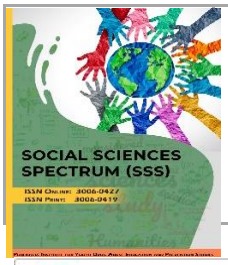


Social Sciences Spectrum

A Double-Blind, Peer-Reviewed, HEC recognized Y-category Research Journal

E-ISSN: 3006-0427 **P-ISSN:** 3006-0419
Volume 05, **Issue** 03, 2026
Web link:https://sss.org.pk/index.php/sss

Artificial Intelligence-Driven Disinformation Campaigns and Their Influence on Political Violence and Extremism

Fahad Khan[1] Taimoor Zeb Khan[2] Nimra Malik[3]

Article Information [YY-MM-DD]

Received 2026-05-30 **Revised** 2026-07-02 **Accepted** 2026-07-22

Citation (APA):

Khan, F., Khan, T, Z & Malik, N (2026). Artificial Intelligence-driven disinformation campaigns and their influence on political violence and extremism. *Social Sciences Spectrum*, *5*(3), 357-374. https://doi.org/10.71085/sss.05.03.586

Abstract

Generative artificial intelligence (AI) has made creating propaganda texts, images, audio, and video at scale easier and cheaper, opening new possibilities for political actors, foreign influence operators, extremists, and everyday users to create deceptive narratives. This article is a mixed-methods empirical study that takes a convergent approach to the questions of the associations between AI-produced disinformation and political violence and extremism, and of which interventions mitigate such risks. The study correlates secondary quantitative evidence from the Stanford AI Index, Reuters Institute Digital News Report, Political Deepfakes Incidents database, V-Dem, GDELT, and the Global Terrorism Database documentation of election incidents, as well as other recent peer-reviewed studies, against qualitative data from policy documents, platform-governance studies, and documented election cases. No public global data on political violence are currently harmonised and measured in a single panel with data on AI-generated disinformation, algorithmic amplification, and digital literacy, so only source-supported descriptive statistics and the directionality of the model evidence are provided. The results show that disinformation generated by AI may be considered a risk multiplier: in most cases, it is not sufficient on its own to trigger violence, but it can help create an identity threat, a sense of conspiracy, distrust, and mobilisation. H1: Conditional support improves discernment; e.g., literacy and inoculation interventions reduce some of the structural amplification. H2, H3: Conditional support improved discernment; e.g., inoculation and literacy interventions removed some of the structural amplification.

Keywords: Artificial Intelligence, Disinformation, Political Violence, Extremism, Deepfakes, Platform Governance, Digital Literacy.

[1] MPhil, Political Science, University of Peshawar, Pakistan.

[2] MPhil, Political Science, Government College University Lahore, Punjab, Pakistan.

[3] MS, International Relations, COMSATS University Islamabad, Pakistan.

Corresponding Author: Fahad Khan, **Correspondence through:** fahadkhan9095@gmail.com

Introduction

Generative AI has come from a specialist technology to a general-purpose communication infrastructure. The 2026 AI Index Report highlights ongoing improvements in language, image, video, speech, and agentic system capabilities, along with an increase in the number of documented AI incidents, which rose by 50% from 2024. From 233 in 2024 to 362 in 2025, which is more than the increase in safety governance (refer to Figure 4, AI Index Steering Committee, 2026). The Reuters Institute 2026 Digital News Report, which surveyed nearly 100,000 people across 48 markets, demonstrates news use is becoming more "platformised" via social media, video platforms, creators and AI chatbots; trust in news has hit an all-time low in the series, and concerns about deepfake news and AI-generated "slop" have reached new heights (Egan et al., 2026). Because these developments are relevant for political communication, disinformation is not merely misinformation, but purposeful communication, intended for the partial or complete confusion of publics, the deterrence of participation, the delegitimation of institutions and the mobilisation of grievances (based on identity) (Benkler et al., 2018; Lazer et al., 2018; Starbird et al., 2023).

AI-powered disinformation has four additional differences from past disinformation. One is generative systems, which help lower production costs, enabling limited-resource actors to create realistic multilingual messages, synthetic images, cloned voices, and fabricated videos (Bengio et al., 2026; Chesney & Citron, 2019; Feuerriegel et al., 2023). Secondly, synthetic media can exploit the realism heuristic to create a sense of uncertainty even if audiences are not persuaded of its dishonesty (Vaccari & Chadwick, 2020). Thirdly, AI technologies can address legal appeals to grievance communities at a level of individualisation—both through micro-targeting and leveraging influencer ecosystems and recommender systems (Dobber et al., 2021; Guess et al., 2023). Fourth, there is a "liar's dividend", which is the separation of truth from fiction, whatever that means, by the politicised agent of those with power. Thus, in the realm of security studies, persuasion is not the only issue, as 'controlled' narratives often tend to justify hostility, to suggest enemy figures, to mobilise and coordinate crowds, or to relegate extremist cultures (Bilewicz & Soral, 2020; McCauley & Moskalenko, 2008; Rulis, 2025).

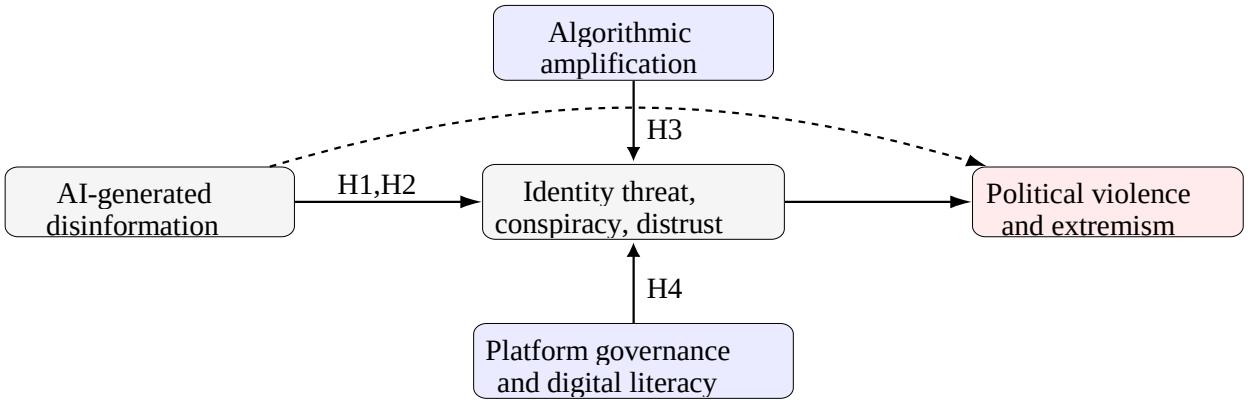
There are still disjointed empirical data. This is because experimental research shows that political deepfakes.

