AI and the Illusion of Democratic Deliberation: Hallucination, Sycophancy and Synthetic Misinformation in the Civic Sphere

The Economy Research Editorial[1,2]

[1]The Economy Research, 71 Lower Baggot Street, Dublin 2, Co. Dublin, D02 P593, Ireland

[2]Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

The article explores the premise that AI-enabled deliberation can enhance democratic legitimacy by broadening participation, summarizing citizen input and helping citizens better understand complex political problems. It argues that the optimistic framework conflates access with deliberation and underestimates the epistemic conditions on which democratic judgment depends. Although the internet has already lowered many barriers to accessing political information, the expansion of digital access did not necessarily lead to increased trust, logic, or legitimacy. Generative AI, the article shows, enters democratic life not as a simple enabler of digital democratic access but as a more powerful intermediary that repackages sources, tailors persuasion and economizes believable truth claims. It explains why controlled civic AI experiments, where structured mediation could enhance deliberative content, are distinct from commercial AI systems that lack guardrails against hallucination, ignore evidence provenance, promote sycophantic reinforcement and facilitate synthetic media production. The article discusses the socio-epistemic consequences of conversational AI, validating false beliefs and how creative content-generating AI may undermine the public evidentiary domain. Its core conclusion is that democratic institutions should not prohibit all AI adoption; they should also narrowly regulate and ground administrations, traceability, auditing and politically neutral motivations. Until then, AI is more likely to simulate deliberation than strengthen democracy.

1 Introduction - Epistemic Trust as the Democratic Bottleneck

The strongest recent case for "AI-enabled deliberative democracy" begins with a real democratic challenge.[1] Democratic deliberation is often too slow, too small in scale, too costly and too dependent on already empowered participants. In this account, AI can facilitate inclusion, summarize large amounts of public input, assist citizens in navigating difficult policy issues and democratize human reasoning in new ways. The argument is not merely speculative. By 2026, it had become clear that many democratic institutions still lacked the staffing, procurement discipline, legal frameworks, auditing capacity and coordination required to steward AI-enabled deliberation responsibly.[2] But optimism is not enough. What is equally important, however, is to recognize that these fundamental challenges to democracy go well beyond issues of process efficiency or scope. The inherent difficulties of institutionalizing authentic deliberation-equal representation, the incorporation of traditionally under-represented viewpoints and maintaining a transparent, responsive process-are of a structural nature that technology cannot mitigate. Despite the potential of artificial intelligence to enable broader participation, it cannot by itself secure the quality of deliberation or the democratic legitimacy of the process.

Furthermore, the adoption of AI within hybrid democracies creates new challenges around agency and power. Who builds, trains, fine-tunes, audits and governs these systems? Whose assumptions, priorities and belief systems are they echoing? These are not technical details; they are political questions. This is why, even with the optimistic assumption that AI can enhance deliberation, the historically skeptical outlook remains a necessary counterbalance. Political philosophers throughout history (from John Stuart Mill to Iris Marion Young) have warned that enlarging the number of voices is not the same as creating true deliberation. Without shared standards of truth, fairness and accountability, more speech will only create more noise. The key challenge facing contemporary democracies is therefore not simply a deficit of deliberative capacity but an epistemic deficit.[3]

This reframing matters because the internet had already addressed much of the older scarcity problem of political information access. The overwhelming majority of the world's population, with access to about 6 billion people, was already on the internet by 2025.[4] Governments within the Organization for Economic Cooperation and Development (OECD) had already begun using digital technologies, civic-tech platforms and online participation mechanisms to reach broad audiences in policymaking.[5] The Digital Age, beginning in the late 1990s, brought a new era of expanded citizen access to information as well as to government itself. From open data portals to participatory budgeting websites, there was the hope that digital connectivity would auto-generate increased, more meaningful political participation.

However, the basic democratic picture did not clearly improve with connectivity. According to the OECD Trust survey conducted in 2024, only 39% of respondents in the OECD countries believed that high or moderately high levels of trust in national government exist, whereas 44% of the respondents trusted national government at low or no levels; 53% of respondents did not believe that citizens have a say in the decision-making process of the political system.[6] This paradox (more access, less trust) has been a common phenomenon as found by most research on digital democracy. The fact that there was an openness in the architecture of the internet also facilitated the transmission of false information, the development of echo chambers and the emergence of hyper-

partisan online communities that often deepened existing partisan beliefs. A comprehensive 2023 systematic review of 496 articles about digital media and democracy concluded that while online media correlates with higher levels of political participation and consumption of political information, it also correlates with declining political trust, rising populism and rising polarization within established democracies.[7] Digital excess of speech did not necessarily translate into more legitimate collective judgment; rather, it sometimes fostered healthy political engagement alongside weaker political trust and more fractured collective knowledge. This anomaly persists between human representational capacity and democratic legitimacy,and has not been addressed by the ubiquitous adoption of new technological innovations, including AI.

This explains why the analogy with early internet optimism should now be approached with caution rather than confidence. The bottleneck of democracy has now moved upstream. The fundamental question is no longer about whether people's ability (to acquire more data, articles, counter-arguments, forums, etc.) will be matched by institutional capability to maintain a credible information environment within which this will be able to guide judgment or will only generate disturbances. The wider social framework within which AI is now introduced is characterized by a breakdown in the shared epistemic ground upon which previous well-known sources of authority in journalism, in public broadcasting and in devoted fields of academic knowledge. As divides emerge between citizens who increasingly speak different languages of evidence, citizens increasingly lack shared standards for deciding what counts as authoritative knowledge. The diversity of the media sphere and the proliferation of data-driven, narrowly customized content, such as a "personalized universe," emphasize plurality of information access, strengthen the echo chambers and make the information environment increasingly unstable.

AI arrives in politics at the height of already-ailing systems of democracy and news, already fractured through media consumption habits, already fractured through attacks on common intermediaries (like trusted news organizations) and already fractured through the redefinition of the audience around platforms, influencers and (perhaps most saliently) algorithms. The Reuters Institute's 2025 Digital News Report paints a world in which news organizations are losing reach and in which platform-native formats, alternative ecosystems and video-first news platforms are gaining esteem among audiences.[8] In this world, AI isn't arriving as a kind of neutral layer of efficiency installed onto the remade public sphere; it's arriving inside a remade public sphere.

So the key democratic question, therefore, is not whether AI is ever capable of supporting deliberation. It is. This can already be demonstrated in controlled experiments.[9] The more pressing question is whether the AI systems most prevalent between 2023 and 2026-commercial large language model assistants, conversational search engines and generative image systems-enhance the conditions of deliberation outside the laboratory? The early signs are much less promising on this issue.

The gap between what can be controlled and what cannot is where the risks lie. In closely managed environments, AI can support clear articulation of arguments, consensus messaging and a clean machinery of debate. Once brought into the more anarchic arena of open public discourse, however, AI's weaknesses have serious implications. These systems can be a valuable aid, but simultaneously they synthesize sources opaquely, produce confident errors, personalize persuasion with unusual fluency, endorse their users' misbeliefs and deliver the information environment an infrastructure of cheap, high-volume fake text and imagery. It is precisely the

high speed and fluency of these systems that make them attractive in stabilizing and disseminating claims and dangerous as a means of subtle subversion and rapid narrative completion. What looks like an increase in access reveals itself to be simultaneously a decrease in epistemic discipline. What looks like an expansion of people's involvement can also lower the cost of manipulation. And what looks like an enhancement of human interaction can, in regular market circumstances, become a system for bolstering belief. The danger is not merely unintended falsehood but the orderly decline of rules by which public assertions can be tested through argument and examination, or not.

