Can Generative Artificial Intelligence Improve Social Science?

Christopher Bail[1,✉]

[1]Duke University, Department of Sociology, Department of Political Science, & Sanford School of Public Policy

Generative Artificial Intelligence 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. In the first section of this article, I assess the potential of Generative AI to improve survey research, online experiments, automated content analyses, agent-based models, and other techniques commonly used to study human behavior. I also discuss the potential of these tools to perform literature reviews, identify novel research questions, and facilitate routine research tasks such as writing, data cleaning, and software programming. In the second section of this article I discuss the many limitations of Generative AI, and whether these tools can be deployed by researchers in an ethical manner. 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 internal and external validity, reproducibility, efficiency, and the proliferation of low-quality research. I conclude by highlighting the need for increased collaboration between social scientists and artificial intelligence researchers. Such community building is not only necessary to ensure broad access to high quality research tools, I argue, but also because the progress of artificial intelligence will require deeper understanding of the social forces that guide human behavior.

Generative AI | Computational Social Science | Agent-Based Models | Survey Research | Ethics | Algorithmic Bias | Social Studies of Science

Correspondence: ***christopher.bail@duke.edu***

Introduction

Generative Artificial Intelligence— technology capable of producing realistic text, images, music, and other creative forms— continues to captivate large audiences. *ChatGPT*, the conversational chatbot generated by *OpenAI*, recently became one of the fastest-growing consumer applications in U.S. history. There is widespread speculation that such Generative AI will have considerable impact on a range of different industries and scientific disciplines—from creative and legal writing to computer science and engineering. 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 evaluate whether these new tools will enhance conventional research methods—and whether they may enable new forms of scientific inquiry altogether. At the same time, I assess the many limitations of Generative AI for social science research—and discuss how scholars interested in exploring such technologies can mitigate risks associated with these largely untested technologies.

In the first section of this article, I provide a brief history of Generative AI and how it came to the attention of a small group of social scientists. In the second section, I ask ask whether Generative AI can effectively simulate human behavior for the purposes of social science research. I examine whether these tools may 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 these models interact with each other to create emergent group behaviors. In the final part of this section I examine how Generative AI might transform the broader practice of social science research. I ask whether these tools can serve as a "virtual research assistant" capable of performing tasks typically assigned to humans—such as coding large groups of documents, or performing literature reviews. Finally, I assess whether Generative AI may help researchers identify novel research questions or untested hypotheses, communicate our research findings more effectively, and expand access to software programming skills.

In the third section of this article I turn to the various risks and potential dangers associated with Generative AI. Much of the public discourse surrounding this new technology focuses on the possibility of a "singularity" where models achieve general artificial intelligence that could threaten human interests and well-being. Many experts believe such concerns eschew well-documented social harms that are already occurring in the short term. These include the tendency of Generative AI to exhibit strong bias against certain populations, exacerbate social inequality, and spread misinformation—among others. I discuss how these issues may negatively impact the quality, efficiency, interpretability, and reproducibility of social science research and generate pressing 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.

My principal goal is to provoke in-depth conversation between social scientists, computer scientists, ethicists, and AI practioners about how to design research that can take advantage of the considerable promise of Generative AI without accelerating its potential for social harm. The most natural place for this conversation to occur, in my view, is the field of computational social science—an interdisciplinary field that leverages tools from data science and machine learning to

develop theories of human behavior using the increasingly voluminous amount of data generated online each day (1–3). Computational social science has already experienced its own share of ethical controversies—long before the advent of Generative AI (3, 4) Excitement about the ability to embed research studies within online ecosystems, for example, may have compromised the safety of dissidents in authoritarian regimes (5), violated privacy (6), and led to the study of large populations without adequate consent (7). These concerning developments within the field most likely to adopt Generative AI suggest there may be a range of "unknown unknowns" that will require careful reflection within an increasingly competitive research environment.

Several caveats are in order. First, my analysis of how Generative AI might transform research is strictly limited to social science and thus does not engage with the many different ways this technology might shape other fields. Second, 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. Therefore, I urge the reader to take caution in evaluating the potential of the research techniques described below, which may yet be judged scientifically unsound, unethical, or both. Third, I do not provide a technical discussion of how Generative AI models work, since these are broadly available elsewhere (8). Instead of a "user's guide" for Generative AI in social science research, I hope to inspire ongoing dialogue among researchers about whether or 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. 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— towards the generation of free-form text, images, and video, via algorithms that are trained on large datasets, usually 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 artificial intelligence researchers to train such models on increasingly large amounts of text, technologies 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 "auto-complete" 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.

