2

Easy to Produce, Hard to Persuade:

The Asymmetric Effects of AI on the Online Information Ecosystem

Brendan Nyhan, Jennifer Pan, Alexandra Siegel, Yamil Velez

Abstract: The hype around large language models and other forms of generative artificial intelligence (AI) has led to widespread concern about their effects on information exposure and persuasion, but these fears are likely to outstrip reality. As we show, AI is more likely to reinforce existing patterns of exposure and behavior than it is to transform how people understand and relate to the political world. Even with widespread AI use, we expect that most people will continue to consume relatively little political news and that it will be difficult to durably change public opinion at scale. We argue instead that the effects of AI on politics are likely to be greatest in the way it relaxes constraints on monitoring and production for lower-capacity actors and in how state control over AI development in contexts like China risks embedding regime-preferred narratives into the technology. We conclude by discussing how AI enables new approaches to studying fundamental questions about information exposure and behavior that go beyond studies that treat AI output as a black box treatment or outcome.

We thank Nathan Persily, Joshua Tucker, and Rachel Gillum for helpful comments. AI tools were used to find and cite sources, to provide feedback on draft versions of the manuscript, and for suggestions on the title. All errors are our own.

The rapid increase in the usage and capacity of artificial intelligence (AI) has raised widespread concerns about information manipulation. Leading figures worry that large language models (LLMs) like ChatGPT will make it possible to create seemingly authentic false or misleading content that is tailored for personalized persuasion and that exposure to this content will shape public beliefs and attitudes at scale.

For instance, in U.S. Senate testimony, OpenAI CEO Sam Altman said, "It's one of my areas of greatest concern: the more general ability of these models to manipulate, to persuade, to provide sort of one-on-one, you know, interactive disinformation" (Altman 2023). Anthropic CEO Dario Amodei similarly worried in a recent essay that, "Much more powerful versions of these models… would likely be capable of essentially brainwashing many (most?) people into any desired ideology or attitude" (Amodei 2026). Political figures have expressed similar concerns. US Senator Richard Blumenthal asserted before the 2024 election that "a deluge of deception, disinformation, and deepfakes are about to descend on the American public" via "political ads and other forms of disinformation that are made possible by artificial intelligence" (Blumenthal 2024). Former South Korean President Yoon Suk Yeol warned that, "Fake news and disinformation based on artificial intelligence and digital technology not only [violate] individual freedom and human rights but also [threaten] democratic systems" (Park and Kim 2024). And United Nations Secretary-General António Guterres told the Security Council that, "The advent of generative AI could be a defining moment for disinformation and hate speech — undermining truth, facts and safety, adding a new dimension to the manipulation of human behaviour and contributing to polarization and instability on a vast scale" (Guterres 2023).

Governments, technology platforms, and civil society should, of course, guard against potential threats from the misuse of AI, but we urge caution among those rushing to conclusions about its effects on the public's beliefs and attitudes. The fears being expressed about the effects of AI mirror those offered after the 2016 U.S. presidential election about the harmful political effects of social media and other forms of online information exposure. Many of the claims made at the time did not hold up to scrutiny. After almost a decade of research, scholars have found that algorithms do not appear to cause mass polarization (e.g., Guess et al. 2023); that most people in the U.S. and other Western countries are not in so-called "echo chambers" dominated by pro-attitudinal sources (e.g., Guess 2021, Arguedas et al. 2022); and that relatively few people consume most of the false or extremist content seen online (Budak et al. 2024). Similarly, despite extensive warnings about the potential role of deepfakes and other forms of AI-based misinformation in the 2024 U.S. election, they played little role in the campaign (Schneier and Sanders 2024). Instead, we have seen the proliferation of "AI slop" that is readily recognized as fake and instead seeks to appeal to the emotions of partisan audiences (Jingnan 2024, Bond and Brumfiel 2025).

Importantly, however, these conclusions may differ outside of the West, where less is known about the prevalence and effects of online misinformation (Blair et al. 2024, Budak et al. 2024). The harms from AI may be greater, for instance, in countries with lower media and technological literacy or higher levels of ethnic tension. We are most concerned about the potential harms that AI tools can create by relaxing monitoring, production, and other capacity constraints that previously limited the reach and scope of information influence and control efforts. First, AI lowers barriers for state and non-state actors to conduct information campaigns that were previously too costly to undertake. Second, in authoritarian countries like China, LLM companies operating under de facto state control are developing models that

reflect the regime's political objectives, risking the embedding and spread of these perspectives as emergent properties of the technology.

In this chapter, we examine how AI affects exposure to information and persuasion, including on social media. We begin at the individual level by examining how exposure to political information may be affected by AI platforms and content, as well as the potential persuasive effects of AI content on public beliefs and attitudes. We next consider how AI may change the incentives and activities of technology platforms as well as state and non-state actors who seek to influence the political and social world. Finally, we describe how large language models and other forms of AI enable new research designs that can answer fundamental questions about information exposure and persuasion.

Information exposure

As generative AI use expands rapidly around the world, particularly in middle- and low-income countries, it is reshaping how people are exposed to digital information (Chatterji et al. 2025). However, changes in information exposure matter for political processes and outcomes only to the extent that they facilitate attitudinal or behavioral change among those exposed to it. We expect AI to accelerate existing patterns in these respects, rather than transform them. Most people will continue to consume relatively little political content, regardless of how it is delivered. Similarly, the challenges of persuasion at scale persist whether content is curated by humans, traditional machine-learning algorithms, or AI-enhanced systems.

First, AI is changing patterns of information exposure when it comes to search. Specifically, online search is shifting from a process of searching with keywords to find relevant sources to interacting with an AI intermediary that provides summaries and serves as a gatekeeper to original sources. For example, AI-generated summaries on search engines like Google have dramatically increased "zero-click searches" – when users obtain answers without clicking through to an external website. Between May 2024 and May 2025, the percentage of zero-click searches jumped from 56 percent to 69 percent (Thekkethil 2025). Organic traffic to news and media websites dropped by 600 million visits from 2024–2025 (Similarweb 2025). In addition, an increasing number of people are turning directly to AI platforms to get answers via chat interfaces. As of July 2025, approximately 10 percent of the global adult population (700 million users) was using ChatGPT weekly (Chatterji et al. 2025). Compared to traditional search, AI search draws on a narrower range of sources, produces less varied responses, and disproportionately surfaces lower-credibility sources (Aral, Li, and Zuo 2026).

On social media, the integration of generative AI expands the set of behavioral signals used to inform platform operations. Social media platforms have long used machine learning with behavioral signals, such as likes, shares, watch time, scroll speed, and follow patterns to support operational decisions ranging from content moderation and spam detection to ranking and recommendation. Even prior to generative AI, platforms followed TikTok in shifting from follower, graph-based feeds to algorithmically curated feeds, which expanded the scope of potential exposure by surfacing content outside users' social networks (Sanderson, Messing, Tucker 2024). The integration of generative AI creates new sources of behavioral data, such as users' conversations with platform-embedded AI chatbots, and allows for richer analysis of content. For example, Meta announced in December 2025 that it would use data from

interactions with AI chatbots across Facebook, Instagram, and WhatsApp to personalize content and ads targeting the user (Meta 2025). TikTok uses multi-modal LLMs to extract specific claims from videos to curate content, and Instagram indexes video, text, and audio content to improve reach (TikTok 2024, Wilson 2026). However, these changes are largely incremental. There is little reason to expect them to overcome fundamental barriers – such as attention paid to politics and the persuasiveness of content – that would lead to attitudinal and behavioral shifts at scale.

Beyond these more direct impacts on information exposure, generative AI may also affect downstream information exposure by shifting patterns of information production. AI tools lower creation costs, potentially enabling more content to be produced with less effort by humans. Such a change could increase the productivity of journalistic organizations, though many observers worry about inaccuracy and degradation of the quality of news reporting (Setty 2026). In addition, autonomous AI agents can now contribute to open-source code, publish content online, and communicate with other AI agents as well as humans without direct human oversight. In February 2026, an AI agent wrote and disseminated a personalized attack piece on Scott Shambaugh, a matplotlib developer, by researching Shambaugh's commit history on GitHub after Shambaugh rejected its code contribution (Shambaugh 2026). Platforms like Moltbook, a social network exclusively for AI agents, which claims over 1.4 million registered agents as of early 2026 (Yildiz 2026), represent an emerging space where AI-to-AI interaction generates its own information ecosystem largely outside human oversight.

