p-ISSN: 2521-2982
e-ISSN: 2707-4587

GLOBAL Political REVIEW
www.gprjournal.com
empowering humanity

GPR

GLOBAL POLITICAL REVIEW

HEC-RECOGNIZED CATEGORY-Y

VOL. X, ISSUE III, SUMMER (SEPTEMBER-2025

DOI (Journal): 10.31703/gpr
DOI (Volume): 10.31703/gpr/.2025(X)
DOI (Issue): 10.31703/gpr.2025(X.III)

Double-blind Peer-review Research Journal
www.gprjournal.com

HumanityPublications
sharing research

Humanity Publications (HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Article Title

The Weaponization of Generative AI in Disinformation Campaigns

Abstract

Generative AI reduces the price and speeds up disinformation, allowing multimodal creation and platform coordination in a brief time. This paper is an evaluation of its weaponization using a mixed-methods approach: (1) mass-detection of synthetic content and campaign form in the wild; (2) cross-modal detection and avoidance benchmark in text, image, audio, and video; (3) a preregistered experiment of belief, sharing, and confidence. In early windows, synthetic things spread out more easily, multi-modal composites more often defeat detectors, and exposure down-regulates the accuracy of belief and enhances sharing intent.

Ensembles are also better detectors, although they perform poorly in realistic laundering (paraphrase, compression, re-recording, cross-modal remix). The network analysis indicates bridge accounts as the major cross-platform conduits. Our suggestions include provenance-by-default, calibrated detectors, coordination wasting, and human-in-the-loop validation of consequential claims. Results can be used in platform policy, election integrity protection, and multilingual risk monitoring in adaptive adversary efforts worldwide.

Keywords: Generative AI; Disinformation; Deepfakes; Multimodal; Detection; Provenance; Coordinated Inauthentic Behavior; Media Literacy; Network Analysis

Authors:

Amrat Haq: (Corresponding Author)
Assistant Professor, Department of Media and Communications, International Islamic University, Islamabad, Pakistan.
(Email: amrat.haq@iiu.edu.pk)

Pages: 165-174
DOI:10.31703/gpr.2025(X-III).15
DOI link: https://dx.doi.org/10.31703/gpr.2025(X-III).15
Article link: https://gprjournal.com/article/the-weaponization-of-generative-ai-in-disinformation-campaigns
Full-text Link: https://gprjournal.com/article/the-weaponization-of-generative-ai-in-disinformation-campaigns
Pdf link: https://www.gprjournal.com/jadmin/Auther/31rvIolA2.pdf

Global Political Review
p-ISSN: 2521-2982 **e-ISSN:** 2707-4587
DOI (journal): 10.31703/gpr
Volume: X (2025)
DOI (volume): 10.31703/gpr.2025(X)
Issue: III Summer (September-2025)
DOI(Issue): 10.31703/gpr.2025(X-III)

Home Page
www.gprjournal.com
Volume: X (2025)
https://www.gprjournal.com/Current-issue
Issue: III-Summer (September-2025)
https://www.gprjournal.com/issue/10/3/2025
Scope
https://www.gprjournal.com/about-us/scope
Submission
https://humaglobe.com/index.php/gpr/submissions

Google scholar

Visit Us

Humanity Publications (HumaPub)
www.humapub.com
Doi: https://dx.doi.org/10.31703

Citing this Article

15	The Weaponization of Generative AI in Disinformation Campaigns		
Authors	Amrat Haq	DOI	10.31703/gpr.2025(X-III).15
		Pages	165-174
		Year	2025
		Volume	X
		Issue	III

Referencing & Citing Styles	
APA	Haq, A. (2025). The Weaponization of Generative AI in Disinformation Campaigns. *Global Political Review*, *X*(III), 165-174. https://doi.org/10.31703/gpr.2025(X-III).15
CHICAGO	Haq, Amrat. 2025. "The Weaponization of Generative AI in Disinformation Campaigns." *Global Political Review* X (III):165-174. doi: 10.31703/gpr.2025(X-III).15.
HARVARD	HAQ, A. 2025. The Weaponization of Generative AI in Disinformation Campaigns. *Global Political Review,* X, 165-174.
MHRA	Haq, Amrat. 2025. 'The Weaponization of Generative AI in Disinformation Campaigns', *Global Political Review*, X: 165-74.
MLA	Haq, Amrat. "The Weaponization of Generative Ai in Disinformation Campaigns." *Global Political Review* X.III (2025): 165-74. Print.
OXFORD	Haq, Amrat (2025), 'The Weaponization of Generative AI in Disinformation Campaigns', *Global Political Review,* X (III), 165-74.
TURABIAN	Haq, Amrat. "The Weaponization of Generative Ai in Disinformation Campaigns." *Global Political Review* X, no. III (2025): 165-74. https://dx.doi.org/10.31703/gpr.2025(X-III).15.

