

The impact of advanced AI systems on democracy

Received: 9 July 2024

Accepted: 13 August 2025

Published online: 1 October 2025

 Check for updates

Christopher Summerfield^{1,2}✉, Lisa P. Argyle³, Michiel Bakker^{4,5}, Teddy Collins⁶, Esin Durmus⁷, Tyna Eloundou⁸, Iason Gabriel⁴, Deep Ganguli⁷, Kobi Hackenburg^{2,9}, Gillian K. Hadfield^{10,11,12}, Luke Hewitt¹³, Saffron Huang^{6,7}, H el ene Landemore¹⁴, Nahema Marchal⁴, Aviv Ovadya¹⁵, Ariel Procaccia¹⁶, Mathias Risse¹⁷, Bruce Schneier¹⁷, Elizabeth Seger¹⁸, Divya Siddarth⁶, Henrik Skaug S etra¹⁹, Michael Henry Tessler⁴ & Matthew Botvinick^{4,20}✉

Advanced artificial intelligence (AI) systems capable of generating humanlike text and multimodal content are now widely available. Here we ask what impact this will have on the democratic process. We consider the consequences of AI for citizens' ability to make educated and competent choices about political representatives and issues (epistemic impacts). We explore how AI might be used to destabilize or support the mechanisms, including elections, by which democracy is implemented (material impacts). Finally, we discuss whether AI will strengthen or weaken the principles on which democracy is based (foundational impacts). The arrival of new AI systems clearly poses substantial challenges for democracy. However, we argue that AI systems also offer new opportunities to educate and learn from citizens, strengthen public discourse, help people to find common ground, and reimagine how democracies might work better.

In 2024, half the global population—including India, the USA and several other of the world's richest and most populous nations—went to the polls. In the short intervening time since these voters last cast their ballot, a step change occurred in the development of advanced artificial intelligence (AI) systems. Large generative models can now produce text, images, audio and audio–visual outputs that closely resemble those produced by humans. The impact that these powerful, publicly available AI systems may have on the political process has been widely debated in the media, often with a focus on the potential of AI to disrupt or corrode democracy. Here we situate this discussion in a growing academic literature across both AI research and social science^{1–7}.

Language has an indispensable role in democracy. It allows information about candidates and policies to be shared with voters, lawmakers to create legislation, and citizens and representatives to collectively discuss, deliberate and decide on which course of action to pursue. In liberal democracies, elected leaders use oratory (the art of formal speaking in public) to explain and justify their decisions and actions to the larger public, interest groups use persuasive messaging to lobby for their preferred policies, and the general public engages in debate, either informally or through organized events such as town halls and citizens' assemblies^{8,9}. Given the primacy of linguistic exchange in the political process, the arrival of conversational machines brings

¹Department of Experimental Psychology, University of Oxford, Oxford, UK. ²UK AI Security Institute, London, UK. ³Department of Political Science, Purdue University, West Lafayette, IN, USA. ⁴Google DeepMind, London, UK. ⁵Massachusetts Institute of Technology, Cambridge, MA, USA.

⁶Collective Intelligence Project, Wilmington, DE, USA. ⁷Anthropic, San Francisco, CA, USA. ⁸OpenAI, San Francisco, CA, USA. ⁹Oxford Internet Institute, University of Oxford, Oxford, UK. ¹⁰Faculty of Law, University of Toronto, Toronto, Ontario, Canada. ¹¹Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA. ¹²School of Government and Policy, Johns Hopkins University, Baltimore, MD, USA. ¹³Stanford Center on Philanthropy and Civil Society, Stanford, CA, USA. ¹⁴Department of Political Science, Yale University, Yale, NH, USA. ¹⁵AI & Democracy Foundation, Covina, CA, USA. ¹⁶School of Engineering and Applied Sciences, Harvard University, Allston, MA, USA. ¹⁷Harvard Kennedy School, Harvard University, Cambridge, MA, USA. ¹⁸Demos, London, UK. ¹⁹Department of Informatics, University of Oslo, Oslo, Norway. ²⁰Yale Law School, New Haven, CT, USA.

✉e-mail: christopher.summerfield@psy.ox.ac.uk; botvinick@google.com

Table 1 | What do we know, and what do we need to know, about AI and democracy?

Theme	What do we know so far?	What do we need to know?
Political bias (epistemic)	Most current AI models are politically biased towards progressive/libertarian views when assessed with multiple-choice questions ¹²⁻¹⁴ , but will typically provide a balanced political view when responding freely ¹⁷	How does interacting with AI models affect the user's political attitudes? Will AI tip the political balance in one direction or another?
Persuasive messaging and dialogue (epistemic)	Current AI models can be highly persuasive in experimental settings ^{31,32} , especially when generating lots of fact-checkable claims ³³	How are persuasive AI models being deployed in the wild, and how persuasive are they outside of an experimental context?
Political polarization (epistemic)	Current AI models are prone to 'sycophancy', aligning their expressed views or preferences with those of the user ⁴⁷ , and this may increase as models become more personalized ⁴⁹	Will increasing personalization of AI models lead to increasing political polarization, creating new opportunities for 'filter bubbles'?
Deliberation and consensus (epistemic)	AI models can be deployed to help people to find common ground, by summarizing arguments and suggesting compromise positions ^{58,62,133}	Can AI be used to meaningfully advance policy negotiation or conflict resolution in a fair and beneficial way?
Information and misinformation (epistemic)	AI models often generate misleading content that is hard to detect ^{72,73} , and this can be exploited by malicious actors to distort the political process ⁹¹ , but AI systems are also useful for fact-checking ^{97,134}	Will the widespread introduction of AI models ultimately help or harm the information environment? Will people prove resilient to fake AI-generated content? Will the digital commons be overwhelmed by 'slop'?
Electoral disruption (material)	AI models can assist with 'hack and leak' and spear-phishing campaigns ^{93,94}	Will we see widespread AI interference in electoral processes in the future, especially from agentic systems? What safeguards should we put in place to prevent this from happening?
Voter suppression (material)	AI models are being used to automate the rejection of voting applications, often based on unreliable evidence ⁹⁷ , and to profile and target journalists and activists ¹⁰³	Can AI agents be used to identify likely voting preferences and target individuals in ways that decrease their opportunity for democratic participation?
Augmenting political decision-making (material)	AI is already being deployed in the public sector, including to write policy and streamline governance mechanisms ¹⁰⁴ , as well as for expert polling ¹⁰⁷	Will the political writing and delegated decision-making of AI models distort the preferences of policymakers in detrimental ways?
Power concentration (foundational)	AI development is dominated by a few large technology firms ^{116,125} ; in some countries, AI is already being widely deployed for surveillance and population control ¹³⁵	Will we eventually see a unipolar or multipolar AI developments landscape? Does the arrival of advanced AI fundamentally favour authoritarian over democratic regimes, by offering new opportunities for state control?
Labour market disruption (foundational)	AI models are already affecting the graduate labour market ¹²⁰	Will AI create substantial growth? If so, will it benefit everyone, or will it dramatically exacerbate inequality? Will any changes to the labour market arrive rapidly or slowly?

The table summarizes the evidence presented in this Perspective, and outlines suggestions for future research that will move the field forwards.

the potential for a dramatic and far-reaching impact on democracy worldwide.

In this Perspective we propose that AI creates three classes of challenge for democracy, but argue that each is tempered by corresponding potential opportunities. First, there are epistemic impacts—those that affect citizens' ability to make educated and competent choices about both representatives and policies—such as risks from misinformation or AI persuasion, and opportunities for citizen upskilling. Second, there are material impacts. AI could be misused to attack the machinery of democracy, but also deployed to improve the efficacy of governance processes. Finally, there are foundational impacts, by which AI may weaken or strengthen the very principles on which democracy is based, or affect its opportunity to flourish worldwide. In this Perspective we discuss each type of impact in turn. We focus on the impact of advanced or 'frontier' AI systems that are capable of generating highly realistic natural language and multimodal outputs (images, audio and video), as well as (via a scaffold) taking actions or executing computer code in digital environments. We summarize the extant evidence and outlook for each section in Table 1. We note that, as this is a fast-moving research area, some of the cited papers are unreviewed preprints.