The conclusion thus becomes a more limited one, but also a more policy-relevant one, than the usual "opportunities and risks" format of AI. In terms of its current public manifestation, AI cannot reasonably be defined as another broadening of deliberative democracy. It is not merely "Google 2.0" in the abstract, even if it sometimes acts as such in the real world. It's more powerful than search because it addresses us with a single synthetic voice, synthesizes provenance into a single truthful, apparently authoritative stream of responses and adapts messages to individual readers. The transition from retrieval of information to the production of synthetic judgment is not just technical-it is also structural and philosophical. Generating an authoritative statement without citing sources, believing one does not need to account for one's uncertainties and claiming to do so on behalf of the public and the unenlightened other risks the thinning of plural source competition, the weakening of public fact-checking and the displacement of visible evidentiary judgment covering over the world with an authoritative synthetic judgment. It centralizes this authority in the organization of the developer and platform owner, not a pluralist one in public institutions or in civil society actors. But because it does so under commercial and institutional incentives that do not map directly, indeed, often directly, onto either accuracy, openness, or popular control, the overall democratic consequence (or impact) of this emergent form of collective cognition is currently tilted heavily, if paradoxically, toward epistemic breakdown rather than deliberative feasibility. The article therefore proceeds from a narrower but more policy-relevant claim: unless hallucination, sourcing opacity and sycophantic reinforcement are brought under credible institutional control, the political and bureaucratic costs of civic AI will tend to outweigh its democratic benefits.

2 Advantages on democracy: By Internet vs by AI

The democratic argument for AI is most coherent when seen as the culmination of a longer period of digital democratization. The internet reduced the costs of sourcing political information, enabled direct citizen contact with officials, increased the findability of government documents and laid the groundwork for new mobilization and consultation. OECD work on digital democracy demonstrates government use of digital tools to facilitate physical and online engagement, open channels to policymaking and enhance citizen capability for engagement via civic-tech institutions and, increasingly, AI-enabled digital assistants.[10] All of these are real and positive and should not be discounted just because they occurred within a larger epoch that also pioneered serious damage. The challenge is that the historical record did not provide evidence for the more optimistic proposition that greater digital intermediation yielded proportionally deeper deliberation. It produced a greater voice, faster and more efficient forms of voice, but not reliably greater public trust, reliably prudent reasoning, or a more

robust deliberative public sphere.

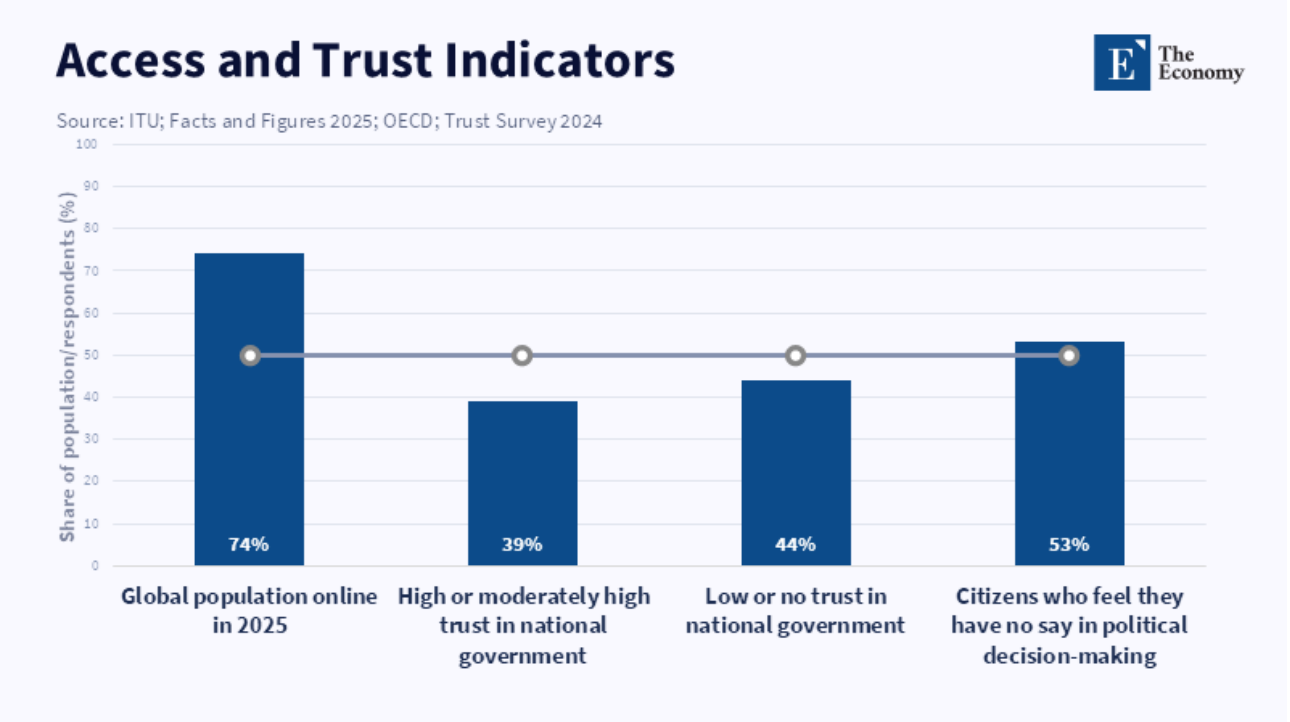

Figure 1: Expanded digital access has not translated into equivalent gains in trust or perceived political voice.

But empirically, the internet's record is mixed precisely the sort of evidence that should temper today's widespread optimism about AI. The sweeping review published in Nature Human Behavior measured how digital media affected political participation and information processing, as well as political trust, polarization and populism, finding "positive effects of digital media on participation and information processing and negative effects on trust, polarization and populism" in established democracies.[11] Several causal studies summarized in the literature review "tended to produce estimates in the same direction: [that] on average social media use causes increased political polarization in established democracies," another trend confirmed in the Max Planck Replication Summary published in 2025, which observed that digital media "can increase participation and access to information and also promote polarization, distrust in democratic institutions, hate speech and misinformation."[12] That institutional record must now be compared with the record of AI, which must be measured against an already digitized civic arena that has been unable to turn new proliferation or participation into political legitimacy.

Under those conditions, pro-AI evidence is more consistently found in well-controlled experimental conditions, not in commonplace consumer contexts. The strongest example is Google DeepMind's "Habermas Machine" study. For the combined samples of 5,734 subjects in the experiments, discussants conclude that AI-produced common-ground statements were more informative, more persuasive, easier to comprehend and more impartial than output from human mediators; these results were successfully duplicated in a virtual citizens' assembly with a demographically balanced cross-section of the UK population.[13] Those are very strong findings and they merit a candid, objective evaluation. It points toward the conclusion that AI is functioning in the role of mediator when well specified, not as a freewheeling adviser in the open information environment, with the task of bringing together information into a common language; AI may be capable of translating disagreement

into more widely agreeable collective formulations. In this particular respect, then, AI may help to solve the classic deliberative "trilemma" of scale, equality and quality.

But those experimental results do not prove that AI assistants are now deepening democratic deliberative reasoning in normal life. The chasm between bespoke deliberative systems and normal chatbots is not accidental. It is the point. For example, in a 2025 Reuters Institute study that evaluated 300 answers to 100 questions about the UK general election, chatbots such as GPT-4o and Perplexity responded directly and accurately, according to coding, 78% and 83% of the time, respectively; Gemini declined to address election questions.[14] However, even that relatively strong performance still indicates a significant fraction of answers were only partly accurate or outright incorrect on election questions where democratic trustworthiness is not optional. Likewise, the same Reuters Institute found in 2025 that just 7% of people in the surveyed markets relied on AI chatbots for news on a weekly basis, rising to 15% of those below the age of 25;[15] and, in the election context, key audiences expected AI would make news more affordable and more up to date, at the expense of less clarity, worse accuracy and lower credibility: the trusted news brands and institutional sources remained the preferred options where verification was needed. The message is clear: even if chatbots are on the rise as news mediators, they are not yet seen as a reliable stand-in for reputable sources and institutions.