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. Finally, a new class of 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 (such as a picture of the psychologist Daniel Kahneman riding an elephant across the campus of Princeton University).

Though the quality of texts and images produced by Generative AI continues to impress many, the fidelity of these models also have well-known limitations. Due to space limitations, I shall mention four in this introduction that are particularly germane to social science research. First, content generated by Generative AI includes the full panoply of human flaws that exist within the training data used to create them. Early LLMs, for example, could be goaded into making racist or sexist comments with minimal effort. Though newer tools have more sophisticated safeguards, it is often trivial to circumvent them via subtle rephrasing of text prompts. Second, LLMs lack the capacity to perform basic problem solving—let alone causal inference—despite the fact that they can produce high scores on standardized tests such as the GRE or Bar Exam (9) and arguably demonstrate "theory of mind" (10). Third, the capacity of Generative AI to perform well on standardized tests may come at the cost of its performance in other areas— a phenomenon often described as "overfitting." That is, the more Generative AI models are trained to succeed at one type of task, the less well they are able to perform others— though these problems may become less significant with scale. Finally LLMs frequently "hallucinate" or create realistic sounding statements— often delivered with confidence— that are patently untrue or misleading.

Opportunities for Social Science with Generative AI

Despite—or perhaps because of—their significant flaws, Generative AI tools appear capable of impersonating humans in some settings. The computer scientist Alan Turing was among the first to propose evaluating artificial intelligence 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. They found GPT-2 could successfully impersonate humans, and that it 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 GPT-

3, the penultimate large language model created by *OpenAI* (11) In a series of experiments, these scholars show that humans are largely unable to determine whether such texts are authored by humans or GPT-3. Finally, Zhou et al. show that GPT-3 can easily produce misinformation about COVID-19 that can escape detection by the type of detection techniques used by social media platforms (12). More recent studies indicate AI-generated content can influence human attitudes, even if it is false or misleading (13, 14).

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 a 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 that describes a job applicant 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 its capacity to impersonate a human. Yet there is also evidence that Generative AI can perform reasonably well at more complex human behaviors. For example, Argyle et al. demonstrate GPT-3 can accurately impersonate respondents to large, nationally-representative public opinion surveys from a range of different demographic backgrounds (15). Prompting such tools with details about the characteristics of a respondent, for example, makes them produce fairly accurate predictions about how a real respondent with such characteristics might respond to a public opinion survey. Some even 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 (16). Understanding how LLMs respond to survey may be doubly important given recent reports that these tools are being trained to impersonate survey respondents by malicious actors seeking to game the industry.

Though silicon samples will not soon displace survey research with human respondents, they may still be very useful for pre-testing surveys, imputing missing data— or perhaps even survey experiments— before they are dispatched (at considerable cost) to large groups of human respondents (15, 17). Horton, for example, argues synthetic research respondents created using GPT-3 can be used to reproduce several classic studies in behavioral economics (18). Similarly, Aher et al. show that GPT-3 can also reproduce classic social psychology experiments—including the infamous Milgram experiment—though interestingly it is not capable of reproducing the "wisdom of crowds" phenomenon(19). Still other studies indicate GPT-3 can replicate classic experiments in cognitive science and the study of morality (20, 21). Others have proposed Generative AI is also a useful tool for creating survey questions, or designing multi-item scales to measure abstract social concepts (22).

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 and increasingly powerful deep neural networks).

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, in other words, demonstrate 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* (23, 24).

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 (25). 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. This may be because 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 that automated agents that are trained to perform poorly at collaborative tasks such as games can actually *improve* the behavior of human team members (26). 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 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. For example, many social science theories indicate group-level processes shape human behavior. But recruiting large groups of people to interact is often logistically impossible, prohibitively expensive—or both. Though Generative AI will probably never perfectly replicate the spontaneous behavior of human groups, researchers may nevertheless be able to dispatch groups of bots in online spaces to approximate such behavior. To give one of many possible examples, researchers could create groups of bots that hold different types of opinions about a given issue, and determine when individual research participants are influenced by majority and minority views after observing conversations between automated agents in an interactive setting. Setting aside the many ethical issues which such research— an issue which I discuss in detail below— studies with simulated agents would reduce costs and forestall the challenge of recruiting large groups of people to participate in research at the same time (27). At the same time, there is not yet a "gold standard" study that shows that groups of automated agents can accurately simulate humans. This—combined with the potential for yet unidentified risks that may occur when LLMs interact with each other—suggest social scientists should proceed with great caution in this area.