Epistemic impacts

Even before powerful large language models (LLMs) became available, algorithms were responsible for shaping the flow of information and misinformation on digital platforms¹⁰. However, the advent of LLMs presages entirely new challenges and opportunities for global epistemic health (the tendency for people worldwide to believe true rather than false information). We consider how democracies may be weakened

by political bias in AI systems, automated persuasion, polarization from personalized content, or the scaling of misinformation, but also how they may be strengthened by AI systems that allow fact-checking, increased mutual intelligibility, deliberative upskilling and automated tooling for political participants to find common ground.

Political bias

Publicly available LLMs already have wide user bases, totalling hundreds of millions of monthly users. If AI is being used to provide information about current affairs, political controversies and electoral choices, then even weak biases in their outputs could substantially impact the distribution of political beliefs among citizens. Several studies have quantified LLM political biases by administering multiple-choice survey questions (such as the **Political Compass test**) and measuring their output probabilities (for example, option A versus option B). These studies have revealed that after safety fine-tuning (a post-training method designed to minimize harmful or illegal outputs¹¹), models favour more libertarian (for example, favouring deregulation) and progressive (for example, supporting civil rights) opinions¹²⁻¹⁴. However, LLMs are highly malleable, and when prompted to play the part of characters with different political worldviews, they will readily adopt diverse opinions^{15,16}.

Nevertheless, this stylized evaluation method may yield results that are unrepresentative of everyday user interactions with an LLM, because multiple-choice items do not offer respondents the opportunity to voice balanced or equivocal replies. In fact, when responding freely to user queries in everyday settings, proprietary models typically preface replies with reminders that they do not hold political opinions, and give scrupulously balanced answers to direct enquiries about the

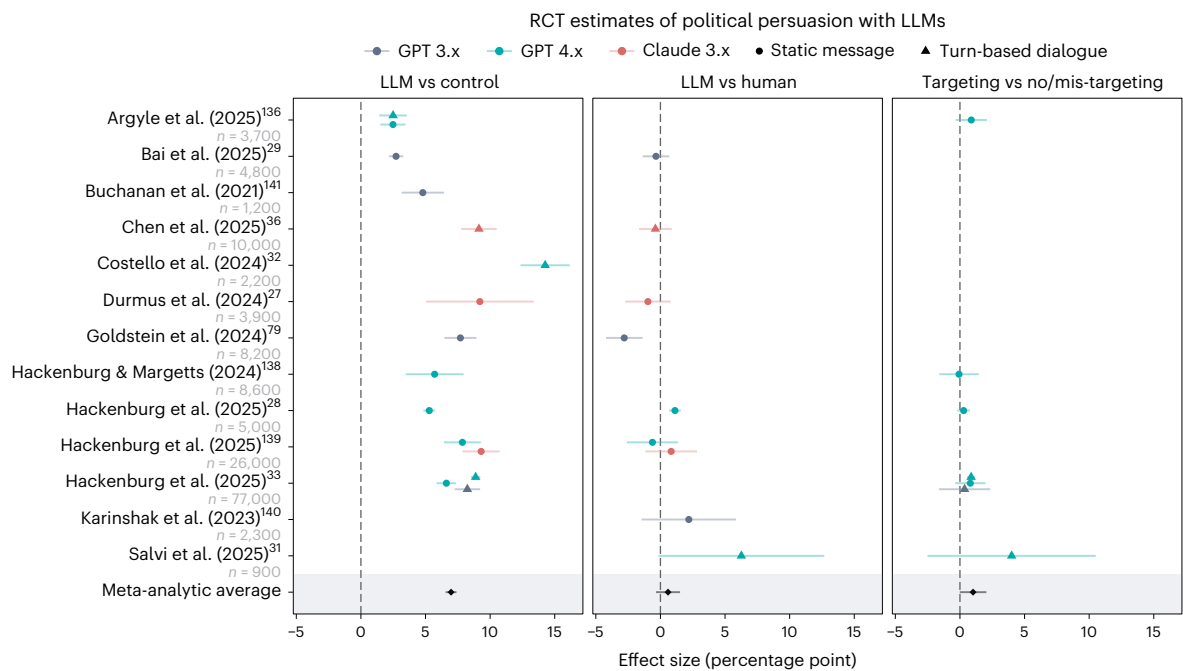


Fig. 1 | Randomized controlled trial estimates of political persuasion with LLMs. This figure includes all studies known to the author team that randomized participants to LLM-generated messages and compared post-treatment attitudes between participants^{27–29,31–33,36,136–141}. For each study, we rescaled the outcome to [0, 100] and then calculated the simple difference in mean outcomes by condition (with 95% confidence intervals, CIs), adjusting for pretreatment attitudes where measured. This approach aims to maximize analytic consistency

across studies, but note that this may differ from the authors’ original analyses. The studies vary in the model used (GPT and Claude), treatment format (static messages and chatbot conversations), reference conditions (experts, laypeople and so on), as well as the political issues considered. We include a meta-analytic average, but also emphasize the large amount of heterogeneity we observed across studies.

relative merits of political representatives or policies. They also remind the user about the limits of their knowledge, and refer them to sources on the internet for the most up-to-date information. In fact, under normal usage conditions, the models may be much less left-leaning than previously argued¹⁷. Developers face difficult choices when choosing the training protocols that define which views models express, being obliged to decide when models should take sides, and when it should acknowledge potentially conflicting perspectives. This has led some to seek public input to these questions^{18,19}. More generally, in politically fraught times, there is a vibrant conversation about whether models (including those built in non-liberal democracies such as China²⁰) should freely express political views or remain studiously neutral.

Even if models are politically biased, it remains uncertain what the impact may be on democracy. We could imagine scenarios where biases in the models tip the political consensus towards views favoured by AI developers, especially in democracies (like the USA) where only a few percentage points typically separate rival parties. However, extant research has focused on evaluating the outputs of models, and we currently lack human interaction evaluations²¹ that directly measure how they may affect the beliefs of the user. Thus, as AI use becomes more pervasive, we will need large-scale studies of how the political views expressed by AI systems may influence the distribution of political views among citizens.

Persuasive messaging and dialogue

Political candidates, parties and interest groups attempt to shape voters’ beliefs through advertisements, media engagement, public events and door-to-door canvassing. In the near future, frontier AI systems may be deployed on digital platforms (such as Reddit or Discord, or social media sites such as X, Bluesky or Truth Social) to persuade citizens towards a particular issue, candidate or party. These influence operations could be mounted en masse²² by state or non-state actors whose

goal is to undermine democracy^{23,24}, with citizens potentially believing that they are interacting with a real human²⁵. It is thus important to assess just how persuasive AI systems can be in an experimental simulacrum of this context. Some researchers have even proposed that future AI systems may have persuasive powers that exceed those of humans, being capable of ‘hypersuasion’²⁶. In what follows, we review evidence concerning the persuasive capabilities of advanced AI systems.

Several recent studies, focused on the US and UK electorate, have aimed to directly measure the impact of LLM-generated messaging on political attitudes through randomized controlled trials. We summarize their results in Fig. 1. First, the magnitude of persuasion varies greatly between studies, due to the political issues considered (for example, immigration and gun control) as well as the models and prompting techniques used. Second, across issues, studies have consistently found that LLMs are able to write messages to persuade on political issues with roughly the same persuasiveness of human writers^{27–29}. More recent studies have focused on the conversational abilities of AI as a potential means of persuasion, given that dialogue is a potent method for persuasion among humans (for example, during doorstep canvassing³⁰). These studies have observed persuasive gains ranging from 2 to 14 percentage points when an AI engages a human in persuasive multi-turn dialogue³¹, with the strongest effects on stances where the human initially disagrees (for example, when debunking the views of conspiracy theorists³²) and when AI was prompted to include a high volume of evidence or many informative claims^{32,33}. Importantly, these persuasive effects were observed despite participants in most studies being fully aware their conversation was with an AI model, and studies aiming to manipulate this have typically found little persuasive cost to identifying messages as AI-generated^{25,34–36}. Overall, this research hints at the potential power of future LLMs that use persuasive dialogue for political ends. Developers should consider mitigations such as increased interpretability and scalable forms of model oversight³⁷.