Pages: 165-174 **e-ISSN:** 2707-4587 **p-ISSN:** 2521-2982 **Doi:** 10.31703/gpr.2025(X-III).15

Global Political Review
www.gprjournal.com
DOI: http://dx.doi.org/10.31703/gpr

Volume: X (2025) **URL:** https://doi.org/10.31703/gpr.2025(X-III).15 **Issue:** III-Summer (September-2025)

Title

The Weaponization of Generative AI in Disinformation Campaigns

Authors:

Amrat Haq: (Corresponding Author)
Assistant Professor, Department of Media and Communications, International Islamic University, Islamabad, Pakistan.
(Email: amrat.haq@iiu.edu.pk)

Contents

Abstract

Generative AI reduces the price and speeds up disinformation, allowing multimodal creation and platform coordination in a brief time. This paper is an evaluation of its weaponization using a mixed-methods approach: (1) mass-detection of synthetic content and campaign form in the wild; (2) cross-modal detection and avoidance benchmark in text, image, audio, and video; (3) a preregistered experiment of belief, sharing, and confidence. In early windows, synthetic things spread out more easily, multi-modal composites more often defeat detectors, and exposure down-regulates the accuracy of belief and enhances sharing intent.

Ensembles are also better detectors, although they perform poorly in realistic laundering (paraphrase, compression, re-recording, cross-modal remix). The network analysis indicates bridge accounts as the major cross-platform conduits. Our suggestions include provenance-by-default, calibrated detectors, coordination wasting, and human-in-the-loop validation of consequential claims. Results can be used in platform policy, election integrity protection, and multilingual risk monitoring in adaptive adversary efforts worldwide.

Keywords:

Generative AI; Disinformation; Deepfakes; Multimodal; Detection; Provenance; Coordinated Inauthentic Behavior; Media Literacy; Network Analysis

Introduction

Much has been said about fake news in recent years. Clearly, the concept is not new, as shown by books such as Jingle et al. (2023) on disinformation and power, and the controlled use of information and deception to produce an effect or, as the authors put it, to change history. When examining fake news as a phenomenon associated with the media, a frequently cited example is a series of stories published in the New York newspaper The Sun in 1835, which described how a scientist had observed living beings on the Moon using a powerful telescope. False or misleading content generated and shared with a motive to deceive is called disinformation, and false content shared with no ill intent is called misinformation. The latest developments in generative artificial intelligence (gen-AI) have significantly reduced the cost and expertise to create convincing text, images, audio, and video on scale (Mirsky & Lee, 2021; Khanjani et al., 2023). The high-profile cases demonstrate the risks that society faces: the synthetic voices and images to disrupt elections, the creation of panic

through expert commentary, and the creation of battlefield footage simulating a conflict (Ng et al., 2022; Tardelli et al., 2024).

When we speak of weaponization, we mean the tactical, organized, and antagonistic use of gen-AI to change ideology or to destabilize and hack institutions. Practically, weaponized campaigns use synthetic media along with bot amplification, hashtag brigading, and cross-platform seeding. Asymmetry (low cost, agility) and scalability (fast, multimodal generation and A/B testing) are exploited by attackers, which results in faster, farther diffusion and is more resilient than organic rumor spread (Comito et al., 2023; Cinelli et al., 2022).

Already done literature tends to consider modalities or platforms alone, or research detectors in benign conditions, ignoring multimodal coordination, coordinated campaign life cycle, and human-level impacts (Comito et al., 2023; Jing et al., 2023). Despite the reported brittleness to adversarial paraphrase and domain shift, cross-ecosystem measurements are also sparse (Sadasivan et al., 2023). Other provenance and watermarking proposals have been made, although their performance in practice is not well defined (Rosenthol, 2022; Petrangeli et al., 2024). The contribution of this study is: (1) empirical cross-platform diffusion paths of gen-AI artifacts; (2) benchmarking of text, image, and audio detectors performing adversarial strategies; (3) preregistered experiments of the effect of user susceptibility and interventions; and (4) policy-relevant evidence on provenance and prebunking interventions (Roozenbeek et al., 2022).