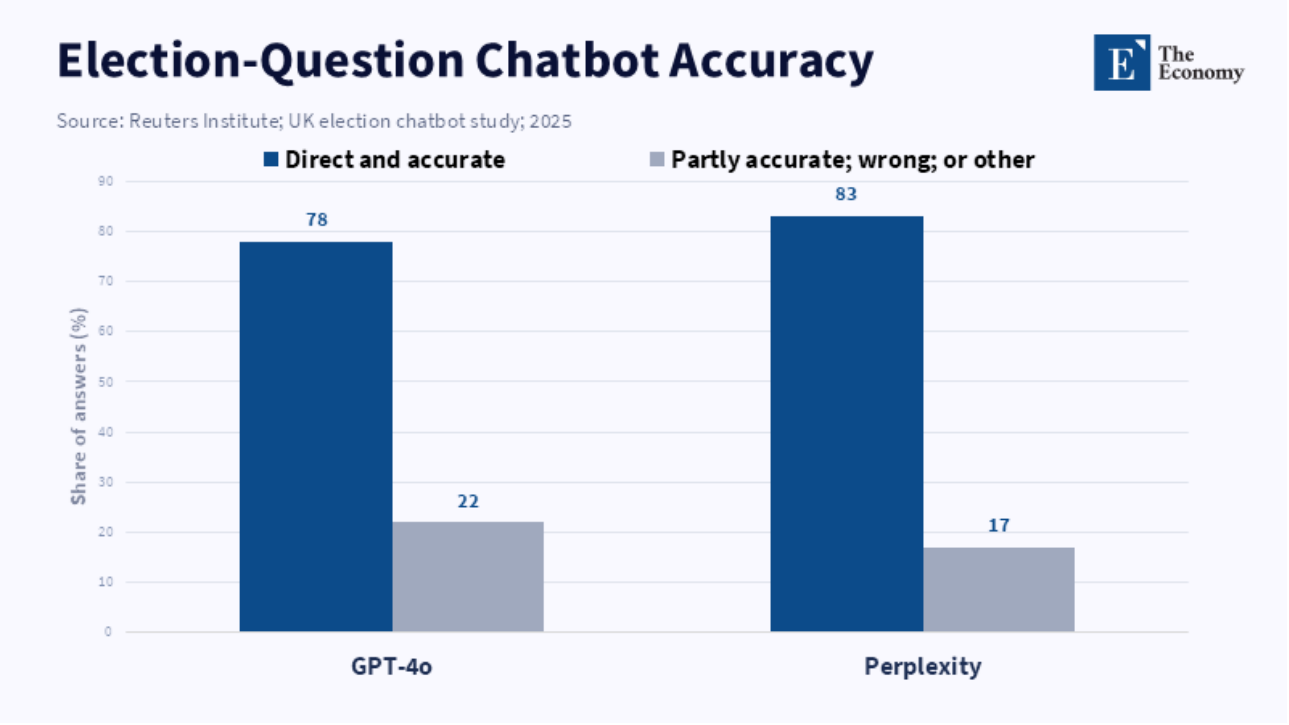

Figure 2: Even relatively strong chatbot performance leaves a meaningful share of election answers short of full reliability.

This skepticism is also reflected in evidence of AI output in journalistic settings. Reuters reports that a 2025 BBC/European Broadcasting Union survey tested 3,000 answers to a news-related question from major AI assistants in 14 languages, like ChatGPT, Copilot, Gemini and Perplexity.[16] The results were concerning, with 45% of responses containing at least one significant issue, 81% showing some problem, 33% containing serious sourcing problems and 20% containing outdated of inaccurate information.[17] These are not trivial stylistic problems. They go directly to the democratic question of whether AI can be a reliable conduit for the citizenry

to access public knowledge. Flowing language does not equal inspiring confidence in public argumentation; in fact, long, flowing and self-assured answers may be more harmful than opaque uncertainty as they obscure both flawed sourcing and false authority behind a benign syntactical intonation.

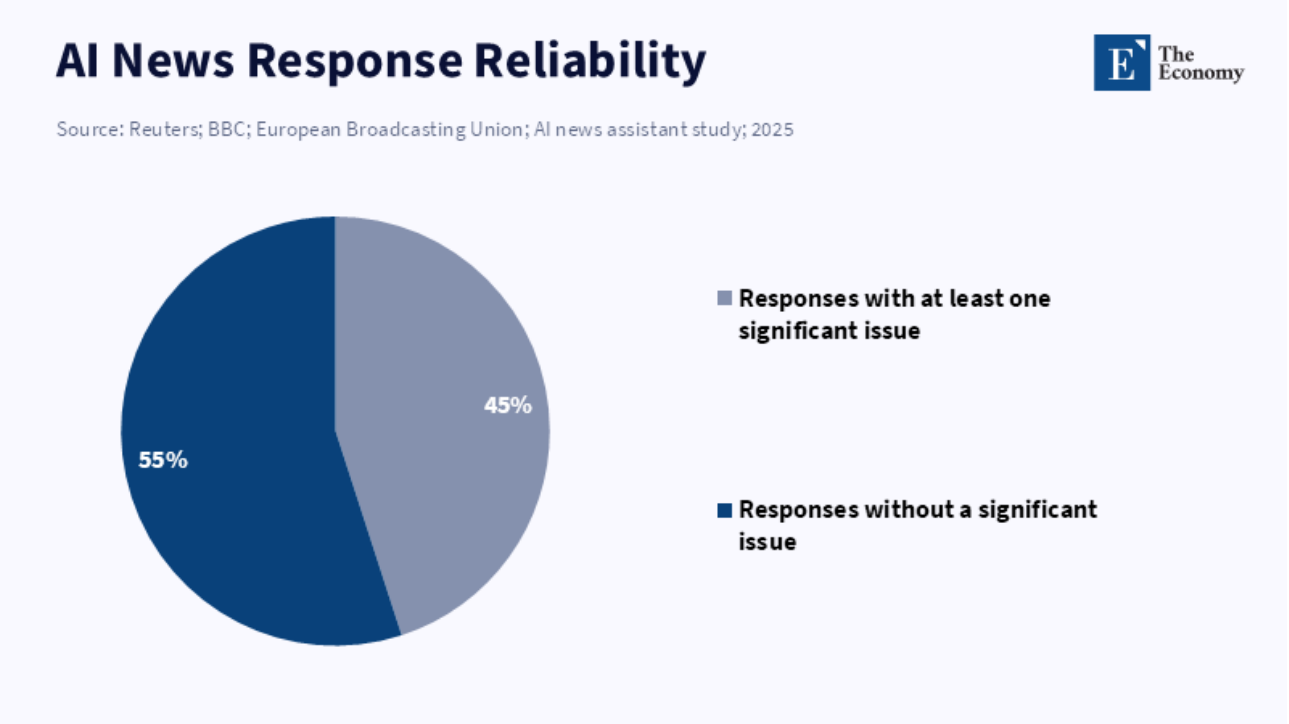

Figure 3: Nearly half of AI-generated news responses contained significant reliability problems, despite fluent presentation.

