How Malicious AI Swarms Can Threaten Democracy *

Daniel Thilo Schroeder[1†], Meeyoung Cha[2], Andrea Baronchelli[3], Nick Bostrom[4], Nicholas A. Christakis[5], David Garcia[6], Amit Goldenberg[7], Yara Kyrychenko[8], Kevin Leyton-Brown[9], Nina Lutz[10], Gary Marcus[11], Filippo Menczer[12], Gordon Pennycook[13], David G. Rand[14], Frank Schweitzer[15], Christopher Summerfield[16], Audrey Tang[17], Jay Van Bavel[11], Sander van der Linden[8], Dawn Song[18], & Jonas R. Kunst[19†]

[1]SINTEF, [2]Max Planck Institute for Security and Privacy, [3]City St George's University of London, [4]Macrostrategy Research Initiative, [5]Yale University, [6]University of Konstanz, [7]Harvard Business School, [8]Cambridge University, [9]University of British Columbia, [10]University of Washington, [11]New York University, [12]Indiana University, [13]Cornell University, [14]Massachusetts Institute of Technology, [15]ETH Zürich, [16]University of Oxford, [17]Digital Minister & Cyber Ambassador Taiwan, [18]University of California Berkeley, [19]BI Norwegian Business School

Abstract

Advances in AI portend a new era of sophisticated disinformation operations. While individual AI systems already create convincing—and at times misleading—information, an imminent development is the emergence of *malicious AI swarms*. These systems can coordinate covertly, infiltrate communities, evade traditional detectors, and run continuous A/B tests, with round-the-clock persistence. The result can include fabricated grassroots consensus, fragmented shared reality, mass harassment, voter micro-suppression or mobilization, contamination of AI training data, and erosion of institutional trust. With increasing vulnerabilities in democratic processes worldwide, we urge a three-pronged response: (1) platform-side defenses—always-on swarm-detection dashboards, pre-election high-fidelity swarm-simulation stress-tests, transparency audits, and optional client-side "AI shields" for users; (2) model-side safeguards—standardized persuasion-risk tests, provenance-authenticating passkeys, and watermarking; and (3) system-level oversight—a UN-backed AI Influence Observatory.

***K*eywords** AI Swarms · Artificial Intelligence · Democracy · Disinformation · Influence Operations · Misinformation

The Rise of Automated Influence

Public opinion manipulation has entered a dangerous new phase, amplifying and complicating its traditional foundations in rhetoric and propaganda. Rapid advances in large language models (LLMs) and autonomous agents [1, 2] now let influence campaigns reach unprecedented scale and precision. Leading AI researchers caution that AI could foster or magnify mass manipulation [3], and cross-national studies confirm that generative tools dramatically expand propaganda output without sacrificing credibility [4, 5, 6]. Controlled experiments likewise show most LLMs can inexpensively create election falsehoods that appear more human than real human text [5, 7, 8, 6]. Techniques meant to refine AI reasoning, such as chain-of-thought prompting [9, 10], can just as effectively polish convincing lies. In short, automated, scalable influence operations are reshaping the information landscape and pose significant risks for society.

Enabled by these capabilities, another disruptive shift is emerging: swarms of collaborative, malicious AI agents. Fusing LLM reasoning with multi-agent architectures [1], these systems are capable of coordinating autonomously, infiltrating communities, and fabricating consensus at minimal cost. Where legacy botnets acted like megaphones, repeating one script, AI swarms behave like adaptive conversationalists with thousands of distinct personas that learn from feedback, pivot narratives, and blend seamlessly into real discourse. By mirroring human social dynamics using adaptive tactics, they threaten democratic discourse.

* *Citation*: **Schroeder et al. How Malicious AI Swarms Can Threaten Democracy. 1-8.... DOI:000000/11111.**

Corresponding author: daniel.t.schroeder@sintef.no; †These authors contributed equally to this work;
The authors are listed in alphabetical order by last name, starting from the third and ending with the second-to-last.

Urgent action is needed before such systems become fully entrenched in our global information ecosystem, threatening the integrity of democratic processes worldwide. This policy brief situates the emerging AI swarm era within the broader evolution of digital influence operations, diagnoses its unprecedented risks, and sets out a policy agenda.