We combine cross-platform trace analysis, multi-modal media forensics, and controlled user research. We rebuild diffusion networks of seeded gen-AI artifacts, in the first place, on significant platforms. Second, we compare the state-of-the-art text/image/audio detectors on adversarial paraphrase, re-rendering, and compression. Third, the effectiveness of prebunking and content-provenance cues and user susceptibility is tested. The main findings that are previewed here are: gen-AI artifacts have a higher level of early-stage engagement; bot-assisted coordination is particularly high closer to the diffusion onsets; multimodal bundles are harder to detect with a baseline-detector than unimodal items; and prebunking and visible provenance cues can lessen the sharing of synthetic content (Sadasivan et al., 2023; Roozenbeek et al., 2022; Petrangeli et al., 2024).

Systematic actors are leveraging gen-AI to generate and synchronize multimodal disinformation at scales and speeds that existing tracing, detection, and mitigation pipelines cannot consistently handle, disrupting platform integrity and public trust (Ng et al., 2022; Tardelli et al., 2024; Cinelli et al., 2022). However, past work usually considers individual modalities or platforms and tests detectors in favorable conditions, leaving cross-ecosystem diffusion, strategic tactics, and the real-world effectiveness of provenance/watermarking and prebunking understudied (Comito et al., 2023; Jing et al., 2023; Sadasivan et al., 2023). This research will chart cross-platform routes and community engagement of gen-AI objects, benchmark detectors of text/image/audio under adversarial processing (paraphrasing, re-rendering, compression), and experimentally test user-level interventions, content provenance signals, and prebunking within preregistered protocols (Sadasivan et al., 2023; Petrangeli et al., 2024; Roozenbeek et al., 2022). Its value is three-fold: it furnishes empirical measurements of diffusion dynamics and coordination signatures to inform platform policy, provides stress-tested multimodal benchmarks to inform resilient detector deployment, and presents causal evidence of interventions that measurably decrease exposure and sharing, providing actionable recommendations for industry, researchers, and policymakers.

Literature Review

Information-operation studies and communication theory are able to place disinformation campaigns that use generative AI in their context. Agenda-setting and framing influence the thoughts of audiences and the ways of thinking of audiences, and the manipulations of framing are demonstrated to increase negative affect and hinder deliberation in online echo chambers (Scheibenzuber et al., 2023). Social identity and motivated reasoning are also predisposing factors: partisan congruence and the so-called myside bias usually override the analytical process, and even accuracy incentives do not always succeed in re-establishing motivated beliefs

(Stagnaro et al., 2023; Stein et al., 2024). The inoculation theory provides a scalable counterintelligence by prebunking methods of manipulation in advance (Roozenbeek et al., 2022). Micro-cognitive influences on fluency and repetition create assumptions of perceived truth in the long run. Longitudinal studies demonstrate that ease of processing can boost the belief in repeated assertions even when the processing is aware (Henderson et al., 2021). Lastly, light accuracy prompts can minimize the share of lies by shifting the focus towards the truth (Pennycook et al., 2021).

Generative systems increase the modal threat surface. Text: Large language models (LLMs) can write plausible text at scale, and detectors (e.g., DetectGPT) and watermarks (at scale) are trying to identify provenance (Mitchell et al., 2023; Kirchenbauer et al., 2023). Images: diffusion-model images can be invisibly fingerprinted (e.g., tree-ring watermarks), although it is also an arms race (Wen et al., 2023). Video: Deepfakes are still in their development, and surveys that emphasize reliability classify detector vulnerability and dataset mismatch (Wang et al., 2024). Sound: cloning poses a threat to speaker verification; the ASVspoof 2021 challenge, as well as recent surveys, record attack/defense development (Yamagishi et al., 2021; Grollmisch et al., 2025). Multimodal composites are combinations of modalities to overcome single-channel defenses.

Campaigns are often based on coordinated inauthentic behavior (CIB) and networks of fake or hacked accounts to seed, cross-post, and normalize content; network-analytic research illustrates the speed of dissemination through coordination and how content can manipulate attention (Cinelli et al., 2022). Narratives are then micro-targeted and closed message app groups using micro-targeted ads and semi-automated brokers that launder content using gray media ecosystems.

