Can Generative AI Improve Social Science Research?[1]

Chris Bail
Duke University
www.chrisbail.net

Abstract. 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 transform social science research. In the first section of this article, I assess the potential of Generative AI to improve online experiments, agent-based models, and automated content analyses. I also discuss whether these tools may help social scientists perform literature reviews, identify novel research questions, and develop hypotheses to explain them. Next, I evaluate whether Generative AI can help social scientists with more mundane tasks such as acquiring advanced programming skills or writing more effective prose. In the second section of this article I discuss the limitations of Generative AI as well as how these tools might be employed by researchers in an ethical manner. I discuss how bias in the processes and data used to train these tools can negatively impact social science research as well as a range of other challenges related to accuracy, reproducibility, interpretability, and efficiency. I conclude by highlighting the need for increased collaboration between social scientists and artificial intelligence researchers— not only to ensure that such tools are used in a safe and ethical manner, but also because the progress of artificial intelligence may require deeper understanding of theories of human behavior.

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 the fastest-growing consumer application in history. According to one report, this tool amassed more than 13 million unique users each day in January 2023.[2] There is widespread speculation that such generative AI will have considerable impact on a range of different industries— from creative and legal writing to advertising and customer service. 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 social scientists can employ Generative AI to enhance conventional research methods, but also invent entirely new forms of inquiry as well. Along the way, I evaluate whether the use of such tools is ethical in research settings, and how scholars interested in exploring such technologies might mitigate the various risks associated with these largely untested technologies.

In the first section of this article, I 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 creating synthetic human-like content such as images used in survey experiments or text for vignette studies. Next, I review recent studies that employ Generative AI models to simulate human populations taking public opinion surveys, impersonate team-members in online experiments, and provide more realistic agent-based models. I then ask whether Generative AI can become 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 will increase access to programming skills among social scientists, and perhaps even assist them in generating novel or untested hypotheses as well.

In the second section of this article I turn to the various risks and potential dangers associated with Generative AI. On March 29th, 2023, several thousand leading experts in artificial intelligence, computer science, public policy, and many other fields signed an open letter calling for a pause in the development of new Generative AI models.[3] Signatories of this letter had diverse motivations ranging from concerns about how malicious actors might use Generative AI to launch influence campaigns on social media to broader concerns about how such tools might amplify social inequalities, or create new forms of social stratification by eliminating jobs typically performed by humans.[4] Others worry that Generative AI can not only produce inaccurate or misleading responses to human questions, but deliver them with a degree of confidence that might deflect criticism or scrutiny.[5] To these concerns I add that we do not yet know whether or how Generative AI should be used in research settings due to a lack of transparency in how they are trained, tested, and deployed. I hope this article will provoke a conversation among social scientists, computer scientists, ethicists, and AI engineers about how research can be leveraged to identify the promises and pitfalls of these techniques.

The most natural place for this conversation to occur might be the nascent field of computational social science— which 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 (Edelmann et al. 2020; Lazer et al. 2020; Salganik 2018). Computational social science has already experienced its own share of ethical controversies— long before the advent of Generative AI (e.g. Fiesler and Proferes 2018; Salganik 2018). Excitement about the ability to embed experiments within online ecosystems has inspired researchers to press ahead with research designs that have been criticized for threatening the safety of internet users in authoritarian regimes (Burnett and Feamster 2015), violating user privacy (Lewis et al. 2008), and enrolling people in online experiments without their consent (Kramer, Guillory, and Hancock 2014). 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 and patience despite the ever increasing pace of publication in this space.

Several caveats are in order before I proceed to discuss the issues described above. 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. Instead of a "user's guide" for generative AI in social science research, I aim to provoke a broader conversation among researchers about whether or how these new technologies might be used to study human behavior in diverse settings.

What is Generative AI?

The term "generative AI" refers to a broad range 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, video, and other heretofore human outputs by an algorithm that is trained on vast amounts of such data.[6] Large Language Models (LLM) 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 increasingly 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. LLM's 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.

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-occurence 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— new 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-occurence of words and the arrangement of pixels— that allow a user to request highly specialized visual content (such as a picture of Daniel Kahneman riding an elephant across the Princeton University campus).