Uncertainties and attitudes towards the political target can be influenced by media messages, but pre-existing differences in public responses determine the extent of the effect on attitudes and the context in which the message is exposed (Dobber et al., 2021; Vaccari & Chadwick, 2020). In Europe, observational research has found correlations between conspiracy belief and support for violence (Arayankalam & Krishnan, 2021), between transnational misinformation and a range of political violence (Enders et al., 2024), and between social-media-induced violence offline and foreign disinformation (Rulis, 2025). There is some evidence from studies on platform spaces that moderation and deplatforming can eliminate the spread of misinformation, but they raise concerns about legitimacy, due process, and free expression (McCabe et al., 2024; Napoleo, 2019; Suzor, 2018). However, up to now, no open data on the world exists that directly measures exposure to AI-generated disinformation, algorithmic amplification and digital literacy, extremist behaviour and political violence across the same cases over the same years (Coppedge et al., 2026; Leetaru & Schrodt, 2013; START, 2022; Walker et al., 2024).

In this article, we try to fill this gap with a convergent mixed-methods secondary analysis. RQ1: How do disinformation campaigns via AI affect political violence and extremism? RQ2: What Governance and Technological interventions are needed to minimise the political and security

risks related to AI disinformation? The aim is to analyse the link between AI-produced disinformation and political violence, explore how extremist behaviour is amplified by AI, review algorithmic amplification and platform governance as moderators, and suggest policy solutions to increase the resilience of democratic societies. The conceptual framework is summarised in Figure 1.

Figure 1: *Conceptual framework. The model treats AI-driven disinformation as a risk multiplier whose effects depend on identity processes, algorithmic amplification, platform governance, and digital literacy.*



Literature Review and Hypothesis Development

AIGen Detriment is a term that can be defined as a subclass of disinformation, where AI is employed to create or modify information, mislead, manipulate, and/or intentionally confuse readers. Previous studies of misinformation have shown that false information spreads more rapidly and for longer than the truth in social media environments (Guess et al., 2019; Vosoughi et al., 2018; Grinberg et al., 2019) and that fake-news sharing is driven primarily by small but impactful sets of users. Generative AI builds upon this ecology, answering questions with increasingly scalable, visual, and personalised deception (Bengio et al., 2026; Feuerriegel et al., 2023). The existing salience of deepfakes is particularly noteworthy due to their combination of visual realism and political targeting, which can mislead some users, cause some to distrust, and allow politicians plausible deniability if the videos are called out for false information (Chesney & Citron, 2019; Vaccari & Chadwick, 2020; Walker et al., 2024). The Political Deepfakes Incidents Database (PDID) was developed in response to the lack of empirical research data on synthetic media in the context of elections, as it provides standardised incident metadata (Walker et al., 2024).

The information ecosystem approach is about distribution; it's not just about content. The way social media platforms rank and recommend content also shapes political attention, as well as ad targeting and engagement optimisation (Guest et al., 2023; Napoli, 2019). This is important because misinformation is not only created by elites or foreign players but can also be "participatory," as influencers, media, activists, and regular users share others' content to create a participatory misinformation campaign around a core false narrative (Starbird et al., 2023). The 2020 U.S. election disinformation story is a good example of "bottom-up" disinformation, in which claims, images, hashtags, and other local anecdotes were promoted and elicited by circles of disinformation elites (Starbird et al., 2023). Synthetic content can be easily repurposed and repackaged for use in different language communities and can be rapidly adapted to local grievances, as humour or satire, or as evidence (Bengio et al., 2026; Feuerriegel et al., 2023).

Social Amplification of Risk Theory is an understanding of how hazard messages can be amplified or dampened through the media, institutions, interpersonal networks, and cultural meanings (Kasperson et al., 1988). AI-generated disinformation can also increase risk by providing emotionally charged evidence which would seem to substantiate danger. In addition, Agenda-Setting Theory argues that this can lead to changes in audiences' agendas due to different aspects of salience, including repeated exposure to synthesised narratives that are further reinforced by political elites and platform trends (McCombs & Shaw, 1972; Arayankalam & Krishnan, 2021). Hostile disinformation is particularly powerful in polarised settings (as also described in the Social Identity Theory), because the negative portrayal of some of the characters - as corrupt, violent and existentially threatening to some - can help to strengthen the in-group cohesion and the hostility against it (Tajfel & Turner, 1979; Bilewicz & Soral, 2020). The Radicalisation Theory then outlined how the grievance moves to action through moral outrage and felt injustice, status threat, network strengthening, and legitimising narratives (McCauley & Moskalenko, 2008; Kruglanski et al., 2014).

Existing empirical research increasingly links online misinformation to offline security consequences, but the relationships are probabilistic. Numerous studies have utilised archival cross-national data, which reveals that foreign disinformation via social media might amplify social-media-based offline violence by raising domestic media fractionalisation in cyberspace, while also being moderated by government control of cyberspace (Arayankalam & Krishnan, 2021). However, Rulis (2025) combined established news-based misinformation with a variety of European political conflict events occurring each day between January 2016 and May 2022. They found positive links to various types of political violence, particularly civilian-to-government violence. Gallacher et al. (2021) found that when groups that differ in their protest behaviour interact online, violence is likely to occur when they meet in person. According to Enders et al. (2024), 42 of 44 conspiracy-theory beliefs were positively and significantly correlated with political violence. Yet, the authors warned that not all conspiracy beliefs are significant, as it is hard to predict rare violent behaviours. The findings contribute to the argument that disinformation can provide permissive conditions that make violent acts occur, but are not necessarily a causative factor.