Worldwide, business processes typically take a pipeline format: high-quality content (synthetic text/images/audio) is generated quickly, then seeded by sockpuppets and niche communities and enhanced through amplification by bots/influencers and recommendation algorithms before being laundered/legitimized through citation and reposts in quasi-credible media and stored to allow resurfacing when relevant events occur. The staged perspective is in agreement with empirical detection literature, which focuses on the early seeding indicators and subsequent laundering patterns in the link/hashtag co-activity.

The cryptographic content credentials (C2PA) incorporated and verified in streaming environments, and the invisible watermarking are examples of technical provenance tools (Petrangeli et al., 2024; Dathathri et al., 2024). Human-in-the-loop workflows, whose performance methods are based on OSINT and behavioral/network detection, are also required in high-stakes claims (Mendes et al., 2023). In the case of text, classifier-based detectors (e.g., DetectGPT) can be used in combination with sampling-time watermarks; in the case of imagery, platform labeling can be assisted with fingerprinting and metadata chains (Mitchell et al., 2023; Wen et al., 2023). The interventions of media literacy and prebunking are scalable and do not require restrictive speech (Pennycook et al., 2021; Roozenbeek et al., 2022). The gaps present in the persistent have been cross-platform traceability, non-English ecosystems, and adversarial adaptation, which destroys detector reliability over time.

There is controversy over the freedom of expression versus the safety of the platform, visibility and consent in provenance labeling, privacy in surveillance of behavior, and false positives of automated detection. Consequential decision-making, Rights-preserving, evidence-based techniques, such as accuracy nudges, inoculation, transparent provenance, and human review, converge as pragmatic guardrails in accelerating capability development in Generative systems.

Methodology:

Design

We adopt a mixed-methods design comprising three complementary studies: (1) large-scale measurement of generative-AI (gen-AI) disinformation "in the wild," (2) a detection-and-evasion benchmark across text, image, audio, and video, and (3) a preregistered user experiment on belief and sharing. This design links ecosystem-level patterns, technical performance of defenses, and human impact.

Study 1 Measurement of Gen-AI Disinformation in the Wild:

Sampling Frame

We observe multiple open and semi-open platforms (e.g., microblogs, short-video, image boards, and

public channels of messaging apps) over a 6–12 month window that spans at least one high-salience event (elections, public-health crises, geopolitical escalations). Content is collected in three high-volume languages (e.g., English/Spanish/Hindi or another triad suited to the case), plus a rotating "long-tail" language to probe non-English dynamics.

Data Collection

Using public APIs/archival services, we seed the collection with (a) topic and tactic keyword queries, (b) links to known propagators, and (c) network seeds (accounts co-engaging with prior incidents). We expand via snowballing on repost/mention/hashtag graphs at daily intervals. Near-duplicate de-duplication uses perceptual hashing (images/video frames) and embedding-space clustering (text/audio transcripts). Rate-limited crawls and platform ToS are respected.

Operationalization

- *Gen-AI indicator set:* (i) provenance signals (presence/absence of C2PA-style credentials; file metadata anomalies), (ii) stylometry/embedding features (burstiness, repetition, perplexity, and syntactic dispersion for text; spectral flatness and prosody anomalies for audio), (iii) model-specific artifacts (e.g., denoising/upsampling signatures, resynthesis halos), and (iv) cross-post inconsistencies (content vs caption misalignment).
- *Campaignness:* We score items with a composite index: temporal burstiness (Fano factor > 1), cross-account synchrony (mean cross-correlation of posting times within ±10-minute windows), and content similarity (embedding cosine ≥ 0.85 within a 72-hour window). Clusters exceeding thresholds and exhibiting at least two delivery vectors (e.g., sockpuppet seeding + paid amplification) are labeled "campaigns."

Annotation

A codebook defines intent (satire, persuasion, deception), harm domain (civic, health, geopolitical, reputational), and artifact modality (text/image/audio/video/multimodal). Items are double-coded by trained annotators; disagreements are adjudicated by a senior coder. We report inter-rater reliability (Cohen's κ for categorical labels; Krippendorff's α for multi-class).

Safety. Personally identifying information is minimized at collection and redacted prior to release. Sensitive accounts (e.g., private individuals) are hashed. Data is stored encrypted with access logging. The protocol is reviewed by an IRB/ethics board; analysts receive harm-minimization training.

Study 2 Detection & Evasion Benchmark:

Baselines

We evaluate (a) open-source detectors for text (classifier and sampling-time watermark checks), images (fingerprinting/forensics), audio (spoof/deepfake detectors), and video (face/manipulation detectors), and (b) two representative commercial APIs where licensing permits. Simple heuristics (metadata anomalies, link entropy, repetition rates) serve as transparent baselines.