One oft-cited concern is that LLMs (which can learn rich conditional distributions) could be used to tailor political messaging to a specific individual, for example, based on features obtained from their social media activity^{38,39}, a practice known as ‘microtargeting’. On the one hand, microtargeting has the potential to increase participation⁴⁰ or heighten interest in topics relevant to minority voters⁴¹. However, there is concern that LLMs could distort campaigning by inferring voter preferences from open-source data (perhaps a single tweet⁴²), and rewording messages to match a citizen’s perspectives, demographics or personality traits^{39,43}. However, as shown in Fig. 1, thus far there is very limited evidence for an effect of targeted messaging on participants’ attitudes²⁸, consistent with existing reports that microtargeted messages are rarely more effective than the average most persuasive message⁴⁴.

Together, these results imply that, in controlled experimental settings, AI models can be highly persuasive. However, we do not know whether or how these results translate to the wild. The persuasive impact of political messaging is dramatically attenuated in the real world, where half the battle is for users to attend to messages in the first place (unlike laboratory studies, which involve a captive audience). The literature would benefit from more field studies that measure the impact of AI on human beliefs in the wild, such as the deployment of persuasive AI on social media or the public’s use of AI assistants for purposes such as writing and research^{45,46}.

Political polarization

In many modern democracies, opinions have become highly polarized, with politics dominated by opposing groups who reject each other’s views and values outright. This takes the form of both issue polarization (highly divergent political perspectives) and affective polarization (animosity between people with different political affiliations). Polarization is often blamed on algorithms that trap users in ‘filter bubbles’ or ‘echo chambers’, where they are insulated from the discomfort of contrary views. LLMs could exacerbate this issue by generating replies that flatter the user’s preconceptions or beliefs. Most current LLMs are relatively generic (rather than personalized to suit the tastes of individual users) and thus provide neutral or diplomatic replies designed to have broad appeal. Nevertheless, there is evidence that even generic LLMs tend to be ‘sycophantic’, or preferentially express views that may be shared by the user, even if these are untrue. This occurs because human feedback provided during the feedback process tends to reward LLM replies that echo user sentiments⁴⁷. Moreover, models are increasingly using long-term memory to explicitly tailor outputs based on users’ expressed demographics, interests and tastes to make AI more appealing⁴⁸.

Despite these concerns, it is unclear whether AI will heighten political polarization⁴⁹. The view that filter bubbles and echo chambers increase partisanship has come into question^{50,51}, and affective polarization may instead occur because social media algorithms often encourage the most divisive content to be viewed, attended to and shared⁵⁰. Moreover, publicly available LLMs, when fine-tuned to give equanimous perspectives on issues of debate, offer the opportunity to expose users to a spectrum of legitimate opinions, and could nourish public discourse in ways that social media platforms have systematically failed to do⁵². An alternative explanation for heightened partisanship is that voters are encouraged to form highly stereotyped perceptions of their political opponents (for example, in the USA, Republicans believe that 32% of Democrats identify as LGBTQ, and Democrats believe that 38% of Republicans earn in excess of US \$250,000 per year, where in reality the figures are 6% and 2%)⁵³. Without careful prompting, LLMs tend to generate caricatured outputs that may exaggerate stereotypical features in exactly this way (for example, assuming that all Republicans are rich and all Democrats are gay)⁵⁴, and thus risk contributing to partisanship by erasing the nuance in the way that people see each other. Thus, as more personalized AI systems come into widespread use, we need to study their potential impact on the fragmentation of

political consensus, and attempt to discern whether they could drive increased polarization or partisanship in a similar way to content curation algorithms on social media.

Deliberation and consensus

In a healthy democracy, people can express diverse opinions, and deliberate in ways that foster better decisions, for example, by listening to each other in an atmosphere of mutual respect. Although some fear that AI may be used to weaken or suppress political discussion⁵⁵, there is also hope that AI could be used to create healthier spaces for deliberation among citizens. In one study, LLMs were prompted to intervene in political discussions between US voters with opposing views on gun control, by proposing less adversarial message rephrasing. Discussants accepted the proposed wording about two-thirds of the time, and when they did so, improvements in perceived conversation quality and democratic reciprocity (the extent to which political opponents report respecting each other’s right to hold contrary views) were observed⁵⁶. Another possibility is that LLM intervention might help to amplify voices that are at risk of being marginalized in a discussion. For example, inserting LLMs into mixed-gender groups of Afghan citizens discussing contentious political issues was shown to increase the range of ideas contributed by female group members⁵⁷. LLMs also offer new opportunities for improving interactions among citizens on social media or debate platforms, by summarizing opinions and optimizing the routing of comments between discussants⁵⁸, and potentially helping humans themselves to become more effective conversation partners⁵⁹. LLMs can thus potentially promote tolerance and respect in the public sphere to facilitate more epistemically productive dialogue.

Another promising opportunity is for LLMs to be used to directly assist humans in finding common ground, by facilitating deliberation over issues of legitimate debate. Currently, formal citizens’ assemblies allow representative groups of people to gather and debate policy issues or provide collective input into new laws or constitutions⁸. However, organizing large-scale in-person events is costly and time-consuming, and face-to-face debate can be susceptible to social desirability biases, where interlocutors are motivated by a desire to win the argument, rather than to reach a mutually acceptable outcome⁶⁰. In one study, LLMs were trained to generate collective statements that maximized endorsement from a group providing private written opinions⁶¹. These statements were preferred over those written by humans, and helped people to converge to a common side of the argument. In recent studies^{62,63}, LLMs were used in conjunction with a carefully designed process to generate a slate of statements to represent the diversity of opinions in a group, according to a rigorous notion of representation. Building such mathematical guarantees into the outcomes of AI-augmented democratic processes may help to alleviate concerns about the biases encoded in LLM summaries, which may otherwise omit essential details or distort the intended meaning⁶⁴. Methods from the field of computational social choice can be used to help LLMs to generate more reliable collective outcomes⁶⁵. More recently, AI has been deployed to generate consensus over community notes on social media platforms⁶⁶, and to help to reconcile the views of peacebuilders from opposing sides of a conflict^{67,68}.

Information and misinformation

LLMs are prone to inadvertently generate content that is poorly sourced, untruthful or over-confident⁶⁹ (‘hallucinate’). Moreover, recent breakthroughs in multimodal generative AI have greatly expanded opportunities for malicious actors to create and manipulate highly realistic audio and video from simple text descriptions, or to alter media to produce propaganda. In a political context, this means manipulating multimodal content to portray political rivals in compromising or defamatory ways, generating deceptive campaign videos, and even counterfeiting entire news websites. Even by 2024, deepfake videos have been deployed with the obvious intent to shift

the electoral calculus in India, Indonesia, Mexico, Pakistan, Slovakia, Ukraine, the USA and Taiwan (documented at <https://restofworld.org/2024/elections-ai-tracker/>), with many fake videos being viewed hundreds of thousands of times. The arrival of hyper-realistic generative content robs news media of its ‘epistemic backstop’—the decisive authority that was previously provided by a video or audio recording of a news event. LLMs may also be used as ‘social bots’ on digital platforms, tasked with spreading false or hyper-partisan content rapidly through networks while disguising its AI-based origin⁷⁰. Unfortunately, evidence suggests that AI systems may be more convincing when they are prompted to produce deceptive content²⁷.

AI-generated media is becoming harder to spot. Recent (as yet unreviewed) work suggests that AI-generated audio and video⁷¹, images of human faces⁷², and tweets⁷³ may now be largely undetectable as such. Developers are working on machine-learning methods to distinguish synthetic and human-generated content, and testing ‘watermarking’ techniques that add an invisible signature identifying digital content as generated or altered by AI⁷⁴. However, many current techniques are quite easy to circumvent⁷⁵, and researchers acknowledge that watermarking is not a silver bullet for avoiding misinformation⁷⁶. Strong watermarking has been argued to be provably impossible, so these safeguards may only deter relatively unsophisticated threat actors^{77,78}.