Can Generative AI Improve Simulation-Based Research?. Because previous studies only examine AI agents in relatively simplistic multiplayer games or problem-solving tasks governed by clear rules, it is largely unknown whether Generative AI can successfully mimic emergent behavior among large groups of people. This is a key goal of the "agent-based modelling" paradigm, in which researchers create synthetic societies to study social processes. This tradition, which dates back to the 1970s, typically involves the creation of a facsimile of a social setting (such as a social network, neighborhood, or marketplace). The researcher then creates individual agents who interact with each other in such settings according to a set of rules determined by the model (28). 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 micro-level patterns (such as in-group bias) that can create macro-level patterns (e.g. residential segregation). But these models also have many well documented weaknesses. First, the agents within such models are usually overly simplistic; making binary decisions from a range of different input parameters that belie the complexity of most human interactions— especially those where the consequences of such choices may not be immediately clear. Second, many agent-based models create *de novo* social settings (such as early human civilizations or social networks where connections between agents are randomly wired) that are seldom observed in real-world settings. Thus, the external validity of research that employs agent-based models is often quite low, and the entire research tradition is sometimes dismissed as meaningless artificial behavior within equally artificial settings that belie the blooming, buzzing, confusion of everyday life—to paraphrase the psychologist William James.

Generative AI tools such as LLMs provide new occasion to revisit the social simulation paradigm. Park et al. (29) created a social simulation where several dozen agents—independently powered by multiple instances of *ChatGPT*—interacted with each other in a fictitious small-town setting. 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 create a renaissance in social simulation research.

Park et al.'s study is not designed to test a social science theory. But it may be easily repurposed to do so. For example, a scholar interested in examining how social media echo chambers might hasten the spread of misinformation could seed a false statement within a network of agents powered by LLMs that are prompted using the characteristics of real social media users—or a corpus of their past messages. By experimentally varying the size and rigidity of the echo chamber—

that is, the heterogeneity of political beliefs that agents are exposed to—researchers could examine how far dangerous misinformation spreads before it is challenged or corrected. What is more, researchers might even be able to simulate what might happen if the people spreading misinformation are confronted with such counter-arguments. Needless to say, such simulations might be far from how real-world events might unfold in such dynamic settings. But they could represent a major improvement upon previous models where political beliefs are condensed into simplistic binary rules that compel agents to act without language, memory, or knowledge of social context (via prompt engineering).

A further advantage of research with Generative AI and agent-based models is that it could be used to study topics that would be dangerous to study in real life (such as violent extremism on social media), or to study populations that are very difficult to survey (e.g. violent extremists) (16). Such simulations might inform what little observational research we have on such topics or populations—and could be calibrated using these data as well. Emergent group behaviors identified through simulation research could inform observational data collection in turn—or, potentially—social interventions designed to prevent such behavior. Far more research is needed, however, to determine whether social simulations have the fidelity 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 Serve as a Virtual Research Assistant?. Regardless of whether Generative AI can effectively simulate human behavior, it may also be useful to social scientists for more menial research tasks. Perhaps the most promising task that could be outsourced to Generative AI is content analysis of text-based data. Even before the emergence of transformer models such as *ChatGPT*, the field of natural language processing produced a series of tools that were widely adopted by social scientists who studied large amounts of text-based data (30). These included topic models and word embeddings that could identify patterns in large corpora, even though researchers often struggled to interpret the output of such "unsupervised" methods. A parallel group of "supervised" models were also widely applied to identify patterns in large corpora by training algorithms using text annotated by human coders.

A series of new papers suggests GPT-3 and more recent models can produce surprisingly accurate analyses of text-based data, with minimal training. For example, Wu et al. demonstrate GPT-3 can produce accurate classifications of the ideology of U.S. elected officials by analyzing their public statements (31). This team 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 can accurately code the credibility of media sources (32). Gilardi et al. argue GPT-3 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 (33). In addition to passing GPT-3 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 that GPT-3 performs better than 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 GPT-3's coding performance to highly trained coders who were instructed to analyze statements about British Elections (34). They find the model produced the same classification 95% of the time.

Ziems et al. offer a more systematic analysis of the capabilities of LLMs for coding texts (35). Using an array of hand-coded datasets from sociology, political science, and psychology—as well as non-social science fields such as history, literature and linguistics—they compare the capabilities of LLMs to reproduce hand-coded labels. Overall, they find LLMs perform reasonably well—particularly in coding data created by political scientists and sociologists. Unsurprisingly, they find the latest models (e.g. *ChatGPT* and Google's FLAN-UL2) perform better than earlier models, and in some cases even surpass supervised models trained on a particular dataset. LLMs also appear to assign more accurate codes for some topics (e.g. misinformation) than others—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. (2023) warn that the most effective usage of LLMs will still require some degree of human supervision, and familiarity with task-specific prompt-engineering. Usefully, these authors also identify best practices for both of these tasks and present a reproducible data analysis pipeline for ongoing evaluation of future models and other datasets. Törnberg provides further practical instructions about best practices for automated coding with LLMs as well (36).