Evolution of Information Operations

Information operations long predate social media: Cold-War propaganda ranged from Soviet claims that AIDS was a US bioweapon to documented covert U.S. efforts against Latin-American leaders. Online manipulation then accelerated in the 2000s–2010s, targeting Brexit, the 2016 US election, and campaigns in Brazil and the Philippines, and even fueling violence such as the Rohingya genocide [11, 12, 13, 14, 15, 16]. Field experiments still find modest, if any, attitude shifts [17, 18], yet sustained state and private investment, along with documented links between hate campaigns and offline harm [19], ensure tactics will keep evolving.

The Russian Internet Research Agency's 2016 Twitter operation shows the limits of manual botnets: one percent of users saw 70% of its content, with no measurable effects on opinions or turnout [20, 17, 18]. Human labor, blunt messaging, and slow iteration capped its reach. But large LLMs remove those caps. Readily available and frequently jailbreak-prone [21, 22], LLMs can now generate persuasive, tailored text at scale and have shifted deeply held beliefs in laboratory settings [23, 24]. Open-source releases by firms such as Meta and DeepSeek further lower access barriers.

Consequently, AI-supported election interference is no longer hypothetical. Taiwan's and India's 2024 campaigns saw AI-generated deepfakes, and entire fabricated news outlets now steer debates on topics from climate policy to foreign aid [25]. Absent guardrails, LLM-driven swarms could transform sporadic mis- and disinformation into persistent, adaptive manipulation of democratic discourse.

Swarm Capabilities

Recent breakthroughs in multi-agent systems (MAS) (i.e., computational environments where multiple AI agents interact, communicate, and collaborate to solve complex problems) have fused LLM reasoning with agentic memory, planning, and communication [26, 27, 28, 29]. Five technical advances now matter most for influence operations (see Figure 1).

A first capability is the shift from central command to a fluid swarm that coordinates in real time. A single adversary can operate thousands of AI personas, scheduling content and updating narrative frames instantaneously across an entire fleet of agents. Local adaptation plus periodic hub-sync blurs the line between command-and-control and emergent "hive" behavior. If these agent swarms evolve further into loosely governed "societies," capable of internal norm formation and division of labor, the challenge shifts from tracing commands to understanding emergent group cognition [30, 31, 32, 29]. This evolution creates profound defensive challenges. Rather than simply tracing and blocking command sources, defenders must confront unpredictable collective behaviors, where new biases may emerge during social coordination that are not detectable at the individual agent level. More concerning, these societies may experience spontaneous or adversarially-induced norm shifts, abandoning their original engineered constraints for unpredictable new behavioral patterns through critical mass dynamics [33].

Second, agents now can map social graphs and slip into vulnerable communities with tailored appeals. Malicious agent swarm tactics include infiltrating vulnerable online communities to get humans to follow them and generating high volumes of content that appeals to those communities [34]. AI agents can employ systems that map social graphs at scale, identify key communities and beliefs, and track trending topics [35]. This can also work in a decentralized fashion while still having global, network–wide efficacy [36]. Equipped with such capabilities, agent swarms can position themselves for maximum impact and tailor messages to the beliefs and cultural cues of each community, enabling more precise targeting than previous botnet approaches.

Third, human-level mimicry lets swarms evade detectors that once caught copy-paste bots. Methods for detecting coordinated inauthentic behavior, such as influence operations, generally rely on activity patterns that are suspiciously similar between accounts and therefore statistically unlikely to result from independent behaviors [37, 38]. High-fidelity photorealistic avatars, context-aware slang, and heterogeneous posting rhythms circumvent the tell-tale synchrony that older detectors flag [39, 40].