Over the longer term, repeated exposure to substantial volumes of realistic deepfake materials may have a systemic effect on the population’s epistemic health. Users are more likely to believe information that is repeated, independent of its plausibility⁷⁹, and AI systems offer new and more targeted ways to deluge users with misleading content. Educating people about deepfakes may help, but it can also lead to legitimate content being widely questioned—a phenomenon called the ‘liar’s dividend’⁸⁰. Widespread exposure to deepfakes may thus render people more vulnerable to misinformation campaigns, or breed broader scepticism regarding online media, as well as traditional news outlets and journalists, across the population⁸¹. In the worst case, irreversible contamination of the digital knowledge commons—vital repositories of shared and publicly accessible data—with fake, deceptive or partisan content may lead to cumulative erosion of public trust in information, and distort our collective understanding of socio-political reality or scientific consensus⁸². However, it has been widely commented that the widespread availability of high-quality deepfakes has so far had much less impact than was previously predicted⁸³.

On a more hopeful note, LLMs also provide new opportunities to help users of online platforms to discern truth from falsehood. One example is progress towards the automation of fact-checking on digital platforms. This includes chatbot websites themselves: many publicly available LLMs already offer embedded citation, where model replies are augmented with hyperlinks that allow users to verify the source of a claim, providing a form of assurance against confabulation^{84,85}. LLMs can also be deployed to check the provenance and veracity of claims made on external news or social media sites, a task that is extremely laborious for humans, with some facts taking hours to verify⁸⁶. In one unreviewed preprint, ChatGPT was asked to classify more than 20,000 pieces of previously fact-checked news, and was found to agree with human raters ~70% of the time. Interestingly, this rate of agreement was maintained beyond its training cutoff (that is, to events that it could not possibly have known about), betraying that the model was relying on an estimate of *prima facie* plausibility to make this judgement rather than actually checking facts as a human might⁸⁷. More reliable fact-checking services may soon be available, but when they arrive, we need to find ways to ensure that they have impact. For example, one study found that participants were just as likely to believe and share content that ChatGPT had flagged as false as that which it had supposedly verified⁸⁸, and according to opinion research⁸⁹, people generally mistrust ChatGPT as a source of political information and are less likely to change their view after exposure to AI-generated content when they are aware of the source than when they think it is written by a human expert²⁵.

Material impacts

Democracy in modern societies is intimately linked to digital information and communication technologies^{3,90}. Such technologies increasingly mediate the relationships between citizens and their environment, and the various interactions between and among citizens, organizations, politicians and the state. This encompasses collective decision-making processes—such as debating, protesting, lobbying, polling, funding or voting—and how policy is implemented by the bureaucratic machinery of the state. AI is the transformative technology of the twenty-first century, and so it naturally has an impact on the materiality of democracy—the infrastructure that supports the democratic process in society.

Electoral disruption

The advent of LLMs has brought an increase in concern that AI could be deployed to disrupt elections⁹¹. This includes the use of AI for ‘hack and leak’ operations, whereby private accounts (for example, email) may be compromised to portray political rivals in an unflattering light, or to otherwise unfairly tip the political balance. For example, the Russian spy agency Star Blizzard is thought to have targeted Western politicians and journalists over a sustained period, as well as directly attacking the UK Electoral Commission, as reported by the UK National Centre for Cyber Security (NCSC)⁹². A common strategy is to use spear-phishing campaigns, in which individuals are maliciously targeted with highly personalized messages, typically over email. LLMs may be able to assist with this process by crafting bespoke deceptive messages based on personal details (for example, scraped from social media profiles). For example, one unreviewed preprint shows how ChatGPT can be used to generate spear-phishing messages tailored to British parliamentarians⁹³. Another preprint has reported that LLMs prompted to assist with spear-phishing attacks can produce click-through rates of ~50%, comparable with human experts, on unknowing participants, suggesting their likely effectiveness in the real world⁹⁴.

Voter suppression

The material practice of democracy relies on elections and referenda being free and fair, and eligible voters should be able to cast their vote unhindered. As voters turn to AI with questions about elections, developers need to ensure that LLMs provide accurate, up-to-date information about eligibility and voter registration, polling station access, voter ID and other election rules. At present, this is not always the case. For example, in one unreviewed report, researchers tested the accuracy of leading proprietary and open-source models on practical queries about electoral participation in the USA (for example, is it OK to wear a MAGA hat to vote in Texas?), finding that over half of replies were inaccurate⁹⁵. The problem may be particularly acute for those models (like the free-to-use version of ChatGPT) that do not use real-time internet queries to obtain up-to-date information for replies, and thus risk providing outdated advice (although deployed models are increasingly fine-tuned to encourage users to seek information from authoritative sources).

Unfortunately, there is increasing evidence that AI systems are being weaponized for voter suppression, especially in a US context. In one well-publicized example, generative AI was used to synthesize an automated telephone message (or robocall) that appeared to feature former President Joe Biden discouraging voters from participating in the 2024 New Hampshire primary⁹⁶. Given the ease with which such deepfake materials can be generated unless models are properly safeguarded—using a short snippet of genuine audio and a few dollars—the use of generative AI to sow confusion among voters and officials seems set to grow. In the near future, heightened personalization of messages to individuals could exacerbate this risk, for example, with robocalls that feature tailored disinformation about eligibility to vote (for example, based on past felony convictions). Tracking and disabling the tools that allow these malicious activities is becoming increasingly difficult as the tools become more sophisticated and widespread.

Another major vulnerability is voter registration, which is already a battleground issue in many US states. According to a news article, a tool called EagleAI, which purports to identify fraudulent voter activity, was deployed by activists to query or reject legitimate registrations (especially from voters from minority groups in contested wards) on the basis of unreliable evidence⁹⁷. EagleAI has been approved for voter roll maintenance in at least one county in the US state of Georgia, potentially giving it the power to arbitrate over thousands of registration challenges⁹⁸. A related risk is that AI's ability to generate content at scale is used to deliberately overwhelm electoral infrastructure, undermining the credibility of the democratic process or suppressing voter participation en masse. According to one news article, in 2024, many US states saw a huge surge in voter records requests (sometimes running to millions of documents) made under freedom of information laws, in an apparent attempt to disrupt legitimate election audit processes⁹⁹. AI can be used to accelerate this sort of disruptive activity. For example, EagleAI also allows partisan groups to file mass voter challenges (attempts to strip large numbers of registrants of their vote) on the basis of limited evidence. By automating this process, activists using AI can lodge an overwhelming number of challenges immediately before the review deadline, rendering officials powerless to overturn them.

In a well-functioning democracy, citizens are free to engage politically as journalists, activists and politicians, free from the threat of online violence and abuse. However, hyper-partisan groups often spread disinformation with a view to discouraging some constituencies from political participation, and there is evidence that women and historically marginalized groups are disproportionately targeted by these attacks¹⁰⁰. For example, female Democrats are reportedly ten times more likely to receive online abuse than their male counterparts¹⁰¹. In one unreviewed article, it is claimed that generative AI is used to craft and disseminate increasing volumes of gendered disinformation and defamation, and to synthesize deepfake pornographic images, which are maliciously targeted at political leaders or activists in order to intimidate and silence them¹⁰². This kind of violence is becoming remarkably common¹⁰³.

Augmenting political decision-making

Democratic representatives are empowered to make choices on behalf of citizens, but to do so they need to access and process relevant information. Currently, politicians rely heavily on experts and polling to brief them on relevant issues (such as the Congressional Research Service in the USA, and other stakeholders and advocates). One promise of LLMs is that they could augment human political decision-making, helping politicians to summarize vast bodies of data, brainstorming policy initiatives or writing draft legislation¹⁰⁴. This could allow legislatures to write, debate and pass more effective bills, or aid in the insertion of 'micro-legislation', minor and subtle text that changes the effect of laws¹⁰⁵, as well as aiding in the detection of loopholes. AI may even start to draft entire pieces of legislation (even if based on human desiderata). For example, in November 2023, a news article reported that the legislature of Porto Alegre, Brazil, had passed the first law written entirely by an LLM¹⁰⁶. If AI can help politicians to respond better to citizens' needs, and generally be more judicious, then this could increase their perceived legitimacy as representatives, and perhaps help to dispel the idea that they are self-serving or out of touch.