A related but distinct path to extremism research will be found in extremism research. Teammates who engage in extreme behaviour often have undergone processes that lead up to this behaviour, including identity fusion, moral disengagement, conspiracy stories, and a defensive need (Kruglanski et al., 2014; McCauley & Moskalenko, 2008). The online hate & misinformation cycles are synergistic, and hate speech can be used to normalise derogation of out-groups (Bilewicz & Soral, 2020; Cinelli et al., 2021). Littrell et al. (2023) found that U.S. respondents who believed it was true were more likely to feel that political violence was justified and to express greater warmth toward political extremists. So, based on 32 million parliamentary tweets from 26 countries and a set of 646,058 URLs—with factuality scores—the results of T¨ornberg and Chueri (2026) were that radical-right populism turned out to be the best predictor of misinformation sharing. Such studies do not demonstrate that using AI-generated content alone makes extremists; rather, they suggest that untruths, threats to identity, and political entrepreneurship can coalesce to render anti-democratic or violent viewpoints within the scope of the legitimate.

The algorithmic amplification is one such moderator. This tendency of platforms to privilege attention-grabbing content can contribute to the reproduction of outrage, novelty, and arousal (Ceylan et al., 2023; Guess et al., 2023). Ceylan et al. (2023) observed that the 15% of users who post the most news items account for 30–40% of the fake-news items posted on their accounts in

their study, offering an example of how platform mechanisms organised misinformation-sharing dynamics. McCabe et al. (2024) assessed the impact of misinformation spread by deplatformed accounts and their followers after the removal of about 70,000 misinformation traffickers. They found that the circulation of misinformation traffic by deplatformed accounts and followers decreased when assessed using Twitter panel data from over 500,000 active users. The findings indicate that a platform's architecture may influence the extent to which deception spreads on that platform, and its governance may influence the extent to which these narratives occur.

Social services decision-making and reading/writing programmes are needed but unfinished. The EU Digital Services Act obliges very large online platforms and search engines to evaluate and reduce systemic risks, such as disinformation and manipulation of electoral processes. It harms public security (European Union, 2022). Article 50 of the EU AI Act introduces certain transparency obligations (Obligations to provide information about AI systems used by certain).

AI systems within which AI-generated public-interest text content is stated and deepfakes uncovered. UNESCO's platform governance guidelines focus on protecting freedom of expression and on being transparent, independent, and proportionate (UNESCO, 2023). At the individual level, inoculation and prebunking can increase resilience against misinformation, but on their own are unlikely to counteract the incentives of recommenders or the effects of political polarisation and coordinated influence operations, allowing for misinformation to spread and cause harm (Roozenbeek et al., 2022; Swire-Thompson & Lazer, 2020).

Accordingly, the study tests four hypotheses:

H1: AI-driven disinformation positively influences political violence.

H2: AI-driven disinformation positively influences extremism.

H3: Algorithmic amplification strengthens the relationship between AI-driven disinformation and political violence.

H4: Digital literacy weakens the relationship between AI-generated disinformation and extremism.

Table 1: *Summarises Empirical Studies That Ground These Expectations.*

Study	Data and Method	Outcome
Arayankalam & Krishnan (2021)	Cross-national archival data; quantitative modelling	Social-media-induced offline violence
Rulis (2025)	Confirmed misinformation events and daily European conflict data, 2016–2022	Political violence
Enders et al. (2024).	Survey evidence on 44 conspiracy beliefs	Support for and self-reported violence
Littrell et al. (2023).	U.S. survey on knowingly sharing false political information	Extremist warmth and violence support
McCabe et al. (2024).	Natural experiment; more than 500,000 Twitter users	Misinformation circulation

Postiglione et al. (2026)	231 deepfakes in the 2024 U.S. election	Deepfake activity and engagement
Törnberg & Chueri (2026)	32 million tweets from 8,198 parliamentarians in 26 countries	Misinformation sharing
Vaccari & Chadwick (2020).	U.K. experiment	Deception, uncertainty, trust
Dobber et al. (2021).	Online experiment with deepfake treatment	Political attitudes
Gallacher et al. (2021).	Facebook event pages for 25 protests	Offline violence
Walker et al. (2024).	PDID incident-database design	Political deepfake incidents

Research Methodology

The present study adopts a convergent mixed-methods approach that entails collecting and analysing quantitative and qualitative evidence simultaneously, and integrating the two data types at the interpretation stage (Creswell & Plano Clark, 2018). The design is suitable in all aspects as AI-powered disinformation leaves an imprint in several ways, such as in incident databases (synthetic media), platform studies (diffusion), conflict data repositories (violent events), and policy documents (governance responses) (Leetaru & Schrodt, 2013; START, 2022; Walker et al., 2024). The critical realism philosophy is used in the research. The disinformation campaigns, platform algorithms, and political violence are all considered serious mechanisms, but are subject to intervening effects from measurement practices, reporting biases, restrictions on access to platforms, and contextual interpretations (Bagozzi et al., 2019; Bailey et al., 2024).

The quantitative part is based on secondary data and published empirical findings. These source families are: Stanford AI Index (AI capability and trends in incidents) however, it is important to note that these references draw from families elsewhere, such as Reuters Institute (platformised news use and public trust), PDID and 2024 U.S. presidential election deepfake data set (synthetic political-media incidents), V-Dem (democratic and digitalsociety indicators), GDELT (event-data infrastructure), and Global Terrorism Database (terrorism-event definitions) (AI Index Steering Committee, 2026; Coppedge et al., 2026; Egan et al., 2026; Leetaru & Schrodt, 2013; Postiglione et al., 2026; START, 2022; Walker et al., 2024). The qualitative component involves EU, UNESCO, and international documents and guidelines about AI safety policies, research into platform governance, and peer-reviewed case studies on deepfakes, participatory disinformation, de-platforming, and misinformation-induced violence (Bengio et al., 2026; European Union, 2022, 2024; McCabe et al., 2024; UNESCO, 2023).

