initiative on Intermediaries and Information. https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/combining-ad-libraries-fact-checking-increase-transparency-misinformation

120 Kirk, N., & Teeling, L. (2022). A review of political advertising online during the 2019 European Elections and establishing future regulatory requirements in Ireland. Irish Political Studies, 37(1), 85–102. https://doi.org/10.1080/07907184.2021.1907888; European Partnership for Democracy. (2020). Virtual Insanity? The need to guarantee transparency in digital political advertising. https://epd.eu/content/uploads/2023/08/Virtual-Insanity-synthesis-of-findings-on-digital-political-advertising-EPD-03-2020.pdf; Who Targets Me (2023). Google's "election ads" Policy and Ad Library are a failure of transparency. https://whotargets.me/en/googles-election-ads-policy-and-ad-library-are-a-failure-of-transparency/

121 Bouchard, P., Faddoul, M., & Cetin, R. B. (2024). No Embargo in Sight: Meta Lets Pro-Russia Propaganda Ads Flood the EU. AI Forensics. https://aiforensics.org/uploads/No_Embargo_in_Sight_AI_Forensics_Report_ad7ede416b.pdf

122 Ivanova, I. (October 2019). Twitter announces ban on all political ads. CBS News. https://www.cbsnews.com/news/twitter-political-ads-will-be-banned-ceo-jack-dorsey-announced-2019-10-31/; Chandlee, B. (October 2019). Understanding our policies around paid ads. TikTok. https://newsroom.tiktok.com/en-us/understanding-our-policies-around-paid-ads

123 Mozilla Foundation. (April 2024). Full Disclosure: Stress testing tech platforms' ad repositories. https://foundation.mozilla.org/en/research/library/full-disclosure-stress-testing-tech-platforms-ad-repositories/

124 Ricks, B., Geurkink, B., & Mozilla Foundation. (2021). Th€se Are Not Po£itical Ad$: How Partisan Influencers Are Evading TikTok's Weak Political Ad Policies. Mozilla Foundation. https://foundation.mozilla.org/en/campaigns/tiktok-political-ads/

125 Global Witness. (June 2024). Ticked off: TikTok approves EU elections disinformation ads for publication in Ireland. https://www.globalwitness.org/en/campaigns/digital-threats/ticked-tiktok-approves-eu-elections-disinformation-ads-publication-ireland/

126 Global Witness (October 2022). TikTok and Facebook fail to detect election disinformation in the US, while YouTube succeeds. https://www.globalwitness.org/en/campaigns/digital-threats/tiktok-and-facebook-fail-detect-election-disinformation-us-while-youtube-succeeds/

127 Coalition for Content Provenance and Authenticity. C2PA Technical Specification. https://c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html

128 Coalition for Content Provenance and Authenticity. (May 2024). OpenAI Joins C2PA Steering Committee. https://c2pa.org/post/openai_pr/

129 European Parliament. (April 2024). Corrigendum: Artificial Intelligence Act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

130 Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B. and Forsyth, D . (2024). Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv. https://arxiv.org/abs/2402.04249

131 European Commission. (n.d.). Trusted flaggers under the Digital Services Act (DSA). https://digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa

132 Jahangir, R., Vialle, E., & Moses, D. (May 2024). More Questions Than Flags: Reality Check on DSA's Trusted Flaggers. Tech Policy Press. https://www.techpolicy.press/more-questions-than-flags-reality-check-on-dsas-trusted-flaggers/

133 Lenoir, T. (May 2024). The Difficult Life of Trusted Flaggers. Tech Policy Press. https://www.techpolicy.press/the-difficult-life-of-trusted-flaggers/

134 Appelman, N., & Leerssen, P. (2022). On "Trusted" Flaggers. Yale Journal of Law & Technology, 24. https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf

135 For example, see recommendations by Sedova, K., McNeill, C., Johnson, A., Joshi, A., & Wulkan, I. (2021). AI and the Future of Disinformation Campaigns Part 2: A Threat Model. Center for Security and Emerging Technology (CSET), Georgetown University. https://cset.georgetown.edu/publication/ai-and-the-future-of-disinformation-campaigns-2/; European Digital Media Observatory (EDMO). (June 2022). 10 Recommendations by the Taskforce on Disinformation and the War in Ukraine. https://edmo.eu/publications/10-recommendations-by-the-taskforce-on-disinformation-and-the-war-in-ukraine/; and Polyakova, A., & Fried, D. (2019). Democratic defense against disinformation 2.0. Atlantic Council. https://www.atlanticcouncil.org/in-depth-research-reports/report/democratic-defense-against-disinformation-2-0/