AI systems can also potentially enhance conduits of communication between legislators, public servants and the electorate. For example, LLMs can produce well-structured texts or oratory that could help politicians to communicate ideas more clearly to their constituents. In turn, AI may open new avenues for people to feed back their views to governments. LLMs are already being used for more effective election polling, harnessing social media data to make microscale predictions about voting intentions that match or exceed those from statistical models used by professional pundits^{107–109}. AI systems may also facilitate civic education by helping voters to inform themselves about the

issues that most concern them and which parties best represent their interests, or provide easier routes to learn about their rights and help them to navigate state bureaucracy and legal processes. For example, the UK has launched a conversational agent that responds to queries about topics covered on government webpages, including the details of forthcoming elections (<https://www.govgpt.uk/>).

Much more is possible. AI systems have already been shown to provide balanced summaries of the opinions expressed by small groups of people^{61,62}, but in newer LLMs with longer context lengths (the number of input tokens on which they can condition their output), this automated opinion digest could potentially be scaled to groups of thousands or more, providing a new, LLM-based mechanism for governments to ascertain what citizens think and want. The summarization process could even be conditioned on demographics, allowing insight into how both majority and minority groups may respond to a political decision¹¹⁰. In theory, if participants' beliefs are modelled accurately, it could be possible to simulate an 'election' for every decision, in which AI agents vote on behalf of stakeholders. This approach has been successfully piloted in the context of sensitive decisions involving food allocation, which must balance considerations of fairness and efficiency¹¹¹, although whether AI systems should serve as human proxies in high-stakes settings remains a controversial topic¹¹². Although we should be wary of naive techno-solutionism¹¹³, advanced AI systems invite us to reimagine how democracies might work for the better.

Foundational impacts

Democracy is based on a set of shared social values and principles that undergird democratic institutions, and offset the burden of democratic participation¹¹⁴. If these norms are eroded, democracies may 'backslide', or gravitate towards authoritarianism¹¹⁵. Aside from the epistemic and material impacts discussed above, AI could either corrode or bolster the foundations of democracy—to either accelerate, or guard against, democratic backsliding. These effects may be partly co-extensive with those already discussed, but differ by virtue of being more diffuse, systemic or secondary to the deployment of AI.

Concentration of wealth and power

One major concern is that AI will serve to concentrate excessive power in the hands of political leaders or parties¹¹⁶. According to one view, democracy flourished in the twentieth century because the technological landscape favoured decentralized economies and polities, where power is distributed across diverse groups and individuals¹¹⁷. However, if AI allows us to dramatically streamline twenty-first-century state bureaucracies, and to implement more effective centralized control, this could tip the balance back in favour of more authoritarian regimes³ or intrinsically favour 'tyranny'¹¹⁸.

Faith in democratic institutions can be undermined by a perception that the political process is rigged to create winners and losers. For example, many elected governments are perceived as being unresponsive to the demands of the majority, and catering selectively to the few. Populist parties with anti-democratic agendas are poised to exploit these grievances for their own political advantage. There are fears that AI will accelerate this trend by increasing inequality, especially in developed nations¹¹⁹. For example, new capabilities exhibited by LLMs could lead to the displacement of entire sectors within the labour market, including administrative and creative industry jobs that were previously thought likely to be spared automation¹²⁰. However, the precise impact that AI will have on the economy and composition of the workforce remains uncertain, with some unreviewed blog posts forecasting new opportunities for job creation^{121,122}. For example, emerging studies show that equipping workers with LLMs tends to bring the skill levels of lower skilled workers in line with their more highly trained counterparts^{123,124}, which could imply that AI will help to create a more level playing field in the workforce. A related concern is that without appropriate governance, the wealth generated by this technological

revolution could become concentrated in the hands of a few multinational corporations, in a handful of countries, who are building AI and distributing its services^{125,126}. Moreover, governments may struggle to keep up with the pace and complexity of private technology development and so critical choices about the way AI shapes societies may be made by corporations instead of democratic polities¹²⁷. If so, this could reinforce the perception that democracy's cherished liberal principles have evolved to serve elites rather than society as a whole.

In a democracy, representatives and public officials need to be accountable for their actions. If politicians fail to deliver, they can be voted out of office. AI could undermine this principle by blurring the lines of accountability when policies fail, because it is unclear whether human or machine made the final decision. In consumer settings, we have already seen an airline attempt to designate an LLM as 'a legal entity with responsibility for its own decisions', to pass the blame to AI for erroneous customer advice¹²⁸. As AI systems become embedded in the machinery of government, this creates new opportunities for blame to be deflected and accountability obscured. Increased automation of the levers of state could also have other unwanted secondary effects, such as disempowering citizens with inflexible bureaucracy, subjecting everyone to a relentless world of 'computer says no'¹²⁹. We should be wary that advances in AI do not presage a descent into 'algocracy'—government by algorithm—in which humans are wholly or partly removed from the loop¹³⁰.

However, if administered properly, AI holds the potential not only to conserve but also to enhance the foundational elements of democracy. AI might well bolster confidence in democratic government by creating wealth, enhancing productivity, and improving health, education and digital infrastructure. Cultural commitments to democracy could also be strengthened by the epistemic and material impacts described earlier, including improvements in government service delivery, facilitation of communication between constituents and elected officials, and scaling of public democratic deliberation. In the limit, AI may provide an opportunity to update the social contract underlying representative democracy, by expanding the role of citizen input in policymaking, up to and including the design of democratic institutions and processes^{61,111,131}. Such reinventions of our democratic institutions may be critical if our existing systems are unprepared for handling the dramatic changes that unfold as advanced AI becomes ever more ubiquitous in our society¹³².

Conclusions and outlook

AI is arguably the most important technology of our times. However, as its impact expands, it is becoming clear that it will increasingly impinge on another of humanity's most important inventions: democratic self-governance. In this brief Perspective, we have surveyed the possible implications of this encounter for the future of democracy.

One important conclusion from our review is that most work has so far focused on understanding the outputs that AI models may generate or (at best) the way that these outputs may affect people in controlled laboratory experiments, or on online studies run with compensated crowdworkers. Although useful, this research does not directly answer the question of how AI systems are being deployed in the real world, and what their likely effects on individuals and society will be in a naturalistic context, particularly given potential emergent feedback loops between AI systems, motivated actors and democratic institutions. For example, although research has shown that AI models can successfully affect people's beliefs about political issues, it is hard to forecast what impact this may have on the integrity of democracy in the future. AI systems are becoming increasingly embedded in systems and infrastructure, paving the way for technology to have an important role in the material implementation of democracy, for example, in elections. However, we also lack information about how AI may shift the balance of defence and offence in society—between those who strive to uphold our democratic freedoms and those who seek to undermine

them. Finally, it is hard to isolate the effect of AI on the foundations of democracy—for example, freedom of the press, and appropriate checks on the power of politicians and oligarchs—from those of other forms of technology.

It seems likely that AI will present specific challenges to democracy at multiple levels: epistemic, material and foundational. However, AI also holds out potential affirmative opportunities. Our analysis suggests that neither exuberant optimism nor despairing pessimism is an appropriate stance. Instead, what is called for is clear-eyed and persistent efforts to shape both the design of AI technology and the design of democratic institutions so that they fit together well, yielding democratic benefits from AI while preventing democratic harms. If we plan carefully, we should be able to assure—and even enhance—our democratic future.