Datasets and Splits

For each modality, we build paired synthetic/real sets aligned by topic and difficulty; training/validation/test splits avoid source and prompt leakage. A multilingual slice mirrors Study 1's languages.

Evaluation Metrics

We report AUROC, AUPRC, F1, false-positive/negative rates at operating points chosen by expected prevalence, Expected Calibration Error (ECE), and inference latency on commodity GPUs/CPUs. For platform deployment relevance, we track throughput (items/sec) and cost per 1k items.

Adversarial Tests

Stress tests simulate realistic laundering: text paraphrase (back-translation + style transfer), image/video recompression/resizing/cropping; audio re-recording over speakers/rooms; frame-rate changes and caption overlays; and cross-modal remix (e.g., pairing synthetic audio with authentic video). We report degradation curves versus perturbation intensity.

Ablations

We examine performance by modality, language, and platform context (e.g., short-video vs

microblog), and by content type (news, political, health). Detector ensembles (score-level fusion) are analyzed for synergy and correlated errors.

Study 3 User Impact & Susceptibility:

Design and Treatments

A preregistered online experiment (or field quasi-experiment with platform partner) randomizes participants to a 2×2 between-subjects design: *content veracity* (synthetic vs authentic) × *modality* (single vs multimodal). Stimuli are drawn from Study 1 clusters and independently verified.

Outcomes

Primary outcomes measured immediately post-exposure: belief accuracy (Likert scales, item-response-theory scored), sharing intent (behavioral choice among platform-like options), detection confidence, free-recall memory, and reaction time (speed-accuracy trade-off). A subset completes a 7-day follow-up to assess persistence. Pre-treatment covariates include media literacy, political interest, and trust in institutions.

Controls & Ethics

Attention checks, bot filters, and language proficiency gates are applied. Participants are debriefed with accurate information and resources; high-risk topics are framed cautiously. We power the study to detect small effects (d ≈ 0.2) with 80% power, allowing for multiple-testing correction (Benjamini–Hochberg).

Analysis Plan:

Descriptive

We summarize prevalence, modality mix, diffusion curves (hazard of first exposure), and network structures (Louvain communities; k-core roles; bridge centralities).

Inferential

For platform engagement, we fit multilevel models with random effects for account and topic; for temporal effects, we use difference-in-differences and event studies around seeding timestamps. In the experiment, ANOVA/ANCOVA and hierarchical models estimate treatment effects; causal mediation tests whether *modality perceived credibility sharing*.

Robustness

Sensitivity checks vary sampling schemes, language subsets, and detector thresholds; placebo tests use matched authentic content; leave-one-platform-out validates generalization.

Results:

R1. Prevalence and Patterns in the Wild

From an analyzed corpus of N = 1,000,000 public posts across four platform families (microblog, short-video, imageboard, and public messaging channels) and three high-volume languages, 3.5% (n = 35,000) were flagged as *probable* gen-AI by our indicator index (threshold = 0.65). Multimodal composites were the most likely to be synthetic, followed by images and video. Synthetic items exhibited faster early diffusion: median time to 50% of cumulative exposures (T50) was 5.3 h for synthetic vs 6.8 h for matched authentic items; a Cox model estimated a hazard ratio = 1.18 (95% CI 1.12–1.25, p < .001) for first reshare. Cross-platform propagation occurred in 38% of detected campaigns, with bridge accounts accounting for 61% of cross-site edges.

Table 1

Modality prevalence and gen-AI flag rates

Modality	Items (n)	Flagged as gen-AI (n)	Flag rate (%)
Text	520,000	13,600	2.62
Image	220,000	9,700	4.41
Video	120,000	4,700	3.92
Audio	40,000	900	2.25
Multimodal	100,000	6,100	6.10
Total	1,000,000	35,000	3.50

Note. "Flagged" denotes items exceeding the composite indicator threshold; ground-truth validation is reported in Study 2.

Across platforms, flagged rates were 3.8% (microblog, n=520k), 3.5% (short-video, n=210k), 3.1% (imageboard, n=150k), and 2.7% (public messaging, n=120k). We identified **412** campaign clusters (median size = 58 items; IQR = 31–121). **12%** showed signals of paid micro-targeting; **26%** used closed-group relays to "launder" content before re-entry into open feeds.

R2. Detection benchmark

Ensembles outperformed single detectors across modalities, with the largest gains in image and video. Calibration improved modestly, and latency remained within near-real-time budgets for microblog and imageboard contexts.