Together, these early studies indicate Generative AI has considerable potential to serve as a virtual research assistant. Though its accuracy may remain significantly lower than human coders for many tasks, these accuracy trade-offs must be weighed against other factors such as the speed with which LLMs can code, or their potential to code texts in many different languages. There is also preliminary evidence that Generative AI may assist researchers with other rudimentary tasks typically assigned to research assistants such as data coding, or data entry (37). As I discuss below, there is also some indication these tools might be useful for performing preliminary literature reviews or meta-analyses, or systematically extracting findings (or effect sizes and research designs) from large groups of studies– tasks which are also typically assigned to human research assistants.

Can Generative AI Help Social Scientists Acquire Programming Skills?. Once data poor, social scientists now face an overwhelming amount of new data from social me-

dia sites, administrative records, and digitized historical archives—among other new digital data sources (2). The number of new technologies available to analyze these data has also expanded dramatically—not only Generative AI, but a range of new tools for analyzing observational data and entirely new forms of technology such as apps that can allow social scientists to collect data innovatively (38). Unfortunately, social science pedagogy has struggled to keep up with increased demand for the skills necessary to create or analyze these new wellsprings of data. Most PhD-granting social science departments do not require students to learn basic programming skills—apart from those necessary for data cleaning or basic statistical analyses.

One of the most important contributions of Generative AI to social science may therefore be expanding access to programming skills. Code-writing assistance using Generative AI has already become widespread using GitHub *CoPilot*, which offers software developers an "auto-complete" for writing code that is powered by OpenAI's Codex. One can also ask *ChatGPT* to write code using natural language prompts. For example, a researcher could ask this tool to write code in Python that creates a simulation to study how social networks shape intraveneous drug use. Though the resulting code may be rather generic, it could nevertheless allow social scientists who are unable to write such code from scratch to tweak the model to serve their purposes. Perhaps even more importantly, Generative AI tools can help novices understand how code works. A *ChatGPT* user, for example, can ask the model to explain what is happening in a single line of code, or how a function operates. Though such interpretation is not always accurate, it may allow social scientists with little programming training to develop a better sense of what is possible with software engineering, or identify what new technologies they need to learn to accomplish their goals.

By facilitating broader access to technical skills, Generative AI could theoretically reshape the value of different skills within social science. Grossman et al. argue Generative AI could decrease the value of those with significant quantitative skills and increase the relevance of those with deep knowledge about theory, concept, and measurement(16). Such individuals may be critical for the progress of prompt engineering, for example, which often requires deep understanding of the context and style of human language. At the other extreme, the advent of Generative AI could make social scientists with technical prowess even more influential—particularly if such tools need to be carefully fine-tuned or re-designed to enable more rigorous research on human behavior.

Can Generative AI Help Social Scientists Write?. Numerous Generative AI tools are now available to help people write. These tools can respond to prompts (e.g., "read this page and write a summary of its contents"), or they can be used in an iterative fashion (e.g., "make the style of the prose in the following sentence more scholarly"). Many faculty members view these tools with skepticism—and warn their students not to use them. Yet Generative AI tools are already transforming the writing process in many different fields (37). They may be most useful for scholars whose first language is not English, or for English-speaking scholars who wish to translate their work for new audiences. LLMs may also be useful to those with more well-worn pens. Korinek proposes social scientists should consider asking LLMs to evaluate the weaknesses in their arguments, for example, or identify counter-arguments (37). Though researchers should not accept recommendations from LLMs blindly, they may encourage us to reflect upon our own blind spots—or to assume the perspective of others who read our prose.

Can Generative AI Help Social Scientists Perform Literature Reviews?. In recent decades, social scientists have enjoyed access to large corpora that compile very large amounts of peer-reviewed research such as the Web of Science. Social studies of science that employ network analysis and natural language processing to study the growth of fields have expanded accordingly (39–41). Proponents of Generative AI have naturally become interested in whether these new tools can expand our capacity to map science—or even guide the trajectory of scholarly inquiry. Early attempts to use Generative AI for this purpose largely failed. For example, Meta's LLM, *Galactica*, was designed to help scientists navigate scholarly literatures more efficiently. It produced such inaccurate responses, however, that it was taken offline after only three days. The debut of Google's BARD chat assistant was similarly marred when it provided an inaccurate response to a question about astronomy during its world-wide launch event.