Fourth, continuous A/B testing refines messages at machine speed. In the near term, human operators will still steer agent swarms while letting the agents perform limited in-context adaptation. However, over time, the balance is likely to tilt toward self-optimizing swarms that harvest their own reward signals: real-time engagement data, structural cues from recommender systems, or even natural-language critiques [41]. Even these systems cannot improve without such

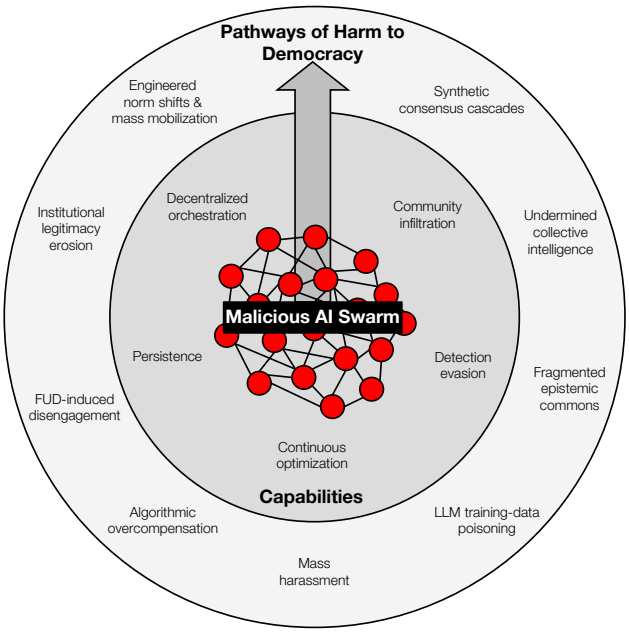

Figure 1: **From Malicious AI Swarm to Democratic Harm.** Concentric-circle schematic linking threat mechanics to civic outcomes. At the core, a Malicious AI Swarm supplies at least five key capabilities—decentralized orchestration, community infiltration, detection evasion, continuous optimization, and persistence (inner ring). These capabilities radiate outward into at least nine democratic-impact pathways, including synthetic consensus cascades, undermined collective intelligence, fragmented epistemic commons, LLM training-data poisoning, mass harassment, algorithmic overcompensation, FUD-induced disengagement, institutional legitimacy erosion, and engineered norm shifts and mass mobilization (outer ring). The upward arrow signals the causal flow from inner mechanics to systemic harm. Note that this flow is likely reciprocal, as impaired democratic functioning increases a society's vulnerability to malicious AI swarms. *FUD = Fear, Uncertainty, Doubt.*

measurable feedback. Given sufficient signals, though, swarms can run millions of micro-A/B tests, propagate the winning variants at machine speed, and iterate far faster than human campaigners.

Finally, round-the-clock presence turns influence into a long-term, low-friction infrastructure. Unlike transient influence operations, agent swarms can persist for protracted durations, embedding themselves within communities over long timescales and gradually shifting discourse. This persistent presence amplifies the effectiveness of other mechanisms described above. In the realm of cognitive warfare, AI's infinite, expendable machine time becomes a strategic weapon against the finite, emotionally taxing efforts of human engagement.

Pathways of Harms to Democracy

The unprecedented capabilities of AI swarm-driven influence campaigns pose profound threats to democracy by shaping public opinion formation [42, 43]. These capabilities convert into at least five cascading harms (see Figure 1).

The immediate effect is a synthetic consensus that exploits social-proof heuristics. Swarms may seed harmonized narratives across disparate niches, creating the illusion of majority agreement within days. Citizens update opinions based on perceived peer norms, not evidence. This chorus of seemingly independent voices creates a mirage of grassroots consensus with enhanced speed and persuasiveness. The result is a deeply embedded manipulation strategy that allows operators to nudge public discourse almost invisibly over extended periods of time.

This coordinated chorus also erodes the independence that collective intelligence requires, and which democracies rely on. While the previous effect exploits the influence of social norms, this second impact directly undermines human cognitive information processing systems. Specifically, the "wisdom of crowds" phenomenon, where aggregated judgments outperform individual experts, depends critically on independence between judgments [44, 45, 46]. When AI agents in a swarm coordinate to simulate diverse voices across platforms–appearing as distinct citizens, experts, and community leaders while actually representing a singular strategic objective–citizens can overestimate the informational value of this artificial consensus.

Collaborating AI agents can tailor streams of misleading information to each sub-community's linguistic, cultural, and emotional markers, weaving partially overlapping but strategically segmented realities. Unlike passive "filter bubbles," these engineered realities are designed to keep groups apart except when not desired, making cross-cleavage consensus increasingly unfeasible. Once initiated, such streams can also spread via social contagion, with the effect of bots potentially cascading beyond the people to whom the bots are directly connected, to two or three degrees of separation, as is often empirically the case [47].