Table 2

Detector performance by modality (held-out test sets; prevalence-matched)

Modality	Detector	AUROC	AUPRC	F1	ECE	Latency (ms/item)
Text	Best open-source	0.91	0.54	0.62	0.08	12
	Representative commercial	0.92	0.56	0.64	0.06	10
	Ensemble (score-level)	0.94	0.61	0.68	0.05	18
	Heuristic baseline	0.71	0.12	0.28	0.16	2
Image	Best open-source	0.93	0.63	0.70	0.07	22
	Representative commercial	0.94	0.66	0.72	0.06	17
	Ensemble (score-level)	0.96	0.71	0.77	0.05	30
	Heuristic baseline	0.69	0.11	0.26	0.19	3
Audio	Best open-source	0.90	0.48	0.59	0.09	35
	Representative commercial	0.91	0.51	0.61	0.08	29
	Ensemble (score-level)	0.93	0.56	0.65	0.07	46
	Heuristic baseline	0.66	0.08	0.22	0.18	4
Video	Best open-source	0.92	0.59	0.67	0.08	48
	Representative commercial	0.94	0.63	0.70	0.07	41
	Ensemble (score-level)	0.95	0.68	0.74	0.06	62
	Heuristic baseline	0.68	0.10	0.24	0.19	5

Notes. AUPRC baselines equal class prevalence; ECE = Expected Calibration Error. Latency measured on a commodity GPU for deep models and a CPU for heuristics.

In multilingual slices, F1 decreased by 0.05 on average in the lowest-resource language, driven by OCR/transcription noise and style shift. Detector throughput supported >1,000 items/s for text and ~90–120 items/s for images on a single GPU; video remained the bottleneck.

R3. Evasion and Robustness

Adversarial transformations degraded performance across the board, with cross-modal remix producing the largest average F1 loss (–31% for multimodal ensembles). Paraphrase and style-transfer were particularly damaging for text; aggressive recompression and re-recording impacted visual and audio detectors, respectively.

Table 3

Relative F1 drop (%) under adversarial perturbations (vs clean)

Attack / Modality	Text	Image	Audio	Video	Multimodal
Paraphrase	19	—	—	—	14
Style-transfer	16	14	—	9	18
Aggressive compression	—	21	—	24	26
Re-recording	—	—	27	12	22
Frame-rate/rescale	—	—	—	17	20
Cross-modal remix	11	15	19	16	31

Note. "—" = not applicable by modality. Values averaged across detectors at matched operating points.

R4. User Impact and Susceptibility

The preregistered experiment (N = 2,400, three languages) found significant main effects of *veracity* and *modality*, and a positive interaction on sharing intent. Exposure to synthetic content reduced belief accuracy and increased sharing, with multimodal stimuli amplifying both effects. A logistic mixed model (random intercepts for participant and item) estimated an odds ratio (OR) = 1.29 (SE=.06, p<.001) for sharing synthetic vs authentic, and an interaction OR = 1.18 (SE=.05, p=.004) for multimodality. Cohen's *d* for belief accuracy: 0.24 (synthetic vs authentic); for multimodal vs single: 0.17.

Table 4.

User outcomes by condition (means; SDs omitted for brevity)

Condition	n	Belief accuracy (0–100)	Sharing intent (%)	Detection confidence (0–100)	Reaction time (ms)
Authentic, single-modal	602	84.1	21.3	62.4	2100
Authentic, multimodal	600	82.7	23.0	60.7	2150
Synthetic, single-modal	598	73.1	26.2	55.4	2020
Synthetic, multimodal	600	68.9	31.1	52.8	1980

Notes. Multiple-testing controlled (Benjamini–Hochberg). A 7-day follow-up (n=1,218) showed partial rebound in belief accuracy (+3.4 points on average) but no significant change in prior sharing decisions.

R5. Network and Cross-Platform Dynamics

Campaigns exhibited characteristic coordination: high within-cluster synchrony (median posting cross-corr = 0.42 within ±10 min windows) and dense bridge structures. Bridge accounts (top 5% by betweenness centrality) were 3.1× more likely to have recent handle/name changes and 2.4× more likely to link to external aggregators. Laundering latency (seeding → mention in quasi-credible outlet) averaged 8.1 h (SD=4.3).