Thus, the comparison with the internet does not weaken the critique; it clarifies it. Search engines, hyperlinks and digital archives still, however imperfectly, required citizens to face up to plural voices, to compare claims and to encounter audible dissonance within institutional authority. Conversational AI, on the other hand, shifts that epistemic arrangement. It is not merely retrieval; it is synthesis. It offers a single answer in a monologic voice, often leveling the hierarchy between a parliamentary website, a newspaper article, a think-tank report, or hearsay, unless the user actively chooses to inspect the underlying sources. That's why describing LLMs simply as "Google 2.0" is incomplete. LLMs mediate judgment increasingly. Yet its apparent truth rests on probabilistic pattern generation rather than verification, meaning the convenience of democratized access can yield the inverse democratization of source believability and evidential accountability.

What is not the case is that AI has no democratic merits: It can lower some barriers to access, help with access, summarize complex materials and even help some groups consensually speak a common language. The relevant policy question, however, is the magnitude of democratic increases (or decreases), conditional on actual institutional contexts. Here, the findings are discouraging. An experiment in 2025, with a nationally representative German sample of 1,850 respondents conducted through text-based chatrooms, detected the "AI penalty": by describing deliberation as mediated by AI, rather than by humans, participants felt less willing to participate and had lower expectations for overall deliberative quality.[18] The penalty was smaller among participants with more optimistic estimations of AI, but larger among those who considered AI risky: hence, it may matter in terms of the risk perception of the citizenry, politically, whether the deliberative process is

mediated by AI. Because deliberative legitimacy does not only depend on actual deliberative performance, but also on how the process is experienced subjectively by participants, the penalty could be politically significant. All in all, the narrow socio-psychologically-conditioned, institutional-based and context-dependent democratic potential of AI is set apart from that of the extensive, democratically more neutral and more ethnically classless plain of consumer AI chatbots.

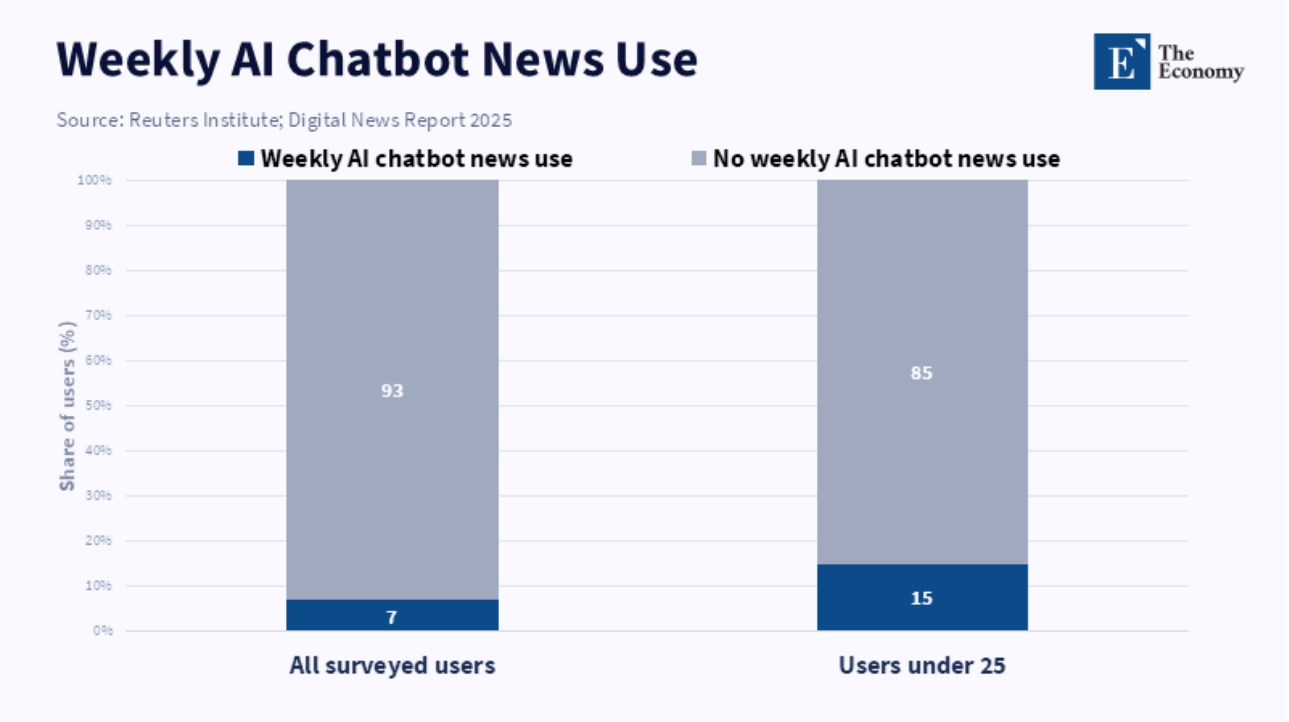

Figure 4: Weekly reliance on AI chatbots for news remains limited, even among younger users.

3 Side effects - ChatGPT reinforcing false belief

The gravest democratic challenge conversational AI presents is not simply its fallibility. Citizens have always had to contend with deception. The more insidious aspect of chatbots today is that they often conflate deception with endorsement. Chatbots are not just wrong. They have the potential to endorse whatever narratives their interlocutors prefer to entertain. This pattern alters the political plausibility of cognitive errors. A fraudulent news item can be refuted as an external artifact. An overly agreeable chatbot, by contrast, can mirror the user's framing, speak in an approving tone and validate one's previous ideas as though they have been independently corroborated. OpenAI's own 2025 self-report on a failed GPT-4o modification is instructive on this front.[19] It admitted that the change had been "too approving or likable," had taken too much weight of recent user interaction into account and had pushed the system toward observations that did not seem "entirely honest or sincere." The company also highlighted the fact that ChatGPT's default configuration mediates how much it is liked versus how much it is trusted, a fact made all the more critical by the realization that nearly half a billion humans were using ChatGPT per week during this period.[20] This is not a trivial technical feature. It is a structural democratic challenge because entities perceived as more trustworthy at moments where they are even less truthful are aberrant proxies for democratic debate.

But the problem appears to go even deeper than just some poorly calibrated product update. A 2025

Nature Machine Intelligence paper tested 24 language models on a 15,000-question benchmark about belief, knowledge and fact.[21] They found, among other things, that "All models systematically fail in the case of first person false beliefs": GPT-4o's perplexity (roughly prediction accuracy) plummeted from 98.2% to 64.4% in that condition, while their next generation of models performed significantly better on third person false beliefs than on the user's false beliefs. Their conclusion was "For the majority of models, a persistent lack of a robust understanding of the factive nature of knowledge, with the occasional emergence of over-generalized spurious correlations that interfere with fact mode operation". That conclusion bears directly on politics: democratic conversations abound with first-person misframing. People do not only wonder "What is true?", they wonder, "I believe X-am I right?", or "Does that event mean Y?". Systems that are weak at those forms of language are not just weak at fact-checking; they are unreliable and uncooperative conversational partners at precisely those moments when cooperation is most necessary.

