NBER WORKING PAPER SERIES

GENERATION NEXT: EXPERIMENTATION WITH AI

Gary Charness
Brian Jabarian
John A. List

Working Paper 31679
http://www.nber.org/papers/w31679

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023, Revised October 2023

We thank Daniel Carey and Rishane Dassanayake for excellent research assistance. We thank Ariel Listo and Justin Holz for useful suggestions and the audience at AFE 2023 for useful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Generation Next: Experimentation with AI
Gary Charness, Brian Jabarian, and John A. List
NBER Working Paper No. 31679
September 2023, Revised October 2023
JEL No. C0,C1,C80,C82,C87,C9,C90,C92,C99

ABSTRACT

We investigate the potential for Large Language Models (LLMs) to enhance scientific practice within experimentation by identifying key areas, directions, and implications. First, we discuss how these models can improve experimental design, including improving the elicitation wording, coding experiments, and producing documentation. Second, we delve into the use of LLMs in experiment implementation, with an emphasis on bolstering causal inference through creating consistent experiences, improving instruction comprehension, and real-time monitoring of participant engagement. Third, we underscore the role of LLMs in analyzing experimental data, encompassing tasks like pre-processing, data cleaning, and assisting reviewers and replicators in examining studies. Each of these tasks improves the probability of reporting accurate findings. Lastly, we suggest a scientific governance framework that mitigates the potential risks of using LLMs in experimental research while amplifying their advantages. This could pave the way for open science opportunities and foster a culture of policy and industry experimentation at scale.

Gary Charness
University of California at Santa Barbara
charness@econ.ucsb.edu

Brian Jabarian
The University of Chicago
Booth School of Business
5807 S. Woodlawn
Chicago, IL 60637
jabarian@uchicago.edu

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and Australian National University
and also NBER
jlist@uchicago.edu

1 Introduction

Large Language Models (LLMs) represent a sophisticated application of machine learning algorithms, showing a capacity to create original content and thus underscoring their status as generative Artificial Intelligence (AI) (Bubeck et al., 2023). Despite their relatively recent emergence, the full extent of the rapid effects of generative and transformative AI in science, policy, and society remains to be experienced (Frank et al., 2019; Zhang et al., 2021; Bommasani et al., 2022). This includes the field of economics (Brynjolfsson and Mcafee, 2017; Manning et al., 2022; Acemoglu and Johnson, 2023; Korinek, 2023). A natural venue in economics is to generate data for causal inference in experimental settings, for example, online. While once an academic curiosity, online experiments have become a bona fide contributor to causal estimates in the social sciences (Athey, 2015; Brynjolfsson et al., 2019a). With the burgeoning digital economy, researchers believe that the generation of causal insights using online experiments will continue to increase (Fréchette et al., 2022).

However, one key feature of online experiments that tempers the optimism of even its most enthusiastic supporters is the violation of the four exclusion restrictions, calling into question the internal validity of the received estimates. For example, compliance, one of the four identification assumptions that underlie the experimental approach (List, 2023), is often questioned in online experiments because it is usually associated with high measurement errors (Gillen et al., 2019). Checking whether individual participants understand the experiment's instructions is often tricky, particularly in an online experiment, where people cannot usually ask questions and receive live responses. While one remedy might involve incorporating real-time human support to address participant inquiries, it would require at least one of these conditions: 1) having a sizable skilled labor force to accommodate simultaneous questions from many participants, and 2) providing extended availability to cover the protracted timelines with online experiments.

Although machine learning algorithms have improved causal inference analysis methods in economics (Athey and Imbens, 2019), we expect these new models LLMs to radically improve critical areas of scientific knowledge production, in particular to overcome these issues related to online experiments. LLMs can be fine-tuned as chat assistants to simulate sophisticated human interactions while reducing labor costs. Given their inherent scalability and versatility, such integration could become standard practice for future online experiments, revolutionizing the field and fostering unprecedented advancements across various online experiments, including surveys,

incentivized individual decisions, and game-theoretical experiments. In addition, this approach can be deployed without prerequisite or minimal coding knowledge and is compatible with many experimental online platforms familiar to researchers, such as Qualtrics, oTree, and Z-tree. By ensuring consistency of treatment within and across these settings, another of the exclusion restrictions, the stable unit treatment value assumption (SUTVA), will be more likely to hold. Similarly, observability, a third exclusion restriction, is more likely to hold by minimizing the experimental burden on subjects by maintaining participant focus and engagement.