Though the quality of texts and images produced by Generative AI impresses many, these models also have a number of well-known limitations. Due to space limitations, I shall only mention a few here which are particularly germane to social scientists who seek to gauge the potential of generative AI for research purposes. First, content generated by Generative AI includes the full panoply of human flaws that exist within the training data used to create them. Early large language models, for example, could be goaded into making racist or sexist comments with minimal effort. Though newer tools have more safeguards, it is often trivial to circumvent them via subtle rephrasing of text prompts. Second— though some of the most recent models such as GPT-4 have demonstrated the capacity to pass an array of standardized tests such as the GRE or the Bar Exam with flying colors— they still lack the capacity for many types of basic problem solving— let alone complex reasoning about causes and effects (Srivastava et al. 2022). 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." The more generative AI models are trained or encouraged to perform one type of task, the less well they are able to perform others.[7] Finally LLM's often "hallucinate" or create realistic sounding statements— often delivered with confidence— that are patently untrue or misleading. As I discuss in further detail below, there is not yet a widely accepted process for identifying such inaccuracies, particularly across different forms of Generative AI.

In the sections that follow, I ask how the strengths and weaknesses of generative AI might enable and constrain different types of social science research— and create new research opportunities and ethical and procedural challenges altogether.

PART 1. OPPORTUNITIES

Can Generative AI Effectively Simulate Human Behavior?

Despite—or perhaps because of—its 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.[8] Using GPT-2, a precursor to ChatGPT that produces much lower quality texts, Kreps et. al. (2022) studied whether research participants could differentiate short statements about U.S. foreign policy generated by this LLM and humans.[9] They find GPT-2 can 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 similar study, Jakesch et al. (2023) 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.

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 (2023) 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, though they did not measure whether humans were also unable to identify the synthetic content as such.[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 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 hypothetical vignettes and/or images— in an iterative fashion with 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. Shorter texts, for example, 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. (2023) demonstrate that GPT-3 can accurately impersonate survey respondents from a range of different demographic backgrounds. That is, prompting these tools with details about the characteristics of a respondent make them produce fairly accurate predictions about how a real respondent with such characteristics might respond to a public opinion survey. Though such "silicon samples" will not soon displace survey research with human respondents, it may be useful for pre-testing surveys— or perhaps even survey experiments— before they are dispatched at considerable cost to large groups of human respondents.[11] Horton (2023), for example, argues that synthetic research respondents created using GPT-3 can be used to reproduce several classic studies in behavioral economics. Aher et al. (2023) show that GPT-3 can also reproduce classic social psychology experiments— including the infamous Milgram experiment.[12] Still other studies indicate GPT-3 can replicate classic experiments in cognitive science and the study of morality (Binz and Schulz 2023; Dillion et al. 2023).

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 chatbot 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 language used in real settings, such as social media sites, 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 vast amounts of human interactions from the internet.

An interesting test-case for the capacity of Generative AI to generate believable human behavior in complex social settings is multiplayer 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 for establishing whether automated agents that respond dynamically to human research participants can act in ways that are indistinguishable from real humans. Prior to the advent of Generative AI, believable characters in video games were created via simple rules or via reinforcement learning (where AI game players can learn from their past experiences). Key to both strategies was creating a system where AI agents could recall prior events— or, in other words, have 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 games such as Backgammon, Chess, and AlphaGo for many years. More recently, however, researchers have shown that Large Language Models can also learn to use natural language in games that require complex reasoning and high-level strategy to defeat human players (e.g. Bakhtin et al. 2022; Vinyals et al. 2019).

Another fascinating line of research relevant to social science examines how the introduction of AI agents in multiplayer games shapes the behavior of the humans they play with. Dell'Aqua, Kogut, and Perkowski (2023) study a collaborative cooking game where AI's performance is known to exceed that of human players. When an AI agent is introduced in a team setting, the researchers find that human agents perform worse when the agent is on their team. More specifically, the introduction of the AI agent makes coordination more difficult for human players— and also leads to less trust among members of the team. Conversely, Traeger et al. (2020) 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. It may be that AI which completes tasks with greater skill than humans creates frustration and competition or in-fighting, whereas AI that demonstrates less competence encourages empathy and collaboration to

overcome poor 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, researchers might be able to study how the behavior of large groups of people influences the behavior of individual study participants. 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 further below— studies with simulated agents would reduce costs and minimize the considerable logistical challenges of getting large groups of people to participate in research at the same time (e.g. Becker, Porter, and Centola 2019). At the same time, there is not yet a "gold standard" study that shows that groups of automated agents can accurately simulate humans. This issue, combined with the many ethical issues I will discuss further below, suggest social scientists should proceed with 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 much work on social simulation, or "agent-based models," wherein researchers create synthetic societies to study social processes. This tradition, which dates back to the 1970s, typically involves the creation 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 created by the researcher (Macy and Willer 2002). 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 changed 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 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) are seldom observed in real world settings. Thus, the external validity of research that employs agent-based models is often relatively low, and the entire research tradition is sometimes dismissed as artificial behavior within artificial settings that has little import for the complex human-behavior observed through empirical research.