The growing volume of AI-generated information can displace human knowledge creation. Within six months of ChatGPT's release, activity on Stack Overflow decreased 25 percent in countries where ChatGPT was accessible compared to countries where access to ChatGPT was limited (del Rio-Chanona et al. 2024). A similar pattern was observed on Stack Exchange, where questions dropped by 18 percent (Quinn and Gutt 2025). This decrease may reflect efficiency gains, as users get answers immediately without needing to wait for community responses, but these shifts mean the incentive to create or contribute original human-generated material declines, potentially diminishing the publicly available base of human-generated information, even as the volume of content in circulation is greatly boosted by generative AI.

Such changes in the information environment will not necessarily have any particular political effects. However, as information volume grows, people will also increasingly rely on AI systems to curate information for them, making LLMs a gatekeeper as well as a producer. When users turn to ChatGPT or other models instead of social media or news sites, LLMs decide what is relevant, what to summarize, and what to omit from an information environment that these models are increasingly populating.

Importantly, though, these dynamics may have different effects across linguistic contexts. Generative AI systems perform worse for low-resource languages (e.g., Burmese, Armenian, Yoruba) than for high-resource languages (e.g., English, French, Chinese) across a range of tasks, including translation, factual question answering, text classification, and summarization, primarily due to the lack of training data (Robinson et al. 2023; Li et al. 2025). Multilingual LLMs display information biases when queried in low-resource languages and, instead, draw on high-resource language sources which reinforce those linguistic perspectives and marginalize low-resource language viewpoints (Sharma, Murray, and Xiao 2025). In addition, LLMs are more likely to produce harmful or unsafe output in low-resource languages

due to less robust safety training and alignment (Yong et al. 2023; Deng et al. 2024). While newer models have expanded their language coverage, most major LLMs underperform for low-resource languages (Pava et al. 2025). The state of language resources depends on historical patterns related to institutions and development. Languages of former colonial powers and wealthier countries have far more digital text available for AI training than languages spoken primarily in lower-income countries or by marginalized populations. For example, Hindi is not a high-resource language despite the large population of speakers because English dominates India's digital sphere. In addition, the digital data that LLMs are trained on reflect the information environment of source countries. Research shows that text scripted for the Chinese media by the Chinese Communist Party appears in LLM training datasets and that prompting models that were developed outside of China in Chinese generates more positive assessments of China's political institutions than the same prompts in English (Waight et al. 2026). The global shift to AI-influenced information exposure will reinforce these existing inequalities. The quality and relevance of information may be systematically lower for populations who already face political and economic barriers to reliable information, while those who can produce digital content at scale may see their influence strengthened. These disparities are rooted in institutions and socio-economic structures. Understanding and addressing these dynamics requires the engagement of social science that purely technical approaches risk overlooking.

Persuasion and belief/attitude change

New technologies ranging from radio to television to video games have frequently inspired panic around their potential harms, including fears of brainwashing and propaganda (e.g., Starker 1989, Drotner 1999, Orben 2020). In this sense, the fears being expressed about the persuasive powers of AI are familiar. As with those technologies, AI-generated content can indeed be persuasive and lead to durable attitude changes, but these effects are likely to be smaller than people fear and to vary with exposure, as we discuss below.

We first distinguish between AI source effects and AI content effects. In general, identifying as a source, AI has mixed or inconsistent effects that may vary by domain or other contextual factors. Lu, Tormala, and Duhachek (2025) find people are more open to counter attitudinal persuasion from AI rather than from human sources because AI is seen as more informative and less biased. By contrast, Boissin et al. (2025), Gallegos et al. (2025), and Spearing et al. (2025) find no effect of AI versus human attribution in persuasive messages (including misinformation). And Altay and Gilardi (2024), Nanz et al. (2025), and Wittenberg et al. (2025) find people are more skeptical of journalistic content, media outlets, and images labeled as AI-generated, respectively.

Some studies examine political slant and bias in AI content, instead of, or in addition to, testing the causal effects of exposure to it. Though definitional and measurement concerns persist (Fisher et al. 2025a), most studies find content from major LLMs leans left, with the possible exception of xAI's Grok, which has shifted to the right (e.g., Rozado 2024, Motoki et al. 2025, Rozado 2025, Westwood, Grimmer, and Hall 2025, Wirtschafter and Nadgir 2025). Exposure to LLM content may, accordingly, have some effect on people's political attitudes. Potter et al. (2024) find, for instance, that engaging in a conversation with an LLM about former U.S. President Joe Biden and current President Donald Trump shifted people's

views toward Biden. However, absent a neutral or non-AI baseline, experimental designs like these do not isolate the causal effect of slant in LLM content or the information provided about the candidates.

The effects of AI-generated content are most precisely estimated in studies that manipulate exposure to AI content on a topic of interest. These studies often find that exposure to AI content or participation in AI dialogues can generate durable changes in people's political beliefs and attitudes – changes that often match or outperform those created by exposure to human-created static content (e.g., Costello et al. 2024, Hackenburg and Margetts 2024, Bai et al. 2025, Costello et al. 2025a, Costello et al. 2025b, Czarnek et al. 2025, Lin et al. 2025, Fisher et al. 2025b). However, these findings differ in magnitude from the claims being made about the persuasive powers of AI. For instance, the most high-profile study in this area is Costello et al. (2024) which finds that belief in the conspiracy theory debunked by an AI in a dialogue with online survey participants was reduced by approximately 14 percentage points (and other studies like Summerfield et al. 2025 have found smaller effects, averaging around 7 percentage points).

Findings differ on the factors that make AI content or dialogue exposure persuasive. One theory is that LLMs can tailor the content they produce to attributes of the person receiving the information. Two recent studies find that personalized content from LLMs is more persuasive than non-personalized content (Matz et al. 2024, Salvi et al. 2025). However, two other studies find few differences resulting from personalization (Hackenburg and Margetts 2024, Argyle et al. 2025), and a third finds the effect is small relative to other factors (Hackenburg et al. 2025).

An alternative account emphasizes the role of novel facts (or fact-like claims) in AI persuasion. Notably, Costello et al. (2025c) uses an ablation design to probe the mechanism for the belief change findings in Costello et al. (2024), which shows how conversations with AI can reduce conspiracy belief. Of the potential mechanisms tested, the only one that eliminated the observed effect was an experimental condition in which the LLM was told not to provide rational, evidence-based arguments. Persuasive effects were also much smaller in a no-facts condition evaluated by Bai et al. (2025). The use of facts and evidence by LLMs was, in turn, associated with greater participant belief change in Bai et al. (2025), Costello et al. (2025a), and Lin et al. (2025). The facts provided by the LLMs in studies like these have been found to be largely accurate when evaluated (Costello et al. 2024, Lin et al. 2025). In some cases, such facts may be lacking; Costello et al. (2025c) find that LLM instead urged caution before facts were known in challenging conspiracy theories about the 2024 assassination attempts against Donald Trump. However, if pushed further, models may make dubious claims that are persuasive but not accurate. Hackenburg et al. (2025) shows that post-training of LLMs that is intended to increase persuasion tends to degrade accuracy, suggesting a key tradeoff in which demands for more factual or fact-like information to support a claim causes models to veer into inaccuracy.

Importantly, the key role of facts and evidence in AI persuasion does not mean that AIs cannot misuse information or mislead people. Model accuracy has tended to improve over time, but such increases are not necessarily linear in time. Notably, Hackenburg et al. (2025) found that GPT-4.5 was less accurate than prior GPT models tested despite being much larger and having more post-training. Accuracy may also degrade in certain contexts. For instance, Lin et al. (2025) found that LLMs made more inaccurate claims when prompted to make an argument for conservative candidates.

In particular, it has long been clear that LLMs can be used to create propaganda, misinformation, and other forms of malign persuasive content (Goldstein et al. 2024). Indeed, they already appear to be in use to expand the output of Russian-backed state media (Wack et al. 2025). The potential for misuse is substantial. Hackenburg et al. (2025) shows that post-training intended to increase persuasion is feasible and can make a smaller open weight model (Llama-8B) equally persuasive to a larger proprietary model (GPT-4o). Costello et al. (2026) similarly shows that LLMs are as effective in convincing people to believe in conspiracy theories as they are in debunking them. Measuring and mitigating these risks is an important topic for future research (Kowal et al. 2025).