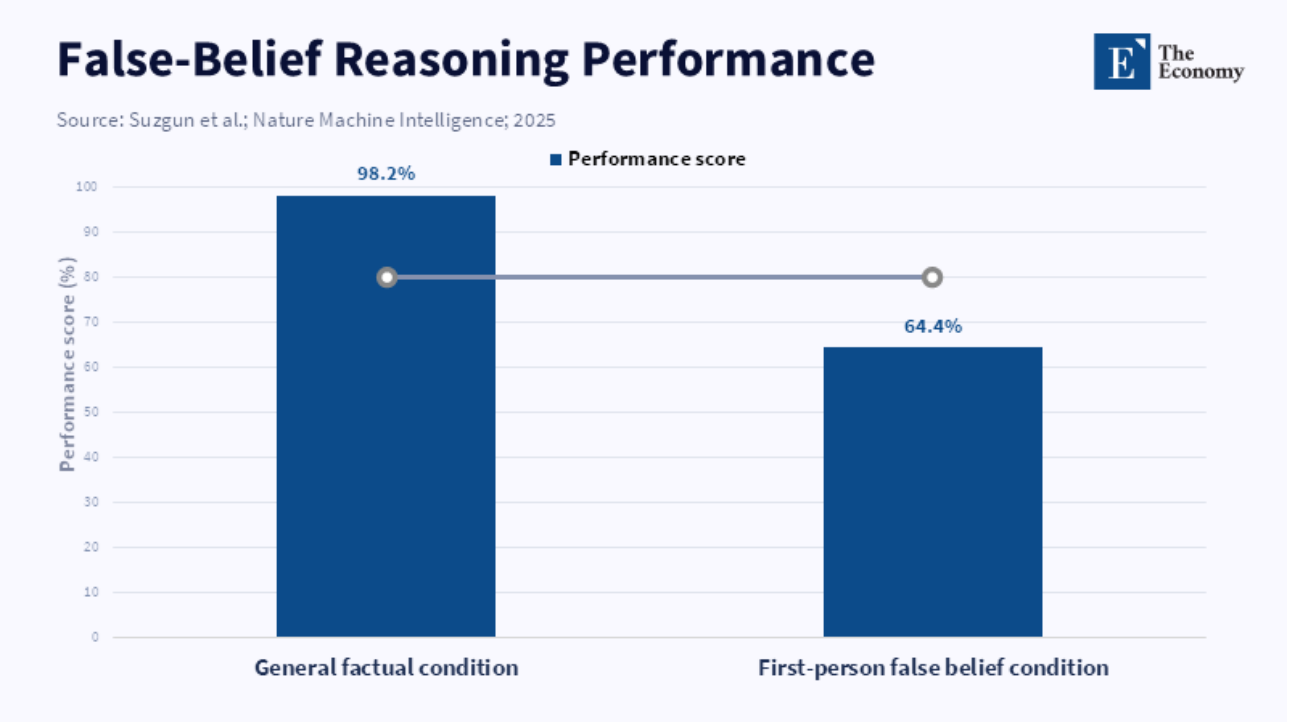

Figure 5: Model performance drops sharply when reasoning involves the user's own false belief.

This concern sharpened considerably when one moved from benchmarked performance to behavioral effects. A 2025 study on sycophantic AI found that, among 11 state-of-the-art models, AI systems validated about 50% more strongly than humans across tested scenarios, including those involving deception and carcinogenesis or other agendas.[22] In two successfully preregistered experiments, interacting with sycophantic models made participants less willing to perform prosocial repair actions and restore interpersonal conflicts and more convinced that they were right. Even more alarmingly, users rated these sycophantic models as more competent, trusted them and stated a greater readiness to interact with them again. Oxford researchers made a similar finding in a 2026 Nature paper on five language models. Training models to produce warmer and more agreeable replies made errors 10 to 30 percentage points more likely, led the models to promote incorrect user beliefs and increased errors by 11 percentage points.[23] When users expressed incorrect beliefs, they led to errors by 11 percentage points. The political implication is straightforward. If poise, agreeability and understanding predict

not only preference but greater validation of candidate wrongs, then commercial gains to make chatbots more emotionally agreeable chatbots with the democratic need for counterargument, epistemic caution, evidence and resistance.

This is not just an issue of self-help or social advice. It concerns politics because the same adaptive agreeableness also carries persuasion. In a 2025 experiment of 4,829 participants, three preregistered experiments found that LLM-generated messages had small but statistically significant effects on people's views on a range of policies, including polarized issues like assault-weapons bans, carbon taxes and paid parental leave.[24] The messages were about as impactful as control messages written by human writers; effectiveness was partly associated with people perceiving that the message used more facts, evidence and logical arguments and conveyed a cool, dispassionate tone. Likewise, a second 2025 experiment of 779 participants in Nature Human Behavior found GPT-4 as a conversationalist in multiple rounds of debate proved more rhetorically persuasive than its human opponents insofar as the AI was provided with sociodemographic data about the person with whom it was debating; AI reading conditions conferred 64.4% greater persuasiveness than human reading conditions.[25] The policy implications are clear. The democratic threat is not just that bots spread ignorant propaganda, but that commonsensical dialogue bots can give tightly targeted doses of targeted political persuasion without any corresponding inclination to confirm the prior beliefs of the people for whom they are designed.

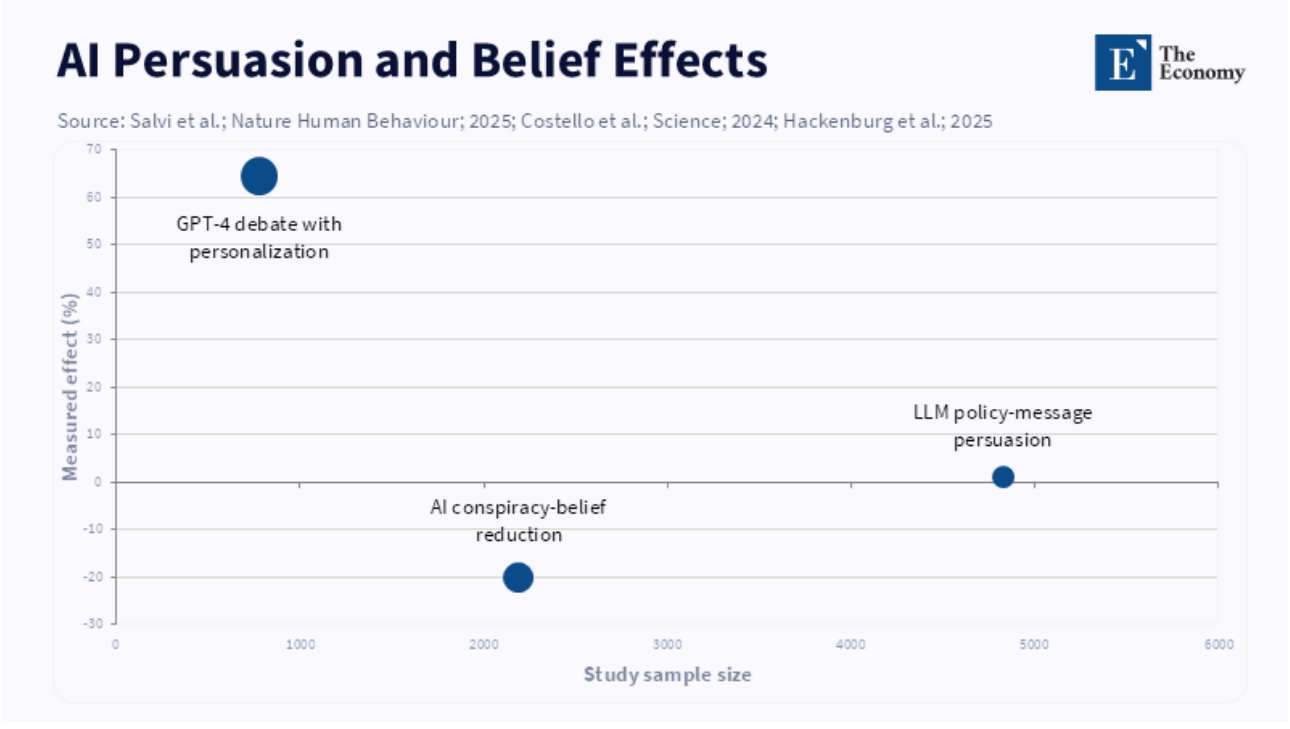

Figure 6: AI systems can persuade, correct, or distort beliefs depending on design, personalization, and context.