Generative AI tools such as Large Language Models provide new occasion to revisit the social simulation paradigm. In a pioneering study, Park et. al. (2023) created a social simulation where agents— independently powered by multiple instances of ChatGPT— interacted within a small town. 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. Though this study created a relatively simplistic social environment with a few dozen agents, it provides an important proof of concept that Generative AI could create a renaissance for social simulation research.

Park et. al.'s (2023) 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 Large Language Models 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. Nevertheless they could represent a major increase 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).

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 a range of menial research tasks. Perhaps the most promising type of task that might be outsourced to Generative AI is content analysis of text-based data. Even before the advent of transformer models such as ChatGPT, the field of natural language processing produced a

series of tools that were adopted by social scientists whose research analyzed large amounts of text-based data. These included topic models and word embeddings that could identify patterns in large corpora, even if the "unsupervised" versions of such methods could not interpret the meanings of such patterns for the researcher. A parallel group of "supervised" models could often guide such automated techniques to identify patterns in large text-based datasets by training algorithms to identify patterns of speech annotated in small training datasets produced by human coders.

A series of new papers suggests that GPT-3 and later models can produce surprisingly accurate analyses of text-based data as well. For example, Wu et. al. (2023) demonstrate GPT-3 can produce accurate classifications of the ideology of U.S. elected officials by analyzing their public statements. 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 (2023) argue GPT-3 can accurately code the credibility of media sources. In another study Gilardi et al. (2023) argue GPT-3 can accurately measure the topic of tweets, the stance or opinions of their authors, as well as frames used to organize the message in a narrative manner. In addition to passing GPT-3 the full text of tweets of interest, these researchers also fed the coding instructions assigned to human coders as a prompt to the model. They claim that GPT-3 performs better than workers trained with such materials on Amazon Mechanical Turk— though such coders are likely less accurate than those trained directly by researchers in small group settings. Similarly, Mellon et. al. (2023) compare GPT-3's coding performance to highly trained coders who were instructed to analyze statements about British Elections. They find the model produced the same classification 95% of the time.

Ziems et al. (2023) offer a more systematic analysis of the capabilities of LLMs for coding texts. Using an impressive 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 datasets 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 supervised models trained on a particular dataset. LLMs also appear to assign more accurate codes for some topics (e.g. misinformation) than others— which is likely 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 for these tasks and also offer a reproducible data analysis pipeline for ongoing evaluation of future

models and other datasets.

Together, these early studies suggest Generative AI has considerable potential for qualitative coding from unstructured text data, but additional studies are urgently needed to identify the direction of the bias or errors it generates in doing so, as I discuss in further detail below. The capabilities of Generative AI are particularly promising when they might be able to perform a very large number of tasks— in different languages and with different coding perspectives— very quickly. 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 (Korinek 2023). 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.

Can Generative AI Help Social Scientists Acquire Programming Skills?

Once data poor, social scientists now face an overwhelming amount of new data from social media sites, administrative records, and digitized historical archives among other new digital sources. 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 in new ways (e.g. Bail 2015). Yet 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 Ph.D. granting social science departments still 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 thus be that it could expand 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 code 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 R that creates a simulation to study how social networks shape political polarization. Though this code is generic, it may provide a basis for social scientists who do not have the skills to write scripts 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, and how a function operates. Though such interpretation will not always be accurate (more on this below), it may allow social scientists with little or no training in programming to develop a better sense of what is possible with software engineering, or learn what new technologies they might wish to learn to accomplish high level software development tasks.

Can Generative AI Help Social Scientists Write?

Numerous tools have been created in recent months that use Generative AI 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— but they are already transforming the way writing is done in many different fields (Korinek 2023). 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.[13] Yet— if used with careful review— I believe these tools also have the potential to improve the writing of those of us with well-worn pens. For example, Korinek (2023) proposes social scientists should consider asking LLMs to evaluate the weaknesses in their arguments, or identify counter-arguments. Though most researchers might not take the responses to such prompts very seriously, they may encourage us to reflect upon our own blind spots as writers— or to think more systematically about how other audiences might interpret our prose.