Countermeasures such as inoculation may not be enough to undo the effects of exposure to AI-generated misinformation (Spearing et al. 2025). First, accurate identification of such content at scale is difficult even for platforms. People may be even worse at making such distinctions, including for images (DeVerna et al. 2024). In addition, reminders of the prevalence of AI images can increase general skepticism about news and digital media (Campante et al. 2025, Sanderson, Zhong, and Tucker 2025).

It is also important to note that parasocial relationships with AI companions could expand the scope of AI persuasion by deepening attention and trust, which often constrain persuasive effects at scale. Contemporary LLM chatbots are responsive, remember information, and can sustain long, emotionally salient conversations, which can increase disclosure, perceived intimacy, and reliance over time (Smith et al. 2025). This sustained engagement can create opportunities for iterated persuasion – testing messages, adapting to objections, and calibrating framing to the user's goals. As a result, even if average "one-shot" exposure effects are small, influence effects may compound over time among high engagement users who treat the chatbot as a trusted social partner (Qi et al. 2025).

We conclude by noting two challenges for research in this area. The first is that the temporal validity of these findings may be low due to rapid change in the state of the art in LLMs and other forms of generative AI (Munger 2019, Munger 2023). Human performance at detecting AI-generated images, for instance, is likely to degrade as technology improves. Second, we may observe second-order effects that alter the dynamics of persuasion in this domain. People may come to distrust AI content more or adjust their information diets to rely more on trusted sources (Campante et al. 2025). AI may also create or contribute to a so-called "liar's dividend" in which people use AI to deny the validity of true information (Sanderson, Messing, and Tucker 2024, Schiff et al. 2025). However, so far, people still do not instinctively disbelieve the images and videos they encounter as a result of AI awareness. Videos and images from Minneapolis, Minnesota, were widely circulated during the period that a large number of Immigration and Customs Enforcement (ICE) and Customs and Border Protection (CBP) agents were deployed there and treated as credible (though, of course, AI-manipulated images and videos were also posted online). The federal government did not attempt to claim that authentic videos and images were AI creations. One constraint may be the number of videos that were created of the same incidents. In addition, news organizations are increasingly devoting resources to verifying digital media and certifying it as legitimate (e.g., Lum and Willis 2026).

Platforms, state actors, and non-state actors

Social media reshaped political information environments largely by lowering distribution costs and barriers to entry, making it cheap and easy to reach large audiences, coordinate supporters, and exploit platforms for amplification (Zhuravskaya, Petrova, and Enikolopov 2020; Papathanassopoulos and Giannouli 2025). By contrast, generative AI lowers the costs of production and interpretation (Sanderson, Messing, and Tucker 2024). Creating tailored text, images, and narratives and classifying and distilling meaning from large quantities of content is faster and less costly than ever. These features of generative AI can be exploited by state and non-state actors to manufacture, adapt, and summarize information at scale (Marcellino et al. 2023). Below, we describe how generative AI is changing how state and nonstate actors can influence political information.

First, LLMs facilitate continuous production and adaptation of targeted information. In the social media era, even sophisticated influence operations often struggled to generate enough plausible, locally tailored content across languages and personas to reach their strategic goals. Generative AI reduces that constraint by enabling high-volume, stylistically diverse, rapidly iterated messaging, which is especially valuable when actors must constantly adapt to moderation and shifting narratives (Marcellino et al. 2023). Analyses of a recent Russian state influence campaign suggest that the adoption of generative AI coincided with expanded output while maintaining apparent persuasiveness, consistent with the emergence of more "industrialized" production pipelines (Wack et al. 2025). Similarly, experiments show that current models can generate high-quality components of election disinformation operations, lowering the time and skill requirements for running campaigns that previously demanded larger teams and specialized expertise (Williams et al. 2025). These changes benefit both state and non-state actors but especially smaller, lower-resourced groups who can now run continuous real-time operations with limited human staffing (Pauwels 2024).

Second, state control over LLM development risks embedding regime narratives into the models themselves. This differs from social media-era information control, which operated primarily through censorship and propaganda directed at content after its creation. LLM companies operating under de facto state control develop models that reflect the regime's political objectives from the outset. Studies of Chinese, Russian, and US-based LLMs show how this manifests as both hard refusals and soft omission or steering of tailored content to domestic audiences and languages (Noels et al. 2025; Pan and Xu 2026), with more fine-grained audits documenting semantic suppression in responses to sensitive prompts (Qiu, Zhou, and Ferrara 2025).

Importantly, state influence on model behavior is not limited to authoritarian regimes. Democratic governments are increasingly using procurement, regulation, and political pressure rather than direct ownership or formal party-state control to shape model development. In the US, federal acquisition guidance directs agencies to manage risks in procured AI systems and requires vendor transparency and testing/disclosure practices, which can incentivize providers to modify model behavior and safety features to meet contracting requirements. In 2025, the executive branch explicitly tied federal procurement to contested "ideological bias" standards to prevent "woke AI in the federal government" (Trump 2025). The next year, the Trump administration went further, threatening to use the Defense Production Act to compel Anthropic to remove safeguards from its Claude AI or to designate it as a "supply chain risk" if it refused to do so (Lawler and Curi 2026).

Third, the data used to train LLMs create the opportunity for political influence beyond national borders. Authoritarian regimes influence their domestic information environments, which then shape the training data used by models developed both inside and outside authoritarian settings (Yang and Roberts 2023). Waight et al. (2026) demonstrate that propaganda from political institutions already influences the output of US-based large language models via their training data. Regime narratives can thus spread as emergent properties of the technology – not just through direct manipulation of foreign models but also via upstream contamination of the data ecosystem on which all models depend. State actors have also attempted more direct interventions, including industrial-scale content flooding designed to contaminate the web and the materials that retrieval-augmented generation (RAG) systems rely on, such as campaigns associated with pro-Kremlin ecosystems like the "Pravda" network (Danet 2025; Châtelet 2025). Some work cautions that apparent "poisoning" may instead reflect data voids (Boyd and Golebiewski 2018) – when high-quality information is scarce, low-quality or partisan sources can dominate what retrieval systems surface without requiring coordinated manipulation (Alyukov et al. 2025). These dynamics could benefit non-state actors as well; even if small groups cannot poison closed training data pipelines, they may flood the open web or exploit data voids in ways that influence what systems retrieve and cite (Alyukov et al. 2025; Pauwels 2024).

Fourth, LLMs make personalization cheap and interactive in a way social media rarely did at scale. Social media enabled microtargeting and experimentation, but sustained, conversational persuasion or recruitment typically required humans in the loop (call centers, troll farms, volunteer brigades) or relatively rigid scripts. LLMs support interactive messaging that can respond to objections, maintain persona consistency, and tailor appeals across channels (Pauwels 2024). In this way, non-state actors and extremist networks using language models can emulate interactive radicalization content and generate persuasive text cheaply, raising recruitment and mobilization risks even when platforms attempt enforcement (McGuffie and Newhouse 2020; Puczyńska et al. 2024). These changes also affect state actors because the same interactivity can be used to frame narratives and to scale harassment and intimidation efforts both domestically and transnationally (Sen and Farooq 2025). However, attention and trust still constrain the effects of these efforts. As a result, they may not result in opinion change even as they create expanded capacity for persistent engagement, testing, and targeting (Williams et al. 2025, Pauwels 2024, Sen and Farooq 2025). While limited attention may constrain average opinion change, it does not similarly constrain the effectiveness of harassment. Even when targets do not believe the message, persistent, personalized abuse can impose costs, such as time, stress, and safety concerns, that deter activism and participation. While coordinated trolling has long been a common occurrence on social media (Bradshaw and Howard 2017), LLMs reduce the labor required to sustain high-volume, adaptive harassment – including rapid replies, persona consistency, and cross-platform variation – potentially increasing prevalence and persistence. As with coordinated trolling operations, such harassment campaigns can be used to target a wide range of individuals from politicians to activists and everyday citizens.