References

1. Coeckelbergh, M. *Why AI Undermines Democracy and What to Do About It* (Polity, 2024).
2. Jungherr, A. Artificial intelligence and democracy: a conceptual framework. *Soc. Media Soc.* <https://doi.org/10.1177/20563051231186353> (2023).
3. Risse, M. in *The Cambridge Handbook of Responsible Artificial Intelligence* (eds Voeneke, S. et al.) 85–103 (Cambridge Univ. Press, 2022).
4. Duberry, J. Artificial intelligence and democracy: risks and promises of AI-mediated citizen-government relations. *IP* **28**, 435–438 (2023).
5. Kreps, S. & Kriner, D. How AI threatens democracy. *J. Democracy* **34**, 122–131 (2023).
6. Seger, E. Generative AI and democracy impacts and interventions. *demos.co.uk* https://demos.co.uk/wp-content/uploads/2024/04/Generative-AI-and-Democracy_Briefing-Paper.pdf (2024).
7. Landemore, H. in *Conversations in Philosophy, Law and Politics* (eds Chang, R. & Srinivasan, A.) 39–69 (Oxford Univ. Press, 2024).
8. Landemore, H. *Open Democracy: Reinventing Popular Rule for the Twenty-First Century* (Princeton Univ. Press, 2020).
9. Fishkin, J. S. *Democracy and Deliberation: New Directions for Democratic Reform* (Yale Univ. Press, 1993).
10. Aral, S. *The Hype Machine* (Currency, 2020).
11. Ziegler, D. M. et al. Fine-tuning language models from human preferences. Preprint at <https://arxiv.org/pdf/1909.08593> (2019).
12. Motoki, F., Pinho Neto, V. & Rodrigues, V. More human than human: measuring ChatGPT political bias. *Public Choice* **198**, 3–23 (2024).
13. Santurkar, S. et al. Whose opinions do language models reflect? In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 29971–30004 (JMLR, 2023).
14. Perez, E. et al. Discovering language model behaviors with model-written evaluations. In *Proc. Findings of the Association for Computational Linguistics: ACL 2023* (eds Rogers, A. et al.) 13387–13434 (Association for Computational Linguistics, 2023).
15. Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 819–862 (Association for Computational Linguistics, 2022).
16. Park, J. S. et al. Social simulacra: creating populated prototypes for social computing systems. In *Proc. 35th Annual ACM Symposium on User Interface Software and Technology* 1–18 (Association for Computing Machinery, 2022).
17. Röttger, P. et al. Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 15295–15311 (Association for Computational Linguistics, 2024).

18. Moats, D. & Ganguly, C. Bringing AI participation down to scale. *Patterns* **6**, 101241 (2025).
19. Huang, S. et al. Collective constitutional AI: aligning a language model with public input. In *Proc. 2024 ACM Conference on Fairness, Accountability and Transparency* 1395–1417 (Association for Computing Machinery, 2024).
20. DeepSeek-AI et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. Preprint at <https://doi.org/10.48550/arXiv.2501.12948> (2025).
21. Ibrahim, L., Huang, S., Ahmad, L. & Anderljung, M. Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks. Preprint at <http://arxiv.org/abs/2405.10632> (2024).
22. Schroeder, D. T. et al. How malicious AI swarms can threaten democracy. Preprint at <https://doi.org/10.48550/arXiv.2506.06299> (2025).
23. Goldstein, J. A. et al. Generative language models and automated influence operations: emerging threats and potential mitigations. Preprint at <http://arxiv.org/abs/2301.04246> (2023).
24. Marcellino, W., Beauchamp-Mustafaga, N., Kerrigan, A., Navarre Chao, L. & Smith, J. *The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0* (RAND Corporation, 2023).
25. Goel, N. et al. Artificial influence: comparing the effects of AI and human source cues in reducing certainty in false beliefs. Preprint at <https://doi.org/10.31219/osf.io/2vh4k> (2024).
26. Luciano, F. Hypersuasion—on AI’s persuasive power and how to deal with it. *Philos. Technol* **37**, 64 (2024).
27. Durmus, E. et al. Measuring the persuasiveness of language models. *Anthropic* <https://www.anthropic.com/news/measuring-model-persuasiveness> (2024).
28. Hackenburg, K., Ibrahim, L., Tappin, B. M. & Tsakiris, M. Comparing the persuasiveness of role-playing large language models and human experts on polarized US political issues. *AI Soc.* <https://doi.org/10.1007/s00146-025-02464-x> (2025).
29. Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C. & Willer, R. LLM-generated messages can persuade humans on policy issues. Preprint at https://doi.org/10.31219/osf.io/stakv_v8 (2025).
30. Broockman, D. & Kalla, J. Durably reducing transphobia: a field experiment on door-to-door canvassing. *Science* **352**, 220–224 (2016).
31. Salvi, F., Horta Ribeiro, M., Gallotti, R. & West, R. On the conversational persuasiveness of GPT-4. *Nat. Hum. Behav.* <https://doi.org/10.1038/s41562-025-02194-6> (2025).
32. Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, eadq1814 (2024).
33. Hackenburg, K. et al. The levers of political persuasion with conversational AI. Preprint at <https://doi.org/10.48550/arXiv.2507.13919> (2025).
34. Boissin, E., Costello, T. H., Alonso, D. M., Rand, D. G. & Pennycook, G. AI reduces conspiracy beliefs even when presented as a human expert. Preprint at https://doi.org/10.31234/osf.io/apmb5_v1 (2025).
35. Gallegos, I. O. et al. Labeling messages as AI-generated does not reduce their persuasive effects. Preprint at <https://doi.org/10.48550/ARXIV.2504.09865> (2025).
36. Chen, Z. et al. A framework to assess the persuasion risks large language model chatbots pose to democratic societies. Preprint at <https://doi.org/10.48550/ARXIV.2505.00036> (2025).
37. El-Sayed, S. et al. A mechanism-based approach to mitigating harms from persuasive generative AI. Preprint at <http://arxiv.org/abs/2404.15058> (2024).
38. Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J. & Rand, D. G. Quantifying the potential persuasive returns to political microtargeting. *Proc. Natl Acad. Sci. USA* **120**, e2216261120 (2023).
39. Matz, S. C. et al. The potential of generative AI for personalized persuasion at scale. *Sci. Rep.* **14**, 4692 (2024).
40. Matthes, J. et al. Understanding the democratic role of perceived online political micro-targeting: longitudinal effects on trust in democracy and political interest. *J. Inf. Technol. Politics* **19**, 435–448 (2022).
41. Zuiderveen Borgesius, F. J. et al. Online political microtargeting: promises and threats for democracy. *ULR* **14**, 82–96 (2018).
42. Heseltine, M. & Clemm Von Hohenberg, B. Large language models as a substitute for human experts in annotating political text. *Res. Politics* <https://doi.org/10.1177/20531680241236239> (2024).
43. Simchon, A., Edwards, M. & Lewandowsky, S. The persuasive effects of political microtargeting in the age of generative AI. *PNAS Nexus* **3**, pgae035 (2024).
44. Coppock, A. *Persuasion in Parallel: How Information Changes Minds about Politics* (Univ. of Chicago Press, 2022).
45. Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L. & Naaman, M. Co-writing with opinionated language models affects users’ views. In *Proc. 2023 CHI Conference on Human Factors in Computing Systems* 1–15 (Association for Computing Machinery, 2023).
46. Williams-Ceci, S. et al. Bias in AI autocomplete suggestions leads to attitude shift on societal issues. Preprint at <https://doi.org/10.31234/osf.io/mhjn6> (2024).
47. Sharma, M. et al. Towards understanding sycophancy in language models. In *Proc. Twelfth International Conference on Learning Representations* <https://openreview.net/forum?id=tvhaxkMKAn> (2024).
48. OpenAI. Memory and new controls for ChatGPT. *Open AI* <https://openai.com/blog/memory-and-new-controls-for-chatgpt> (2024).
49. Kirk, H. R., Vidgen, B., Röttger, P. & Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat. Mach. Intell.* **6**, 383–392 (2024).
50. Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. How social media shapes polarization. *Trends Cogn. Sci.* **25**, 913–916 (2021).
51. Bail, C. A. et al. Exposure to opposing views on social media can increase political polarization. *Proc. Natl Acad. Sci. USA* **115**, 9216–9221 (2018).
52. Google Jigsaw Team. Announcing experimental bridging attributes in perspective API. *Medium* <https://medium.com/jigsaw/announcing-experimental-bridging-attributes-in-perspective-api-578a9d59ac37> (2024).
53. Ahler, D. J. & Sood, G. The parties in our heads: misperceptions about party composition and their consequences. *J. Politics* **80**, 964–981 (2018).
54. Cheng, M., Piccardi, T. & Yang, D. CoMPosT: characterizing and evaluating caricature in LLM simulations. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* 10853–10875 (Association for Computational Linguistics, 2023).
55. Sætra, H. S. A shallow defence of a technocracy of artificial intelligence: examining the political harms of algorithmic governance in the domain of government. *Technol. Soc.* **62**, 101283 (2020).
56. Argyle, L. P. et al. Leveraging AI for democratic discourse: chat interventions can improve online political conversations at scale. *Proc. Natl Acad. Sci. USA* **120**, e2311627120 (2023).
57. Hadfi, R. et al. Conversational agents enhance women’s contribution in online debates. *Sci. Rep.* **13**, 14534 (2023).
58. Small, C. T. et al. Opportunities and risks of LLMs for scalable deliberation with Polis. Preprint at <https://arxiv.org/abs/2306.11932> (2023).
59. Lin, I. W. et al. IMBUE: improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. Preprint at <http://arxiv.org/abs/2402.12556> (2024).

