PNAS PERSPECTIVE OPEN ACCESS

Can Generative AI improve social science?

Christopher A. Bail[a,b,c,1]

Edited by David Lazer, Northeastern University, Boston, MA; received September 7, 2023; accepted April 5, 2024, by Editorial Board Member Mark Granovetter

Generative AI that can produce realistic text, images, and other human-like outputs is currently transforming many different industries. Yet it is not yet known how such tools might influence social science research. I argue Generative AI has the potential to improve survey research, online experiments, automated content analyses, agent-based models, and other techniques commonly used to study human behavior. In the second section of this article, I discuss the many limitations of Generative. I examine how bias in the data used to train these tools can negatively impact social science research—as well as a range of other challenges related to ethics, replication, environmental impact, and the proliferation of low-quality research. I conclude by arguing that social scientists can address many of these limitations by creating open-source infrastructure for research on human behavior. Such infrastructure is not only necessary to ensure broad access to high-quality research tools, I argue, but also because the progress of AI will require deeper understanding of the social forces that guide human behavior.

Generative AI | computational social science | agent-based model | survey research | algorithmic bias

Generative AI—technology capable of producing realistic text, images, music, and other creative forms—continues to captivate large audiences. Many speculate such technology will impact a range of industries and scientific disciplines—from creative and legal writing to computational biology. Yet sociologists, political scientists, economists, and other social scientists are only beginning to explore how Generative AI will transform their research. In this article, I argue these tools may advance the scale, scope, and speed of social science research—and may enable new forms of scientific inquiry as well. At the same time, I assess the many limitations of Generative AI for social science research—and discuss how scholars can mitigate risks while exploring this promising new technology.

In the first section of this article, I provide a brief history of Generative AI for social scientists. In the second section, I ask whether Generative AI can effectively simulate human behavior for the purposes of social science research. I assess whether these tools can be useful for survey research, or creating experimental primes within online experiments. Next, I review recent studies that employ Generative AI models to simulate dynamic human behaviors. These include experiments where human respondents interact with Generative AI, or simulations where researchers prompt models to interact with each other to study emergent group behaviors. I argue such research may help social scientists begin to reverse engineer the "social sense" of human beings—or how we create shared understandings of acceptable behavior in different social milieux. Finally, I argue Generative AI has the potential to transform automated text analysis. Since Generative AI tools can analyze very large groups of documents in many different languages with great speed, I propose they may significantly expand the range of research questions that social scientists can study.

In the third section of this article, I turn to the various limitations and potential dangers associated with Generative AI. Much of the public discourse surrounding this new technology focuses on the possibility of a "singularity" where AI models supersede human intelligence and threaten our well-being. Many scholars believe such concerns eschew well-documented social harms that are already occurring in the short term (1). These include the tendency of Generative AI to exhibit strong bias against stigmatized groups, spread misinformation, and potentially exacerbate social inequality or climate change—among other negative outcomes. I discuss how these issues may negatively impact the quality, efficiency, interpretability, and replicability of social science research as well—and generate new questions about ethics and the protection of human subjects. I also evaluate the potential of these models to generate and disseminate "junk science" which could impede scientific inquiry for years to come. Mitigating each of these risks is challenging, I argue, because the processes used to train Generative AI are largely opaque—and accurate tools for detecting AI-generated content are not yet effective at scale.

In the final section of this article, I argue that social scientists can address many of the challenges of research with Generative AI by creating our own open-source infrastructure (2). By developing our own Generative AI models, social scientists can more effectively diagnose how the model training process impacts scientific analysis of human behavior and ensure these new tools evolve according to the interest of science, and not only the corporations that produce many of the most popular models at present. Most importantly, I argue open-source infrastructure could create a community of scholars that work to identify best practices

Author affiliations: [a]Department of Sociology, Duke University, Durham, NC 27708; [b]Department of Political Science, Duke University, Durham, NC 27708; and [c]Department of Public Policy, Duke University, Durham, NC 27708

Author contributions: C.A.B. designed research; performed research; contributed new reagents/analytic tools; and wrote the paper.

The author declares no competing interest.

This article is a PNAS Direct Submission. D.L. is a guest editor invited by the Editorial Board.

[1]Email: christopher.bail@duke.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2314021121/-/DCSupplemental.

Published May 9, 2024.

for research with Generative AI, prevent these tools from reproducing the academic caste system, and allow social scientists to develop solutions to future challenges and prevent these tools from being repurposed for malicious purposes.

Several caveats are in order. First, my analysis is limited to social science and thus does not engage with the many different ways Generative AI might shape other fields. Second, I focus on the impact of Generative AI on scientific research, and not its broader impact on social life—a topic that is certainly worthy of another analysis. Third, the field of Generative AI research is changing so rapidly that any attempt to take stock of its potential will become out of date quickly—as well as information about its possible risks or dangers. Indeed, many of the studies I discuss below are preprints that have not yet undergone rigorous peer review, and may therefore fail to replicate. I therefore urge readers to take caution in evaluating the potential of the research techniques described below, which may yet be judged scientifically unsound, unethical, or both through a more systematic future review. Third, I do not provide a technical discussion of how Generative AI models work, since these are broadly available elsewhere (3). Instead of a "user's guide" for Generative AI in social science research, I hope to inspire ongoing dialog among researchers about how this new technology should be used to study human behavior in different settings.

What is Generative AI?

The term "Generative AI" describes a broad set of tools developed by researchers in statistics, computer science, and engineering that are sometimes called "Foundation Models." At a high level, the term demarcates a shift in the use of machine learning technology from pattern recognition—where tools are created to identify latent patterns in text, images, or other unstructured datasets—toward the generation of free-form text, images, and video, via algorithms that are trained on large datasets, often collected from online sources. Large language models (LLMs) such as ChatGPT ingest vast amounts of text-based data, and identify the probability that a word (or set of words) will occur given the presence of other language patterns within a passage of text. As technology progressed to allow AI researchers to train such models on progressively larger amounts of text—and with powerful new "transformer" architectures—tools such as GPT-3 became more adept at predicting the language most likely to follow different "prompts"—short pieces of text designed to shape the LLM's outputs, such as a question. LLMs thus resemble the "autocomplete" technologies that have become pervasive on search engines, apps, and other digital spaces over the past decade, but with considerably greater scale and more sophisticated training processes that are described in additional detail below. Though scholars debate whether LLMs "understand" the output they produce, many are impressed by their capacity to mimic humans in conversational settings, synthesize disparate sources of information, and perform basic reasoning (4–6).