Fifth, generative AI changes detection and attribution, allowing for greater plausible deniability by both state and non-state actors. Social media manipulation often leaves traces of coordination in metadata that platforms and researchers could detect. LLM-enabled operations can generate endless paraphrasing and stylistic variants, making “copy-pasta” detection more difficult and raising the verification burden for defenders (Marcellino et al. 2023). These challenges interact with manipulation at the training data level. Even if individual posts are removed, the broader information environment may still be saturated, and

retrieval systems may keep rediscovering variants (Danet 2025, Alyukov et al. 2025). As a result, those seeking to combat influence operations in the age of generative AI must focus not only on removing content but on maintaining the integrity of the evidence and reference ecosystems that models consult.

Finally, generative AI may intensify the cat-and-mouse nature of authoritarian and anti-authoritarian uses. Social media already empowered citizen journalism and mobilization while exposing activists to surveillance, repression, harassment, and doxxing. Generative AI perhaps intensifies this dynamic by improving surveillance, repression, and propaganda, as well as verification and documentation. On the one hand, AI tools can amplify state surveillance capacity by lowering the costs of monitoring vast quantities of social media posts, messages, images, and geolocation traces so that security services can identify networks and flag individuals quickly (CEIP 2019). AI-enabled biometric identification (especially facial recognition) can also facilitate the tracking of protesters and dissidents, increasing arrest and retaliation risks even when participation is brief or decentralized (Freedom House 2023). Importantly, these tactics are not limited to authoritarian contexts and have been increasingly utilized by democratic regimes. For example, recent reporting demonstrates that US Immigration and Customs Enforcement (ICE) is using AI-enabled tools to analyze driver's license scans, extract information from seized phones, cross-search location data, and link records from federal, state, and commercial databases (Hubbard 2025). Similarly, local US law enforcement including the Miami-Dade Sheriff's Office (MDSO) and the Los Angeles Police Department (LAPD) have bought access to GeoSpy, an AI tool that can almost instantly geolocate the location where a photo was taken using features in the image, such as architecture and vegetation, and plans to use it in criminal investigations (Cox 2026).

On the other hand, AI can also make human rights work more agile by automating repetitive tasks, such as translation, summarization, and pattern identification, and by supporting digital tools for advocacy and monitoring (Espiritusanto et al. 2024). By lowering the costs of producing public facing content, generative AI can also help small organizations and low-resource political candidates or activists generate polished materials quickly, ranging from press releases to multilingual outreach. This shift could expand participation and coalition-building by lowering the costs of adaptation (Tomić et al. 2023, Kerley et al. 2024). Open-Source Intelligence (OSINT) research emphasizes that AI can help process large volumes of public information for extraction and knowledge discovery, which can support verification and accountability work (Evangelista et al. 2021). Human rights practitioners argue for "authenticity infrastructure" and resilient witnessing practices to fortify the evidentiary value of digital media in an era of deepfakes (Gregory 2023). More broadly, scholarship on AI and content moderation stresses that algorithmic systems can be used either to suppress speech or to help create conditions for robust democratic exchange, depending on governance, accountability, and institutional incentives (Llansó 2020). Recent work on democracy movements similarly emphasizes tactical uses of AI for counter-messaging and real-time logistical support in contentious politics (Chenoweth 2025). In this way, while LLMs do appear to be facilitating greater digital repression and surveillance, they are also accelerating an adaptation race in which states, challengers, platforms, and civil society organizations continually update both offensive and defensive techniques to achieve their political goals (Feldstein 2019; Gregory 2023; Llansó 2020, Chenoweth 2025).

New approaches to research

Generative AI may have not only important impacts at the micro and macro levels but also the power to change social science research on these topics. For instance, traditional experimental research designs that use static stimuli may be insufficient for testing theories about information exposure or understanding the policy implications of those findings. However, the relevant stimuli in the real world (e.g., political ads or news articles) often vary in both systematic and idiosyncratic ways that are not captured by a single researcher-chosen stimulus, which often trade off ecological validity for analytical ease (Clifford, Leeper, and Rainey 2024, Clifford and Rainey 2024). This gap between research and practice may increase as politics, marketing, medicine, and education increasingly embrace personalization. Adaptive research designs are needed that reflect participants' information diets, backgrounds, and preferences. By adaptive, we mean research designs that tailor treatment or measurement to individual participants based on their characteristics, prior responses, or other contextual factors.

Current research on social media algorithms is already pushing in this direction. As public concern about the polarizing effects of algorithms has intensified, scholars have collaborated with social media companies to assess the political effects of recommendation algorithms, which vary at the individual level. As discussed above, Guess et al. (2023) randomly assign Facebook and Instagram users to either a default algorithm or a reverse-chronological feed. In contrast to traditional survey experiments employing fixed vignettes, the content of each feed varies at the individual level. This design recovers a policy effect rather than the effect of a single experimental stimulus.

Newer designs also leverage personalized stimuli to study attitudinal backfire (Velez and Liu 2024), the relationship between beliefs and attitudes (Velez, Liu, and Clifford 2026), and corrections of conspiracy beliefs (Costello et al. 2024). For example, Velez and Liu (2024) use "tailored experiments" to assess whether motivated reasoning is activated when voters' deeply held issue stances are at stake. While backfire is rare in the literature (e.g., Coppock 2022), Velez and Liu use OpenAI's GPT-3 to generate personalized counterarguments targeting political convictions disclosed in an open-ended question. Across five studies, they find more evidence of moderation than polarization, with reliable evidence of backfire only emerging when counterarguments are presented using a vitriolic tone. This ability to align interventions with theoretically optimal conditions becomes possible with generative AI, which makes it possible for high-quality stimuli to be generated on the fly with little researcher supervision.

Tailored designs powered by LLMs can also enable adaptive measurement of rare or unexpected beliefs and attitudes. For example, Velez (2025) develops a crowdsourced adaptive survey method (CSAS) to convert open-ended responses from participants about misinformation and issue priorities into properly formatted survey questions using large language models such as OpenAI's GPT-3 and GPT-4, leveraging multi-arm bandit algorithms to prioritize the most promising user-submitted questions. This method allows survey batteries to adapt to participants and surfaces new questions that might escape the notice of survey researchers.

Tailored approaches also offer a new way to test canonical theories. Many models in political science and psychology are built around inherently heterogeneous constructs. In theories of motivated reasoning, for example, strong issue attitudes are expected to activate defensiveness in response to counter attitudinal

information (Taber and Lodge 2006). Proximity voting models similarly emphasize "issue weights" as voters choose between candidates (e.g., Grynaviski and Corrigan 2006). Yet in both cases, researchers often operationalize these constructs using fixed issue batteries that feature a subset of policies. But if only a small share of the population is "activated" by the interventions (i.e., has a strongly held, stable attitude about the issue in question), such studies may be underpowered to detect theoretically meaningful relationships.

To provide an example, Velez, Liu, and Clifford (2026) use tailored experiments to resolve a long-standing puzzle about why belief change seldom produces attitude change. While numerous studies show positive effects of information on factual beliefs (Wood and Porter 2019; Hopkins, Sides, and Citrin 2019; Nyhan, Porter, Reifler, and Wood 2020), attitudes and behavior are remarkably resistant. Velez, Liu, and Clifford propose that the failure of counterarguments to shift attitudes hinges on "belief relevance" – the extent to which beliefs bear on attitudes. In a conversation with OpenAI's GPT-4o, participants are encouraged to provide reasons for their issue position ("focal beliefs"). The LLM then summarizes this justification, generates unmentioned justifications that support the participant's views ("distal beliefs"), and constructs counterarguments targeting both along with a placebo argument. The authors find that counterarguments targeting focal beliefs outperform those targeting distal beliefs, suggesting that belief relevance may condition the effects of counter attitudinal information on attitudes.

Despite their promise, the move from fixed to tailored stimuli introduces a unique set of challenges. Stimuli such as text are already high-dimensional and personalization adds an additional layer of complexity. Scholars will need to carefully design their studies to recover the effects of bundled treatment consisting of model versions, parameters, participant inputs, and user interfaces. Such studies are not without precedent. Tailored experiments are not dissimilar from experiments that hinge on dynamic stimuli such as peer-to-peer conversations, where instantiations of the treatment vary across people. Still, it is worth recognizing that the mapping between this bundled treatment and the latent theoretical construct requires the (often strong) assumptions that the intended treatment is being manipulated and no stimulus-level confounding is present.