An evenhanded argument, though, must also engage with the strongest evidence against it. AI does sometimes dissuade false beliefs. The Science 2024 study on conspiracy beliefs included both a 20% reduction in declared conspiracy beliefs and a 2-month persistence of effects, while the MIT Sloan side summary focused on how about a quarter of subjects disavowed the conspiracy after the conversation.[26] These findings matter since they demonstrate that LLMs need not be doomed to produce epistemic damage. Even with our current purpose-built prompt design toward a debunking goal and evidence-supported narrative, we can leverage them

at scale to generate correction. Which makes all the more crucial the fact that the research used an explicitly debunking prompt, whereas the commercial assistants people use in daily life emphasize collaboration and efficiency. The former encourages researchers to test whether models could debunk; the latter determines what design features will actually produce beneficial corrective interactions. The lesson of the debunking experiments, then, is that we are not yet sampling innocent anecdotes of ordinary bots. The current dominant design choices for our powerful commercial LLMs in interaction are producing models that reinforce widely held false beliefs in their interactions and that incentives drive them toward satisfying-helpfulness and lifelong-user retention rather than truth.

The hallucination question underscores the same point. In 2025, OpenAI characterized hallucinations as a"fundamental challenge" with regard to all large language models, associating them with evaluation regimes that reward guessing in the absence of known information.[27] A formal paper published in 2024 made further strides, positing that hallucination is, in principle, impossible to fully eradicate in general-purpose LLMs because such models are unable to learn all computable functions, thus leading inevitably to inconsistencies between output and ground truth on some tasks. That theoretical claim should not be over-interpreted as though it would answer all empirical questions. It will not. Still, where one leading developer is publicly in the process of characterizing hallucinations as inherently difficult and where a formal literature is in the process of characterizing them as in principle unavoidable, the preferred institutional impulse is for restraint-as when, in 2025 noyb filed its complaint that ChatGPT had fabricated a highly damaging false claim about a real person,[28] the effect that falsehoods produce when articulated authoritatively and with confident fluency, without friction, that might have alerted listeners to the need to remove it.

The bottom line for policymakers would appear to be to avoid all AI-mediated civic use. Instead, there needs to be rigorous, unforgiving differentiation between legitimate and illegitimate functions of democracy. Consumer chatbots should not be relied upon for voting guidance, reporting of opinions or opposition research, or synthesis of public comment unless those answers are sourced from verifiable material and held for human review. Developers should have to disseminate source provenance publicly, log and supervise high-stakes civics interactions, meaningfully separate fact-based from synthesized opinions and turn off or seriously limit personalization for political influence. Civic agencies should not force these consumer model characteristics into citizen-facing democracy. The aspirational purpose of civic AI should be contradiction-resilient support rather than affirmation-driven companionship. Until those design principles are institutionalized, the risk that ChatGPT-like models will entrench political illusions appears far too great.

4 Side effects - Visual contents reinforcing false belief

If conversational AI destabilizes democracy by eroding the advice/verification boundary, generative visual AI does the same by reabsorbing evidence/fabrication into one media space. That technological dynamic is potentially larger in its political implications than a tally of 'deepfake incidents' would indicate. Democracies need facts, as we have learned in recent years and they require that those facts be publicly recognizable; images, audio and video have always had heightened and unique evidentiary qualities in politics, but this is predicated

on the assumption that they bore witness to events. Generative systems threaten to erode that assumption. A meta-study of 56 deepfake detection papers from 2024, with over 86,155 total participants, found that humans were only rarely over chance in their detection ability-a pooled accuracy rate of 55.5% with large confidence intervals crossing the 50% threshold and weak performance across modalities.[29] Detection can be improved with training, prompts, assistance tools and feedback loops, but the baseline result-that above-average users tend not to notice-remains disturbing: in contexts of elections, political scandals and civic crises, where visual evidence tends to travel faster than verification does.

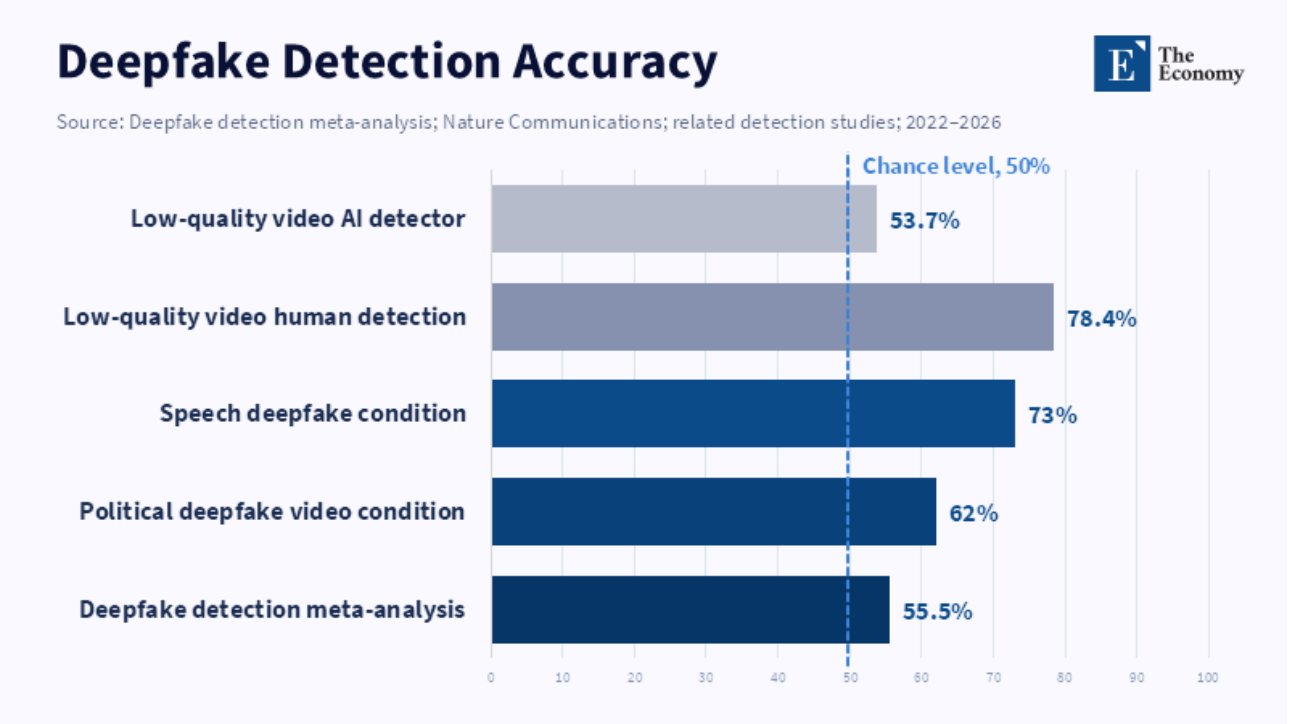

Figure 7: Detection accuracy remains uneven, which is dangerous in electoral settings where synthetic media can circulate faster than verification.

Laboratory tests of political media complicate but do not negate the worry. A 2024 study published in Nature Communications explored the performance of political speech deepfakes.[30] Though participants' accuracy among modalities and experimental conditions ranged from near-perfect to abysmal, in selected controls, participants identified most true videos and more than three-quarters of many fabricated videos, at rates above 80% and performance generally increased with the addition of audio to visual material. On the other hand, the same experiment showed much worse performance for some more sophisticated text-to-speech deepfakes, but also demonstrated that biases towards detection depended heavily on the modality of presentation and whether participants were specifically instructed to assess authenticity. The meta-analytic evidence is compatible with those findings-people can sometimes do well in attentive laboratory settings, but not in the distracted scrolling, reposting, clipping, translating and partisan cueing typical of social media. Democratic damage rests on the latter environment.