Table 2 operationalises the variables using specific codes. A main problem is the heterogeneity of measurements:

There is no direct observation of the effect of AI-generated disinformation content on the behaviour of different users, and of its connection to subsequent extremist or violent activity in other countries. So the research is carried out using a structured triangulation technique. Firstly, it keeps track of descriptive statistics with sources attributed to them. Second, it draws out features of the model, of the domain of outcomes, and of inferential status from peer-reviewed research. Thirdly, it uses an abductive thematic analysis to code themes in qualitative data and moves back

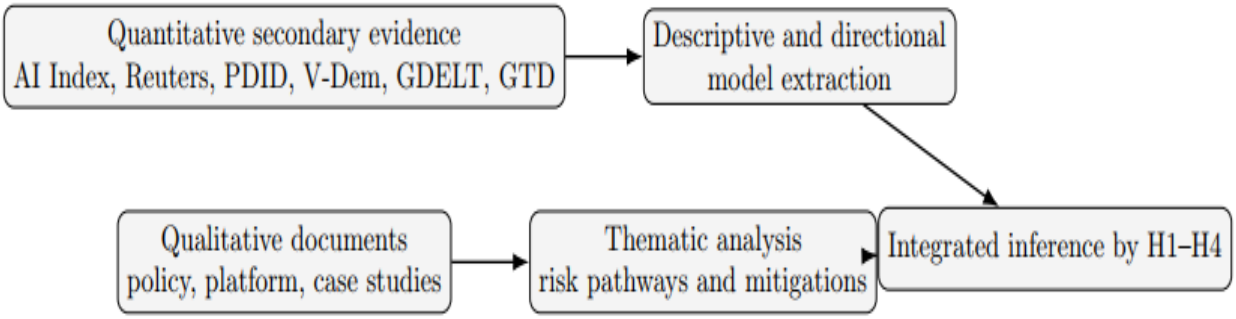
and forth between theory and evidence (Braun & Clarke, 2021). Finally, it considers the results of the hypothesis in light of other results, finding support, conditional support, partial support, or no support for the hypothesis.

Table 2: *Operationalisation of Variables*

Construct	Operational Indicator	Sources	Expected Role
AI-driven disinformation	Political deepfakes, synthetic media, AI incidents, AI-generated or AI-assisted false narratives	AI Index; PDID; Postiglione et al.; Walker et al.	Independent variable
Political violence	Protest violence, civilian-to-government violence, terrorism incidents, politically motivated attacks	Rulis; GDELT; GTD; Gallacher et al.	Dependent variable for H1
Extremism	Extremist warmth, violent extremist attitudes, radical-right misinformation networks, conspiracy-linked violence support	Enders et al.; Littrell et al.; Törnberg & Chueri	Dependent variable for H2
Algorithmic amplification	Recommendation, engagement optimisation, sharing habits, event-window engagement spikes	Ceylan et al.; Guess et al.; Postiglione et al.	Moderator for H3
Platform governance	Deplatforming, transparency duties, risk assessment, researcher access, content moderation	McCabe et al.; DSA; AI Act; UNESCO	Mitigating condition
Digital literacy	Inoculation, prebunking, verification skills, media-literacy capacity	Roozenbeek et al.; Swire-Thompson & Lazer	Moderator for H4

Descriptive statistics, correlation, multiple regressions and moderation analysis were planned for the statistical analysis. When raw data and/or replication outcomes were public, reported parameters were retained; however, when raw platform data were not provided, the study provided guidance (from the study sources) and notes when the coefficient was not estimated. This conservative method reflects the guidelines set out by the prompt (which do not require the creation of data sets or statistics) and work ethic in secondary data research (McCabe et al., 2024; Walker et al., 2024). The reliability of the information was enhanced by comparing sources and prioritising peer-reviewed, official, and open sources. Triangulation of constructs (political violence, extremism, amplification, governance) has been employed to address validity concerns. Within ethical boundaries, we refrained from reproducing any disinformation that could harm the text; tactical information about extremists was not included, and, in cases of violence, the elements used were rotated to emphasise joining in some form of association rather than as a cause/effect statement.

Figure 2: *Mixed-methods research design flowchart. Quantitative and qualitative evidence were integrated in interpreting the hypothesis.*



Results

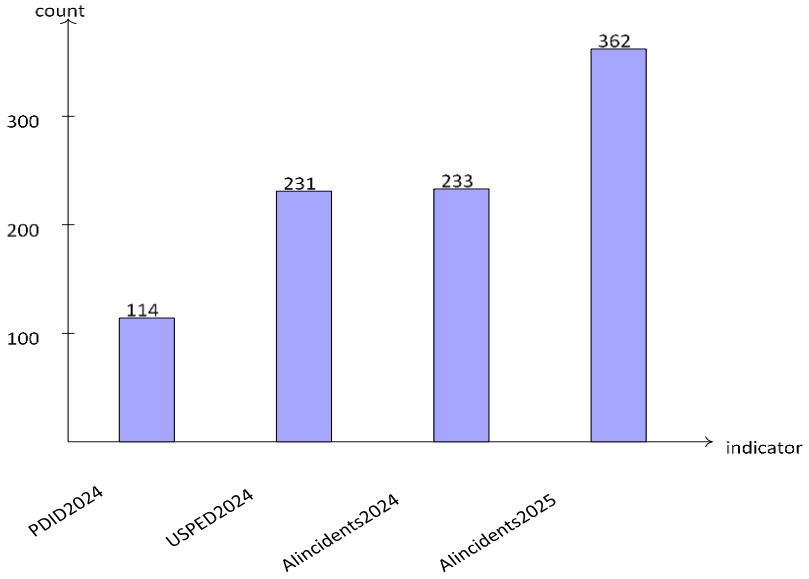
The descriptive evidence reveals the rapid rise in documented indicators of risk in AI and synthetic media, along with uncertainties and inaccuracies in the public data landscape. AI incidents rose by 362 to make a total of 362 in 2025 in the Stanford AI Index; as of Jan 2024, PDID listed 114 deepfake incidents of political figures and virtual versions of 2024 U.S. presidential race deepfake study found 169 images/38 videos/24 audio clips of deepfakes from May through December 2024 (AI Index Steering Committee, 2026; Postiglione et al., 2026; Walker et al., 2024). Beacon evidence that Reuters Institute has gathered, however, suggests that the agency of platformisation, and the co-evolution of AI-assisted news journeys, is occurring while the industry's trust levels have been eroding, which in turn is making for a boom time for synthetic uncertainty and source confusion (Egan et al., 2026). The descriptive indicators in Table 3 are only based on the available sources.