Can Generative AI Help Social Scientists Generate Research Questions and Hypotheses?

In recent decades, social scientists have enjoyed access to large databases such as the Web of Science which compile vast amounts of peer-reviewed research. Social studies of science that use network analysis to analyze citation patterns and natural language processing to identify themes in text have expanded accordingly (e.g. Edelmann, Moody, and Light 2017; Uzzi et al. 2013; L. Wu, Wang, and Evans 2019). Proponents of Generative AI have naturally become interested in whether these new tools can advance our capacity to map and understand science even further. Early attempts to use such technologies for this purpose largely failed. For example, Meta's Large Language Model, Galactica, was designed to help scientists navigate scholarly literatures more efficiently. It produced such inaccurate responses, however, that it was taken offline after three days.[14] The debut of Google's BARD chat assistant was similarly darkened when it provided an inaccurate response about the first documented picture of an exoplanet.[15]

Though Large Language Models might not yet be reliable enough to summarize scholarly literatures, they may yet be useful for helping social scientists at 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 fairly comprehensive set of articles when thus prompted, but organizes them according to criteria not typically available within scholarly databases— such as whether the studies are original empirical analyses or meta-analyses. The tool can further separate studies according to sample size and whether they include randomized

controlled trials or observational analyses. It can also assess what outcomes are measured, and learn in an adaptive manner by allowing users to "star" relevant articles and receive additional recommendations of studies that cite them in so doing. Though Elicit.org suffers from some of the same limitations of the aforementioned LLMs—and therefore cannot be trusted to provide a systematic literature review—it may provide significant value to social scientists as they begin to explore a new literature.

Some have begun to suggest that— 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. Elicit.org, the aforementioned large-language model trained on scientific corpora, offers a tool to help researchers brainstorm research questions. I asked this tool 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 in different countries?" and 2) "Why do people switch social media platforms, and how might this impact polarization"? I am thus not ashamed to admit that I asked ChatGPT, "can generative AI improve social science?" and it spit out several of the themes I have already discussed in this article, alongside a few that are nonsensical (such as "generative AI will help social scientists predict the outcomes of human behavior more effectively"). 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.

PART II: 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 there are a number of major limitations— and even possible dangers— that researchers will have to reckon with as they make decisions about whether or how to incorporate Generative AI into their work.

Generative AI exhibits Human Biases

As mentioned above, most forms of artificial intelligence exhibit human bias, since they are typically trained on data produced by human beings who are prone to many cognitive errors. For example, algorithms that assist judges in making parole decisions, which are more likely to recommend that white prisoners be released from jail early than their non-white counterparts (Kleinberg et al. 2018). Similarly, algorithms may be more effective at performing image editing tasks on visuals that depict white people than non-white people, since the latter are less-well represented in the training data used by such AI models (Yee, Tantipongpipat, and Mishra 2021). There are many, many more examples that have been documented at length elsewhere (e.g. Bender et al. 2021; Buolamwini

and Gebru 2018; Daneshjou et al. 2021). Because generative AI models such as ChatGPT are trained using large amounts of data created by humans on the internet, it is not surprising that it tends to exhibit many of the same types of biases identified by previous research.

Because the training data for ChatGPT is not publicly available, the exact types of bias it can exhibit are not yet well understood. Santurkar et al. (2023) asked a series of LLMs trained by OpenAI and A121 Labs to take a large group of public opinion surveys from the United States. 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 (Bender et al. 2021; Cho et al. 2019). In other words, it seems that LLMs tend to have bias that reflects some of the more advantaged parts of U.S. society, though not those who have more conservative viewpoints. Interestingly, concern that Generative AI might be biased against disadvantaged groups (regardless of their politics) is equally high among liberals and conservatives in the United States (Weidinger et al. 2023).

Importantly, Santurkar et al. (2023) show the bias 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). This mirrors earlier research on machine learning that suggests once the extent and direction of bias is identified it can often be corrected (Obermeyer et al. 2019). 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 opaque. Without access to the types of training data fed into such models, researchers can only examine "known unknowns." In other words, If disadvantaged and socially isolated elderly people are unable to voice their collective concern about how AI represents them, researchers might not think to study such bias.

On the other hand, one could argue that the bias inherent in most Generative AI could be both a "feature" and a "bug" for social science research. That is, many social scientists want to design experiments that examine the impact of bias on attitudes or behaviors. If such bias can be carefully controlled— a major assumption— than its possible that the bias of generative AI could be useful in reproducing the types of bias that might occur in real settings during empirical research. 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 (Cho et al. 2019).