Careful prompting provides a path to manipulating latent constructs while minimizing known sources of confounding like length or readability. For example, researchers can guide an LLM to generate a counterargument of roughly the same length, valence, or tone across arms. Moreover, researchers can surgically vary certain instructions or remove them to assess potential mechanisms (e.g., Costello et al. 2025c, Argyle et al. 2025). However, while these approaches minimize known sources of confounding, LLMs might inherently bundle certain features in its internal representation of the text. For example, LLM-generated arguments about immigration may systematically co-vary with a "nationalism" feature, rendering it difficult to identify which text feature is producing the observed effect.

Imai and Nakamura (2024) propose a method of text deconfounding that leverages the internal representations of LLMs to minimize latent unobserved text confounding. This method exploits the fact that internal representations provide information about how LLMs encode semantic dimensions of generated text. Obtaining this internal representation requires accessing a model with open weights and extracting the hidden state from the final model layer before text generation and estimating two models via double machine learning: a deconfounding model that predicts treatment status using latent text features and an outcome model that predicts outcomes using treatment status and these covariates.

Conditioning on latent text features provides stronger evidence that estimated effects are not driven by unobserved features that covary with the treatment of interest (see also Roberts, Stewart, and Nielsen 2020).

Beyond text confounding, there is also the risk of "model drift." As Barrie, Palmer, and Spirling (2025) demonstrate, replicability challenges emerge when using proprietary LLMs that do not fix model weights. If papers use LLMs to classify text, and those LLMs are subsequently deprecated or model weights are modified, it becomes impossible to procedurally replicate classification procedures. Indeed, the authors discuss a number of models that entered a deprecated state after data collection. The authors caution against the use of proprietary models in social science research, advocating for "locally versioned open models" that possess fixed weights and can be downloaded onto consumer devices.[1]

Despite replicability limitations, use of proprietary models may be preferable if the interest is in studying how generative AI affects information seeking and persuasion in ecologically valid settings, such as elections, wars, or routine use by citizens. Given that political actors are likely to depend on high-performing models, assessing the persuasive capabilities and political biases of lesser-known, less effective open weights models risks studying the wrong question. If the goal is to inform public discourse about LLMs, prioritizing replicability over how political agents use these tools in real-world situations may limit the applicability of our research. Proprietary models may also be preferred over open source models if they are more effective in accomplishing research objectives including not just classification accuracy but, for instance, experimental designs where a strong "first stage" relationship between the experimental manipulation and theoretical construct is critical for a fair test of the theory. If an open weights model is not capable of generating relevant stimuli, such a test will be less informative. Finally, it is worth recognizing that many real-world political constructs are stochastic and thus not completely replicable. It is impossible to faithfully reproduce every aspect of a deliberation experiment, for instance, since the dynamics of conversations hinge on group composition and emergent interactions among participants. LLM interventions may be similarly classified as "stochastic latent treatments" that cannot be held perfectly constant across replications.

We also anticipate generative AI models themselves becoming subjects of research. A long tradition in computer science has evaluated fairness, bias, and alignment to ensure that models are producing helpful and harmless outputs. Though several providers have undertaken audits of political bias, these audits depart from how political scientists tend to think about this construct. For example, OpenAI's audit of their models focused on dimensions of bias related to validation or invalidation of political beliefs, refusals, and whether political content was balanced (OpenAI 2025). Though these are sensible dimensions from a user experience perspective, they depart from the measurement models that have characterized the literature on ideology. Aldahoul et al. (2025) use item response models to measure political ideology across different LLMs, whereas Westwood, Grimmer, and Hall (2025) present participants with pairwise comparisons of model outputs and ask them to gauge bias. Such studies aim to measure political positions on a continuum, better approximating political bias than the "balanced responding" dimension that existing audits have captured.

[1] With few exceptions, such as the Ai2's Olmo model, most "open source" large language models, such as Meta's Llama, OpenAI's OSS, and Google's Gemma, release model weights without training instructions or training data, departing from standard definitions of open source. The principal advantages of these models are that model weights are fixed and can be inspected.

Despite their advantages as research tools and subjects of research, generative AI models raise distinct ethical challenges. Chief among them is “hallucination,” or the lack of factual grounding in model outputs. Though newer models have improved significantly by drawing on web searches, the quality of LLM outputs is uneven, with some referring to LLMs as possessing a “jagged intelligence,” excelling at difficult tasks while inexplicably failing at simple ones (Karpathy 2024). While models may further improve, there are alternative architectures, such as retrieval-augmented generation (RAG), that can improve factual responses. RAG supplements the responses of LLMs to user queries with information drawn from curated external databases. For example, Velez, Green, and Sevi (2025) use RAG to develop an AI-powered voting advice application that provides young independent voters with verified information about political parties. Though hallucinations may still persist (Magesh et al. 2025), such approaches – including newer agentic search techniques that retrieve and synthesize information from structured databases – can reduce the likelihood of unsupported or fabricated claims, while retaining the interactive qualities of generative AI.

Another ethical dimension lies in the research transparency and replicability challenges mentioned above. Demonstrating the capabilities of proprietary LLMs can be interpreted as “carrying water” for platforms. However, avoiding the systematic study of proprietary models creates a vacuum that can be filled only by the providers themselves, who are trying their hand at estimating political bias in ways that deviate from how social scientists measure these concepts. Social scientists assessing the persuasive impact of proprietary models can help temper warnings of LLMs as hyper-persuasive agents. Such warnings likely do more to encourage malicious use than carefully conducted studies that provide precise effect sizes.

Though proprietary LLMs are often characterized as "black boxes" that lack transparency, it is worth comparing them against other information sources, such as social media platforms. In contrast to most social media algorithms, LLM APIs are publicly available, allowing researchers to probe how they respond to different queries and measure their behavior over time. For example, researchers can measure how models respond to contested political issues, job resumes of applicants from protected backgrounds, and other socially relevant queries. While the underlying weights are not available, public APIs allow for interventions and analyses that are not typically possible with social media algorithms absent bespoke tools developed by researchers (Piccardi et al. 2025). Compared to other algorithms shaping information environments, LLMs are unusually amenable to systematic study. That said, broader transparency concerns would be significantly addressed if LLM providers were to open-source more models and "freeze" specific versions for academic research, committing to maintain them over time.

Conclusion

This review suggests that the effects of AI on persuasion and the online information ecosystem will be more subtle than many observers expect. AI will change what news and information people consume substantially, in part by displacing search engines from their key role in online information acquisition. But differences in information exposure are unlikely to transform mass opinion, which is constrained by public inattention and pre-existing attitudes. The most substantial effects of AI could be the ways in which it transforms the capabilities of platforms, state actors, and non-state actors to engage in content production, monitoring, and surveillance at unprecedented scale, especially in authoritarian regimes. But

large-scale targeted harassment and "flooding" attacks on information spaces are now cheap and easy to conduct even for individuals. Importantly, these tactics can all be effective without changing minds.

We caution that these conclusions apply unevenly throughout the world and may change over time. In particular, second-order effects of AI ubiquity on trust in online information and the effectiveness of "liar's dividend"-style denials of unwelcome fact could dominate AI's first-order effects on what people see and believe. For this reason, it is essential to leverage all the research tools available to us, including AI, to understand the effects it is having on the world.

References

Aldahoul, N., Ibrahim, H., Varvello, M., Kaufman, A., Rahwan, T., & Zaki, Y. (2025). Large language models are often politically extreme, usually ideologically inconsistent, and persuasive even in informational contexts. arXiv preprint arXiv:2505.04171. https://doi.org/10.48550/arXiv.2505.04171

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, *3*(10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403

Altman, S. (2023, May 16). Testimony before the U.S. Senate Committee on the Judiciary, Subcommittee on Privacy, Technology, and the Law. Hearing: "Oversight of A.I.: Rules for Artificial Intelligence." 118th Congress. https://www.congress.gov/event/118th-congress/senate-event/LC71543/text

Alyukov, M., Makhortykh, M., Voronovici, A., & Sydorova, M. (2025). LLMs grooming or data voids? LLM-powered chatbot references to Kremlin disinformation reflect information gaps, not manipulation. *Harvard Kennedy School Misinformation Review*, *6*(5). https://doi.org/10.37016/mr-2020-187

Amodei, D. (2026, January 26). The adolescence of technology: Confronting and overcoming the risks of powerful AI. https://www.darioamodei.com/essay/the-adolescence-of-technology

Aral, S., Li, H., & Zuo, R. The Rise of AI Search: Implications for Information Markets and Human Judgement at Scale. *arXiv preprint arXiv:2602.13415* (2026).