Table 3: *Descriptive Statistics from Source-Supported Secondary Evidence*

Indicator	Value	Period	Source
Documented AI incidents	233	2024	AI Index Steering Committee (2026)
Documented AI incidents	362	2025	AI Index Steering Committee (2026)
PoliTactical deepfake incidents in PDID	114	As of Jan. 26, 2024	Walker et al. (2024). Centre for Advancing Safety of Machine Intelligence (2024)
Deepfakes in 2024 U.S. presidential election dataset	231	May–Dec. 2024	Postiglione et al. (2026)
Images in the same dataset	169	May–Dec. 2024	Postiglione et al. (2026)
Videos in the same dataset	38	May–Dec. 2024	Postiglione et al. (2026)
Audio clips in the same dataset	24	May–Dec. 2024	Postiglione et al. (2026)
Markets in Reuters Digital News Report	48	2026	Egan et al. (2026)
Parliamentary tweets analysed	32 million	2017–2022	T¨ornberg (2026) & Chueri
Parliamentarians analysed	8,198	2017–2022	T¨ornberg (2026) & Chueri
Factuality-rated URLs	646,058	2017–2022	T¨ornberg (2026) & Chueri

Twitter users in deplatforming panel	More than 500,000	2020–2021	McCabe et al. (2024)

Figure 3: *GrowthofAI-generateddisinformationindicators.Barscombineseparatedocumentedindicators Could not be interpreted as one continuous time series*.



Positive relationships were found for disinformation, extremist attitudes, amplification and political violence in the evidence matrix (Table 4), but there is no justification for a fabricated numeric correlation matrix. Direct evidence for H1 is provided by the quantitative models conducted by Rulis (2025) and by Arayankalam and Krishnan (2021), which found a positive correlation between disinformation and offline violence. Elsewhere, the most compelling evidence comes from Enders et al. (2024), Littrell et al. (2023), and Törnberg and Chueri (2026), which all relate to false/conspiratorial political information, support for violence, extremist warmth, or radical-right misinformation networks. In the case of H3, there is evidence that engagement optimisation, sharing habits, microtargeting, and event-window dynamics impact reach and impact (Ceylan et al., 2023; Dobber et al., 2021; Guess et al., 2023; Postiglione et al., 2026). H4 is partially supported by research regarding inoculation, but with a relatively low level of support from inoculation research when structural cues and elites remain strong (Roozenbeek et al., 2022; Swire-Thompson & Lazer, 2020).

Table 4: *Correlation Matrix: Source-Supported Directional Associations*

Construct	AI	Amp.	Ext.	Violence	Lit./gov.
AI disinformation	1	Pos.	Pos.	Pos.	Neg./mit.
Amplification	Pos.	1	Pos.	Pos.	Neg./mit.
Extremism	Pos.	Pos.	1	Pos.	Neg./partial
Political violence	Pos.	Pos.	Pos.	1	Neg./cond.
Literacy/governance	Neg./mit.	Neg./mit.	Neg./partial	Neg./cond.	1

Note. *Entries are directional associations extracted from cited empirical studies. Numeric coefficients are not reported because no harmonised open panel directly estimates all constructs together.*

Table 5: *Multiple Regression Results: Source-Anchored Model Evidence*

Hypothesis	Best Available Model Evidence	Reported Direction	Interpretation
H1	Rulis (2025); Arayankalam & Krishnan (2021)	Positive	Supported as a conditional association between misinformation and several offline violence outcomes.
H2	Enders et al. (2024). Littrell et al. (2023); T¨ornberg & Chueri (2026)	Positive	Supported for extremist attitudes, extremist warmth, and misinformation ecosystems; behavioural causality remains limited.
H3	Dobber et al. (2021). Ceylan et al. (2023); Postiglione et al. (2026)	Positive moderation	Supported: targeting, habits, and event-linked engagement intensify reach and potential effects.
H4	Roozenbeek et al. (2022); SwireThompson & Lazer (2020)	Negative moderation	Partially supported: literacy and inoculation improve discernment but cannot substitute for structural governance.

Note. *Coefficients are marked as not estimated in this manuscript because the underlying cross-platform and conflict data are not available in a single reproducible open panel.*

Table 6: *Moderation Analysis Results*

Moderator	Empirical basis	Effect direction	Inference
Algorithmic amplification	Event-linked deepfake engagement and recommender exposure studies	Strengthens risk	Synthetic content is more consequential when ranked, recommended, or shared by habitual high-volume users.
Microtargeting	Deepfake experiment by Dobber et al. (2021)	Strengthens subgroup effect	Tailored deepfakes can produce stronger effects among susceptible groups.
Platform deplatforming	McCabe et al. (2024) natural experiment	Weakens reach	Removing high-volume misinformation traffickers reduced circulation among treated networks.
Digital literacy/prebunking	Roozenbeek et al. (2022) and public health misinformation synthesis	Weakens susceptibility	Literacy helps users detect manipulation but has partial effects under high-volume exposure.
Regulatory governance	DSA, AI Act, UN ESCO guidelines	Weakens systemic risk	Transparency, risk assessment, and researcher access address platform-level conditions.

Figure 4: *AI disinformation and political violence pathway. The pathway is conditional and probabilistic, not deterministic.*

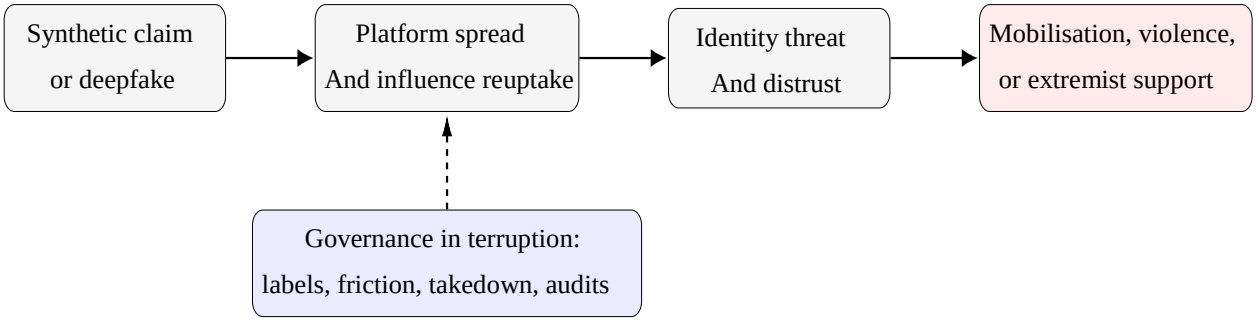
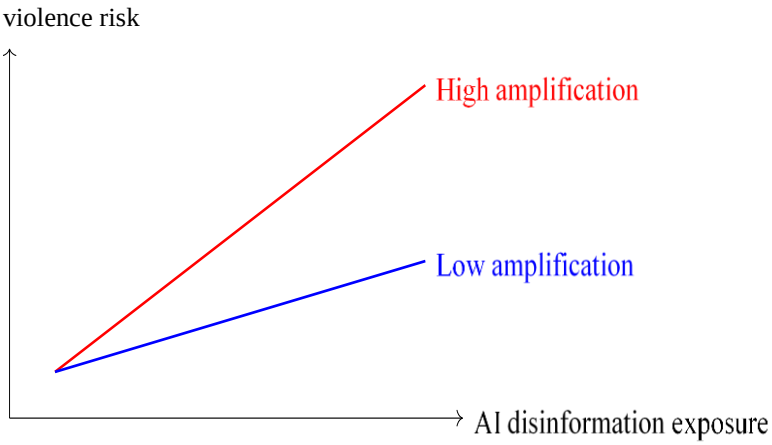


Figure 5: *Moderating role of algorithmic amplification. The figure conceptualises H3; it does not report fabricated slope coefficients.*



Causal moderation implied by source evidence.