By flooding the web with fabricated chatter, agent swarms can contaminate the data from which future models learn. Malicious operators spin up faux publics that flood the web. LLMs then ingest this chatter; at the next retraining cycle, fabricated narratives calcify inside the model weights [48]. Thus, AI swarms can rig the epistemic substrate on which future deliberation and future AI tools will rely, compounding democratic erosion.

Separate from fragmentation, swarms can at a low cost unleash coordinated synthetic harassment campaigns (i.e., "synthetic shitstorms") that relentlessly target politicians, dissidents, whistleblowers, journalists, and their family and social networks with overwhelming, psychologically tailored abuse. Unlike conventional trolling, these operations appear as spontaneous public outrage while actually representing orchestrated action by thousands of AI personas adapting dynamically to target responses. By the time monitoring teams distinguish these AI campaigns from organic criticism, targets may have already withdrawn from public life, delivering significant victories for campaign operators while systematically excluding critical voices from democratic discourse.

Generally, as trust collapses, fear-uncertainty-doubt (FUD) can drive users into gated channels and silence. When citizens realize that vast portions of online speech may be AI-generated, baseline trust in social platforms and their users could sharply decline. Some threat actors may even welcome their synthetic interventions being exposed, reasoning that the mere revelation of manipulation can sow as much confusion as successful deception. Compounding this, genuine users will inevitably be misidentified as bots, weaponizing false accusations to discredit individuals and intensify FUD. This "epistemic vertigo" may mesh with a flood of low-cost LLM spam that overwhelms timelines, making genuine human conversation harder to find. Together, FUD and content saturation could drive users to disengage or retreat into gated channels, shrinking the shared public sphere on which deliberative democracy relies.

Algorithmic over-compensation can then elevate celebrity voices while sidelining ordinary citizens. When timelines flood with AI-authored posts, both ranking algorithms and users may retreat to obvious trust proxies—big follower numbers, blue checks, and legacy clout. Attention may concentrate around celebrities and major brands, while ordinary participants fade from view. The public sphere contracts from many-to-many dialogue to a few-to-many broadcast, a shift that erodes democratic pluralism and encourages further cynicism or migration to closed groups.

Swarms can even tip latent norms into action [49] or dampen conformity, accelerating anti-democratic political action [30]. Rather than occupying central or influential positions, these agents could operate on the periphery of social networks, where early mobilization often begins [50]. Similar strategies can be weaponized for micro-targeted voter suppression or mobilization on behalf of specific parties and causes. Reinforcement-learning agents run thousands of messaging experiments per hour, iteratively adjusting content while classifiers mine engagement metrics and textual responses to infer voting intent and gauge each tactic's success.

Taken to extremes, coordinated doubt may corrode institutional legitimacy and invite "emergency" rule. By coordinating subtle yet increasingly profound doubts about electoral commissions, courts, or statistics bureaus, swarms could corrode procedural trust. As confidence falters, previously unthinkable "emergency" measures (e.g., postponing elections, rejecting certified results) may become palatable, especially when deep fake endorsements from fabricated civic leaders amplify the call [51, 5].

Governance Measures and Technical Defenses

The emergence of sophisticated, collaborative, and increasingly autonomous swarms of AI agents capable of orchestrating influence operations marks a critical juncture. The causal arrow runs both ways: while malicious swarms endanger democratic norms, the quality of democratic governance will in turn shape how potent, or containable, those swarms become. A full shift to continuous detection, simulation, and proactive, defense-oriented AI must therefore be implemented and operational before AI swarm capabilities reach mass availability.

The first line of defense is always-on detection of anomalous coordination, backed by public audits. To defend the public information sphere in real time, platforms and regulators require always-on detectors that scan live traffic for statistically anomalous coordination patterns—the digital fingerprints of inauthentic influence swarms. Here, advanced network analytics methods are needed that can (i) identify emergent clusters by surfacing camouflaged indicators of coordinated activity; and (ii) spot narrative-alignment drifts with streaming topic models and change-point analysis that

flag lock-step shifts in sentiment [52]. Deployments require explicit democratic mandates, independent audits, and public transparency dashboards to guard against misuse [53].