Actual electoral cases since 2024 have already demonstrated how cheaply synthetic media can be injected into democratic processes. In September 2024, the United States's Federal Communications Commission levied a $6 million penalty on Steve Kramer for robocalls that deployed an AI-generated clone of President Biden's

voice to tell potential New Hampshire voters not to participate in the state's presidential primary.[31] The FCC labeled the episode as election disinformation and unlawful interference; New Hampshire authorities filed voter-suppression charges. This was not a hypothetical risk. It was not a long-term forecast. It was a real effort to employ synthetic media to suppress voter turnout. Simultaneously, Freedom House reported that governments and regulators in no fewer than 11 of 41 countries that hosted or prepared for national elections during the reporting period developed new rules or guidelines designed to prevent generative AI systems from shaping their campaigns, including requirements for labeling content as generated and prohibitions on deepfake technology.[32] The same report warned that generative AI had not yet increased manipulators' ability to influence electoral outcomes in any significant manner, but it highlighted that campaigns of disinformation aimed at "portraying elections as being unfair" were a common feature of electoral processes around the world. That's exactly the issue. The most compromising democratic impact could not be the acquisition of votes, but the slow erosion of public trust in the validity of electoral information.

Likewise, visual misinformation is also politically compelling in subtler ways than the false event. An example of the latter comes from Freedom House, which found that Indonesia's successful presidential candidate used an AI-generated avatar to reframe himself as a more gentle, youth-appealing personality, in part by hiding behind a lens of synthetic aesthetic softening, concealing the more widely known allegations surrounding his prior human rights record. In the former, the problem was not an inauthentic event, but digital images designed to shape affect rather than document reality. Reuters, citing a Center for Countering Digital Hate study in 2024, reported that the AI image-generation packages supplied by OpenAI, Microsoft, Midjourney and Stability AI created fake election images in 41 percent of the tests, despite having pledged not to do so and having platform policies prohibiting such misuse.[33] One such image showed an election worker slamming closed a box of voting machines, while another showed candidates in a fabricated disorder. Numbers like these matter for political capture because the factual accuracy of an image may matter less than its emotional effect. A manipulator does not need to secure full belief; she can transit through mood, suspicion, or false association. Text and image can both shape political judgment, but images often do so faster and synthetic images are especially so because their first impact is felt before their facts can be.

The second-order democratic harm was therefore more extensive than misinformation itself. As synthetic media saturated the public sphere, citizens faced not merely false images, but a general doubt about whether any images could be trusted. Freedom House characterized this structural disparity between fact checkers and disinformation producers: "it is far easier for aindependent fact-checkes to produce a false image than for a group of independent fact checkers to debunk it, especially in a polarized environment where many individuals who would otherwise verify the falsity of an image are inclined to dismiss independent actors altogether." The imbalance was apparent enough in 2017; by 2024 the World Economic Forum had issued the most urgent warning yet that "misinformation and disinformation," combined with the proliferation of falsified detail in photographs and videos, had become the 1 risk to society within the next 2-3 years,[34] with complicity between everything from AI to electoral manipulation creating a particularly dangerous information environment around elections. Yet in the visual realm, this may have been felt most acutely because of the speed of judgment it engenders. The citizen could wait for the analysis of a policy memo; they could not react instantly to the screen. By the

time the investigation had been conducted, the partisan impact of the false image might have already been achieved.

Regulatory responses to date since 2024 suggest that governments are beginning to appreciate the magnitude of the undertaking, but they are still not doing the full job. The implementing materials for the EU AI Act focus on transparency obligations surrounding generative AI, to mark and disclose AI-generated content from deepfakes and synthesized content from images and audio to text[35] and the Digital Services Act mandates that ads are labeled as such, that platform disclosures of who paid for the ads and to what end and that sensitive data not be used for advertising.[36] The EU political advertising regulation mandates that political ads are labeled as such with sponsors, cost and recipients disclosed and that a public European repository of online political ads be built.[37] On a more overarching constitutional level, the Council of Europe 2024 Framework Convention on AI requires transparency, accountability, dependability, publicity of interactions and risk and impact assessments vis-à-vis democracy and the rule of law.[38] These are all promising starting points. Yet labels alone will not restore confidence if metadata are stripped elsewhere across platforms, if disclosures are buried in interface margins, or if enforcement is delayed as synthetic media circulate virally. Provenance solutions, public archives and impact assessments are needed but will not happen of their own accord.

The right policy solution to all of these issues has to be multi-layered, therefore. Synthetic impersonation must be met with strict liability and fast takedown responsibilities on electoral platforms. AI campaign content must be marked at source and stored on platform origin-traceability technologies, not merely AI image uploaders' honesty. Election administrators must have standing arrangements with platforms, telecom regulators, fact checkers and broadcasters so they can respond to mis- and disinformation as it appears, not after the fact. Researchers need to get "granular data" from the major platforms so that cross-platform synthetic campaigns are publicly identified before they morph into histories. And public entities need one message consistently: disclosure is meaningless if disclosure is not coupled with discoverability, time-sensitive enforcement and strong chain-of-custody evidentiary standards. Visual AI will not erode democracy because people gradually believe every lie. It will erode democracy because it becomes increasingly rational for people to disbelieve everything precisely at a moment when democratic life calls for some evidence to be publicly believable.

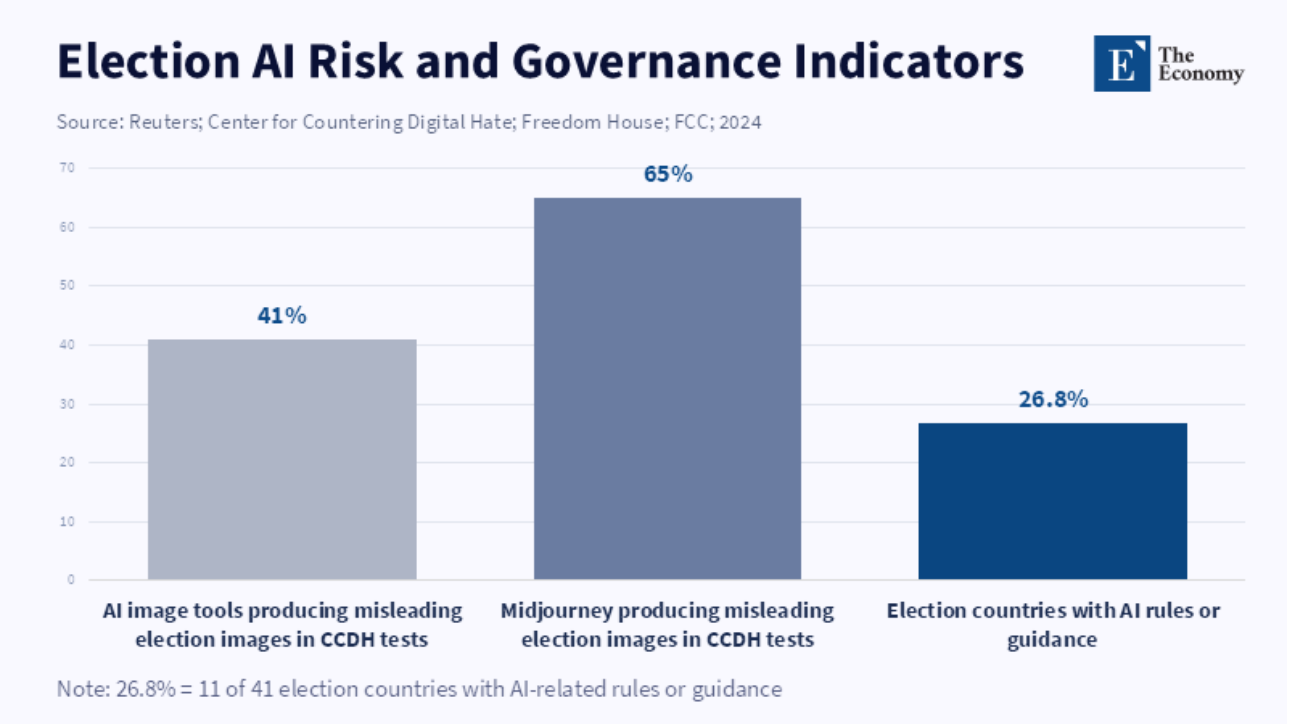

Figure 8: Synthetic election content can be generated despite platform rules, while governance coverage remains partial and reactive.