136 Global Internet Forum to Counter Terrorism (GIFCT). (n.d.). Content Incident Protocol. https://gifct.org/content-incident-protocol/

137 European Commission. (June 2022). 2022 Strengthened Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

138 Ozawa, J. V., Lukito, J., Bailez, F., & Fakhouri, L. G. (2024). Brazilian Capitol attack: The interaction between Bolsonaro's supporters' content, WhatsApp, Twitter, and news media. Harvard Kennedy School Misinformation Review. https://misinforeview.hks.harvard.edu/article/brazilian-capitol-attack-the-interaction-between-bolsonaros-supporters-content-whatsapp-twitter-and-news-media/

139 Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. Journal of Management Information Systems, 36(3), 931–968. https://doi.org/10.1080/07421222.2019.1628921

140 Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. CHI Conference on Human Factors in Computing Systems (CHI '20), April 25–30, 2020, Honolulu, HI, USA. ACM, New York, NY, USA. https://doi.org/10.1145/3313831.3376213

141 Wojcik, T., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker M.B.F., Coleman, K., and Baxter, J. (October 2022). Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation. Social and Information Networks (2210.15723v1). https://arxiv.org/pdf/2210.15723

142 Czopek, M. (30 June 2023). "Why Twitter's Community Notes feature mostly fails to combat misinformation" Poynter. https://www.poynter.org/fact-checking/2023/why-twitters-community-notes-feature-mostly-fails-to-combat-misinformation/

143 Kaiser, B., Mayer, J., Matias, J.N. (July 2023). Warnings That Work: Combating Misinformation Without Deplatforming. Lawfare. https://www.lawfaremedia.org/article/warnings-work-combating-misinformation-without-deplatforming; Constine, J. (December 2019). Instagram hides false content behind warnings, except for politicians. Tech Crunch. https://techcrunch.com/2019/12/16/instagram-fact-checking/; Bond, S. (October 2020). Facebook And Twitter Limit Sharing 'New York Post' Story About Joe Biden. NPR. https://www.npr.org/2020/10/14/923766097/facebook-and-twitter-limit-sharing-new-york-post-story-about-joe-biden

144 Clayton, K., Blair, S., Busam, J.A. et al. (2020). Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. Political Behaviour 42, 1073–1095. https://doi.org/10.1007/s11109-019-09533-0.

145 Ross, B., Jung, A., Heisel, J., & Stieglitz, S. (2018). Fake news on social media: The (in) effectiveness of warning messages. Thirty Ninth International Conference on Information Systems. https://www.researchgate.net/publication/328784235_Fake_News_on_Social_Media_The_InEffectiveness_of_Warning_Messages ; Capraro, V., & Celadin, T. (2023). "I think this news is accurate": endorsing accuracy decreases the sharing of fake news and increases the sharing of real news. Personality and Social Psychology Bulletin, 49(12), 1635-1645. https://doi.org/10.1177/01461672221117691

146 Electoral Integrity Partnership. (2020). Repeat Offenders: Voting Misinformation on Twitter in the 2020 United States Election. https://www.eipartnership.net/2020/repeat-offenders

147 Bradshaw, S., & McCain, M. (2022). Lost In Translation: Language Gaps in Social Media Labels. Lawfare. https://www.lawfaremedia.org/article/lost-translation-language-gaps-social-media-labels

148 Kaiser, B., Wei, J.W., Lucherini, E., Lee, K., Matias, J.N., & Mayer, J.R. (2020). Adapting Security Warnings to Counter Online Disinformation. USENIX Security Symposium. https://www.usenix.org/system/files/sec21summer_kaiser.pdf

149 Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2022). Misinformation warnings: Twitter's soft moderation effects on covid-19 vaccine belief echoes. Computers & security, 114. https://doi.org/10.1016/j.cose.2021.102577

150 Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of foreign misinformation. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-45