Arguedas, A. R., Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2022). Echo chambers, filter bubbles, and polarisation: A literature review. Reuters Institute for the Study of Journalism, University of Oxford. https://doi.org/10.60625/risj-etxj-7k60

Argyle, L. P., Busby, E. C., Gubler, J. R., Lyman, A., Olcott, J., Pond, J., & Wingate, D. (2025). Testing theories of political persuasion using AI. *Proceedings of the National Academy of Sciences*, *122*(18), e2412815122. https://doi.org/10.1073/pnas.2412815122

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, *16*, 6037. https://doi.org/10.1038/s41467-025-61345-5

Barrie, C., Palmer, A., & Spirling, A. (2025). Replication for language models: Problems, principles, and best practice for political science. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2024). Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology*, *55*, 101732. https://doi.org/10.1016/j.copsyc.2023.101732

Blumenthal, R. (2024, April 17). ICYMI video: Blumenthal highlights dangers of artificial intelligence in spreading disinformation with election deepfakes [Press release]. U.S. Senate, Subcommittee on Privacy, Technology, and the Law. https://www.blumenthal.senate.gov/newsroom/press/release/icymi-video-blumenthal-highlights-dangers-of-artificial-intelligence-in-spreading-disinformation-with-election-deepfakes

Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. *PNAS Nexus*, *4*(11), pgaf325. https://doi.org/10.1093/pnasnexus/pgaf325

Bond, S., & Brumfiel, G. (2025, December 24). 2025 has seen an explosion of AI-generated slop. *NPR*. https://www.npr.org/2025/12/24/nx-s1-5629169/2025-has-seen-an-explosion-of-ai-generated-slop

Boyd, d., & Golebiewski, M. (2018). Data voids: Where missing data can easily be exploited. Data & Society Research Institute. https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3-1.pdf

Bradshaw, S., & Howard, P. N. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation (Working Paper 2017.12). Computational Propaganda Research Project, Oxford Internet Institute, University of Oxford. https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, *630*(8015), 45–53. https://doi.org/10.1038/s41586-024-07417-w

Campante, F. R., Durante, R., Hagemeister, F., & Sen, A. (2025). GenAI misinformation, trust, and news consumption: Evidence from a field experiment (NBER Working Paper No. 34100). National Bureau of Economic Research. https://www.nber.org/papers/w34100

Carnegie Endowment for International Peace [CEIP]. (2019, September). The global expansion of AI surveillance (S. Feldstein, Author). https://carnegieendowment.org/research/2019/09/the-global-expansion-of-ai-surveillance

Châtelet, V. (2025, April 18). Exposing Pravda: How pro-Kremlin forces are poisoning AI models and rewriting Wikipedia. Atlantic Council, Digital Forensic Research Lab. https://www.atlanticcouncil.org/blogs/new-atlanticist/exposing-pravda-how-pro-kremlin-forces-are-poisoning-ai-models-and-rewriting-wikipedia/

Chatterji, A., Cunningham, T., Deming, D., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use ChatGPT (NBER Working Paper No. 34255). National Bureau of Economic Research. https://doi.org/10.3386/w34255

Chenoweth, E. (2025, February). How AI can support democracy movements: Summary report of a research and practice workshop. Ash Center for Democratic Governance and Innovation, Harvard Kennedy School. https://www.hks.harvard.edu/publications/how-ai-can-support-democracy-movements-summary-report-research-and-practice-workshop

Clifford, S., Leeper, T. J., & Rainey, C. (2024). Generalizing survey experiments using topic sampling: An application to party cues. *Political Behavior*, *46*, 1233–1256. https://doi.org/10.1007/s11109-023-09870-1

Clifford, S., & Rainey, C. (2024). Estimators for topic-sampling designs. *Political Analysis*, *32*(4), 431–444. https://doi.org/10.1017/pan.2024.1

Coppock, A. (2022). *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Costello, T. H., Rabb, N., Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2025a). Reducing belief in conspiracy theories as they unfold using large language models. PsyArXiv preprint. https://doi.org/10.31234/osf.io/jnm9r_v1

Costello, T. H., Pennycook, G., Willer, R., & Rand, D. G. (2025b). Deep canvassing using AI. OSF preprint. https://doi.org/10.31219/osf.io/q7e6u_v1

Costello, T. H., Pennycook, G., & Rand, D. G. (2025c). Just the facts: How dialogues with AI reduce conspiracy beliefs. PsyArXiv preprint. https://doi.org/10.31234/osf.io/h7n8u_v1

Costello, T. H., Pelrine, K., Kowal, M., Arechar, A. A., Godbout, J.-F., Gleave, A., Rand, D., & Pennycook, G. (2026). Large language models can effectively convince people to believe conspiracies. arXiv preprint arXiv:2601.05050. https://doi.org/10.48550/arXiv.2601.05050

Cox, J. (2026, February 12). Cops are buying GeoSpy, an AI that geolocates photos in seconds. *404 Media*. https://www.404media.co/cops-are-buying-geospy-ai-that-geolocates-photos-in-seconds/

Czarnek, G., Orchinik, R., Lin, H., Xu, H. G., Costello, T., Pennycook, G., & Rand, D. G. (2025). Addressing climate change skepticism and inaction using human-AI dialogues. PsyArXiv preprint. https://doi.org/10.31234/osf.io/mqcwj_v1

Danet, D. (2025). LLM grooming: A new cognitive threat to generative AI. Working paper, Centre Géopolitique de la Datasphère. HAL: hal-05241525. https://hal.science/hal-05241525/

del Rio-Chanona, R. M., Laurentsyeva, N., & Wachs, J. (2024). Large language models reduce public knowledge sharing on online Q&A platforms. *PNAS Nexus*, *3*(9), pgae400. https://doi.org/10.1093/pnasnexus/pgae400

Deng, Yue, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. (2024). "Multilingual jailbreak challenges in large language models." *arXiv preprint arXiv:2310.06474*

DeVerna, M. R., Yan, H. Y., Yang, K.-C., & Menczer, F. (2024). Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, *121*(50), e2322823121. https://doi.org/10.1073/pnas.2322823121

Drotner, K. (1999). Dangerous media? Panic discourses and dilemmas of modernity. *Paedagogica Historica*, *35*(3), 593–619. https://doi.org/10.1080/0030923990350303

Espiritusanto, O., Nachawati-Rego, L., & Magallón-Rosa, R. (2024). The role of AI in citizen journalism, human rights activism, and monitoring: Limits and possibilities. In J. Sixto-García, A. Quian, A.-I. Rodríguez-Vázquez, A. Silva-Rodríguez, & X. Soengas-Pérez (Eds.), *Journalism, Digital Media and the Fourth Industrial Revolution* (pp. 211–226). Springer. https://doi.org/10.1007/978-3-031-63153-5_16

Evangelista, J. R. G., Sassi, R. J., Romero, M., & Napolitano, D. (2021). Systematic literature review to investigate the application of open source intelligence (OSINT) with artificial intelligence. *Journal of Applied Security Research*, *16*(3), 345–369. https://doi.org/10.1080/19361610.2020.1761737

Feldstein, S. (2019). The road to digital unfreedom: How artificial intelligence is reshaping repression. *Journal of Democracy*, *30*(1), 40–52.

Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L., Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M. E., Pan, J., Song, D., & Choi, Y. (2025a). Political neutrality in AI is impossible—but here is how to approximate it. arXiv preprint arXiv:2503.05728.

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., Pan, J., Tsvetkov, Y., & Reinecke, K. (2025b). Biased LLMs can influence political decision-making. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6559–6607). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.328

Freedom House. (2023). *Freedom on the Net 2023: The repressive power of artificial intelligence*. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

Gallegos, I. O., Shani, C., Shi, W., Bianchi, F., Gainsburg, I., Jurafsky, D., & Willer, R. (2025). Labeling messages as AI-generated does not reduce their persuasive effects. arXiv preprint arXiv:2504.09865.