Parallel advancements have been made with image—and, to a lesser extent, video. Instead of calculating the probability of words given other words, Generative AI tools that create de novo images use the co-occurrence of pixels of different colors or sizes to weave together a range of synthetic visuals. These include synthetic human faces, reproductions of classic artwork, or surreal—and at times quite innovative—forms of art that have provoked both excitement and concern among people in creative industries (1). Models such as DALL-E and Stable Diffusion create such visual content through text prompts—searching for connections between patterns in the co-occurrence of words and the arrangement of pixels—that allow a user to request highly specialized visual content.

Opportunities for Social Science with Generative AI

Despite—or perhaps because of—their significant flaws, Generative AI models appear capable of impersonating humans in some settings. The computer scientist Alan Turing was among the first to propose evaluating AI by identifying whether humans can distinguish content produced by people or AI. Using GPT-2, a precursor to ChatGPT that produces much lower quality texts, Kreps et al. studied whether research participants could differentiate short statements about U.S. foreign policy generated by this LLM and human respondents (7). They found GPT-2 could successfully impersonate humans, and can even write lengthy news stories about international affairs that are judged to be as credible as those authored by real journalists. In a more recent study, Jakesch et al. examined whether human survey respondents could discern whether texts about job postings and online dating profiles were created by humans or LLMs (5). They show humans are largely unable to determine whether such texts are authored by humans or LLMs in a series of experiments. Finally, Zhou et al. show that GPT-3 can easily produce misinformation about COVID-19 that can escape detection by most social media platforms (8). More recent studies indicate AI-generated content can influence human attitudes, even if it is false or misleading (9, 10).

Despite the obvious potential for harm when Generative AI successfully impersonates humans, these same capabilities may be useful to social scientists for research purposes. For example, social science experiments often include texts or images designed to prime human respondents to behave in a certain manner, or exhibit some type of feeling. A researcher interested in studying how emotions shape responsiveness to political advertising campaigns, for example, may wish to show respondent texts or images designed to create fear before asking them about their voting intentions. Or, a researcher who aims to evaluate racial discrimination in hiring may wish to show research participants two images—one that features a Caucasian job applicant and another that depicts an African American job candidate—and subsequently evaluate participant's perceptions of the employability of the two candidates, ceteris paribus. Generative AI may be useful for creating such vignettes and/or images—especially with iterative feedback from researchers—to increase the external validity and comparability of those primes, or to protect the privacy of real humans whose images might be used in such studies.

Creating a compelling piece of short text or a single image is a relatively low-bar for Generative AI to pass (and one where it still often fails). Shorter texts provide fewer opportunities for Generative AI tools to make errors or hallucinate untruths (or half truths) that decrease their capacity to impersonate a human. Yet there is also evidence that Generative AI performs reasonably well at more complex human behaviors. For example, Argyle et al. indicate GPT-3 can accurately impersonate respondents to a nationally representative public opinion survey from a range of different demographic backgrounds (11). Prompting such tools with details about the characteristics of a respondent, for

example, makes them respond to public opinion surveys in a manner that is very similar to real respondents with the same attributes. Some argue such "silicon samples" could be used to produce more diverse samples than the convenience samples utilized by so many university researchers—and may also allow researchers to administer lengthier survey instruments, since LLMs have potentially unlimited attention spans (12). At the same time, more recent research indicates GPT 3.5 turbo produces accurate mean estimates of attitudes within a population, but understates variances—exaggerating extreme attitudes (13). Another study indicates LLMs exhibit an affirmative bias in yes/no questions (14). Studies also indicate LLMs represent some demographic subgroups more accurately than others (13, 15). Yet these studies do not employ the latest models, and only focus on one country: the United States. Understanding differences between how LLMs and humans respond to surveys may be doubly important because these tools are being trained to impersonate respondents by malicious actors seeking to game the survey industry (16). Though silicon samples will not soon displace survey research with human respondents, they may still be very useful for pretesting surveys before they are dispatched (at considerable cost) to large groups of human respondents, or imputing missing data (11, 17). Some argue Generative AI is also a useful tool for creating survey questions, or designing multi-item scales to measure abstract social concepts (18).

There is also evidence that Generative AI can be used to reproduce experiments. Horton, for example, argues synthetic research respondents created using GPT-3 can be used to reproduce several classic studies in behavioral economics (19). Similarly, Aher et al. show that GPT-3 can also reproduce classic social psychology experiments—including the infamous Milgram experiment—though they argue it cannot reproduce the "wisdom of crowds" phenomenon (20). Still other studies indicate LLMs can replicate classic experiments in cognitive science and the study of morality and replicate human behavior in the Prisoner's Dilemma and other behavioral games (21–24). Ashokkumar et al. find the correlation between treatment effects observed in 482 studies and responses created by GPT-4 is 0.86. Furthermore, they find this correlation holds for both published and unpublished studies, and across demographic subgroups (25). The capacity of Generative AI to impersonate humans may thus improve as model size increases, and as researchers experiment with prompting LLMs with even richer forms of data such as in-depth qualitative interviews or detailed life histories. I discuss the challenge of protecting user privacy in such efforts in further detail below.

Whether Generative AI can successfully impersonate humans in more complex social settings such as interpersonal conversations is much less clear. This is an important question since the Turing test is most often administered in a setting where a human can interact—and ask questions of—both an AI chatbot and a human in order to distinguish them from each other. Early attempts to create chatbots that could pass the Turing test largely failed. Rule-based chatbots such as ELIZA, the 1968 invention that delivered Rogerian psychotherapy by identifying keywords in user input and linking them to sets of responses that encouraged them to self-reflect, lacked the capacity to respond to emergent or dynamic conversational turns in a compelling manner. Chatbots that followed such simple rules were eventually displaced by those which learn from natural language use in the 2000s and 2010s. But until recently, these chatbots also appeared incapable of passing the Turing test, since they struggled to generate original content and frequently redirected conversations—or failed to follow other conventions in human conversation that made them fairly easy to identify. Generative AI holds the potential to create more realistic human-like interactions given that many such tools are trained on larger amounts of data that describe human interactions—and also because of recent technical innovations (e.g., transformer models).

A crude test of the capacity of Generative AI to generate plausibly human behavior in social settings is multiplayer online games. Though such games certainly do not simulate the full range of human behaviors that are of interest to social scientists, they may provide a useful baseline to evaluate the performance of these tools in more complex settings. Prior to the advent of Generative AI, believable characters in video games were created via simple rules, or via "reinforcement learning" where AI characters adapt their behavior based upon past experiences with human players. Key to such behavior was a system where AI agents could recall prior events—or exhibit a working memory. Such AI has been commonplace in video games for some time, and AI systems have even surpassed the capabilities of human players in a variety of more simple games such as Backgammon, Chess, and AlphaGo for many years. More recently, however, researchers have shown that LLMs can also learn to use natural language in games that require complex reasoning and high-level strategy to defeat human players, such as Diplomacy (26, 27).