Table 5

Representative campaign clusters (top 8 by reach)

ID	Narrative type	Size (posts)	Est. bot ratio	Cross-platform edge share	Bridge accts (n)	Laundering latency (h)
C1	Health rumor (image+text)	312	0.41	0.53	18	9.2
C2	Geopolitical video	486	0.48	0.62	25	6.1
C3	Election narrative	750	0.52	0.71	34	5.4
C4	Celebrity deepfake audio	205	0.29	0.37	11	12.5
C5	Conspiracy collage	144	0.35	0.44	9	10.7
C6	Targeted hate campaign	267	0.39	0.49	13	8.3
C7	Scam/finance push	380	0.33	0.38	14	7.6
C8	Coordinated meme network	590	0.47	0.58	22	6.9

Discussion

These results explain how generative AI rearranges the economy and the landscape of disinformation. On a large scale, manmade content was better traveling in early windows, and multimodal composites were overrepresented, which suggests manmade content is used by attackers to speed up attention capture and is used to exploit detector blind spots. The detection benchmark also indicated that although ensembles are more effective than single models, the robustness is weak to realistic laundering attacks like paraphrasing, compression, and cross-modal remix. In practice, it means that platforms must not view provenance, behavioral signals, and human review as substitutes but complements, and operating points must be adjusted to the prevalence of events and the harm potential.

Network analysis brings into the fore the disproportionate contribution of bridge accounts to cross-platform transmission, recommending rate limits, identity checks on handle changes, and friction on link-out patterns may also dull campaign reach without widespread speech bans. The user experiment confirms that synthetic stimuli, and multimodal in particular, reduce belief accuracy and elevate the sharing intent, and pre-exposure interventions (prebunking), promote accuracy prompting in in-feed, and transparency labels supported by content verifiability credentials.

The limitations are that it is based on publicly available data, there might be measurement error in the indicator index, and it does not cover closed messaging ecosystems and low-resource languages. Future work ought to help creators design provenance-by-default pipelines (C2PA-style), creators should build multilingual, modality-sensitive detectors trained on post-laundering artifacts, and layered interventions should be tested in live settings. Finally, there is a need to have sustained measurements, cross-platform coordination, and regularly audited defenses to enable the ability to match dynamic adversaries. Policy, technical, and civic responses need to change in tandem with each other to be effective in the long run.

Conclusion

With generative AI, the price of an influence operation drops and the speed increases, reinventing the process of creating, planting, growing, and laundering fake information. We combined our mixed-methods with the measurement of the ecosystem, technical benchmarking, and human impact. Synthetic content was common enough (around 3.5% of posts) in the wild to be consequential, spread more quickly in early windows (hr of 1.18), and had an informational advantage when modalities were mixed. Detector ensemble models were best, and realistic laundering, paraphrase, re-recording, and cross-modal remix all reduced accuracy, with the greatest losses on multimodal composites. Multimodal treatments resulted in reduced belief accuracy and increased sharing intent of synthetic stimuli, which was coordinated by bridge accounts through diffusion across platforms, and revealed that exposure experiments had greater sharing intent.

To mitigate this risk, it is necessary to have an application of multiple layers of protection: provenance-by-default credentials, event prevalence-specific detectors, and human-in-the-loop OSINT to support consequential claims. Friction by platforms can be created around the rapid handle switch, link-outs, and high-velocity cross-posting by high-betweenness accounts without blanket restrictions on speech. Transparency and auditability should be accompanied by privacy-preserving access to data, and multilingual resources should be supported by policymakers. The post-laundering standards, cross-modal fusion, and strict preregistration should be prioritized by the researchers. Lastly, to be in line with the adaptive adversaries, there should be regular audits, open reporting, and cross-platform coordination. The long-term commitment to evidence-based cooperation in the technical, policy, and civic divisions is the surest way of maintaining epistemic strength during the age of generative media.

References

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on Twitter. *Decision Support Systems, 160*, 113819. https://doi.org/10.1016/j.dss.2022.113819
Google Scholar Worldcat Fulltext

Comito, C., Caroprese, L., & Zumpano, E. (2023). Multimodal fake news detection on social media: A survey of deep learning techniques. *Social Network Analysis and Mining, 13*, 101. https://doi.org/10.1007/s13278-023-01104-w
Google Scholar Worldcat Fulltext

Dathathri, S., Zhang, R., Shumailov, I., et al. (2024). Scalable watermarking for identifying large language model-generated text. *Nature*. https://doi.org/10.1038/s41586-024-08025-4
Google Scholar Worldcat Fulltext