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus*, *3*(2), pgae034. https://doi.org/10.1093/pnasnexus/pgae034

Gregory, S. (2023). Fortify the truth: How to defend human rights in an age of deepfakes and generative AI. *Journal of Human Rights Practice*, *15*(3), 702–714. https://doi.org/10.1093/jhuman/huad035

Grynaviski, J. D., & Corrigan, B. E. (2006). Specification issues in proximity models of candidate evaluation (with issue importance). *Political Analysis*, *14*(4), 393–420. https://doi.org/10.1093/pan/mpl003

Guess, A. M. (2021). (Almost) everything in moderation: New evidence on Americans' online media diets. *American Journal of Political Science*, *65*(4), 1007–1022. https://doi.org/10.1111/ajps.12589

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

Guterres, A. (2023, July 18). Secretary-General urges Security Council to ensure transparency, accountability, oversight, in first debate on artificial intelligence [Remarks to the UN Security Council]. United Nations. https://peacekeeping.un.org/en/secretary-general-urges-security-council-to-ensure-transparency-accountability-oversight-first

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, *121*(24), e2403116121. https://doi.org/10.1073/pnas.2403116121

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, *390*(6777), eaea3884. https://doi.org/10.1126/science.aea3884

Hopkins, D. J., Sides, J., & Citrin, J. (2019). The muted consequences of correct information about immigration. *Journal of Politics*, *81*(1), 315–320. https://doi.org/10.1086/699914

Hubbard, S. (2025, December 18). ICE uses a growing web of AI services to power its immigration enforcement and surveillance. American Immigration Council. https://www.americanimmigrationcouncil.org/blog/ice-uses-ai-immigration-enforcement-surveillance/

Imai, K., & Nakamura, K. (2024). Causal representation learning with generative artificial intelligence: Application to texts as treatments. arXiv preprint arXiv:2410.00903.

Jingnan, H. (2024, October 18). AI-generated images have become a new form of propaganda this election season. *NPR*. https://www.npr.org/2024/10/18/nx-s1-5153741/ai-images-hurricanes-disasters-propaganda

Karpathy, A. [@karpathy]. (2024, July 25). Jagged intelligence [Post]. X (formerly Twitter). https://x.com/karpathy/status/1816531576228053133

Kerley, B., Miller, C., and Campagnucci, F. Leveraging ai for democracy. In National Endowment for Democracy's International Forum for Democratic Studies. 2024. https://www.ned.org/wp-content/uploads/2024/10/NED_Leveraging-AI-for-Democracy-Report.pdf

Kowal, M., Timm, J., Godbout, J.-F., Costello, T., Arechar, A. A., Pennycook, G., Rand, D., Gleave, A., & Pelrine, K. (2025). It's the thought that counts: Evaluating the attempts of frontier LLMs to persuade on harmful topics. arXiv preprint arXiv:2506.02873.

Li, Zihao, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. "Language ranker: A metric for quantifying llm performance across high and low-resource languages." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, pp. 28186-28194. 2025.

Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading voters using human-artificial intelligence dialogues. *Nature*, *648*, 394–401. https://doi.org/10.1038/s41586-025-09771-9

Llansó, E. J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720920686

Lu, L., Tormala, Z. L., & Duhachek, A. (2025). How AI sources can increase openness to opposing views. *Scientific Reports*, *15*, 17170. https://doi.org/10.1038/s41598-025-00791-z

Lum, D. & Willis, H. (2026). Videos Show Moments in Which Agents Killed a Man in Minneapolis. New York Times, January 24, 2026. https://www.nytimes.com/2026/01/24/us/minneapolis-shooting-federal-agents-video.html

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, *22*(2), 216–242. https://doi.org/10.1111/jels.12413

Marcellino, W., Beauchamp-Mustafaga, N., Kerrigan, A., Chao, L. N., & Smith, J. (2023). *The rise of generative AI and the coming era of social media manipulation 3.0*. RAND Corporation, PE-A2679-1. https://www.rand.org/pubs/perspectives/PEA2679-1.html

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, *14*, 4692. https://doi.org/10.1038/s41598-024-53755-0

McGuffie, K., & Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. arXiv preprint arXiv:2009.06807.

Meta. (2025, October 7). Improving your recommendations on our apps with AI at Meta. *Meta Newsroom*. https://about.fb.com/news/2025/10/improving-your-recommendations-apps-ai-meta/

Motoki, F., Pinho Neto, V., & Rangel, V. (2025). Assessing political bias and value misalignment in generative artificial intelligence. *Journal of Economic Behavior & Organization*, *234*, 106904. https://doi.org/10.1016/j.jebo.2025.106904

Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. *Social Media + Society*, *5*(3). https://doi.org/10.1177/2056305119859294

Munger, K. (2023). Temporal validity as meta-science. *Research & Politics*, *10*(3). https://doi.org/10.1177/20531680231187271

Nanz, A., Binder, A., & Matthes, J. (2025). AI in the newsroom: Does the public trust automated journalism and will they pay for it? *Journalism Studies*, *26*(14), 1745–1764. https://doi.org/10.1080/1461670X.2025.2547301

Noels, S., Bied, G., Buyl, M., Rogiers, A., Fettach, Y., Lijffijt, J., & De Bie, T. (2025). What large language models do not talk about: An empirical study of moderation and censorship practices. In *Machine Learning and Knowledge Discovery in Databases: Research Track, ECML PKDD 2025, Lecture Notes in Computer Science* (vol. 16013). Springer. https://doi.org/10.1007/978-3-032-05962-8_16

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, *42*(3), 939–960. https://doi.org/10.1007/s11109-019-09528-x

O'Brien, M., & Toropin, K. (2026, March 5). Pentagon says it is labeling AI company Anthropic a supply chain risk 'effective immediately'. Associated Press. https://apnews.com/article/pentagon-anthropic-supply-chain-risk-ai-030625

OpenAI. (2025, October 9). Defining and evaluating political bias in LLMs. https://openai.com/index/defining-and-evaluating-political-bias-in-llms/

Orben, A. (2020). The Sisyphean cycle of technology panics. *Perspectives on Psychological Science*, *15*(5), 1143–1157. https://doi.org/10.1177/1745691620919372

Pan, J., & Xu, X. (2026). Political censorship in large language models originating from China. *PNAS Nexus*, *5*(2), pgag013. https://doi.org/10.1093/pnasnexus/pgag013

Papathanassopoulos, S., & Giannouli, I. (2025). Political communication in the age of platforms. *Encyclopedia*, *5*(2), 77. https://doi.org/10.3390/encyclopedia5020077

Park, J., & Kim, J. (2024, March 18). South Korea's Yoon warns of tech threat to democracy at summit. *Reuters*. https://www.reuters.com/world/asia-pacific/south-koreas-yoon-warns-tech-threat-democracy-summit-2024-03-18/

Pauwels, E. (2024). *Preparing for next-generation information warfare with generative AI* (CIGI Paper No. 310). Centre for International Governance Innovation. https://www.econstor.eu/handle/10419/311791

Pava, Juan N., Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Vukosi Marivate, and Sanmi Koyejo. "Mind the (language) gap: Mapping the challenges of LLM development in low-resource language contexts." *Stanford Institute for Human-Centered Artificial Intelligence (HAI): Stanford, CA, USA* (2025).

Piccardi, T., Saveski, M., Jia, C., Hancock, J., Tsai, J. L., & Bernstein, M. S. (2025). Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, *390*(6776), eadu5584. https://doi.org/10.1126/science.adu5584

Potter, Y., Lai, S., Kim, J., Evans, J., & Song, D. (2024). Hidden persuaders: LLMs' political leaning and their influence on voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*. arXiv:2410.24190.

Puczyńska, J., Podhajski, M., Wojtasik, K., & Michalak, T. P. (2024). Large language models in jihadist terrorism and crimes. *Terrorism—Studies, Analyses, Prevention*, (5), 351–379. https://doi.org/10.4467/27204383TER.24.012.19400

Qi, T., Liu, H., & Huang, Z. (2025). An assistant or a friend? The role of parasocial relationship of human-computer interaction. *Computers in Human Behavior*, *167*, 108625. https://doi.org/10.1016/j.chb.2025.108625

Qiu, P., Zhou, S., & Ferrara, E. (2025). Information suppression in large language models: Auditing, quantifying, and characterizing censorship in DeepSeek. arXiv preprint arXiv:2506.12349.