To extend protection all the way to end users, platforms should expose the same risk signals through optional "AI shields"—lightweight browser or operating system modules that run on the client side. A shield labels posts that carry high swarm-likelihood scores, lets users down-rank or hide them, and surfaces short provenance explanations in situ. Because all scoring happens locally, the shield preserves privacy while giving citizens agency over their information diets. Aggregated, anonymized feedback from shields can also flow back into the public dashboard, creating a distributed early-warning mesh that strengthens the central detectors.

Complementing live monitors, high-fidelity simulation can stress-test those detectors before each election. Real-time monitors are effective only when they anticipate the tactics they will face. Because defenders rarely have access to the underlying code or even the decision-making logic of malicious swarms, high-fidelity, agent-based simulation may be the only reliable window into how these systems might behave. AI agents seeded into synthetic networks can replicate a platform's graph structure, content cadence, and recommender logic, yielding traces that recalibrate live detectors. Iterative red-teaming then pits hostile swarms against these defenses, mapping persuasion ceilings and long-term persona strategies, and continuously hardening guardrails.

Where manipulation slips through, calibrated defensive agents can supply water-marked counter-narratives. A proactive defense strategy must harness AI technology itself to counteract manipulative influence operations. As threats evolve, defensive AI agents can strategically disseminate accurate information and pro-democratic counter-narratives, engage constructively with and warn targeted communities, and promote media literacy and resilience at scale [23, 54]. That said, counter-messaging must opt for precision over volume: if defensive agents indiscriminately flood a platform, human voices could disappear in a sea of synthetic content, triggering the very collapse we seek to avert. Defensive AI should therefore intervene only where manipulation is detected, clearly watermark its output, and operate under human-in-the-loop oversight. Such calibrated use of AI can help build trust before crisis points emerge and ensure that credible, evidence-based information sources are embedded within public discourse, keeping online communities resilient without drowning out authentic participation.

Oversight must be democratic and global, anchored by a UN-backed AI Influence Observatory that combines early-warning detection methods with public certification of incidents. Building on proposals for technically proficient supervisory bodies [3] and the UN's nascent AI Advisory Body [55], the Observatory could ingest anonymized coordination signals from participating platforms and model providers, fuse them with crowdsourced reports from civil-society labs, and issue rapid advisories ranking each incident by provenance confidence and projected civic harm. The Observatory should maintain and continually update an open, searchable database of verified influence-operation incidents, allowing researchers, journalists, and election authorities to track patterns and compare response effectiveness across countries in real time. To guarantee both legitimacy and skill, its governing board would mix rotating member-state delegates, independent technologists, data engineers, and civil society watchdogs.

Stronger provenance may help societies harden identity without muting speech as persuasive swarms proliferate [52]. Policymakers may incentivize rapid adoption of passkeys, cryptographic attestations, and federated reputation standards, backed by ongoing R&D to harden them against spoofing [56]. However, "proof-of-human" alone is no panacea: millions lack formal IDs, biometrics raise privacy risks, and verified accounts can be hijacked. Therefore, a key priority should be regulating the resulting commercial market that underpins large-scale manipulation, where private sellers offer services ranging from vanity metrics to coordinated influence operations at remarkably low costs.

Finally, frontier models must undergo standardized persuasion-risk evaluations and publish scores in model cards before release. Companies should disclose within 24 hours whether entities interacting with users are human or automated. Pre-bunking campaigns that educate citizens about AI tactics can build cognitive immunity, but paradoxically, the very alertness they foster can still be exploited by threat actors to deepen FUD. "Prosocial media" architectures on interoperable platforms can offer structural resilience [57]. Governments and tech firms must prioritize AI-safety R&D while funding independent assessments of misuse potential and societal impact.