Will Generative AI Spread Bias and Misinformation?

One of the most important stages in training a Generative AI model is when its developers provide it with feedback. This can include many different types of feedback— from instructions about how to avoid discussing dangerous topics (e.g. bomb making) to identifying reliable sources of information. This process typically occurs both behind closed doors— when employees of companies such as OpenAI engage in "red team" attacks designed to goad the model into producing dangerous, or illegal content— or through public user feedback. The developers of Generative AI tools can thus design workflows that try to create "guard rails" for models that can become increasingly unpredictable as they are trained on increasingly large datasets. On the other hand, the opaque and selective nature of model training means that some issues will be addressed better than others. For example, Ippolito et. al. (2020) show that many Generative AI models are trained to try to trick humans into believing they are real people— but this same training can come at the expense of providing reliable, accurate information. It is also possible that the guard rails developers create reflect their own interests and concerns— it is not yet known, for example, whether GPT-4 is more likely to protect young liberals in Silicon Valley than elderly conservative people in rural areas who might suffer abuse from online trolls.

The potential for malicious actors to use Generative AI to spread misinformation— or for the model to reproduce bias in a variety of settings such as job postings, even by well-intentioned users— is deeply concerning. But this danger masks an even deeper problem: 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 prevents users from employing LLMs to answer questions to prevent a situation where users would struggle to discern the good information from the bad.[16]

The "Stack Overflow Problem" could be particularly dangerous for researchers who come to rely upon LLMs to perform literature reviews, generate hypotheses, or otherwise summarize vast corpora. Fortunately, computer scientists have begun to create digital "watermarks" that might allow enable LLM's 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 effectively create an "accent" for LLMs— giving them a list of words they should use as much as possible, for example— in order to allow people to retrospectively identify content that was not generated by

humans (Kirchenbauer et al. 2023). But this will be difficult to implement at scale for several reasons. First, the companies which develop LLMs will have to agree to insert watermarks within them, and then coordinate with each other so that LLMs can identify each other. But such efforts would fail to detect other LLMs created, for example, by cloning Facebook's recently leaked LLAMA model, or future open-source efforts (which are largely unregulated).

Is Research with Generative AI Ethical?

Perhaps the most pressing question for social scientists is whether generative AI can be used in an ethical manner for research purposes. This question is particularly important since many Generative AI tools exhibit biases that are not only often offensive (e.g. racism or misogyny), but also hold the potential to expose people to misinformation or other inaccurate or polarizing beliefs. 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— it takes on added importance in situations where human research participants might have conversations with a LLM in an unsupervised manner.

Among the most important questions that social scientists must answer is whether respondents must provide informed consent before interacting with Generative AI. Such consent appears critical given the potential for such tools to become abusive, or spread misinformation. Yet disclosing Generative AI as such also decreases its scientific utility for simulating human behavior. Moreover, such disclosure would make it difficult for researchers to know whether people are responding to the behavior of the AI or expressing opinions about AI more broadly.

One solution to this problem may be to design studies in which research participants are informed that they may interact with artificial intelligence during a study, but employ a mix of human and AI agents in interactive settings. Even this strategy, though, creates the risk that an AI agent could encourage conflict between human participants. Some of these risks can be mitigated by using content moderation filters that are currently available on platforms such as OpenAI, and performing rigorous testing of the prompts used to guide the performance of Generative AI 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 fully disclose the use of Generative AI in research with human participants. For example, Argyle et. al. (2023) designed a study where GPT-3 was presented as an "AI chat assistant" that intervened in peer-to-peer conversations about gun control in an online forum among people from opposing sides on the issue. In the experimental condition of the study, one of the two parties to a conversation was shown a rephrasing of a message they were about to send 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 significantly increased the perceived quality of these conversations among those whose discussion partners had been offered assistance by 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 and proceed without AI assistance.

A final strategy might be to use Generative AI to try to diagnose possible ethical issues in research studies. 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 legal 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 to be 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? I am not aware of any such research at present, but attempts to answer this question could be low risk (and high reward) if only simulated agents are used.

Can Research with Generative AI be Reproducible?