60. Mercier, H. & Landemore, H. Reasoning is for arguing: understanding the successes and failures of deliberation. *Polit. Psychol.* **33**, 243–258 (2012).
61. Bakker, M. A. et al. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Oh, A. H. et al.) <https://openreview.net/forum?id=G5ADoRKiTyJ> (Curran Associates, 2024).
62. Fish, S. et al. Generative social choice. In *Proc. 25th ACM Conference on Economics and Computation* 985–985 (Association for Computing Machinery, 2024).
63. Boehmer, N., Fish, S. & Procaccia, A. D. Generative social choice: the next generation. Preprint at <https://doi.org/10.48550/arXiv.2505.22939> (2025).
64. Van Opheusden, B. et al. Methodology for analyzing the individual comments to the NTIA's AI RFC. *Imbue* <https://imbue.com/perspectives/ntia-rfc-analysis-individual-methodology/> (2023).
65. Brandt, F., Conitzer, V., Endriss, U., Lang, J. & Procaccia, A. D. *Handbook of Computational Social Choice* (Cambridge Univ. Press, 2016).
66. De, S., Bakker, M. A., Baxter, J. & Saveski, M. Supernotes: driving consensus in crowd-sourced fact-checking. In *WWW '25: Proc. of the ACM on Web Conf. 2025*, 3751–3761 (ACM, 2025).
67. Konya, A. et al. Using collective dialogues and AI to find common ground between Israeli and Palestinian peacebuilders. In *Proc. 2025 ACM Conference on Fairness, Accountability and Transparency* 312–333 (Association for Computing Machinery, 2025).
68. Masood Alavi, D., Wählisch, M., Irwin, C. & Konya, A. Using artificial intelligence for peacebuilding. *J. Peacebuilding Dev.* **17**, 239–243 (2022).
69. Augenstein, I. et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mach. Intell.* **6**, 852–863 (2024).
70. Yang, K. & Menczer, F. Anatomy of an AI-powered malicious social botnet. *JQD* <https://doi.org/10.51685/jqd.2024.icwsm.7> (2024).
71. Cooke, D., Edwards, A., Barkoff, S. & Kelly, K. As good as a coin toss: human detection of AI-generated images, videos, audio and audiovisual stimuli. Preprint at <http://arxiv.org/abs/2403.16760> (2024).
72. Nightingale, S. J. & Farid, H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl Acad. Sci. USA* **119**, e2120481119 (2022).
73. Spitale, G., Biller-Andorno, N. & Germani, F. AI model GPT-3 (dis) informs us better than humans. *Sci. Adv.* **9**, eadh1850 (2023).
74. Christ, M., Gunn, S. & Zamir, O. Undetectable watermarks for language models. In *Proc. Thirty Seventh Conference on Learning Theory* (eds Agrawal, S. & Roth, A.) 1125–1139 (PMLR, 2024).
75. Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W. & Feizi, S. Can AI-generated text be reliably detected? In *Proc. ICLR 2024* <https://openreview.net/forum?id=NvSwR4IvLO> (2024).
76. Dathathri, S. et al. Scalable watermarking for identifying large language model outputs. *Nature* **634**, 818–823 (2024).
77. Leibowicz, C. R., McGregor, S. & Ovadya, A. The deepfake detection dilemma: a multistakeholder exploration of adversarial dynamics in synthetic media. In *Proc. 2021 AAAI/ACM Conference on AI, Ethics and Society* 736–744 (Association for Computing Machinery, 2021).
78. Zhang, H. et al. Watermarks in the sand: impossibility of strong watermarking for language models. In *Proc. 41st International Conference on Machine Learning* 58851–58880 (JMLR, 2024).
79. Hasher, L., Goldstein, D. & Toppino, T. Frequency and the conference of referential validity. *J. Verbal Learn. Verbal Behav.* **16**, 107–112 (1977).
80. Goldstein, J. A. & Long, A. Deepfakes, elections and shrinking the liar's dividend. *Brennan Center* <https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend> (2024).
81. Vaccari, C. & Chadwick, A. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty and trust in news. *Soc. Media Soc.* <https://doi.org/10.1177/205630512090340> (2020).
82. Huang, S. & Siddarth, D. Generative AI and the digital commons. Preprint at <http://arxiv.org/abs/2303.11074> (2023).
83. Chow, A. R. AI's underwhelming impact on the 2024 elections. *Time* <https://time.com/7131271/ai-2024-elections> (30 October 2024).
84. Nakano, R. et al. WebGPT: browser-assisted question-answering with human feedback. Preprint at <http://arxiv.org/abs/2112.09332> (2022).
85. Menick, J. et al. Teaching language models to support answers with verified quotes. Preprint at <http://arxiv.org/abs/2203.11147> (2022).
86. Adair, B., Li, C., Yang, J. & Yu, C. Progress toward 'the Holy Grail': the continued quest to automate fact-checking. *northwestern.edu* <https://cj2017.northwestern.edu/documents/progress-cj2017-paper-18.pdf> (2017).
87. Hoes, E., Altay, S. & Bermeo, J. Leveraging ChatGPT for efficient fact-checking. Preprint at <https://doi.org/10.31234/osf.io/qnjkf> (2023).
88. DeVerna, M. R., Yan, H. Y., Yang, K.-C. & Menczer, F. Fact-checking information from large language models can decrease headline discernment. *Proc. Natl Acad. Sci. USA* **121**, e2322823121 (2024).
89. Sidoti, O. & McClain, C. 34% of U.S. adults have used ChatGPT, about double the share in 2023. *Pew Research Center* <https://www.pewresearch.org/short-reads/2025/06/25/34-of-us-adults-have-used-chatgpt-about-double-the-share-in-2023/#chatgpt-and-the-2024-presidential-election> (2024).
90. Mumford, L. Authoritarian and democratic technics. *Technol. Cult.* **5**, 1–8 (1964).
91. Simon, F. & Altay, S. Don't panic (yet): assessing the evidence and discourse around generative AI and elections. *knightcolumbia.org* <https://knightcolumbia.org/content/dont-panic-yet-assessing-the-evidence-and-discourse-around-generative-ai-and-elections> (2025).
92. Russian FSB cyber actor star Blizzard continues worldwide spear-phishing campaigns. *ncsc.gov.uk* <https://www.ncsc.gov.uk/news/star-blizzard-continues-spear-phishing-campaigns> (2023).
93. Hazell, J. Spear phishing with large language models. Preprint at <http://arxiv.org/abs/2305.06972> (2023).
94. Heiding, F., Lermen, S., Kao, A., Schneier, B. & Vishwanath, A. Evaluating large language models' capability to launch fully automated spear phishing campaigns: validated on human subjects. Preprint at <https://doi.org/10.48550/arXiv.2412.00586> (2024).
95. Angwin, J., Nelson, A. & Palta, R. Seeking reliable election information? Don't trust AI. *Proofnews* <https://www.proofnews.org/seeking-election-information-dont-trust-ai/> (27 February 2024).
96. Matza, M. Fake Biden robocall tells voters to skip New Hampshire primary election. *BBC News* <https://www.bbc.com/news/world-us-canada-68064247> (22 January 2024).
97. Panditharatne, M. Preparing to fight AI-backed voter suppression. *Brennan Center* <https://www.brennancenter.org/our-work/research-reports/preparing-fight-ai-backed-voter-suppression> (2024).
98. Berzon, A. & Corasaniti, N. Georgia County signs up to use voter database backed by election deniers. *The New York Times* <https://www.nytimes.com/2023/12/01/us/politics/georgia-county-election-deniers-trump.html> (1 December 2023).