Another line of research examines how the introduction of AI agents in multiplayer games shapes the behavior of the humans they play with. Dell'Aqua, Kogut, and Perkowski study a collaborative cooking game where AI's performance is known to exceed that of human players (28). When an AI agent is introduced in a team setting, the researchers find that human agents perform more poorly when the agent is on their team, compared to an all-human team. The authors argue the introduction of the AI agent makes coordination more difficult for human players—and also creates less trust among members of the team. Conversely, Traeger et al. find automated agents that are trained to perform poorly at collaborative tasks can actually improve the behavior of human team members (29). It is possible that AI which completes tasks with greater skill than humans creates frustration and in-fighting, whereas AI that demonstrates less competence encourages human empathy and collaboration to overcome poor group performance.

If groups of automated agents can create believable group behavior when dispatched in unison, this may enable new forms of research as well. Many social science theories describe group-level processes that shape individual behavior. But recruiting large groups of people to interact is often logistically impossible, prohibitively expensive—or both. Though Generative AI may never replicate the spontaneous behavior of human groups, researchers may nevertheless be able to dispatch groups of bots in online spaces to approximate such behavior. Allamong et al. provide a proof of concept of such research (30). These researchers were interested in studying how social media users behave when they are surrounded by people who do not share their political views. But recruiting social media users with heterogeneous beliefs to interact with each other is extremely difficult (31). Instead, Allamong et al. built a social media research platform

where respondents were recruited to interact with LLMs that were prompted to impersonate members of the opposing political party for ten minutes. Though respondents were told they might interact with automated accounts during the study's informed consent dialog, most participants expressed uncertainty about whether they interacted with humans or bots. These findings are preliminary due to the study's small sample size, but the research design indicates LLMs may be useful for conducting research on group-level processes provided researchers carefully monitor human–AI interactions for hallucinations or abuse in real time.

Can Generative AI Improve Simulation-Based Research? Recent studies indicate generative AI tools may also be useful for simulating large human populations in and of themselves. This may enrich the "agent-based modeling" (ABM) paradigm, in which researchers create synthetic societies to study social processes (32–34). This decades-old tradition requires researchers to create a facsimile of a social setting (such as a social network, neighborhood, or marketplace) using computer code. Researchers also create agents who interact with each other in such settings according to a set of rules proscribed by theories of human behavior (35). For example, a researcher may assign an agent membership in one of two identity groups and then simulate a contest for control of territory between them. The agents in such a model can be assigned behaviors such as maximizing their own self-interest (or that of a group to which they belong), and these parameters can be systematically varied in order to identify the range of possible outcomes within the broader social setting.

A key strength of agent-based models is that they allow researchers to explore hypothetical scenarios and identify individual-level patterns (such as in-group bias) that can create macrolevel patterns (e.g., residential segregation). Early ABMs employed agents who followed simplistic rules such as moving to new neighborhoods if people from an outgroup move into their neighborhood (34). Recent work employs more sophisticated agents that can have many characteristics—and follow multiple rules using human-like decision-making processes (e.g., bounded rationality) (32, 36, 37). Yet ABMs are often criticized for failing to capture the full spectrum of human behaviors (38). For example, conventional agents within ABMs do not use language, interpret social contexts, or engage in conversations with each other (39).

Recent studies indicate LLMs may be used to address some of the limitations of simulation-based research. Park et al. created a simulacrum where several dozen agents—independently powered by multiple instances of GPT 3.5-turbo—interacted with each other in a fictitious small-town setting (6). The researchers gave the agents personalities and traits (e.g., "a pharmacist who is gregarious"), and developed a software infrastructure which allowed agents to have memories that summarized past interactions with other agents. These agents not only developed daily routines as the simulation progressed (e.g., waking up and eating breakfast), but also demonstrated emergent group properties. For example, one agent announced she was having a party, and the other agents began to discuss whether they would attend. One of the agents even asked one of the others out on a date to attend this event, and others engaged in gossip about this burgeoning romantic relationship. Though this study created a relatively simplistic social environment with a small number of agents, it provides a proof of concept that Generative AI has the potential to advance social simulation research.

More recent studies indicate LLMs can be integrated within ABMs to develop or test more sophisticated theories of human behavior. For example, Törnberg et al. create a simulated social media platform with five hundred agents whose behaviors are calibrated using data from the American National Election Study (ANES) (39). The agents are prompted to read news stories and make posts—or like content—according to information about the social media habits and political opinions of ANES respondents. This study both reproduces known dynamics on social media platforms and simulates what might happen if they used alternative newsfeed algorithms that optimize for consensus, instead of user engagement. Similarly, Gao et al. use real-world social media data to calibrate a model they claim successfully predicts the spread of information and emotions about gender discrimination and nuclear energy (40). Another study indicates LLMs can reproduce social movement dynamics on social media (41).

Because most LLMs are probably trained with large amounts of social media data, reproducing known human dynamics within such platforms may represent a lower bound for evaluating the prospect of these tools to improve ABMs more broadly. Yet some recent studies indicate LLMs can recreate competitive dynamics within simulated economic and labor markets (42, 43), the diffusion of information and decision-making within organizations, and crisis response (41, 44, 45). Finally, there have been several recent efforts to generate software frameworks for agent-based modeling that could reduce the entry-costs for social scientists who hope to further expand this research design to even more research questions (46, 47).

At the same time integrating LLMs within ABMs may reinvigorate preexisting debates about the latter. Scholars often debate whether increasing the complexity of agents is desirable if emergent group dynamics of interest can be created with parsimonious models (33). It is likewise unclear how the performance of LLMs within ABMs should be evaluated. Is it sufficient for LLMs to reproduce known group dynamics within ABMS? Or, should they be assessed based upon their capacity to predict real-world outcomes (37)? Would we have more confidence in results generated by LLMs if they could be confirmed by multiple models (48)? If so, the probabilistic nature of LLMs may complicate scholars' ability to reproduce each other's findings, as I discuss in additional detail below. Future studies are also needed to identify whether LLMs make ABMs more sensitive to stochasticity—or if they provide more realistic representations of the unpredictable nature of so much human behavior.

If such issues can be addressed, integrating LLMs and ABMs could open new lines of inquiry. This approach could be employed to study topics that are very difficult to examine in real life (such as violent extremism on social media), or to study populations that are very difficult to access (e.g., violent extremists) (12). Simulations might also inform what little observational research we have on these topics—and could be calibrated using these observational data as well. Emergent group behaviors identified through simulation research could further inform observational data collection in turn—or, potentially—social interventions designed to prevent such behavior. Far more research is needed to determine whether LLM-based simulations are realistic enough to be useful in such endeavors—especially since many populations that are

difficult to study may not be well represented in the training data used to create Generative AI.