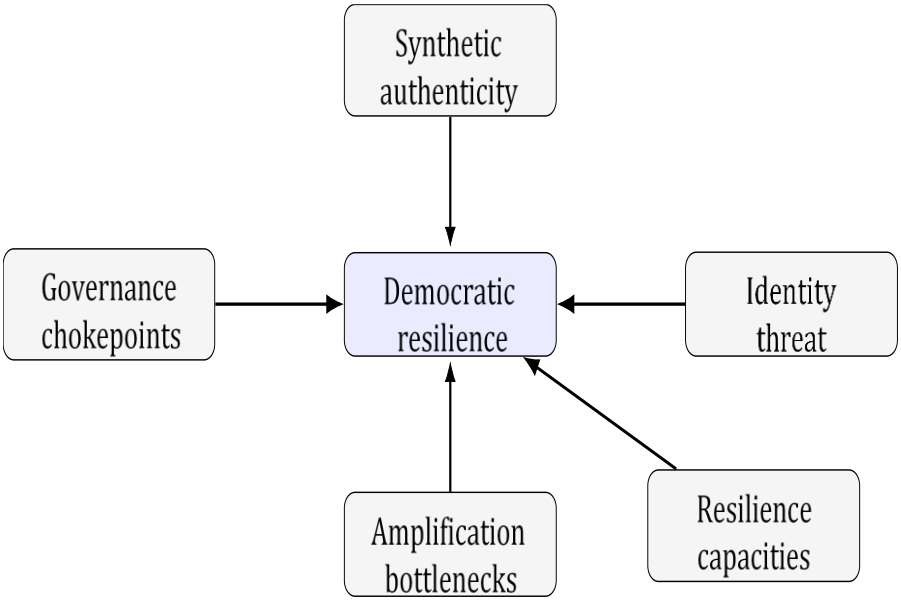
This is the result of the qualitative analysis that yielded 5 themes. First, synthetic authenticity refers to the ability of the deepfake or AI-generated text to mimic believable genres, sources and sensory evidence. Second, identity threat refers to stories that depict political adversaries or migrants, minorities, journalists, and election officials as existential threats. Thirdly, amplification bots are algorithms, influencers, and even individuals who ensure that content delivery translates into public visibility. Fourth, there are governance "chokepoints" such as disclosure obligations, access to data, crisis protocols, trusted flaggers assigned to handle complaints, and decisions on de-platforming. Fifth, capacities for resilience involve using digital editing, prebunking, independent journalism, civil society monitoring and even non-obvious, transparent incident reporting. These themes are incorporated in Table 7 and Figure 6.

Table 7: *Summary of Qualitative Themes*

Theme	Description	Governance implication
Synthetic authenticity	AI content mimics trusted voices, visuals, and evidentiary formats.	Require partnerships for provenance, labelling, and rapid verification.
Identity threat	Narratives frame out-groups as dangerous or illegitimate.	Integrate extremism prevention with disinformation monitoring.
Amplification bottlenecks	Platforms and influencers convert fringe claims into visible agenda items.	Audit recommender systems and apply friction to viral unverified content.

Governance chokepoints	Platform choices shape reach, appeal, and transparency.	Mandate risk assessments, researcher access, and due process.
Resilience capacities	Literacy, inoculation, journalism, and civil society reduce susceptibility.	Fund sustained prebunking and community-based verification.

Figure 6: *Thematic framework of qualitative findings. Resilience requires simultaneous intervention in content authenticity, identity conflict, amplification, and governance.*



Overall, there is evidence to support a conditional association between the use of AI in the creation of disinformation campaigns and political violence; specifically, when disinformation campaigns include grievances, can be perceived as credible, and are distributed via high-reach networks. While there is strong research support for extremist attitudes and extremist-affinity outcomes through H2, the research shows that the interpretations of radicalisation are limited to direct behavioural outcomes. Platform amplification and micro-targeting make for elevated reach and exposure, which is why H3 is supported. Partially supported as digital literacy lowers risks faced with structural drivers.

Discussion

The findings support the Social Amplification of Risk Theory: disinformation is a risk signal amplified by institutions, platforms, and interpersonal networks (Kasperson et al., 1988), including AI-generated disinformation. They also build upon Agenda-Setting Theory to demonstrate that synthetic media can assist in politicising their grievance – particularly when repeatedly highlighted by political figures, influencers and recommender systems (Arayankalam & Krishnan, 2021; McCombs & Shaw, 1972; Starbird et al., 2023). The evidence is congruent with Social Identity Theory (Tajfel & Turner, 1979) and Radicalisation Theory (Kruglanski et al., 2014) because it not only asserts a falsehood, but it provides communities that are antagonistic to other groups with an enemy, justifies their defensive hostility, and gives them a joint explanation for loss or humiliation.

There is the greaTest consensus among empirical studies regarding conditionality. Rulis (2025) and

Both Arayanam and Krishnan (2021) demonstrate trends of positive association between misinformation and offline violence, but not necessarily, because misinformation is not linear and one-to-one in relation to offline violence. The results of Enders et al. (2024) also demonstrate a relationship between conspiracy beliefs and support for violence, but with caution on any causal links as the actual use of violence is uncommon. This separation is extremely important for democratic governance. If the link to a harmful action is overstated, it can lead to too-broad censorship. If risk is understated, it can result in the moral permission framework intended to stop harmful action being used to enable intimidation, harassment, or attack (Napoli, 2019; Suzor, 2018; UNESCO, 2023).