99. Layne, N. Insight: pro-Trump activists swamp election officials with sprawling records requests. *Reuters* <https://www.reuters.com/world/us/pro-trump-activists-swamp-election-officials-with-sprawling-records-requests-2022-08-03/> (3 August 2022).
100. Rheault, L., Rayment, E. & Musulan, A. Politicians in the line of fire: incivility and the treatment of women on social media. *Res. Politics* <https://doi.org/10.1177/2053168018816228> (2019).
101. Di Meco, L. & Brechenmacher, S. Tackling online abuse and disinformation targeting women in politics. *Carnegie Endowment for International Peace* <https://carnegieendowment.org/2020/11/30/tackling-online-abuse-and-disinformation-targeting-women-in-politics-pub-83331> (30 November 2020).
102. Judson, E. Generative AI: the new frontier for gendered disinformation? *Demos* <https://demos.co.uk/blogs/generative-ai-the-new-frontier-for-gendered-disinformation/> (27 March 2024).
103. Murgia, M. *Code Dependent: Living in the Shadow of AI* (Picador, 2024).
104. Janatian, S., Westermann, H., Tan, J., Savelka, J. & Benyekhlef, K. From text to structure: using large language models to support the development of legal expert systems. Preprint at <http://arxiv.org/abs/2311.04911> (2023).
105. Casey, A. & Niblett, A. A framework for the new personalization of law. *Univ. Chicago Law Rev.* <https://doi.org/10.2139/ssrn.3271992> (2019).
106. Jeantet, D. & Savarese, M. Brazilian city enacts an ordinance that was secretly written by ChatGPT. *AP News* <https://apnews.com/article/brazil-artificial-intelligence-porto-alegre-5afd1240afe7b6ac202bb0bbc45e08d4> (30 November 2023).
107. Cerina, R. & Duch, R. Artificially intelligent opinion polling. Preprint at <http://arxiv.org/abs/2309.06029> (2023).
108. Sanders, N. E., Ulinich, A. & Schneier, B. Demonstrations of the potential of AI-based political issue polling. *Harvard Data Sci. Rev.* <https://doi.org/10.1162/99608f92.1d3cf75d> (2023).
109. Cerina, R. & Rouméas, É. The democratic ethics of artificially intelligent polling. *AI Soc.* <https://doi.org/10.1007/s00146-024-02150-4> (2025).
110. Gordon, M. L. et al. Jury learning: integrating dissenting voices into machine learning models. In *Proc. CHI Conference on Human Factors in Computing Systems* 1–19 (Association for Computing Machinery, 2022).
111. Lee, M. K. et al. WeBuildAI: participatory framework for algorithmic governance. *Proc. ACM Hum. Comput. Interact.* **3**, 1–35 (2019).
112. Agnew, W. et al. The illusion of artificial inclusion. In *Proc. CHI Conference on Human Factors in Computing Systems* 1–12 (Association for Computing Machinery, 2024).
113. Sætra, H. S. & Selinger, E. The siren song of technological remedies for social problems: defining, demarcating and evaluating techno-fixes and techno-solutionism. *SSRN J.* <https://doi.org/10.2139/ssrn.4576687> (2023).
114. Rawls, J. *Political Liberalism* (Columbia Univ. Press, 2005).
115. Bermeo, N. On democratic backsliding. *J. Democracy* **27**, 5–19 (2016).
116. Allen, D. & Weyl, E. G. The real dangers of generative AI. *JoD* **35**, 147–162 (2024).
117. von Hayek, F. A. *The Road to Serfdom* (Routledge, 1997).
118. Harari, Y. N. Why technology favors tyranny. *The Atlantic* (October 2018).
119. Acemoglu, D. Harms of AI. In *The Oxford Handbook of AI Governance* (ed. Bullock, J. B.) 660–706 (Oxford Univ. Press, 2023).
120. Acemoglu, D., Autor, D., Hazell, J. & Restrepo, P. *AI and Jobs: Evidence from Online Vacancies* (NBER, 2020).
121. Ekelund, H. *Why There will be Plenty of Jobs in the Future—Even with Artificial Intelligence* (World Economic Forum, 2024).
122. McAfee, A. Generally faster: the economic impact of generative AI. *storage.googleapis.com* https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/Generally_Faster_-_The_Economic_Impact_of_Generative_AI.pdf (2024).
123. Noy, S. & Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).
124. Choi, J. H., Monahan, A. & Schwarcz, D. B. Lawyering in the age of artificial intelligence. *SSRN J.* <https://doi.org/10.2139/ssrn.4626276> (2023).
125. Schaake, M. *The Tech Coup: How to Save Democracy from Silicon Valley* (Princeton Univ. Press, 2024).
126. Acemoglu, D. & Johnson, S. *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity* (Basic Books, 2023).
127. Hadfield, G. K. & Clark, J. Regulatory markets: the future of AI governance. Preprint at <http://arxiv.org/abs/2304.04914> (2023).
128. Yagoda, M. Airline held liable for its chatbot giving passenger bad advice—what this means for travellers. *BBC News* <https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know> (23 February 2024).
129. Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor* (St Martin's Press, 2017).
130. Danaher, J. The threat of algocracy: reality, resistance and accommodation. *Philos. Technol.* **29**, 245–268 (2016).
131. Koster, R. et al. Human-centred mechanism design with Democratic AI. *Nat. Hum. Behav.* **6**, 1398–1407 (2022).
132. Ovadya, A. Reimagining democracy for AI. *J. Democracy* **34**, 162–170 (2023).
133. Tessler, M. H. et al. AI can help humans find common ground in democratic deliberation. *Science* **386**, eadq2852 (2024).
134. Vykopal, I., Pikuliak, M., Ostermann, S. & Šimko, M. Generative large language models in automated fact-checking: a survey. Preprint at <https://doi.org/10.48550/arXiv.2407.02351> (2024).
135. Chin, J. & Lin, L. *Surveillance State: Inside China's Quest to Launch a New Era of Social Control* (St Martin's Press, 2022).
136. Argyle, L. P. et al. Testing theories of political persuasion using AI. *Proc. Natl Acad. Sci. USA* **122**, e2412815122 (2025).
137. Goldstein, J. A., Chao, J., Grossman, S., Stamos, A. & Tomz, M. How persuasive is AI-generated propaganda? *PNAS Nexus* **3**, pgae034 (2024).
138. Hackenburg, K. & Margetts, H. Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl Acad. Sci. USA* **121**, e2403116121 (2024).
139. Hackenburg, K. et al. Scaling language model size yields diminishing returns for single-message political persuasion. *Proc. Natl Acad. Sci. USA* **122**, e2413443122 (2025).
140. Karinshak, E., Liu, S. X., Park, J. S. & Hancock, J. T. Working with AI to persuade: examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum. Comput. Interact.* **7**, 1–29 (2023).
141. Buchanan, B., Lohn, A., Musser, M. & Sedova, K. Truth, lies and automation: how language models could change disinformation. *Center for Security and Emerging Technology* <https://cset.georgetown.edu/publication/truth-lies-and-automation/> (2021).

Author contributions

All authors contributed to conceptualizing, writing, editing and revising this manuscript.

Competing interests

The following authors are full- or part-time remunerated employees of commercial developers of AI technology: M. Bakker, I.G., N.M., M.H.T. and M. Botvinick (Google DeepMind), E.D. and D.G. (Anthropic) and T.E.

(OpenAI), A.P. (Fundamental AI Research (FAIR), Meta). C.S. and K.H. are part-time remunerated government employees (at the UK AI Security Institute). D.S. and S.H. are employees of the non-profit organization Collective Intelligence Project. A.O. is an employee of the AI & Democracy Foundation. E.S. is an employee of Demos. None of these employers had any role in the preparation of the manuscript or the decision to publish. The remaining authors declare no competing interests.

Additional information

Correspondence should be addressed to Christopher Summerfield or Matthew Botvinick.

Peer review information *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2025