Can Generative AI Improve Text Analysis? Regardless of whether Generative AI can effectively simulate human behavior, it may also help social scientists with other common research tasks such as content analysis of text-based data. Wu et al. demonstrate GPT-3.5 can produce accurate classifications of the ideology of U.S. elected officials by analyzing their public statements (49). They passed the names of random pairs of elected officials to the model and asked it to identify which of the two was "more conservative" or "more liberal." The results closely approximate the popular DW-Nominate method for measuring the ideology of elected officials using roll-call voting, but also identified more nuance within moderates who often vote against the extreme wings of their parties. Similarly, Yang and Menczer argue GPT-3.5 can accurately code the credibility of media sources (50). Gilardi et al. argue GPT 3.5-turbo can accurately measure the topic of tweets, the stance or opinions of their authors, and the "frames" used to organize the message in a narrative manner (51). In addition to passing GPT 3.5-turbo the full text of tweets, these researchers also fed the coding instructions that would typically be assigned to human coders as a prompt to the model. They find this model performs better than human workers trained with such materials on Amazon Mechanical Turk—though such coders are known to be less accurate than those trained directly by researchers in small group settings. Mellon et al., however, compared the coding performance of several prominent LLMs to highly trained coders who were instructed to analyze statements about British Elections (52). They find LLMs produced the same classification roughly 95% of the time. Argyle et al. also demonstrate that LLMs have considerable potential for coding the topic of unstructured conversations between multiple people using a mobile chat platform (53).

Ziems et al. offer perhaps the most systematic analysis of the capabilities of LLMs for coding texts to date (54). Using datasets coded by experts from sociology, political science, and psychology—as well as nonsocial science fields such as history, literature, and linguistics—they compare the capabilities of LLMs to reproduce the work of human expert annotators. Overall, they find LLMs perform well—particularly in coding data created by political scientists and sociologists. Unsurprisingly, they find the latest models perform best. LLMs appear to assign more accurate codes for some topics than others, however, which may be an artifact of the way they were trained. That such models can reproduce coding decisions of humans without any specific training is encouraging, but Ziems et al. argue usage of LLMs will still require some degree of human supervision, and familiarity with task-specific prompt-engineering. Usefully, these authors also present a reproducible data analysis pipeline for ongoing evaluation of future models and other datasets. Social scientists have also begun to identify best practices for coding tasks with LLMs. For example, Törnberg provides a practical guide for how to set up coding workflows using APIs, and provides detailed recommendations about how to write prompts for social science coding tasks (55).

The studies reviewed above indicate text analysis may be among the most promising ways Generative AI might improve social science research. That LLMs cannot yet match the accuracy of expert human coders means that they will not soon displace conventional text analysis. But human coders are also prone to a variety of well-documented errors that range from subjective bias to inconsistency and lack of attention—particularly when researchers organize small teams to code documents in a coordinated fashion. LLMs can also exert bias and be inconsistent, as I discuss in further detail below. But LLMs may enable social scientists to examine corpora of unprecedented size with unforeseen speed. Rather than taking a random sample of documents, for example, social scientists now have the potential to code an entire corpus in short order. LLMs also appear capable of performing coding tasks in many of the world's most prominent languages as well as other rudimentary tasks typically assigned to human research assistants such as data coding or data entry (as discussed in greater detail in *SI Appendix*) (56). More studies are needed to evaluate the promise of LLMs for text analysis—and to evaluate possible privacy issues discussed in additional detail below. But for now, they appear poised to have a significant impact upon the range of questions social scientists can ask with text-based data. Multimodal models that can translate content from images into text suggest these coding capabilities may soon apply to other mediums as well.

Limitations and Possible Dangers

I have presented an optimistic view of the potential for Generative AI to improve social science thus far. But these tools have well-documented limitations that could negatively impact social science research. In the following sections, I discuss these limitations in detail.

Generative AI Exhibits Human Biases. Most AI tools are trained using data created by humans, and frequently exhibit a broad range of prejudice and cognitive errors accordingly (1, 57–60). Generative AI has heightened concerns about bias, because these tools are trained on large amounts of data created by humans on the internet—where intergroup prejudice is pervasive. One way to assess the scale and direction of bias in Generative AI is to ask LLMs to complete public opinion surveys. Santurkar et al. asked a series of LLMs trained by OpenAI and A121 Labs to respond to questions within a large group of surveys administered within the United States (15). They compared how the models responded to questions about abortion, gun control, and a range of other topics. They find most LLM's responses are more liberal than the general population, and reflect those who are younger and have more education. LLMs are particularly unlikely to perform the responses of those over sixty-five years old, those who are widowed, or those who identify as Mormon. Other researchers have shown that LLMs tend to exhibit bias against women and racial minorities (59, 61). LLMs also appear to have distinctive personality characteristics—specifically, they are more likely to be extroverted and agreeable than neurotic (62). This may be due to the fact that many LLMs are created with customer service applications in mind.

Santurkar et al. show that bias within LLMs can be partially addressed using prompt engineering—for example, asking the model to perform the role of a specific group (e.g., a wealthy Republican from Texas) (15). This mirrors earlier research which suggests removing bias from AI tools may be easier than removing it from human populations (63). However, such strategies depend critically upon the capacity of researchers to identify bias in the first place. This is no easy task when the processes used to train the most popular

Generative AI models—such as GPT-4—are largely unknown. Without access to the types of training data fed into such models, researchers can only examine "known unknowns." If poor elderly people in rural areas are unable to voice their collective concern about how Generative AI represents them, for example, researchers may be unlikely to identify such bias.

A key question for social scientists is whether the bias of Generative AI is a "bug" or a "feature" for research purposes. We often design experiments that examine the impact of bias on attitudes or behaviors. If bias in Generative AI tools can be carefully controlled—a major assumption—it could allow researchers to study its impact in empirical settings (for example, a survey respondent evaluating a hypothetical applicant for a job). It is further possible that Generative AI might be useful in "reverse engineering" some types of bias. Running experiments on the pronouns produced in response to a broad range of prompts, for example, has the potential to identify new types of gender discrimination—particularly within the online settings that produce the training data for Generative AI tools (61). On the other hand, the inability of Generative AI tools to perform accurate representations of people from marginalized groups could hinder social science research. Those who hope LLMs might help researchers assess the impact of their interventions among more diverse populations might be disappointed by the quality of such impersonations because of insufficient training data.