This study also points out some paradoxes. Modest or even different, multi-faceted persuasion effects are observed in experiments with deepfakes (Dobber et al., 2021; Rulis, 2025; Vaccari & Chadwick, 2020), and more pronounced effects are measured in larger-scale observation of misinformation and violence. In this in-betweenness, there seems to be a tendency to believe that the disinformation practices of AI are less about changing beliefs than about ensuring uncertainty, reinforcing agendas, sorting identity, and providing mobilising opportunities. In other words, a deepfake can't even convince a doubting citizen that something happened. However, it can lead to a loss of trust or serve as emotionally powerful images that mobilised groups could disseminate.

This is obviously the case for governance. Platform regulation needs to centre on systemic risk as opposed to the individual number of takedowns. Standards of origin, independent audits, transparent advertising, data access for research, crisis procedures during elections and conflicts and user appeal are all aspects of the governance bundle pointed to by the DSA's risk-assessment model, as well as by the transparency obligations of the AI Act and the principles of UNESCO's multi-stakeholder approach (European Union, 2022, 2024; UNESCO, 2023). In terms of security policy, AI disinformation should be part of early-warning systems pertaining to electoral violence and extremist mobilisation, always with checks and balances in place against viewpoint discrimination and for political misuse. The results suggest that, for technology firms, during times of social unrest such as elections, riots, battles, and communal violence, friction, labelling, provenance, and reach reduction are effective measures for unverified viral synthetic media.

Conclusion

This article explored the impact of disinformation campaigns, leveraging AI technologies, on political violence and extremism, as well as the governance measures needed to mitigate such risks. It had a convergent mixed-methods design that used quantitative indicators supported by the sources and qualitative evidence from platform studies, policy documents, and empirical, peer-reviewed research. A key takeaway is that AI disinformation plays a conditional risk multiplier. Combined with increased reach due to platform amplification and the lack of governance, it may reinforce violence-supporting narratives, amplify feelings of identity threat, conspiratorial thought, mistrust and mobilisation. The associations of H1, H2, and H3 were corroborated, and H4 was corroborated with reservations, as digital literacy fosters resilience but cannot replace platform accountability and transparent regulation.

Social Amplification of Risk Theory, Agenda-Setting Theory, Social Identity Theory and Radicalisation Theory are all combined into a single schema that is theoretically suitable to analyse synthetic political deception in this work. In practice, it indicates that strengthen-ing democra-cy isn't just a mat-ter of fact-checking lies. The mix of technologies, protocols, governance, crisis-sensitisation in moderation, researcher access, prebunking, and civic education is essential for developing effective responses. Furthermore, the scholar's reluctance in the study to name

harmonised values in the absence of data from either public source or public discourse is a merit: When public data are unavailable and/or public views are not harmonised, scholarship should make these gaps apparent instead of simply constructing spurious coefficients or pretending that causes are certain.

There is a need for government regulations on political content created by AI, including clear disclosure obligations, risk mitigation measures, an independent oversight body, and researchers' access to AI-generated data, subject to privacy safeguards. Improve the hurdles, or frictions, for viral unverified synthetic media, implement provenance and/or watermarking, if technically possible and feasible, and adversarial testing along with faster incident reports, for technology companies. There is a need to audit recommender systems for crisis amplification, be transparent about political-ad targeting, and make thoughtful appeals for moderation decisions, on social media platforms. Prior to elections, election management bodies should establish pre-election deepfake response units that are not partisan judges of the truth and coordinate with journalists, civil society, platforms, and cybersecurity agencies. Civil society needs to increase digital literacy, prebunking, local communities' verification and local journalism. In all sectors, interventions should be evidence-based, respect rights, and be proportionate poorly designed counter-disinformation policy can actually endanger democratic legitimacy.

Future research is needed to develop first-hand data on the association between disinformation content generated by AI and attitudes, behaviours, and the occurrence of offline events, in a privacy- and civil-liberties-protecting manner. Testing labels, provenance cues, friction, and prebunking may be possible through field experiments across languages and platforms. For cross-national longitudinal studies, V-Dem and platform transparency data should be used together with event datasets with granular detail on incidents (as offered in V-Dem) such as economic, contestation, and repression data from either GDELT or ACLED (or another one of the ‘GTD-style’ datasets). Multilingual research is particularly pressing because deepfake datasets are only available in English and do not capture all high-risk environments. Lastly, there is the question of whether the new governance arrangements, such as the DSA and AI Act, can be shown to have a material impact in reducing synthetic-media harms while not impeding any legitimate political speech.

Conflict of Interest

The authors showed no conflict of interest.

Funding

The authors did not mention any funding for this research.

References

AI Index Steering Committee. (2026). *Artificial Intelligence Index Report 2026*. Stanford Institute for Human-Centered Artificial Intelligence.

Arayankalam, J., & Krishnan, S. (2021). Relating foreign disinformation through social media, domestic online media fractionalization, government's control over cyberspace, and social media-induced offline violence: Insights from the agenda-building theoretical perspective. *Technological Forecasting and Social Change, 166*, 120661. https://doi.org/10.1016/j.techfore.2021.120661

Armaly, M. T., Buckley, D. T., & Enders, A. M. (2022). Christian nationalism and political violence: Victimhood, racial identity, conspiracy, and support for the Capitol attacks. *Political Behaviour, 44*(2), 937–960. https://doi.org/10.1007/s11109-021-09758-y

Armaly, M. T., & Enders, A. M. (2024). Who supports political violence? *Perspectives on Politics, 22*(2), 427–444. https://doi.org/10.1017/S1537592722001086

Bagozzi, B. E., Brandt, P. T., Freeman, J. R., Holmes, J. S., Kim, A., Mendizabal, A. P., & Potz-Nielsen, C. (2019). The prevalence and severity of underreporting bias in machine- and human-coded data. *Political Science Research and Methods, 7*(3), 641–649.

Bailey, D. H., Jung, A. J., Beltz, A. M., Eronen, M. I., Gische, C., Hamaker, E. L., Kording, K. P., Lebel, C., Lindquist, M. A., Moeller, J., Razi, A., Rohrer, J. M., Zhang, B., & Murayama, K. (2024). Causal inference on human behaviour. *Nature Human Behaviour, 8*(8), 1448–1459. https://doi.org/10.1038/s41562-024-01939-z

Bengio, Y., Clare, S., Prunkl, C., Andriushchenko, M., Bucknall, B., Murray, M., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., & others. (2026). *International AI Safety Report 2026*. International AI Safety Report.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalisation in American politics*. Oxford University Press.