Though LLMs are not reliable enough to summarize scholarly literatures, they may be a useful muse to social scientists during preliminary stages of research. *Elicit.org* is a Large Language Model trained on scholarly databases that can generate a list of articles that respond to a question such as "Does Immigration Increase Crime?" The tool not only produces a set of articles that address this question, but organizes them according to criteria not typically available within many scholarly databases— such as whether the studies are original empirical analyses or meta-analyses. The tool can further separate studies according to sample size or whether they include randomized controlled trials. *Elicit* can also assess what outcomes are measured, and exploit the network-structure of citations to recommend similar studies in an interactive manner. On the other hand, tools such as Elicit have many well-known limitations. Its training data excludes books, for example, and it misses important references at least as often as an uninitiated human research assistant.

Insofar as Generative AI is capable of providing a broad perspective on many different scientific fields, it may also be useful for identifying novel research questions. I asked *Elicit* to generate a new set of questions about social media and political polarization— a topic which I have studied extensively. Several of the questions it generated were unimpressive or nonsensical. But of the eight questions it proposed, I consider two of them to be fairly good ideas that test the boundaries of the field: 1) "How is the impact of social media on politics different across countries?" and 2) "Why do people switch social media platforms, and how might this impact

polarization"? Though Generative AI will not soon serve as a capable dissertation advisor, it may nevertheless be "good to think with," as the French sociologist Pierre Bourdieu was fond of saying.

Limitations and Possible Dangers

To this point, I have presented a somewhat optimistic view of the potential for Generative AI to improve social science. But these tools have major limitations that could negatively impact the accuracy, interpretation, and practice of social science research. In the following sections, I discuss these limitations in detail.

Generative AI Exhibits Human Biases. Artificial intelligence is routinely criticized for amplifying various types of inter-group bias (42–46). Such bias exists because most AI tools are trained using data created by humans, and therefore often exhibit a broad range of prejudice and cognitive errors. To cite one of many possible examples, technologies for predicting crime may evince bias against African-Americans. This may result from the existence of racial prejudice in training data (e.g. sentencing decisions by judges) or because crime is more often reported within African-American communities because of increased police surveillance in such areas (47). There are many, many more examples that have been documented at length elsewhere (44, 45, 48). Generative AI has heightened concerns about bias, because these tools are trained on large amounts of data created by humans on the internet– where inter-group 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 (49). By comparing how the models responded to questions about abortion, gun control, and a range of other topics, the researchers were able to assess how closely each model resembles 60 different demographic subgroups in the United States. They find most LLM's responses are considerably 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, and those who live alone. Other researchers have shown that LLMs tend to exhibit bias against women and racial minorities (44, 50). In other words, LLMs appear to reflect the interests of the most advantaged part of U.S. society, though not those who have more conservative viewpoints or live in rural areas. Another study indicates LLMs have distinctive personality characteristics— specifically, they are more likely to be extroverted and agreeable than neurotic (51). 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— i.e. when a researcher asks the model to perform the role of a specific group (e.g. a wealthy Republican from Texas) (49). This mirrors earlier research which suggests removing bias from AI tools may be easier than removing it from human populations (52). 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 *ChatGPT*—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 address such bias on their own.

A key question for social scientists is whether the tendency of Generative AI to exhibit bias is a "bug" or a "feature" for research purposes. Social scientists often design experiments that examine the impact of bias on attitudes or behaviors. If such bias 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 (50). 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, for example, might be disappointed by the quality of such impersonations because of insufficient training data.

One of the most important stages in training a Generative AI model is when its developers provide it with feedback through a process known as "fine-tuning." Developers often attempt to train their models to avoid making racist statements or discussing dangerous topics such as how to create weapons, for example. This process typically occurs both behind closed doors. Employees of the companies that create Generative AI engage in "red team" attacks designed to goad the model into producing prejudiced, dangerous, or illegal content. Developers then develop workflows to prevent the models from discussing such content. Though such guardrails arguably improve the safety of Generative AI, they may impede the ability of social scientists to leverage their bias for research purposes (16). Researchers who want to use LLMs to impersonate biased groups, for example, may discover these tools are unable or 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 (16, 53). A recent study also suggests fine-tuning models for safety purposes may degrade their performance on many other tasks—from mathematics to writing code (54). The opaqueness of the fine-tuning process may create other problems that are difficult to detect. Many LLMs are fine-tuned to impersonate humans more accurately– but this training pro-

cess may also make them more likely to share inaccurate information (55).