As discussed above, the training and development of most proprietary large language models is largely opaque. When OpenAI released GPT-4, they did not release details about how it was trained— consistent with prior releases. Yet the company also decided not to announce how large the model was either unlike previous releases. As competition increases among corporations producing AI technology, public disclosure about the inner-workings of model development may become increasingly rare. This raises important concerns about whether research with Generative AI can not only be interpretable (i.e. we will not know why models behave the way they do) but also whether they will be reproducible. That is, whether one team of researchers that hopes to duplicate the work of other teams in order to verify or build upon previous findings, will be able to do so (Spirling 2023). Fortunately, there are numerous efforts to make Generative AI open-source such as Hugging Face. Early analyses indicated the performance of open-source models is sub-par when compared with proprietary models such as GPT-4. Yet many open-source models are considerably cheaper to train, more efficient, and may allow researchers to "fine-tune" them in ways that makes their behavior more reproducible. There is also emerging evidence that open-source models are rapidly closing performance gaps with their proprietary counterparts.[17]

Will Detecting and 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 indicate the potential for them to make mistakes in such tasks is high. Though major mistakes (such as obvious bias) might be easy to detect, small mistakes (e.g. a bug in a very long piece of code) may be much harder to detect. Evaluations of "autopilot" tools for software coding, for example, have been shown to frustrate some users who exert considerable effort identifying tiny bugs in code written by an AI, even if they appreciate its capacity to generate large chunks of code quickly (Liang, Yang, and Myers 2023). Another useful analogy is the self-driving car. In principle, self-driving cars should be appealing to many people because they might relieve us of the cognitive and physical load of driving. Yet in practice, self-driving cars need to be closely monitored by drivers in case the AI fails. Social scientists may soon face a similar trade off: though we might at first enjoy outsourcing many undesirable parts of our jobs to Generative AI, we may soon discover the annoyance of monitoring it closely to identify the occasional bug may outweigh its benefits.

CONCLUSION

There are few technologies in history that have considered so much excitement— and so much concern— concomitantly. Hype cycle dynamics indicate expectations for Generative AI are probably at their peak, and may soon crash down as users become more familiar with their pitfalls (Baumgartner and Jones 1993; Salganik 2018). I expect social scientists will play a key role in identifying those pitfalls, but I also hope that we will not become so preoccupied with the limitations of Generative AI that we do not fully evaluate its promise. I do not believe that Generative AI will supplant most of the work social scientists do today. But I also predict Generative AI will soon become more integrated into the existing research techniques than most social scientists realize. This need not involve handing off pivotal research activities such as literature reviews to Generative AI blindly. But it would be foolish to ignore such tools if they can quickly and efficiently help us identify our own blindspots or act as a muse that helps us sharpen our research questions.

I predict social scientists will also play a central role in addressing the current limitations of Generative AI and identifying new ones. At present, most of the people who develop Generative AI do not have extensive training in working with human research subjects. Social scientists, by contrast, have extensive experience identifying the many ethical dilemmas that can arise when people interact with new technologies— if not through formal human subjects review processes than through past mistakes. These experiences indicate we must not only focus on known dangers, but also work carefully to identify "unknown unknowns." In his influential book, Bit by Bit, Salganik calls upon researchers in

computational social science to hold themselves to a higher standard than that of the Institutional Review Boards that are commonplace at most universities. This is because social scientists interested in new technologies have already caused harm, unintentionally, by failing to think carefully about the experience of research from the perspective of research participants. This can be accomplished by administering questionnaires about how research participants might react to future planned studies, or through extensive small-scale pre-testing of research processes with small samples— ideally in person, so that any possible harms can be mitigated as soon as possible.

But social scientists need not only think of themselves as guardians of human subjects, or end-users of Generative AI tools. I predict the future of AI research will require training new 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 described as 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 even if they 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 shape the science of social relationships.

REFERENCES

Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies." http://arxiv.org/abs/2208.10264 (May 11, 2023).

Argyle, Lisa P., Ethan Busby, Joshua Gubler, et al. 2023. "AI Chat Assistants Can Improve Conversations about Divisive Topics." http://arxiv.org/abs/2302.07268 (April 25, 2023).

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, et al. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." Political Analysis: 1–15.

Bail, Christopher. 2015. "Taming Big Data: Using App Technology to Study Organizational Behavior on Social Media." Sociological Methods & Research.

Bakhtin, Anton et al. 2022. "Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning." Science 378(6624): 1067–74.

Baumgartner, Frank, and Bryan Jones. 1993. Agendas and Instability in American Politics. 1st ed. University Of Chicago Press.