In the coming years, democracies have a brief but crucial opportunity to pull AI-enabled influence operations back from the brink. If platforms deploy always-on swarm detectors, frontier labs submit models to standardized persuasion stress tests, and governments launch an AI Influence Observatory that publishes open incident telemetry, we can blunt the most destabilizing tactics before critical political future events, without freezing innovation. Doing so demands the same choreography already mastered by AI swarms: rapid iteration, transparent data-sharing, and tight coordination across scientific, civil-society, industry, and election-security teams. Like earlier calls to confront emerging bio-risks, success hinges on acting "collaboratively ... without unnecessarily impeding scientific research," while keeping the public square resilient and accountable [58]. By committing now to principled action, upcoming societal decisions could even become a proving ground—rather than a casualty—of democratic AI governance.

References

[1] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18 (6):186345, 2024. doi: 10.1007/s11704-024-40231-1.

[2] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025. doi: 10.1007/s11432-024-4222-0.

[3] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 384 (6698):842–845, 2024. doi: 10.1126/science.adn0117.

[4] Cheng-Liang Hung, Wen-Chen Fu, Chi-Chen Liu, and Hsiao-Ju Tsai. Ai disinformation attacks and taiwan's responses during the 2024 presidential election. *Graduate Institute of Journalism, National Taiwan University*, 2024.

[5] Brad Smith. Meeting the moment: combating AI deepfakes in elections through today's new tech accord - Microsoft On the Issues — blogs.microsoft.com. `https://blogs.microsoft.com/on-the-issues/2024/02/16/ai-deepfakes-elections-munich-tech-accord/`, 2024. [Accessed 16-05-2025].

[6] Morgan Wack, Carl Ehrett, Darren Linvill, and Patrick Warren. Generative propaganda: Evidence of ai's impact from a state-backed disinformation campaign. *PNAS nexus*, 4(4):pgaf083, 2025. doi: 10.1093/pnasnexus/pgaf083.

[7] Angus R Williams, Liam Burke-Moore, Ryan Sze-Yin Chan, Florence E Enock, Federico Nanni, Tvesha Sippy, Yi-Ling Chung, Evelina Gabasova, Kobi Hackenburg, and Jonathan Bright. Large language models can consistently generate high-quality content for election disinformation operations. *PloS one*, 20(3):e0317421, 2025. doi: 10.1371/journal.pone.0317421.

[8] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. How persuasive is ai-generated propaganda? *PNAS nexus*, 3(2):pgae034, 2024. doi: 10.1093/pnasnexus/pgae034.

[9] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. URL `https://arxiv.org/abs/2501.19393`.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[11] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018. doi: 10.1126/science.aao2998. URL `https://www.science.org/doi/abs/10.1126/science.aao2998`.

[12] Samantha Bradshaw and Philip N Howard. The global disinformation order: 2019 global inventory of organised social media manipulation. 2019.

[13] Dan Arnaudo. Computational propaganda in brazil: Social bots during elections. 2017.

[14] Jonathan Corpus Ong and Jason Vincent Cabañes. Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the philippines. 2018.

[15] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017. doi: 10.1257/jep.31.2.211.

[16] Paul Mozur. A Genocide Incited on Facebook, With Posts From Myanmar's Military (Published 2018) — nytimes.com. `https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html`, 2018. [Accessed 16-05-2025].

[17] Christopher A Bail, Brian Guay, Emily Maloney, Aidan Combs, D Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017. *Proceedings of the national academy of sciences*, 117(1):243–250, 2020. doi: 10.1073/pnas.1906420116.

[18] Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet

de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023. doi: 10.1126/science.abp9364. URL https://www.science.org/doi/abs/10.1126/science.abp9364.

[19] Carlos Arcila Calderón, Patricia Sánchez Holgado, Jesús Gómez, Marcos Barbosa, Haodong Qi, Alberto Matilla, Pilar Amado, Alejandro Guzmán, Daniel López-Matías, and Tomás Fernández-Villazala. From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and lgbt communities. *Humanities and Social Sciences Communications*, 11(1):1–14, 2024. doi: 10.1057/s41599-024-03899-1.

[20] Gregory Eady, Tom Paskhalis, Jan Zilinsky, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Exposure to the russian internet research agency foreign influence campaign on twitter in the 2016 us election and its relationship to attitudes and voting behavior. *Nature communications*, 14(1):62, 2023. doi: 10.1038/s41467-022-35576-9.