Quinn, M., & Gutt, D. (2025). Heterogeneous effects of generative artificial intelligence (GenAI) on knowledge seeking in online communities. *Journal of Management Information Systems*, *42*(2), 370–399. https://doi.org/10.1080/07421222.2025.2487313

Roberts, M.E., Stewart, B.M., & Nielsen, R.A. (2020). "Adjusting for Confounding with Text Matching." *American Journal of Political Science* 64(4): 887–903.

Robinson, N., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 392–418). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.40.pdf

Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE*, *19*(7), e0306621. https://doi.org/10.1371/journal.pone.0306621

Rozado, D. (2025). Measuring political preferences in AI systems: An integrative approach. arXiv preprint arXiv:2503.10649.

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, *9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Sanderson, Z., Messing, S. and Tucker, J.A. (2024). Misunderstood mechanics: How AI, TikTok, and the liar's dividend might affect the 2024 elections. Brookings Institution. https://www.brookings.edu/articles/misunderstood-mechanics-how-ai-tiktok-and-the-liars-dividend-might-affect-the-2024-elections/

Sanderson, Z., Zhong, W. and Tucker, J.A. (2025). It Works When It Works: Measuring The Direct And Indirect Effects Of AI Labels On Political Images. https://osf.io/preprints/socarxiv/nf785_v1

Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review*, *119*(1), 71–90. https://doi.org/10.1017/S0003055423001454

Schneier, B., & Sanders, N. (2024, December 2). The apocalypse that wasn't: AI was everywhere in 2024's elections, but deepfakes and misinformation were only part of the picture. *The Conversation*. Republished by Harvard Ash Center. https://ash.harvard.edu/articles/the-apocalypse-that-wasnt-ai-was-everywhere-in-2024s-elections-but-deepfakes-and-misinformation-were-only-part-of-the-picture/

Sen, R., & Farooq, N. (2025). AI-driven digital transnational repression: Past lessons, present challenges, and future directions. In M. Hasan & A. E. Ruud (Eds.), *The long reach of the strong arm: Evolving forms of transnational authoritarianism* (pp. 31–59). Palgrave Macmillan. https://doi.org/10.1007/978-3-032-04940-7_3

Setty, Riddhi (2026). In This Cleveland Newsroom, AI Is Writing (But Not Reporting) the News. Columbia Journalism Review, February 25, 2026. https://www.cjr.org/news/cleveland-newsroom-ai-rewrite-desk-chris-quinn-plain-dealer.php

Shambaugh, Scott (2026). An AI Agent Published a Hit Piece on Me, February 12, 2026. https://theshamblog.com/an-ai-agent-published-a-hit-piece-on-me/

Sharma, N., Murray, K., & Xiao, Z. (2025). Faux polyglot: A study on information disparity in multilingual large language models. arXiv preprint arXiv:2407.05502.

Similarweb. (2025). *GenAI and how it's impacting US publishers*. Similarweb. https://www.similarweb.com/corp/reports/generative-ai-publishers/

Smith, M. G., Bradbury, T. N., & Karney, B. R. (2025). Can generative AI chatbots emulate human connection? A relationship science perspective. *Perspectives on Psychological Science*, *20*(6), 1081–1099. https://doi.org/10.1177/17456916251351306

Spearing, E. R., Gile, C. I., Fogwill, A. L., Prike, T., Swire-Thompson, B., Lewandowsky, S., & Ecker, U. K. H. (2025). Countering AI-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science*, *12*(6), 242148. https://doi.org/10.1098/rsos.242148

Starker, S. (1989). *Evil influences: Crusades against the mass media*. Transaction Publishers.

Summerfield, C., Argyle, L. P., Bakker, M., Collins, T., Durmus, E., Eloundou, T., Gabriel, I., Ganguli, D., Hackenburg, K., Hadfield, G. K., Hewitt, L., Huang, S., Landemore, H., Marchal, N., Ovadya, A., Procaccia, A., Risse, M., Schneier, B., Seger, E., Siddarth, D., Sætra, H. S., Tessler, M. H., & Botvinick, M. (2025). The impact of advanced AI systems on democracy. Nature Human Behaviour, 9(12), 2420–2430. https://doi.org/10.1038/s41562-025-02309-z

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Thekkethil, D. (2025, July 7). Similarweb: Zero-click searches surge to 69% since Google AI Overviews launched. *Stan Ventures*. https://www.stanventures.com/news/similarweb-zero-click-search-surge-google-ai-overviews-3562/

TikTok. (2024). Our approach to content moderation. TikTok Transparency Center. https://www.tiktok.com/transparency/en-us/content-moderation

Tomić, Z., Damnjanović, T., & Tomić, I. (2023). Artificial intelligence in political campaigns. South Eastern European journal of communication, 5(2), 17-28. https://doi.org/10.47960/2712-0457.2.5.17

Trump, D.J. (2025). Preventing Woke AI in the Federal Government. White House executive order, July 23, 2025. https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government

Velez, Y. R. (2025). Crowdsourced adaptive surveys. *Political Analysis*, *33*(4), 284–297. https://doi.org/10.1017/pan.2024.34

Velez, Y. R., & Liu, P. (2024). Confronting core issues: A critical assessment of attitude polarization using tailored experiments. *American Political Science Review*, *119*(2), 1036–1053. https://doi.org/10.1017/S0003055424000819

Velez, Y. R., Green, D. P., & Sevi, S. (2025). Chatbot Voting Advice Applications inform but seldom sway young unaligned voters. *Proceedings of the National Academy of Sciences*, *122*(50), e2515516122. https://doi.org/10.1073/pnas.2515516122

Velez, Y. R., Liu, P., & Clifford, S. (2026). When information affects attitudes: The effectiveness of targeting attitude-relevant beliefs. APSA Preprints. https://preprints.apsanet.org/engage/apsa/article-details/67e7f3f981d2151a02322745

Wack, M., Ehrett, C., Linvill, D., & Warren, P. (2025). Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign. *PNAS Nexus*, *4*(4), pgaf083. https://doi.org/10.1093/pnasnexus/pgaf083

Waight, H., Yang, E., Yuan, Y., Messing, S., Roberts, M., Stewart, B., & Tucker, J. (2026). State Media Control Influences Large Language Models. *Nature* [Forthcoming].

Westwood, S. J., Grimmer, J., & Hall, A. B. (2025). Measuring perceived slant in large language models through user evaluations (Working Paper No. 4262). Stanford Graduate School of Business. https://www.gsb.stanford.edu/faculty-research/working-papers/measuring-perceived-slant-large-language-models-through-user

Williams, A. R., Burke-Moore, L., Chan, R. S.-Y., Enock, F. E., Nanni, F., Sippy, T., Chung, Y.-L., Gabasova, E., Hackenburg, K., & Bright, J. (2025). Large language models can consistently generate high-quality content for election disinformation operations. *PLOS ONE*, *20*(3), e0317421. https://doi.org/10.1371/journal.pone.0317421

Wilson, Alex. 2026. "Instagram Algorithm 2026- What Changed, What Works, What to Stop Doing." OrangeMonke. https://orangemonke.com/blogs/instagram-algorithm/.

Wirtschafter, V., & Nadgir, N. (2025, October 16). Is the politicization of generative AI inevitable? Brookings Institution. https://www.brookings.edu/articles/is-the-politicization-of-generative-ai-inevitable/

Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., & Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus*, *4*(6), pgaf170. https://doi.org/10.1093/pnasnexus/pgaf170

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Yang, E., & Roberts, M. E. (2023). The authoritarian data problem. Journal of Democracy, 34(4), 141–150.

Yildiz, Guney. (2026). Inside Moltbook: The Social Network Where AI Agents Talk And Humans Just Watch. Forbes. https://www.forbes.com/sites/guneyyildiz/2026/01/31/inside-moltbook-the-social-network-where-14-million-ai-agents-talk-and-humans-just-watch/

Yong, Zheng-Xin, Cristina Menghini, and Stephen H. Bach. "Low-resource languages jailbreak gpt-4." *arXiv preprint arXiv:2310.02446* (2023).

Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, *12*(1), 415–438. https://doi.org/10.1146/annurev-economics-081919-050239