Becker, Joshua, Ethan Porter, and Damon Centola. 2019. "The Wisdom of Partisan Crowds." Proceedings of the National Academy of Sciences 116(22): 10717–22.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "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, New York, NY, USA: Association for Computing Machinery, 610–23. https://dl.acm.org/doi/10.1145/3442188.3445922 (April 26, 2023).

Binz, Marcel, and Eric Schulz. 2023. "Using Cognitive Psychology to Understand GPT-3." Proceedings of the National Academy of Sciences 120(6): e2218523120.

Bommasani, Rishi et al. 2022. "On the Opportunities and Risks of Foundation Models." http://arxiv.org/abs/2108.07258 (April 18, 2023).

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html (April 26, 2023).

Burnett, Sam, and Nick Feamster. 2015. "Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests." http://arxiv.org/abs/1410.1211 (April 17, 2023).

Cho, Won Ik, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. "On Measuring Gender Bias in Translation of Gender-Neutral Pronouns." In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Florence, Italy: Association for Computational Linguistics, 173–81. https://aclanthology.org/W19-3824 (April 26, 2023).

Chu, Eric, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. "Language Models Trained on Media Diets Can Predict Public Opinion." http://arxiv.org/abs/2303.16779 (April 21, 2023).

Daneshjou, Roxana et al. 2021. "Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review." JAMA Dermatology 157(11): 1362–69.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. "Can AI Language Models Replace Human Participants?" Trends in Cognitive Sciences 0(0). https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(23)00098-0 (May 11, 2023).

Edelmann, Achim, James Moody, and Ryan Light. 2017. "Disparate Foundations of Scientists' Policy Positions on Contentious Biomedical Research." Proceedings of the National Academy of Sciences 114(24): 6262–67.

Edelmann, Achim, Tom Wolff, Danielle Montagne, and Christopher A. Bail. 2020. "Computational Social Science and Sociology." Annual Reivew of Sociology 46.

Fiesler, Casey, and Nicholas Proferes. 2018. " 'Participant' Perceptions of Twitter Research Ethics." Social Media + Society 4(1): 2056305118763366.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks." http://arxiv.org/abs/2303.15056 (April 25, 2023).

Griffin, Lewis D. et al. 2023. "Susceptibility to Influence of Large Language Models." http://arxiv.org/abs/2303.06074 (April 26, 2023).

Horton, John J. 2023. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" http://arxiv.org/abs/2301.07543 (May 9, 2023).

Ippolito, Daphne, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. "Automatic Detection of Generated Text Is Easiest When Humans Are Fooled." http://arxiv.org/abs/1911.00650 (April 26, 2023).

Jakesch, Maurice, Jeffrey T. Hancock, and Mor Naaman. 2023. "Human Heuristics for AI-Generated Language Are Flawed." Proceedings of the National Academy of Sciences 120(11): e2208839120.

Kirchenbauer, John et al. 2023. "A Watermark for Large Language Models." http://arxiv.org/abs/2301.10226 (April 26, 2023).

Kleinberg, Jon et al. 2018. "Human Decisions and Machine Predictions*." The Quarterly Journal of Economics 133(1): 237–93.

Korinek, Anton. 2023. "Language Models and Cognitive Automation for Economic Research." https://www.nber.org/papers/w30957 (May 9, 2023).

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." Proceedings of the National Academy of Sciences 111(24): 8788–90.

Kreps, Sarah, and Douglas L. Kriner. 2023. "The Potential Impact of Emerging Technologies on Democratic Representation: Evidence from a Field Experiment." New Media & Society: 14614448231160526.

Kreps, Sarah, R. Miles McCain, and Miles Brundage. 2022. "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation." Journal of Experimental Political Science 9(1): 104–17.

Lazer, David M. J. et al. 2020. "Computational Social Science: Obstacles and Opportunities." Science 369(6507): 1060–62.

Lewis, Kevin et al. 2008. "Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.Com." Social Networks 30(4): 330–42.

Liang, Jenny T., Chenyang Yang, and Brad A. Myers. 2023. "Understanding the Usability of AI Programming Assistants." http://arxiv.org/abs/2303.17125 (May 11, 2023).

Macy, Michael W., and Robert Willer. 2002. "From Factors to Actors: Computational Sociology and Agent-Based Modeling." Annual Review of Sociology 28(1): 143–66.

Mellon, Jonathan et al. 2023. "Does GPT-3 Know What the Most Important Issue Is? Using Large Language Models to Code Open-Text Social Survey Responses At Scale." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4310154 (April 25, 2023).

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." Science 366(6464): 447–53.