But there is yet another challenge: one of the most important stages in training a Generative AI model is when its developers provide it with feedback through "fine-tuning" or "reinforcement learning with human feedback." AI companies usually attempt to train their models to avoid making racist statements, for example. This process typically occurs behind closed doors via "red team" attacks designed to goad the model into producing prejudiced, dangerous, or illegal content. Developers then create workflows to prevent the models from discussing such content. Though such guardrails probably improve the safety of Generative AI tools for public use, they may impede the ability of social scientists to leverage bias for research purposes (12). Researchers who want to use LLMs to impersonate biased groups, for example, may discover these tools are unwilling to perform such roles because they have been fine-tuned according to the normative preferences of highly educated liberals who may have more concern about the protection of marginalized groups than others (12, 64). To the extent that most proprietary LLMs are trained to be helpful chat assistants, they may also differ from typical human populations in other difficult-to-detect ways. One study, for example, suggests LLMs exhibit more rational behavior than humans (65). But there is also some evidence that the opposite problem may exist: fine-tuning LLMs to pass the Turing Test may make them more likely to share inaccurate information (66).

Will Generative AI create "Junk Science"? The potential for malicious actors to use Generative AI to spread misinformation in the short term is very concerning because tools such as LLMs are so adept at impersonating humans at scale (67). But the capacity of Generative AI to produce inaccurate information with confidence may also create insidious problems in the long term. As the internet becomes increasingly flooded with biased or inaccurate texts and images generated by AI, what will prevent future models from training themselves on these same flawed data? A recent example of how such a scenario might unfold is Stack Overflow, a popular "question and answer" website that software developers use to help each other write code. As enthusiasm about the capacity of Generative AI to write code peaked, some users created bots that automatically passed people's questions about software to LLMs. Though many of the answers produced by the LLM were high quality, others were completely incorrect. The website quickly announced a new policy banning LLMs to prevent a situation where users would struggle to distinguish the good information from the bad.

Researchers who rely upon LLMs to perform literature reviews, generate new research questions, or summarize large corpora they are unable to read may face similar problems. Journals and funding agencies may find themselves overwhelmed by low-quality "junk-science" created by LLMs. Computer scientists have begun to create digital "watermarks" that flag AI-generated content. Watermarks are already being used in Generative AI models that create images, but they are somewhat more difficult to implement within LLMs. One proposal is to create an "accent" for LLMs—giving them a list of words they should use whenever possible—to allow people to retrospectively identify content that was not generated by humans (68). But even this proposal will be difficult to implement at scale. Each entity that develops LLMs will not only have to agree to use watermarks, but they will also need to coordinate with each other. Large companies might be encouraged to do this through government regulation. But such coordination would be unable to detect LLMs created by individuals skilled enough to develop smaller models on their own.

Is Research with Generative AI Ethical? Among the most pressing questions for social scientists is whether research with Generative AI is ethical (69). This question is particularly important since many Generative AI tools exhibit biases that are not only offensive (e.g., racism or misogyny), but may also hallucinate inaccurate information that could be shared by research participants on social media platforms, or elsewhere. While these questions may be less important for social scientists using Generative AI in a carefully supervised manner—for example, using DALL-E to generate a picture of a person that might be used in a survey experiment—they take on added importance in situations where human research participants might have conversations with a LLM in an unsupervised manner. On the other hand, studies that require humans to interact with each other also risk exposing research participants to offensive language, misinformation, or abuse. Indeed, one might argue that the risks of such behavior with real human populations might be greater than research that deploys carefully prompted Generative AI in interactive settings.

Another important question is whether researchers must always obtain informed consent before exposing study participants to Generative AI. This practice appears critical for any study where a respondent could be exposed to misinformation or abusive language generated by LLMs. Yet disclosing the role of Generative AI in research also decreases its scientific utility for simulating human behavior. This is because disclosing the existence of Generative AI within a research context would make it difficult for researchers to know whether study participants' attitudes and behaviors are shaped by their experiences interacting with synthetic agents, or their attitudes toward AI more broadly (5).

One solution to this problem may be to design studies in which research participants are informed they may interact with AI during a study, but employ a mix of human and AI agents within interactive settings. Even this strategy, though,

creates the risk that an AI agent could encourage conflict between human participants. Some of these risks might be mitigated via content moderation filters that are currently available for some LLMs—and through rigorous testing of the prompts used to guide LLMs in research settings. Yet given the probabilistic nature of these models—and the ever-changing ways abuse and harassment can occur in online settings—such strategies will require great care.

Another strategy is to design studies where Generative AI acts as a mediator between human participants. For example, Argyle et al. recruited a large group of Americans with opposing views about gun regulation to participate in a peer-to-peer chat on an online forum (53). In the experimental condition, one person in each pair was shown a rephrasing of a message they were about to send to their partner created by GPT-3. These rephrasings employed evidence-based insights from social science about how to make conversations about divisive issues less polarizing (e.g., active listening). The researchers found this intervention made conversations about gun control more productive and less stressful for those whose partner used recommendations from GPT-3. This intervention does not require deception since human impersonation is not necessary to evaluate the research question. Furthermore, the researchers did not force human participants to accept the rephrasings proposed by GPT-3; rather, they were allowed to choose from several of them, edit their original message, or reject all of them.

A final strategy might be to use Generative AI to try to diagnose possible ethical issues. Earlier I mentioned that researchers demonstrated that GPT-3 could perform the responses characteristic of participants in the infamous Milgram experiment. In this study, research participants were asked to administer a lethal shock to another participant whom they could not see. Milgram showed that many respondents were willing to do so out of deference to authority, but the study was widely criticized for creating trauma among participants. If a similar experiment were attempted today about an issue that is not yet widely viewed as unethical, could GPT-3 be used to simulate outcomes before the study is launched with human participants? If so, could such simulations help researchers evaluate the likelihood of ethical issues ante facto? Because LLMs are trained using retrospective data, they may be of limited utility in predicting ethical issues on the horizon, but they may nevertheless help researchers learn from each other's mistakes. Similarly, these tools might be useful for detecting plagiarism or data fabrication as well.

Though Generative AI might help us solve some ethical problems—such as using simulations to study dangerous social interventions—it also raises new concerns about privacy and confidentiality. If a researcher uses GPT-4 to code a series of in-depth interviews about a sensitive topic such as intimate partner violence, the full-text of these interviews may be logged inside private corporations that are not beholden to the same standards for protecting human subjects as university researchers. Even worse, such data could potentially be sold to other corporations.

A final ethical concern is the impact of Generative AI on climate change. A 2019 study indicates training a single LLM may generate as much carbon dioxide as the lifetime emission of five automobiles (70). Since the size of Generative AI models has grown considerably since 2019, social scientists must carefully reflect upon the presumably much larger environmental costs of developing such technologies—even if recent engineering advances have made training processes more efficient. The cost of training models must also be weighed against the efficiencies they create, however. One study, for example, suggests the carbon emissions of writing and illustrating are lower for AI than for humans (71).