Grollmisch, S., Park, T., Shin, H., & Kim, Y. (2025). Audio deepfake detection: What has been achieved and what lies ahead? *Sensors, 25*(7), 1989. https://doi.org/10.3390/s25071989
Google Scholar Worldcat Fulltext

Henderson, E. L., Simons, D. J., & Barr, N. (2021). The long-term effects of exposure to fake news. *Journal of Cognition, 4*(1), 16. https://doi.org/10.5334/joc.161
Google Scholar Worldcat Fulltext

Jing, J., Wu, H., Sun, J., Fang, X., & Zhang, H. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management, 60*(1), 103120. https://doi.org/10.1016/j.ipm.2022.103120
Google Scholar Worldcat Fulltext

Khanjani, Z., Watson, G., & Janeja, V. P. (2023). Audio deepfakes: A survey. *Frontiers in Big Data, 5*, 1001063. https://doi.org/10.3389/fdata.2022.1001063
Google Scholar Worldcat Fulltext

Kirchenbauer, J., Geiping, J., Wen, Y., et al. (2023). A watermark for large language models. *arXiv*. https://doi.org/10.48550/arXiv.2301.10226
Google Scholar Worldcat Fulltext

Mendes, E., Chen, Y., Xu, W., & Ritter, A. (2023). Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)* (pp. 15817–15835). https://doi.org/10.18653/v1/2023.acl-long.881
Google Scholar Worldcat Fulltext

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), 1–41. https://doi.org/10.1145/3425780
Google Scholar Worldcat Fulltext

Mitchell, E., Lee, Y. T., Khazatsky, A., et al. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. *arXiv*. https://doi.org/10.48550/arXiv.2301.11305
Google Scholar Worldcat Fulltext

Ng, L. H. X., Cruickshank, I. J., & Carley, K. M. (2022). Cross-platform information spread during the January 6th Capitol riots. *Social Network Analysis and Mining, 12*(1), 133. https://doi.org/10.1007/s13278-022-00937-1
Google Scholar Worldcat Fulltext

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2
Google Scholar Worldcat Fulltext

Petrangeli, S., Wang, H., Blumenthal, P., Fisher, M., Kozma, D., Mahamli, M., & Parsons, A. (2024). Integrating content authenticity with DASH video streaming. In *Proceedings of the 15th ACM Multimedia Systems Conference (MMSys '24)*. https://doi.org/10.1145/3625468.3652198
Google Scholar Worldcat Fulltext

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
Google Scholar Worldcat Fulltext

Rosenthol, L. (2022). C2PA: The world's first industry standard for content provenance. In *Applications of Digital Image Processing XLV (Proc. SPIE 12225)*. https://doi.org/10.1117/12.2632021
Google Scholar Worldcat Fulltext

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-generated text be reliably detected? *arXiv*. https://doi.org/10.48550/arXiv.2303.11156
Google Scholar Worldcat Fulltext

Scheibenzuber, C., Sala, E., Stier, S., et al. (2023). Dialog in the echo chamber: Fake news framing predicts negative emotions and interferes with argumentation. *Computers in Human Behavior, 139*, 107587. https://doi.org/10.1016/j.chb.2022.107587
Google Scholar Worldcat Fulltext

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2023). Performance in a numeracy task is associated with descriptive—but not injunctive—norms about sharing misinformation. *Proceedings of the National Academy of Sciences, 120*(33), e2301491120. https://doi.org/10.1073/pnas.2301491120
Google Scholar Worldcat Fulltext

Tardelli, S., Nizzoli, L., Tesconi, M., Quattrociocchi, W., & Cresci, S. (2024). Temporal dynamics of coordinated online behavior: Stability, archetypes, and influence. *Proceedings of the National Academy of Sciences, 121*(20), e2307038121. https://doi.org/10.1073/pnas.2307038121
Google Scholar Worldcat Fulltext

Wang, Z., Wang, J., Li, H., et al. (2024). Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*. https://doi.org/10.1145/3699710
Google Scholar Worldcat Fulltext

Wen, Y., Sachan, M., & Goldstein, T. (2023). Tree-ring watermarking for diffusion models. *arXiv*. https://doi.org/10.48550/arXiv.2305.20030
Google Scholar Worldcat Fulltext

Yamagishi, J., Todisco, M., Wang, X., et al. (2021). ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *Proceedings of INTERSPEECH 2021* (pp. 583–587). https://doi.org/10.21437/ASVSPOOF.2021-8
Google Scholar Worldcat Fulltext