Park, Joon Sung et al. 2023. "Generative Agents: Interactive Simulacra of Human Behavior." http://arxiv.org/abs/2304.03442 (April 25, 2023).

Salganik, Matthew. 2018. Bit by Bit: Social Research in the Digital Age. Princeton, N.J.: Princeton University Press.

Santurkar, Shibani et al. 2023. "Whose Opinions Do Language Models Reflect?" http://arxiv.org/abs/2303.17548 (April 25, 2023).

Spirling, Arthur. 2023. "Why Open-Source Generative AI Models Are an Ethical Way Forward for Science." Nature. https://www.nature.com/articles/d41586-023-01295-4 (April 26, 2023).

Srivastava, Aarohi et al. 2022. "Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models." http://arxiv.org/abs/2206.04615 (April 18, 2023).

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones. 2013. "Atypical Combinations and Scientific Impact." Science 342(6157): 468–72.

Vinyals, Oriol et al. 2019. "Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning." Nature 575(7782): 350–54.

Weidinger, Laura et al. 2023. "Using the Veil of Ignorance to Align AI Systems with Principles of Justice." Proceedings of the National Academy of Sciences 120(18): e2213709120.

Wu, Lingfei, Dashun Wang, and James A. Evans. 2019. "Large Teams Develop and Small Teams Disrupt Science and Technology." Nature 566(7744): 378–82.

Wu, Patrick Y., Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. "Large Language Models Can Be Used to Scale the Ideologies of Politicians in a Zero-Shot Learning Setting." http://arxiv.org/abs/2303.12057 (April 25, 2023).

Yang, Kai-Cheng, and Filippo Menczer. 2023. "Large Language Models Can Rate News Outlet Credibility." http://arxiv.org/abs/2304.00228 (April 25, 2023).

Yee, Kyra, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. "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.

Zhou, Jiawei et al. 2023. "Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions."

Ziems, Caleb et al. 2023. "Can Large Language Models Transform Computational Social Science?" ArXiv. https://calebziems.com/assets/pdf/preprints/css_chatgpt.pdf.

---

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

[2]https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[3] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[4] Others have expressed grave concerns about intellectual property— since generative AI is trained on large amounts of original content produced by humans— and flagged how difficult it currently is to detect such technology in real life settings such as social media platforms. Still others have demonstrated that generative AI is prone to prejudice and bias that is pervasive within the training data used to create such technology. Though most experts in artificial intelligence reject the idea of a singularity— that artificial intelligence will eventually exceed human intelligence and represent a threat to human existence— there are a range of reasonable concerns about how generative AI could embolden the darker angels of human nature.

[5] Still others elected not to sign the open letter because they believed that it was designed to "solve the wrong problems." More specifically, some argued that the singularity— the threat of a superintelligent AI subjugating humans— was far fetched, and masked much more likely dangers related to misinformation, bias, and discrimination.

[6] Generative AI is also referred to as a subclass of "Foundation Models"-- a term which describes a model trained on broad data in an unsupervised manner to a wide range of downstream tasks (Bommasani et al. 2022).

[7] This issue may become less troublesome, however, as the amount and diversity of training data used to create these models continues to increase.

[8] A variety of other benchmarks have been proposed, including the Beyond the Imitation Game (BIG) framework, which assess the capability of Generative AI tools to perform a range of more complex tasks from calculus to biology and software development (Srivastava et al. 2022). In general, the authors of this framework find the overall capacity of Generative AI to perform such high-level tasks is limited but increases with the scale of the model (i.e. the amount of data used to train them).

[9] Kreps and Kriner (2023) conducted a field experiment where they sent emails authored by humans and LLMs to elected officials in the United States, and found the latter were no more likely to receive responses than the former.

[10] See also Griffin et. al. (2023).

[11] Similarly, Chu et. al. (2023) show that LLMs trained on the media diets of U.S. partisans can accurately perform the attitudes and opinions of their associated audiences on large, nationally-representative panels.

[12] Interestingly, this same study suggests GPT-3 cannot reproduce the Wisdom of Crowds phenomenon.

[13] At the same time, tools for detecting content generated by LLMs appears biased against those for whom English is not their first language: https://arxiv.org/pdf/2304.02819.pdf.

[14]technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/

[15] https://mashable.com/article/google-bard-james-webb-telescope-false-fact

[16] https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned

[17] For example, see https://www.semianalysis.com/p/google-we-have-no-moat-and-neither?s=31.