151 Moravec, P. L., Collis, A., & Wolczynski, N. (2023). Countering State-Controlled Media Propaganda Through Labeling: Evidence from Facebook. Information Systems Research. https://doi.org/10.1287/isre.2022.0305

152 Note that the sample for this research was 54% self-identified as liberal, which potentially does not represent the partisan demographics of the United States. See: Moravec, P. L., Collis, A., & Wolczynski, N. (2023). Countering State-Controlled Media Propaganda Through Labeling: Evidence from Facebook. Information Systems Research. https://doi.org/10.1287/isre.2022.0305; Moravec, P. L., Collis, A., & Wolczynski, N. (2023). Appendices: Countering state-controlled media propaganda through labeling: Evidence from Facebook. https://pubsonline.informs.org/doi/suppl/10.1287/isre.2022.0305/suppl_file/isre.2022.0305.sm1.pdf

153 Arnold, J. R., Reckendorf, A., & Wintersieck, A. L. (2021). Source alerts can reduce the harms of foreign disinformation. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-68

154 Arnold, J. R., Reckendorf, A., & Wintersieck, A. L. (2021). Source alerts can reduce the harms of foreign disinformation. Harvard Kennedy School (HKS) Misinformation Review. https://doi.org/10.37016/mr-2020-68

155 Schliebs, M., Bailey, H., Bright, J., and Howard, P.N. (2021). China's Public Diplomacy Operations: Understanding Engagement and Inauthentic Amplification of PRC Diplomats on Facebook and Twitter. Oxford Internet Institute, University of Oxford. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/05/Chinas-Public-Diplomacy-Operations-Dem.Tech-Working-Paper-2021.1-4.pdf

156 Purnell, N. (July 2023). Meta's Threads Isn't Labeling Propaganda Accounts From Russia, China State Media. Wall Street Journal. https://www.wsj.com/articles/metas-threads-isnt-labeling-propaganda-accounts-from-russia-china-state-media-3f4c6cf8; Center for Countering Digital Hate (CCDH). (February 2022). Facebook failing to label 91% of posts containing Russian propaganda about Ukraine. https://counterhate.com/blog/facebook-failing-to-label-91-of-posts-containing-russian-propaganda-about-ukraine/

157 Bodnar, J. (March 2023). TikTok's Russia Challenge: Kremlin-Funded Media Reaches Millions on the App. GMF Alliance for Securing Democracy. https://securingdemocracy.gmfus.org/tiktoks-russia-challenge/ ; Martin, I., & Baker-White, E. (July 2023). TikTok Has Pushed Chinese Propaganda Ads To Millions Across Europe. Forbes. https://www.forbes.com/sites/iainmartin/2023/07/26/tiktok-chinese-propaganda-ads-europe/

158 European Commission. (April 2024). Commission Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to Article 35(3) of Regulation (EU) 2022/2065. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C_202403014

159 Shakir, U. (May 2024). TikTok is adding an 'AI-generated' label to watermarked third-party content. The Verge. https://www.theverge.com/2024/5/9/24152667/tiktok-ai-generated-label-content-credentials-cai-c2pa

160 Wong, C. M. L., & Wu, Y. (2023). Limits to inoculating against the risk of fake news: a replication study in Singapore during COVID-19. Journal of Risk Research 26(10), 1037-1052. https://doi.org/10.1080/13669877.2023.2249909

161 Lewsey, F. (n.d.). Social media experiment reveals potential to 'inoculate' millions of users against misinformation. University Of Cambridge. https://www.cam.ac.uk/stories/inoculateexperiment

162 Neylan, J., Biddlestone, M., Roozenbeek, J. et al. (2023). How to "inoculate" against multimodal misinformation: A conceptual replication of Roozenbeek and van der Linden (2020). Sci Rep 13. https://doi.org/10.1038/s41598-023-43885-2

163 McPhedran, R., Ratajczak, M., Mawby, M. et al. (2023). Psychological inoculation protects against the social media infodemic. Sci Rep 13. https://doi.org/10.1038/s41598-023-32962-1

164 Harjani, T., Basol, M.-S., Roozenbeek, J., & Linden, S. van der. (2023). Gamified Inoculation Against Misinformation in India: A Randomized Control Trial. Journal of Trial & Error, 3(1). https://doi.org/10.36850/e12