[21] Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39, 2025. doi: 10.1145/371200.

[22] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024. ISSN 2667-2952. doi: 10.1016/j.hcc.2024.100211. URL https://www.sciencedirect.com/science/article/pii/S266729522400014X.

[23] Thomas H Costello, Gordon Pennycook, and David G Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024. doi: 10.1126/science.adq1814.

[24] Sterling Williams-Ceci, Maurice Jakesch, Advait Bhat, Kowe Kadoma, Lior Zalmanson, and Mor Naaman. Biased ai writing assistants shift users' attitudes on societal issues. 2025. doi: 10.31234/osf.io/mhjn6_v2.

[25] Tetiana Schipper. Disinformation by design: leveraging solutions to combat misinformation in the philippines' 2025 election. *Data & Policy*, 7:e39, 2025. doi: 10.1017/dap.2025.18.

[26] Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking llm agents on consequential real world tasks, 2024. URL https://arxiv.org/abs/2412.14161.

[27] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents, 2023. URL https://arxiv.org/abs/2307.14984.

[28] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework, 2024. URL https://arxiv.org/abs/2308.00352.

[29] Shiyang Lai, Yujin Potter, Junsol Kim, Richard Zhuang, Dawn Song, and James Evans. Evolving ai collectives to enhance human diversity and enable self-regulation, 2024. URL https://arxiv.org/abs/2402.12590.

[30] Andrea Baronchelli. Shaping new norms for ai. *Philosophical Transactions of the Royal Society B*, 379(1897): 20230028, 2024. doi: 10.1098/rstb.2023.0028.

[31] Giordano De Marzo, Claudio Castellano, and David Garcia. Ai agents can coordinate beyond human scale, 2025. URL https://arxiv.org/abs/2409.02822.

[32] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019. doi: 10.1038/s41586-019-1138-y.

[33] Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368, 2025. doi: 10.1126/sciadv.adu9368. URL https://www.science.org/doi/abs/10.1126/sciadv.adu9368.

[34] Bao Tran Truong, Xiaodan Lou, Alessandro Flammini, and Filippo Menczer. Quantifying the vulnerabilities of the online public square to adversarial manipulation tactics. *PNAS nexus*, 3(7):pgae258, 2024. doi: 10.1093/pnasnexus/pgae258.

[35] Weihua Li, Yuxuan Hu, Chenting Jiang, Shiqing Wu, Quan Bai, and Edmund Lai. Abem: An adaptive agent-based evolutionary approach for influence maximization in dynamic social networks. *Applied Soft Computing*, 136: 110062, 2023. doi: 10.1016/j.asoc.2023.110062.

[36] Hirokazu Shirado and Nicholas A Christakis. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, 545(7654):370–374, 2017. doi: 10.1038/nature22332.

[37] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. Uncovering coordinated networks on social media: methods and case studies. In *Proceedings of the international AAAI conference on web and social media*, volume 15, pages 455–466, 2021.

[38] Alexander C Nwala, Alessandro Flammini, and Filippo Menczer. A language framework for modeling social media account behavior. *EPJ Data Science*, 12(1):33, 2023. doi: 10.1140/epjds/s13688-023-00410-9.

[39] Kaicheng Yang, Danishjeet Singh, and Filippo Menczer. Characteristics and prevalence of fake social media profiles with ai-generated faces. *Journal of Online Trust and Safety*, 2(4), September 2024. ISSN 2770-3142. doi: 10.54501/jots.v2i4.197. URL http://dx.doi.org/10.54501/jots.v2i4.197.

[40] Kaicheng Yang and Filippo Menczer. Anatomy of an ai-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, 4, May 2024. ISSN 2673-8813. doi: 10.51685/jqd.2024.icwsm.7. URL http://dx.doi.org/10.51685/jqd.2024.icwsm.7.

[41] Angelica Chen, Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R Bowman, Kyunghyun Cho, and Ethan Perez. Learning from natural language feedback. *Transactions on Machine Learning Research*, 2024.