Is Research with Generative AI Replicable? A key pillar of the open-science movement is that researchers should design studies that can be replicated by others. Though Generative AI may help researchers increase the external validity of their research designs, this may come at the cost of internal validity—or the capacity of different groups of researchers to reproduce or replicate each other's results (2, 12). As I mentioned at the outset of this article, many of the studies discussed above are preprints that have not yet been peer-reviewed, much less replicated. Determining how many of these studies will replicate several years from now should therefore be a central concern of anyone evaluating the promise of Generative AI for social science.

There are several reasons why research with Generative AI is difficult to replicate. First, these models are probabilistic in nature. Even a single researcher using identical prompts in a successive manner should expect a LLM to produce different responses. The tendency for LLMs to produce heterogeneous results can be partly controlled via "temperature" parameters that regulate the predictability of model output. At present, however, there are no standards for what values of this parameter are appropriate (72). Lower values might produce more dependable results, but a researcher who employs LLMs to interact with humans in a field experiment may not want them to become overly repetitive. Studies indicate subtle differences in the wording of prompts can also produce very different output within the same LLM (13). Fortunately, researchers are making progress identifying the sensitivity of prompt variations via automated processes that perturb text or vary the amount of context provided (73). But best practices for such sensitivity analyses have not yet been identified in social science—or any other field—to my knowledge.

Most LLMs also produce different results over time. This is because many of them are constantly being fine-tuned to make them more effective or create new safeguards against bias or illicit behavior. This may create "drift" within LLMs, wherein improving model performance in one area might change the outputs they produce in another domain (74). Finally, social scientists must consider broader forms of temporal validity (13, 75). As LLMs evolve in response to user behavior in different ways—as well as ongoing events in the world—this will create another significant challenge for those who strive for reproducible research with Generative AI.

In addition to assessing the reproducibility of findings within the same model, social scientists should also consider the stability of results across multiple models. There is already some evidence that different models will produce substantially different results. Ziems et al., for example, report substantial differences in the ability of ChatGPT and Google's FLAN model to reproduce expert coding of qualitative documents across a range of social science fields (54). Similarly, Santurkar et al. report substantial differences in the capacity of different LLMs to accurately represent demographic groups across surveys (15). The source of such discrepancies across models is very difficult to explain since the processes used to train and fine-tune large proprietary models are so opaque. At the very least, however, scholars should clearly report which versions of LLMs they use, and precisely when they performed the analysis.

Creating Open-Source Infrastructure for Social Science Research

As the sections above describe, Generative AI has many limitations for social science research—most of which we are only beginning to understand. How can social scientists work together to minimize the risks of research with Generative AI without sacrificing the many opportunities it creates? Accomplishing this agenda will require deeper understanding of how Generative AI tools are built and fine-tuned. Yet such information about GPT-4 and other leading proprietary models remain closely guarded industry secrets. Indeed, OpenAI has not even disclosed even the most basic information about GPT-4—such as its size, or number of parameters.

If social scientists become reliant upon proprietary models, we also risk tying our endeavors to the vicissitudes of corporate interests (2). For the time being, both Meta and Google have made two of their models available to the public: Llama and Gemma. These models are well documented and released alongside scientific papers that can help social scientists evaluate their potential. Meta has even released detailed information about the inner workings of the model such as the numeric weights its uses to respond to prompts. This enables researchers to better control the fine-tuning process, and study how LLMs work. Yet Meta recently stopped providing information about the datasets used to train Llama, and does not provide licenses without restriction, as is customary with open-source software. This may portend challenges on the horizon. There is no guarantee that Meta will suddenly restrict access to its Generative AI tools, or begin charging researchers to access it like their peers. Though researchers at wealthy institutions may be able to afford such fees, many others may not—reproducing inequality within the academic caste system (76).

There is precedent for concern about technology companies transitioning from generous data-sharing models to those that are highly restrictive. The example of social media companies is instructive. In the early 2010s, many companies shared generous amounts of data with researchers. This enabled entirely new forms of scholarly research, and social scientists and researchers inside industry frequently collaborated with each other and presented their work at conferences. In more recent years, data-sharing practices at most large social media companies have been discontinued (77). The Academic Developer Program at Twitter that I helped create once allowed social scientists to collect vast amount of data. Access to much smaller amounts of data is now prohibitively expensive for most researchers.

One alternative would be for social scientists to develop their own, open-source Generative AI models (2). Such an effort could build upon the recent proliferation of open-source Generative AI tools catalogued by Hugging Face. In addition to being free to use, most open-source models provide more transparency than their proprietary counterparts. This not only allows researchers to better understand the processes used to train and fine-tune Generative AI, but it could also allow them to control such processes altogether. Some newer models created by Mistral even offer model weights that are "raw," or not fine-tuned. Social scientists could build similar models to better control when and how models become biased—especially in research settings where the values built into models by large corporations may inhibit research. Social scientists could also work together to create training data for Generative AI tools, which would allow us to exert even further control over their behavior. Open-source models also have privacy benefits. Prompts used by researchers could be carefully protected, instead of being potentially resold to third-parties or used to develop future models (72).

Open-source models also often create and sustain a community of people with shared concerns. Rather than guessing when and how proprietary models may exhibit bias—or endlessly testing different prompts to achieve research goals—social scientists could work together to identify the limitations of Generative AI tools for social science research. Transparent, public discussion about Generative AI may also help researchers evaluate some of the other risks above—such as the dissemination of misinformation. Social scientists could also design open-source Generative AI tools to maximize the chances that research can be replicated by running experiments designed to test whether certain training and fine-tuning processes enable scholars to more easily reproduce each other's work.

Another reason for social scientists to consider building their own infrastructure is that the quality and performance of open-source Generative AI tools has increased markedly in recent months. Though benchmarks for evaluating the performance of LLMs are hotly debated, it is noteworthy that Llama 3 and R Command + and other freely available models are beginning to approach the performance of proprietary models in many areas. More importantly, two recent analyses suggest open-source models perform well at some of the types of social science applications described above, such as coding unstructured text. One study indicates open-source models perform better than crowd-workers at annotating news articles and tweets—even though they perform slightly less well than ChatGPT on these tasks (72). Another indicates the open-source model FLAN outperforms proprietary models on a number of text-coding tasks, compared to ground-truth labels created by human experts (54).

On the other hand, open-source models may create a range of new risks. Many worry that publicly releasing model weights—or the process used to train LLMs—will empower malicious actors to build custom LLMs to spread misinformation on social media, conduct personalized phishing campaigns, create nonconsensual pornography, or access dangerous information about biotechnology and weapons. Yet multiple groups of leading scholars and policy experts argue such risks are marginal when compared to information that is already readily available on the internet (78, 79).