Will Generative AI Spread Misinformation about Social Science?. The potential for malicious actors to use Generative AI to spread misinformation—or for these tools to exhibit bias in a variety of settings used by well-intentioned actors in the short term—is deeply concerning. But the capacity of Generative AI to produce inaccurate information or "hallucinate" 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 *Stack Overflow* users created bots that automatically passed people's questions about software to an LLM. Though some of the answers produced by the LLM were high quality, others were completely incorrect. The website quickly announced a new policy that prevented users from employing LLMs to answer questions to prevent a situation where users would struggle to distinguish the good information from the bad—particularly given the tendency of LLMs to deliver inaccurate information in a confident manner, and its capacity to generate thousands of answers in short order.

The "Stack Overflow Problem" could be particularly dangerous for researchers who rely upon LLMs to perform literature reviews, generate new research questions, or otherwise summarize large corpora they are unable to read themselves. Journals and funding agencies may find themselves overwhelmed by low-quality "junk-science" created by LLMs. Fortunately, computer scientists have begun to create digital "watermarks" that may allow LLMs to identify themselves, or other models. 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 (56). 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. The potential for such small-scale LLM development was made much easier by the recent leak of Facebook's LLAMA model, and the rapid progress of other small open-source LLMs.

Is Research with Generative AI Ethical?. Perhaps the most pressing question for social scientists is whether research with Generative AI is ethical (57, 58). 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 assume added importance in situations where human research participants might have conversations with a LLM in an unsupervised manner.

Must researchers 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. Even worse, 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 towards artificial intelligence more broadly.

One solution to this problem may be to design studies in which research participants are informed they may interact with artificial intelligence 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 should not be considered fail-safe.

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 (59). 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 rephrasing used 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 eschews the issue of informed consent, 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. The AI-as mediator approach may also facilitate peer-to-peer mental health interventions and empower women within democratic discourse as well, according to recent studies (60, 61).

A final strategy might be to use Generative AI to try to di-

agnose possible ethical issues in research studies themselves. 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 amongst 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.

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 *ChatGPT* 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. Generative AI tools may also lead social science researchers to infringe upon the intellectual property of others unknowingly. A political scientist who uses an LLM to create a vignette for a survey experiment could accidentally employ language that is very similar to that of a best-selling government thriller.

A final ethical concern is the impact of Generative AI on climate change. A 2019 study indicates training a single large language model may generate as much carbon dioxide as the lifetime emission of five automobiles (62). 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.

Is Research with Generative AI Reproducible?. A key pillar of the open-science movement is that researchers should design studies that can be replicated by others. Given the probabilistic nature of Generative AI, this creates a fundamental challenge (63). What is more, most organizations that develop Generative AI are constantly fine-tuning them to make them more effective or to create new safeguards against bias or illicit behavior. But ongoing development can also cause "drift" within LLMs, such that behavior that was observed at one point could be quite different at a future point in time (54). Put differently, Generative AI may help researchers increase the external validity of their research designs, but this may come at the cost of internal validity (16). Because the process of model development and fine-tuning is so secretive, researchers who attempt to replicate each other's work may not be primed to look for different model behavior across time. Moreover, the lack of transparency across different Generative AI models suggests researchers could observe very different behavior between, for example, *ChatGPT* and Google's *Bard* chatbot. The open-source models I discuss in additional detail below may provide an opportunity to increase the reproducibility of Generative AI models—as well as recent advances in prompt engineering (64). But these strategies may still lack temporal validity. A researcher who attempts to use a model trained on data from 2020 to discuss politics in 2023 should not be surprised if it produces language that assumed Donald Trump is still president of the United States.

Will Fixing Bugs Caused by AI Make us Less Efficient?. Above I argued that Generative AI may assist social scientists in a variety of mundane tasks such as coding, programming, and writing. Yet the many limitations of these tools just discussed apply to these areas as well. Though researchers may be able to detect unwanted racist bias in a text rather easily, small mistakes in a lengthy piece of code authored by an LLM is much more difficult to detect. Indeed, expert coders who use the "autopilot" tools described above report identifying such tiny bugs in code can make the costs of Generative AI for software development outweigh its benefits (65). A useful analogy is the self-driving car. Such vehicles appeal to many because they could reduce the cognitive and physical stresses associated with driving. Yet in practice, many self-driving cars need to be closely monitored by drivers in case the AI fails . In other words, the need to constantly monitor self-driving cars may substantially reduce the benefits of automation. Social scientists may soon face a similar trade off: though we may initial enjoy outsourcing difficult or tedious parts of our jobs to Generative AI, we may soon discover that monitoring its performance may be more trouble than it is worth.