165 Wong, C. M. L., & Wu, Y. (2023). Limits to inoculating against the risk of fake news: a replication study in Singapore during COVID-19. Journal of Risk Research, 26(10), 1037-1052. https://doi.org/10.1080/13669877.2023.2249909

166 Courchesne, L., Ilhardt, J., & Shapiro, J. N. (2021). Review of social science research on the impact of countermeasures against influence operations. Harvard Kennedy School (HKS) Misinformation Review. https://misinforeview.hks.harvard.edu/article/review-of-social-science-research-on-the-impact-of-countermeasures-against-influence-operations/

167 European Digital Media Observatory (EDMO). (May 2022). EDMO releases report on researcher access to platform data. https://edmo.eu/edmo-news/edmo-releases-report-on-researcher-access-to-platform-data/

168 Horowitz, J. (March 2024). Meta to Replace Widely Used Data Tool—and Largely Cut Off Reporter Access. Wall Street Journal. https://www.wsj.com/tech/meta-to-replace-widely-used-data-tooland-largely-cut-off-reporter-access-43fc3f9d; Bundtzen, S. (July 2023). Data Access. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/explainers/data-access/

169 DeLong, L.A. (August 2021). Facebook Disables Ad Observatory; Academicians and Journalists Fire Back. Center for Cybersecurity, New York University. https://cyber.nyu.edu/2021/08/21/facebook-disables-ad-observatory-academicians-and-journalists-fire-back/; Kayser-Bril, N. (August 2021). AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook. Algorithm Watch. https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/

170 Bond, S. (August 2023). Elon Musk sues disinformation researchers, claiming they are driving away advertisers. NPR. https://www.npr.org/2023/08/01/1191318468/elon-musk-sues-disinformation-researchers-claiming-they-are-driving-away-adverti

171 Bundtzen, S. (July 2023). Data Access. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/explainers/data-access/

172 Slovakian Council For Media Services. (n.d.). Digital platforms regulation. https://rpms.sk/en/digital-platforms-regulation; Slovakian Council For Media Services. (March 2024). Quantitative content analysis of narratives surrounding the 2023 election to the National Council of the Slovak Republic. https://rpms.sk/sites/default/files/2024-04/Kvantitativna_obsahova_analyza_EN.pdf

173 Council For Media Services. (March 2024). Quantitative content analysis of narratives surrounding the 2023 election to the National Council of the Slovak Republic.https://rpms.sk/sites/default/files/2024-04/Kvantitativna_obsahova_analyza_EN.pdf

174 European Commission. (March 2024). Approval of the content of a draft Communication from the Commission on Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to the Digital Services Act. https://ec.europa.eu/newsroom/repository/document/2024-13/C_2024_2121_1_EN_annexe_acte_autonome_cp_part1_v3_tpHHZgYyBGFMF8J5rE0OR1GdOis_103911.pdf

175 Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. Nature Human Behaviour, 6(10), 1372-1380. https://doi.org/10.1038/s41562-022-01388-6

176 Courchesne, L., Ilhardt, J., & Shapiro, J. N. (2021). Review of social science research on the impact of countermeasures against influence operations. Harvard Kennedy School (HKS) Misinformation Review. https://misinforeview.hks.harvard.edu/article/review-of-social-science-research-on-the-impact-of-countermeasures-against-influence-operations/; Harjani, T., Basol, M.-S., Roozenbeek, J., & Linden, S. van der. (2023). Gamified Inoculation Against Misinformation in India: A Randomized Control Trial. Journal of Trial & Error, 3(1). https://doi.org/10.36850/e12

177 Courchesne, L., Ilhardt, J., & Shapiro, J. N. (2021). Review of social science research on the impact of countermeasures against influence operations. Harvard Kennedy School (HKS) Misinformation Review. https://misinforeview.hks.harvard.edu/article/review-of-social-science-research-on-the-impact-of-countermeasures-against-influence-operations/

178 Courchesne, L., Ilhardt, J., & Shapiro, J. N. (2021). Review of social science research on the impact of countermeasures against influence operations. Harvard Kennedy School (HKS) Misinformation Review. https://misinforeview.hks.harvard.edu/article/review-of-social-science-research-on-the-impact-of-countermeasures-against-influence-operations/

ISD | Powering solutions to extremism, hate and disinformation

Amman | Berlin | London | Paris | Washington DC

www.isdglobal.org