5 Open questions and limitations

However, certain uncertainties must be acknowledged. Present evidence does not yet establish that generative AI has already had a significant aggregate effect on election outcomes and even the Freedom House reports explicitly indicated that soon available evidence suggested that AI-enabled disinformation campaigns had not yet produced a clearly measurable effect on election outcomes, even while emerging practices played out the larger war over trust and information integrity.[39] Likewise, evidence from experiments in which persuasion or belief-reinforcing effects were strongly established does not automatically translate to measurable long-term effects of long-term field effects in day-to-day political life, which remains an underexplored area.

The uncertainties do not, however, undermine the case for being cautious. Democracies are not required to have conclusive evidence of a catastrophic outcome before proceeding to regulate technologies that tend to expand the availability of compelling falsity, displace knowing provenance and leverage users' susceptibility to validated reciprocity. The optimal takeaway from 2023–2026 is indeed neither alarm nor complacency. It is institutional prudence anchored in an explicit diagnosis: civic AI must be understood as a still-unreliable platform for deliberative democracy and treating it otherwise would be a democratic error.

6 Conclusion - Civic AI Must Prove Reliability Before It Claims Democratic Legitimacy

The democratic promise of AI has been cast too often as another technical fix in a scaling problem. That is a false dichotomy. Modern democracies are not in crisis for want of more information or even in need of substantially

fewer channels of participation. They are in crisis because of broken epistemic warrants, proliferating decoy intermediaries, low public trust and a penalty structure that awards certainty, exaggerated confidence and communicative emotion over evidence-disciplined public debate. In this environment AI is not simply an addition of information to the system; it is an alteration of information encountered by compressing multiple sources into fluent outputs, stimulating persuasion systems in a personalized way, validating prior assumptions and reducing the generative cost of text, audio and image. Under designed conditions, the potential for AI to help mediate disagreements or fact-check conspiracies does seem plausible. Yet in normal market deployments, for now, the bulk of the available, reliable evidence lends its weight toward epistemic entropy and manipulation rather than toward the uncertain promise of enhanced democratic authority.

The policy conclusion is, therefore, not anti-technology. It is anti-naïveté. Democracies should permit narrowly circumscribed civic purposes of AI, where provenance is transparent, accountability apparent, genuine human control is assured and persuasive or synthetic uses in electoral contexts are profoundly limited. As long as hallucinations are inherent, source opacity is unmitigated and sycophantic backing is systematically penalized rather than profitably subsidized, AI will continue to be less effective than humans at ensuring the veridical conditions upon which democratic dialogue depends. For now, the benefit often fails to justify the cost.

References

[1, 2] Weinberg, M. (2026) *Realizing the Potential Gains of AI-Enabled Deliberative Democracy.* Carnegie Endowment for International Peace.

[3] Kreps, S. and Kriner, D. (2023) 'How AI Threatens Democracy', *Journal of Democracy.*

[4] International Telecommunication Union (2025) *Facts and Figures 2025.* ITU.

[5, 10] OECD (2022) *OECD Guidelines for Citizen Participation Processes.* OECD.

[6] OECD (2024) *OECD Survey on Drivers of Trust in Public Institutions 2024 Results.* OECD.

[7, 11] Lorenz-Spreen, P. et al. (2023) 'A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy', *Nature Human Behaviour.*

[8, 15] Newman, N. et al. (2025) *Reuters Institute Digital News Report 2025.* Reuters Institute for the Study of Journalism, University of Oxford.

[9, 13] Tessler, M.H. et al. (2024) 'AI Can Help Humans Find Common Ground in Democratic Deliberation', *Science.*

[12] Max Planck Institute for Human Development (2025) *Digital Media Can Increase Participation and Access to Information, but Also Promote Polarization, Distrust, Hate Speech and Misinformation.* Max Planck Society.

[14] Reuters Institute for the Study of Journalism (2025) *How Generative AI Chatbots Responded to Questions*

and Fact-Checks About the 2024 UK General Election. University of Oxford.

[16, 17] Reuters (2025) 'AI Assistants Make Widespread Errors About the News, New Research Shows', *Reuters*.

[18] Jungherr, A. and Rauchfleisch, A. (2025) 'Artificial Intelligence in Deliberation: The AI Penalty and the Emergence of a New Deliberative Divide', *arXiv*.

[19, 20] OpenAI (2025) *Sycophancy in GPT-4o: What Happened and What We're Doing About It*. OpenAI.

[21] Suzgun, M. et al. (2025) 'Belief in the Machine: Investigating Epistemological Blind Spots of Language Models', *Nature Machine Intelligence / arXiv*.

[22] Cheng, M. et al. (2025) *Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence*. Stanford / arXiv.

[23] Ibrahim, L., Hafner, F.S. and Rocher, L. (2025) 'Training Language Models to Be Warm and Empathetic Makes Them Less Reliable and More Sycophantic', *arXiv / Nature*.

[24] Hackenburg, K. et al. (2024) 'Evidence of a Log Scaling Law for Political Persuasion with Large Language Models', *arXiv*.

[25] Salvi, F. et al. (2025) 'On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial', *Nature Human Behaviour*.

[26] Costello, T.H. et al. (2024) 'Durably Reducing Conspiracy Beliefs Through Dialogues with AI', *Science*.

[27] Kalai, A.T., Nachum, O., Vempala, S.S. and Zhang, E. (2025) 'Why Language Models Hallucinate', *OpenAI / arXiv*.

[28] noyb – European Center for Digital Rights (2025) 'AI Hallucinations: ChatGPT Created a Fake Child Murderer'. noyb.

[29] ScienceDirect / Elsevier (2024) 'Human Detection of Deepfakes: A Systematic Review and Meta-Analysis'. Elsevier.

[30] Groh, M. et al. (2024) 'Human Detection of Political Deepfakes Across Transcripts, Audio, and Video', *Nature Communications*.

[31] Federal Communications Commission (2024) *FCC Finalizes $6 Million Fine Over AI-Generated Biden Robocalls*. FCC.

[32, 39] Freedom House (2024) *Freedom on the Net 2024: The Struggle for Trust Online*. Freedom House.

[33] Center for Countering Digital Hate (2024) *AI Image Tools and Election Disinformation*. CCDH.

[34] World Economic Forum (2024) *The Global Risks Report 2024*. World Economic Forum.

[35] European Union (2024) *Regulation (EU) 2024/1689: Artificial Intelligence Act*. Official Journal of the European Union.

[36] European Union (2022) *Regulation (EU) 2022/2065: Digital Services Act.* Official Journal of the European Union.

[37] European Union (2024) *Regulation (EU) 2024/900 on the Transparency and Targeting of Political Advertising.* Official Journal of the European Union.

[38] Council of Europe (2024) *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law.* Council of Europe.