[42] Andreas Jungherr. Artificial intelligence and democracy: A conceptual framework. *Social Media + Society*, 9(3):20563051231186353, 2023. doi: 10.1177/20563051231186353. URL https://doi.org/10.1177/20563051231186353.

[43] Christopher Summerfield, Lisa Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian Hadfield, Luke Hewitt, Saffron Huang, Helene Landemore, Nahema Marchal, Aviv Ovadya, Ariel Procaccia, Mathias Risse, Bruce Schneier, Elizabeth Seger, Divya Siddarth, Henrik Skaug Sætra, MH Tessler, and Matthew Botvinick. How will advanced ai systems impact democracy?, 2024. URL https://arxiv.org/abs/2409.06729.

[44] Ulrike Hahn. Individuals, collectives, and individuals in collectives: The ineliminable role of dependence. *Perspectives on Psychological Science*, 19(2):418–431, 2024. doi: 10.1177/17456916231198479.

[45] Toby D Pilditch, Ulrike Hahn, and David Lagnado. The problem of dependency. *Synthese*, 205(4):143, 2025. doi: 10.1007/s11229-025-04969-w.

[46] Nicholas A Christakis and James H Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Hachette UK, 2009.

[47] Edoardo M Airoldi and Nicholas A Christakis. Induction of social contagion for diverse outcomes in structured experiments in isolated villages. *Science*, 384(6695):eadi5147, 2024. doi: 10.1126/science.adi5147.

[48] Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Data poisoning in llms: Jailbreak-tuning and scaling laws, 2024. URL https://arxiv.org/abs/2408.02946.

[49] Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119, 2018. doi: 10.1126/science.aas882.

[50] Pablo Barberá, Ning Wang, Richard Bonneau, John T Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. The critical periphery in the growth of social protests. *PloS one*, 10(11):e0143611, 2015. doi: 10.1371/journal.pone.0143611.

[51] Nicholas Diakopoulos and Deborah Johnson. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New media & society*, 23(7):2072–2098, 2021. doi: 10.1177/1461444820925811.

[52] Filippo Menczer, David Crandall, Yong-Yeol Ahn, and Apu Kapadia. Addressing the harms of ai-generated inauthentic content. *Nature Machine Intelligence*, 5(7):679–680, 2023. doi: 10.1038/s42256-023-00690-w.

[53] Margaret E Roberts. Resilience to online censorship. *Annual Review of Political Science*, 23(1):401–419, 2020. doi: 10.1146/annurev-polisci-050718-032837.

[54] Hirokazu Shirado and Nicholas A Christakis. Network engineering using autonomous agents increases cooperation in human groups. *Iscience*, 23(9), 2020. doi: 10.1016/j.isci.2020.101438.

[55] Dame Wendy Hall. Governing ai for humanity. 2024. URL https://www.un.org/en/ai-advisory-body.

[56] Steven Adler, Zoë Hitzig, Shrey Jain, Catherine Brewer, Wayne Chang, Renée DiResta, Eddy Lazzarin, Sean McGregor, Wendy Seltzer, Divya Siddarth, Nouran Soliman, Tobin South, Connor Spelliscy, Manu Sporny, Varya Srivastava, John Bailey, Brian Christian, Andrew Critch, Ronnie Falcon, Heather Flanagan, Kim Hamilton Duffy, Eric Ho, Claire R. Leibowicz, Srikanth Nadhamuni, Alan Z. Rozenshtein, David Schnurr, Evan Shapiro, Lacey Strahm, Andrew Trask, Zoe Weinberg, Cedric Whitney, and Tom Zick. Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online, 2025. URL https://arxiv.org/abs/2408.07892.

[57] E. Glen Weyl, Luke Thorburn, Emillie de Keulenaar, Jacob Mchangama, Divya Siddarth, and Audrey Tang. Prosocial media, 2025. URL `https://arxiv.org/abs/2502.10834`.

[58] Katarzyna P Adamala, Deepa Agashe, Yasmine Belkaid, Daniela Matias de C Bittencourt, Yizhi Cai, Matthew W Chang, Irene A Chen, George M Church, Vaughn S Cooper, Mark M Davis, et al. Confronting risks of mirror life. *Science*, 386(6728):1351–1353, 2024. doi: 10.1126/science.ads9158.