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalisation. *Political Psychology, 41*(S1), 3–33. Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. Oxford Internet Institute.

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE.

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), eadh2132.

Centre for Advancing Safety of Machine Intelligence. (2024). *Tracking political deepfakes: New database aims to inform, inspire policy solutions*. Northwestern University.

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences, 120*(4), e2216614120. https://doi.org/10.1073/pnas.2216614120

Chesney, R., & Citron, D. K. (2019). Deepfakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820.

Cinelli, M., Pelicon, A., Mozetić, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports, 11*, 22083.

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D., Angiolillo, F., Bernhard, M., Cornell, A., Fish, M. S., Gastaldi, L., Gjerløw, H., Glynn, A., Hicken, A., Lu¨hrmann, A., Maerz, S. F., Marquardt, K. L., McMann, K. M., Mechkova, V., & others. (2026). *V-Dem Country-*

Year/Country-Date Dataset v16. Varieties of Democracy Project. https://doi.org/10.23696/vdemds26

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do microtargeted deepfakes have real effects on political attitudes? *The International Journal of Press/Politics, 26*(1), 69–91.

Egan, J., Robertson, C., Ross Arguedas, A., Newman, N., Nielsen, R. K., Mukherjee, M., & Fletcher, R. (2026). *Reuters Institute Digital News Report 2026*. Reuters Institute for the Study of Journalism.

Enders, A., Klofstad, C., & Uscinski, J. (2024). The relationship between conspiracy theory beliefs and political violence. *Harvard Kennedy School Misinformation Review, 5*(6). https://doi.org/10.37016/mr-2020-163

European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council on a single market for digital services (Digital Services Act). *Official Journal of the European Union*.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

Feuerriegel, S., DiResta, R., Goldstein, J. A., & Pröllochs, N. (2023). Research can help to tackle AI-generated disinformation. *Nature Human Behaviour, 7*, 1818–1821.

Gallacher, J. D., Heerdink, M. W., & Hewstone, M. (2021). Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media Society, 7*(1), 1–16. https://doi.org/10.1177/2056305120984445

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*(6425), 374–378.

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences, 119*(1), e2110013119.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(1), eaau4586.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A.,

Dimmery, D., Freelon, D., Gentzkow, M., Gonz´alez-Bail´on, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., & Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behaviour in an election campaign? *Science, 381*(6656), 398–404.

Kasperson, R. E., Renn, O., Slovic, P., Brown, H. S., Emel, J., Goble, R., Kasperson, J. X., & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis, 8*(2), 177–187.

Kruglanski, A. W., Gelfand, M. J., B´elanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The psychology of radicalisation and deradicalisation: How significance quest impacts violent extremism. *Political Psychology, 35*(S1), 69–93.

L abuz, M., & Nehring, C. (2024). On the way to deep fake democracy? Deep fakes in election campaigns in 2023. *European Political Science*. Advance online publication.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson,

E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.

Leetaru, K., & Schrodt, P. A. (2013). *GDELT: Global data on events, location, and tone, 1979–2012*. Paper presented at the International Studies Association Annual Convention.

Littrell, S., Klofstad, C., Diekman, A., Funchion, J., Murthi, M., Premaratne, K., Seelig, M., Verdear, D., Wuchty, S., & Uscinski, J. E. (2023). Who knowingly shares false political information online? *Harvard Kennedy School Misinformation Review, 4*(4). https://doi.org/10.37016/mr-2020-121

McCabe, S. D., Ferrari, D., Green, J., Lazer, D. M. J., & Esterling, K. M. (2024). Post-January 6th deplatforming reduced the reach of misinformation on Twitter. *Nature, 630*, 132–140. https://doi.org/10.1038/s41586-024-07524-8

McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalisation: Pathways toward terrorism. *Terrorism and Political Violence, 20*(3), 415–433.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly, 36*(2), 176–187.

Munger, K., & Phillips, J. (2022). Right-wing YouTube: A supply and demand perspective. *The International Journal of Press/Politics, 27*(1), 186–219.

Napoli, P. M. (2019). *Social media and the public interest: Media regulation in the disinformation age*. Columbia University Press.

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. (2018). The effect of extremist violence on hateful speech online. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1). https://doi.org/10.1609/icwsm.v12i1.15040

Postiglione, M., Gortner, I., Fosdick, L., Gao, C., Kraus, S., & Subrahmanian, V. S. (2026). A nonpartisan study of deepfake activity and engagement around the 2024 U.S. presidential election. *Proceedings of the International AAAI Conference on Web and Social Media, 20*(1), 1891–1908. https://doi.org/10.1609/icwsm.v20i1.42728

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8*(34), eabo6254. https://doi.org/10.1126/sciadv.abo6254

Rulis, M. (2025). The influences of misinformation on incidences of politically motivated violence in Europe. *The International Journal of Press/Politics, 30*(4), 1096–1115. https://doi.org/10.1177/19401612241257873

START. (2022). *Global Terrorism Database 1970–2020* [Data file]. National Consortium for the Study of Terrorism and Responses to Terrorism, University of Maryland.

Starbird, K., DiResta, R., & DeButts, M. (2023). Influence and improvisation: Participatory disinformation during the 2020 U.S. election. *Social Media Society, 9*(2), 1–17.

Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media Society, 4*(3), 1–11.

Swire-Thompson, B., & Lazer, D. (2020). Public health and online misinformation: Challenges and recommendations. *Annual Review of Public Health, 41*, 433–451. https://doi.org/10.1146/annurev-publhealth-040119-094127

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.

T¨ornberg, P., & Chueri, J. (2026). When do parties lie? Misinformation and radical-right populism across 26 countries. *The International Journal of Press/Politics, 31*(2), 367–386. https://doi.org/10.1177/19401612241311886

UNESCO. (2023). *Guidelines for the governance of digital platforms: Safeguarding freedom of expression and access to information through a multistakeholder approach*. UNESCO.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media Society, 6*(1), 1–13. https://doi.org/10.1177/2056305120903408

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

Walker, C. P., Schiff, D. S., & Schiff, K. J. (2024). Merging AI incidents research with political misinformation research: Introducing the Political Deepfakes Incidents Database. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(21), 23053–23058. https://doi.org/10.1609/aaai.v38i21.30349

Woolley, S. C., & Howard, P. N. (Eds.). (2018). *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press