Nevertheless, a suite of open-source tools for research on social science would require careful leadership from a diverse group of scholars, familiar with the risks and advantages of Generative AI. This council could be charged with determining what components of the model and training process should be publicly released—and whether access to some of this information should be restricted to protect against abuse. A council could also weigh the benefits of building large models against the environmental costs related to the amount of electricity currently required to train and power them. Such an effort would require a broader organization to implement the decisions of the Council, and thus considerable financial resources. It would require administrators and other staff responsible for controlling access to the model and creating the infrastructure necessary for scholars to use it (e.g., APIs and cloud services). It would also require technical staff to develop and maintain the tools. If such an organization could be realized, however, it would not only improve access to state-of-the-art technology among a broad group of scholars, but also improve our capacity to build disciplinary standards of open science and ethics into research with Generative AI.

Such an organization could also explore broader common goods, such as the creation of a large silicon sample of human populations that researchers can use to conduct preliminary tests of human subjects, or an open-source codebase for integrating LLMs into agent-based models.

Conclusion

Few technologies have created so much excitement—and so much concern—as Generative AI. Hype cycle dynamics indicate expectations for these tools may soon reach their peak, and crash down rapidly as users become more familiar with their limitations (80). I expect social scientists will continue to play a key role in identifying those pitfalls given their extensive experience studying subjects such as bias and misinformation. But I also hope that social scientists will not become so preoccupied by the limitations of Generative AI that we do not fully evaluate its promise. For every new problem these tools create, they also hold the potential to solve many others. If the capabilities of these tools continue to expand at a fraction of their current pace, Generative AI may become a fixture within the social scientist's toolkit much sooner than many researchers realize.

Above all, I encourage social scientists not to think of themselves as mere "end-users" of Generative AI. I predict the future of AI research will require training models to better understand the science of social relationships—for example, how an AI agent should interact in group settings where the goal is not simply to provide utility for a single user, but to navigate the more complex challenges associated with emergent group behaviors. If I am correct, social scientists may soon find themselves at the center of efforts to "reverse engineer" what the sociologist William H. Sewell Jr. calls the "social sense." That is, the ability for Generative AI to detect and navigate the taken-for-granted social norms and expectations that guide so much human behavior—especially those that are rarely captured by our pens (or keyboards). This will require a much more sophisticated understanding of how the behavior of individual agents is constrained by social networks, institutions, organizations, and other extra-individual factors that are cornerstones of the science of human behavior.

Data, Materials, and Software Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. For helpful comments on previous versions of this manuscript, I am grateful to the Editor, three anonymous reviewers, Lisa Argyle, Tyson Brown, Petter Törnberg, Sunshine Hillygus, Isaac Mehlhaff, Patrick Park, Lynn Smith-Lovin, and Jessi Streib.

1. K. Sayash, A. Narayanan, *AI Snake Oil* (Princeton University Press, 2024).
2. A. Spirling, Why open-source generative AI models are an ethical way forward for science. *Nature* **616**, 413 (2023).
3. A. Vaswani *et al*., Attention is all you need. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.1706.03762 (Accessed 20 December 2023).
4. M. Mitchell, D. C. Krakauer, The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2215907120 (2023).
5. M. Jakesch, J. T. Hancock, M. Naaman, Human heuristics for AI-generated language are flawed. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2208839120 (2023).
6. J. S. Park *et al*., Generative agents: Interactive simulacra of human behavior. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2304.03442 (Accessed 20 December 2023).
7. S. Kreps, R. M. McCain, M. Brundage, All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *J. Exp. Polit. Sci.* **9**, 104–117 (2022).
8. J. Zhou, Y. Zhang, Q. Luo, A. Parker, M. Choudhury, "Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions" in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York, NY, 2023).
9. L. D. Griffin *et al*., Susceptibility to influence of large language models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2303.06074 (Accessed 20 December 2023).
10. E. Karinshak, S. X. Liu, J. S. Park, J. T. Hancock, Working with AI to persuade. *Proc. ACM Hum.-Comput. Interact.* **7**, 116:1–116:29 (2023).
11. L. P. Argyle *et al*., Out of one, many: Using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
12. I. Grossmann *et al*., AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
13. J. Bisbee, J. Clinton, C. Dorff, B. Kenkel, J. Larson, Synthetic replacements for human survey data? the perils of large language models. socArXiv [Preprint] (2023). https://doi.org/10.31235/osf.io/5ecfa (Accessed 20 December 2023).
14. V. Dentella, F. Günther, E. Leivada, Systematic testing of three language models. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2309583120 (2023).
15. S. Santurkar *et al*., Whose opinions do language models reflect? arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2303.17548 (Accessed 20 December 2023).
16. V. Veselovsky *et al*., Prevalence and prevention of large language model use in crowd work. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.15683 (Accessed 19 December 2023).
17. J. Kim, B. Lee, AI-augmented surveys: Leveraging large language models for opinion prediction in nationally representative surveys. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2305.09620 (Accessed 20 December 2023).
18. F. M. Götz, R. Maertens, S. Loomba, S. van der Linden, Let the algorithm speak. *Psychol. Methods*, 10.1037/met0000540 (2023).
19. J. J. Horton, Large language models as simulated economic agents: What can we learn from homo silicus? arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2301.07543 (Accessed 20 December 2023).
20. G. Aher, R. I. Arriaga, A. T. Kalai, Using large language models to simulate multiple humans and replicate human subject studies. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2208.10264 (Accessed 20 December 2023).
21. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
22. D. Dillion, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? *Trends Cognit. Sci.* **27**, 597–600 (2023).
23. Q. Mei, Y. Xie, W. Yuan, M. O. Jackson, A turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2313925121 (2024).
24. C. Xie *et al*., Can large language model agents simulate human trust behaviors? arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2402.04559 (Accessed 20 December 2023).
25. A. Ashokkumar, L. Hewitt, I. Ghezae, R. Willer, Prediction of social science experimental results using large language models (Working Paper, 2024).
26. A. Bakhtin *et al*., Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* **378**, 1067–1074 (2022).
27. O. Vinyals *et al*., Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
28. F. Dell'Acqua, B. Kogut, P. Perkowski, Super Mario Meets AI: Experimental effects of automation and skills on team performance and coordination (SSRN Scholarly Paper, 2020).
29. M. L. Traeger, S. Strohkorb Sebo, M. Jung, B. Scassellati, N. A. Christakis, Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6370–6375 (2020).
30. M. B. Allamong *et al*., Outnumbered online: An experiment on Partisan imbalance in a dynamic social media environment. OSF [Preprints] (2023). https://doi.org/10.31219/osf.io/tygec (Accessed 20 December 2023).
31. J. Becker, E. Porter, D. Centola, The wisdom of partisan crowds. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 10717–10722 (2019).
32. R. Axelrod, *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration* (Princeton University Press, 1997).
33. J. M. Epstein, R. Axtell, *Growing Artificial Societies: Social Science from the Bottom Up* (Brookings Institution Press, 1996).
34. T. C. Schelling, *Micromotives and Macrobehavior* (WW Norton & Company, 1978).
35. M. W. Macy, R. Willer, From factors to actors: Computational sociology and agent-based modeling. *Annu. Rev. Sociol.* **28**, 143–166 (2002).
36. D. Byrne, G. Callaghan, *Complexity Theory and the Social Sciences: The State of the Art* (Routledge, 2022).
37. R. L. Axtell *et al*., Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7275–7279 (2002).
38. R. Conte, M. Paolucci, On agent-based modeling and computational social science. *Front. Psychol.* **5**, 668 (2014).
39. P. Törnberg, D. Valeeva, J. Uitermark, C. Bail, Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2310.05984 (Accessed 20 December 2023).
40. C. Gao *et al*., S3: Social-network simulation system with large language model-empowered agents. arXiv [Preprint] (2023). https://arxiv.org/abs/2307.14984 (Accessed 19 December 2023).
41. X. Mou, Z. Wei, X. Huang, Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2402.16333 (Accessed 20 December 2023).
42. Q. Zhao *et al*., CompeteAI: Understanding the competition behaviors in large language model-based agents. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.17512 (Accessed 19 December 2023).