Conclusion

Few technologies have considered 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(3). 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 *a foritiori* become a fixture within the social scientist's toolkit much sooner than many researchers realize.

A collective effort is needed to ensure that social scientists continue to shape the future of Generative AI. Our capacity to perform high quality research with these tools will require us to learn to identify and control bias—and train models to distinguish high quality research from plausible hallucinations. We may also need to create new open-source re-

search environments designed to enable research that is ethical, reproducible, and broadly accessible. Social scientists must work together to create these collective goods soon, before the architectures used to create Generative AI become so deeply embedded within large corporations that they are only accessible to a handful of researchers. Developing such a framework would also create new responsibilities. Access to a powerful LLM designed for social science research, for example, would have to be carefully controlled to prevent it from being used to study how to create more effective social media influence campaigns.

Above all, social scientists should not 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.

ACKNOWLEDGEMENTS

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

Bibliography

1. Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A. Bail. Computational Social Science and Sociology. *Annual Reivew of Sociology*, 46, 2020.
2. David Lazer, Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062, August 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aaz8170. Publisher: American Association for the Advancement of Science Section: Policy Forum.
3. Matthew Salganik. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, Princeton, N.J., 2018.
4. Casey Fiesler and Nicholas Proferes. “Participant” Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1):2056305118763366, January 2018. ISSN 2056-3051. doi: 10.1177/2056305118763366. Publisher: SAGE Publications Ltd.
5. Sam Burnett and Nick Feamster. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests, July 2015. arXiv:1410.1211 [cs].
6. Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer, and Nicholas Christakis. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30:330–342, 2008. doi: 10.1016/j.socnet.2008.07.002.
7. Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, June 2014. ISSN 0027-8424, 1091-6490.
8. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs].
9. Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2022. arXiv:2206.04615 [cs, stat].
10. Michal Kosinski. Theory of Mind Might Have Spontaneously Emerged in Large Language Models, August 2023. arXiv:2302.02083 [cs].
11. Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120 (11):e2208839120, March 2023. doi: 10.1073/pnas.2208839120. Publisher: Proceedings of the National Academy of Sciences.
12. Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea Parker, and Munmun Choudhury. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. April 2023. doi: 10.1145/3544548.3581318.
13. Lewis D. Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly T. Mai, Maria Vau, Matthew

Caldwell, and Augustine Marvor-Parker. Susceptibility to Influence of Large Language Models, March 2023. arXiv:2303.06074 [cs].

14. Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):116:1–116:29, April 2023. doi: 10.1145/3579592.
15. Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, pages 1–15, February 2023. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2023.2. Publisher: Cambridge University Press.
16. Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. AI and the transformation of social science research. *Science*, 380(6650):1108–1109, June 2023. doi: 10.1126/science.adi1778. Publisher: American Association for the Advancement of Science.
17. Junsol Kim and Byungkyu Lee. AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys, May 2023. arXiv:2305.09620 [cs].
18. John J. Horton. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?, January 2023. arXiv:2301.07543 [econ, q-fin].
19. Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies, February 2023. arXiv:2208.10264 [cs].
20. Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, February 2023. doi: 10.1073/pnas.2218523120. Publisher: Proceedings of the National Academy of Sciences.
21. Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 0(0), May 2023. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2023.04.008. Publisher: Elsevier.
22. Friedrich M. Götz, Rakoen Maertens, Sahil Loomba, and Sander van der Linden. Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, pages No Pagination Specified–No Pagination Specified, 2023. ISSN 1939-1463. doi: 10.1037/met0000540. Place: US Publisher: American Psychological Association.
23. Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, December 2022. doi: 10.1126/science.ade9097. Publisher: American Association for the Advancement of Science.
24. Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, November 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1724-z.
25. Fabrizio Dell'Acqua, Bruce Kogut, and Patryk Perkowski. Super Mario Meets AI: Experimental Effects of Automation and Skills on Team Performance and Coordination. December 2020. doi: 10.2139/ssrn.3746564.
26. Margaret L. Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A. Christakis. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*, 117(12):6370–6375, March 2020. doi: 10.1073/pnas.1910402117. Publisher: Proceedings of the National Academy of Sciences.
27. Joshua Becker, Ethan Porter, and Damon Centola. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22):10717–10722, May 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1817195116.
28. Michael W. Macy and Robert Willer. From Factors to Actors: Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28(1):143–166, 2002. doi: 10.1146/annurev.soc.28.110601.141117. _eprint: https://doi.org/10.1146/annurev.soc.28.110601.141117.
29. Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior, April 2023. arXiv:2304.03442 [cs].
30. J. Grimmer, M.E. Roberts, and B.M. Stewart. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press, 2022. ISBN 978-0-691-20799-5.
31. Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. Large Language Models Can Be Used to Scale the Ideologies of Politicians in a Zero-Shot Learning Setting, April 2023. arXiv:2303.12057 [cs].
32. Kai-Cheng Yang and Filippo Menczer. Large language models can rate news outlet credibility, April 2023. arXiv:2304.00228 [cs].
33. Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30): e2305016120, July 2023. doi: 10.1073/pnas.2305016120. Publisher: Proceedings of the National Academy of Sciences.
34. Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, and Marta Miori. Does GPT-3 know what the Most Important Issue is? Using Large Language Models to Code Open-Text Social Survey Responses At Scale, 2023.
35. Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can Large Language Models Transform Computational Social Science? *ArXiv*, 2023.
36. Petter Törnberg. How to use LLMs for Text Analysis, July 2023. arXiv:2307.13106 [cs].
37. Anton Korinek. Language Models and Cognitive Automation for Economic Research, February 2023.
38. Christopher Bail. Taming Big Data: Using App Technology to Study Organizational Behavior on Social Media. *Sociological Methods & Research*, 2015.
39. Achim Edelmann, James Moody, and Ryan Light. Disparate foundations of scientists' policy positions on contentious biomedical research. *Proceedings of the National Academy of Sciences*, 114(24):6262–6267, June 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1613580114.
40. Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical Combinations and Scientific Impact. *Science*, 342(6157):468–472, October 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1240474.
41. Lingfei Wu, Dashun Wang, and James A. Evans. Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382, 2019. Publisher: Nature Publishing Group UK London.
42. Ruha Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity, Cambridge, UK Medford, MA, 1st edition edition, June 2019. ISBN 978-1-5095-2640-6.
43. Seth Lazar and Alondra Nelson. AI safety on whose terms? *Science*, 381(6654):138–138, July 2023. doi: 10.1126/science.adi8982. Publisher: American Association for the Advancement of Science.
44. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret 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*, FAccT '21, pages 610–623, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922.
45. Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018. ISSN: 2640-3498.
46. Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24, October 2021. ISSN 2573-0142. doi: 10.1145/3479594. arXiv:2105.08667 [cs].
47. Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1):237–293, February 2018. ISSN 0033-5533. doi: 10.1093/qje/qjx032.
48. Roxana Daneshjou, Mary P. Smith, Mary D. Sun, Veronica Rotemberg, and James Zou. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA Dermatology*, 157(11):1362–1369, November 2021. ISSN 2168-6068. doi: 10.1001/jamadermatol.2021.3129.
49. Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect?, March 2023. arXiv:2303.17548 [cs].
50. Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. On Measuring Gender Bias in Translation of Gender-neutral Pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3824.
51. Max Pellert, Clemens Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. AI Psychometrics: Using psychometric inventories to obtain psychological profiles of large language models, December 2022.
52. Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, October 2019. doi: 10.1126/science.aax2342. Publisher: American Association for the Advancement of Science.
53. Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, March 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00458-8. Number: 3 Publisher: Nature Publishing Group.
54. Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time?, August 2023. arXiv:2307.09009 [cs].
55. Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. Automatic Detection of Generated Text is Easiest when Humans are Fooled, May 2020. arXiv:1911.00650 [cs].
56. John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A Watermark for Large Language Models, January 2023. arXiv:2301.10226 [cs].
57. Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from Language Models, December 2021. arXiv:2112.04359 [cs].
58. Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533088.
59. Lisa P. Argyle, Ethan Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, and David Wingate. AI Chat Assistants can Improve Conversations about Divisive Topics, March 2023. arXiv:2302.07268 [cs].
60. Rafik Hadfi, Shun Okuhara, Jawad Haqbeen, Sofia Sahab, Susumu Ohnuma, and Takayuki Ito. Conversational agents enhance women's contribution in online debates. *Scientific Reports*, 13(1):14534, September 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-41703-3. Number: 1 Publisher: Nature Publishing Group.
61. Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Hu-

man–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, January 2023. ISSN 2522-5839. doi: 10.1038/s42256-022-00593-2. Number: 1 Publisher: Nature Publishing Group.

62. Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP, June 2019. arXiv:1906.02243 [cs].
63. Arthur Spirling. Why open-source generative AI models are an ethical way forward for science. Nature, 2023.
64. Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting GPT-3 To Be Reliable, February 2023. arXiv:2210.09150 [cs].
65. Jenny T. Liang, Chenyang Yang, and Brad A. Myers. Understanding the Usability of AI Programming Assistants, March 2023. arXiv:2303.17125 [cs].