43. Y. Li, Y. Zhang, L. Sun, Metaagents: Simulating interactions of human behaviors for LLM-based task-oriented coordination via collaborative generative agents. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.06500 (Accessed 19 December 2023).
44. N. Ghaffarzadegan, A. Majumdar, R. Williams, N. Hosseinichimeh, Generative agent-based modeling: Unveiling social system dynamics through coupling mechanistic models with generative artificial intelligence. arXiv [Preprint] (2023). https://arxiv.org/abs/2309.11456 (Accessed 19 December 2023).
45. B. Xiao, Z. Yin, Z. Shan, Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. arXiv [Preprint] (2023). https://arxiv.org/abs/2311.06957 (Accessed 19 December 2023).
46. A. S. Vezhnevets *et al*., Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. arXiv [Preprint] (2023). https://arxiv.org/abs/2312.03664 (Accessed 19 December 2023).
47. Z. Kaiya *et al*., Lyfe agents: Generative agents for low-cost real-time social interactions. arXiv [Preprint] (2023). https://arxiv.org/abs/2310.02172 (Accessed 19 December 2023).
48. R. Axtell, R. Axelrod, J. M. Epstein, M. D. Cohen, Aligning simulation models: A case study and results. *Comput. Math. Organ. Theory* **1**, 123–141 (1996).
49. P. Y. Wu, J. Nagler, J. A. Tucker, S. Messing, Large language models can be used to scale the ideologies of politicians in a zero-shot learning setting. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2303.12057 (Accessed 20 December 2023).
50. K. C. Yang, F. Menczer, Large language models can rate news outlet credibility. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2304.00228 (Accessed 20 December 2023).
51. F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2305016120 (2023).
52. J. Mellon *et al*., Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, 10.1177/20531680241231468 (2024).
53. L. P. Argyle *et al*., Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2311627120 (2023).
54. C. Ziems *et al*., Can large language models transform computational social science? arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2305.03514 (Accessed 20 December 2023).
55. P. Törnberg, How to use LLMs for text analysis. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2307.13106 (Accessed 20 December 2023).
56. A. Korinek, Language Models and Cognitive Automation for Economic Research (National Bureau of Economic Research, 2023).
57. R. Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity, Cambridge, UK/Medford, MA, ed. 1, 2019).
58. S. Lazar, A. Nelson, AI safety on whose terms? *Science* **381**, 138 (2023).
59. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York, NY, 2021), pp. 610–623.
60. K. Yee, U. Tantipongpipat, S. Mishra, Image Cropping on Twitter. *Proc. ACM Hum.-Comput. Interact.* **5**, 1–24 (2021).
61. W. I. Cho, J. W. Kim, S. M. Kim, N. S. Kim, "On measuring gender bias" in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (2019).
62. M. Pellert, C. Lechner, C. Wagner, B. Rammstedt, M. Strohmaier, AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspect. Psychol. Sci.*, 10.1177/17456916231214460 (2024).
63. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
64. P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, K. Kersting, Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258–268 (2022).
65. Y. Chen, T. X. Liu, Y. Shan, S. Zhong, The emergence of economic rationality of GPT. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2316205120 (2023).
66. D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Automatic detection of generated text is easiest when humans are fooled. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.1911.00650 (Accessed 20 December 2023).
67. S. Feuerriegel *et al*., Research can help to tackle AI-generated disinformation. *Nat. Hum. Behav.* **7**, 1818–1821 (2023).
68. J. Kirchenbauer *et al*., A watermark for large language models. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2301.10226 (Accessed 20 December 2023).
69. L. Weidinger *et al*., "Taxonomy of risks posed by language models" in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT'22)* (Association for Computing Machinery, New York, NY, 2022), pp. 214–229.
70. E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP. arXiv [Preprint] (2019). https://doi.org/10.48550/arXiv.1906.02243 (Accessed 20 December 2023).
71. B. Tomlinson, R. W. Black, D. J. Patterson, A. W. Torrance, The carbon emissions of writing and illustrating are lower for AI than for humans. *Sci. Rep.* **14**, 3732 (2024).
72. M. Alizadeh *et al*., Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. arXiv [Preprint] (2023). https://arxiv.org/abs/2307.02179 (Accessed 19 December 2023).
73. C. Si *et al*., Prompting GPT-3 to be reliable. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2210.09150 (Accessed 20 December 2023).
74. L. Chen, M. Zaharia, J. Zou, How is ChatGPT's behavior changing over time? arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2307.09009 (Accessed 20 December 2023).
75. K. Munger, Temporal validity as meta-science. *Res. Polit.* **10**, 1–10 (2023).
76. A. Clauset, S. Arbesman, D. B. Larremore, Systematic inequality and hierarchy in faculty hiring networks. *Sci. Adv.* **1**, e1400005 (2015).
77. D. Freelon, Computational research in the post-API age. *Polit. Commun.* **35**, 1–4 (2018).
78. R. Bommasani *et al*., *Considerations for Governing Open Foundation Models. Issue Brief Human, Stanford HAI* (Elsevier, 2023).
79. S. Kapoor *et al*., On the societal impact of open foundation models. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2403.07918 (Accessed 20 December 2023).
80. M. Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton University Press, Princeton, NJ, 2018).