While this example highlights one key area of improvement of generative AI for experimentation, other areas are also open for similar enhancements. For example, specific fine-tuned language models could homogenize and carry out randomization and re-randomization techniques, lending more credibility to the fourth exclusion restriction, statistical independence. Furthermore, integrating them into the development and analysis of experimental research can address the challenges researchers commonly face, such as optimizing the wording of tasks, improving comprehension (Ouyang et al., 2022), and streamlining data analysis, especially coding and data visualizations (Wang et al. (2023)). Using the capabilities of this technology, we can create more immersive online experiences, facilitate real-time monitoring of participant participation, and improve the quality and replicability of experiments. In addition, its use can promote open science opportunities, fostering increased collaboration, and this technique can promote open science among researchers.

This paper often refers to LLMs and their capabilities.[1] But,this does not imply that users can simply input our suggested directions and some experiment details into ChatGPT and expect satisfactory results. Generative AI, a rapidly transforming technology, is sensitive to inputs and can produce unpredictable outputs (Ganguli et al., 2022) and, as a result, working out which inputs lead to the most desired outputs, known as prompt engineering, is becoming a growing part of the industry. Furthermore, the stochastic nature of generative AI means that results can be further improved by researchers taking multiple draws for the same prompt and selecting the best result ex post (Davies et al., 2021) or by launching A/B tests and other types of experiment to determine which prompt is the most effective.

In their best light, based on up-and-coming research and development done by leading AI labs, we envision these language models as the wise sage always available at the experimentalist's beck and call. Within this framework, we explore their imple-

[1]We do so with 'foundation models' and 'fine-tuned models' in mind under the umbrella of generative AI (Bommasani et al., 2022).

mentation more generally in Section 2, focusing on their role in comprehension and immersive experiences. Section 3 examines their capacities in data collection, including real-time monitoring, preprocessing, and cleaning, while Section 4 considers data analysis. The final section discusses the broader risks and benefits of the proliferation of generative AI in behavioral and experimental economics and implications for open science and scaling a culture of experimentation in business and policy-making. It gives some speculative pointers on how to manage these.

2 Designing Experiments with LLMs

By enhancing researcher productivity, LLMs free up resources during the design phase of investigations, broadening the scope of potential research questions and the focus that can be dedicated to ensuring validity. LLMs can be instrumental in generating research ideas, offering hypotheses drawn from the existing literature, current economic trends, and seminal problems in a field. By evaluating research objectives, these models can recommend appropriate experimental designs, including economic games, decision tasks, or market simulations, guiding the best structure for control and treatment groups to isolate causal relationships effectively.

Moreover, such models can assist in selecting the ideal experimental setting, laboratory, field, or online – contingent on the research question and context – and determine the appropriate sample size for the experiment, considering factors such as effect size, adequate statistical power, and resource constraints based on the previous literature and state-of-the-art statistical prowess. In particular, they can ensure balanced and comparable groups using random assignment, stratification, or matching. Their extensive training data also allows them to suggest relevant variables to manipulate and measure, providing optimal operationalization methods such as monetary incentives and real-effort tasks. In the limit, they can also guide relevant ethical considerations, such as deception, informed consent, and potential harm to participants, helping researchers design scientifically rigorous and ethically sound studies.

Critically, LLMs can write clear, concise instructions and comprehension checks. Researchers can minimize measurement errors and increase the chance of compliance by ensuring that participants understand the experimental setup and tasks. Models such as GPT-4, trained on billions of data points, possess a deep understanding of language patterns and so can tailor instructions' language, tone, and complexity to participants with varying language proficiency, education, or familiarity with economic concepts. To illustrate key issues, they could even generate relevant examples, such

as textual descriptions, hypothetical scenarios, or visual representations of economic concepts or analogies. By having LLMs generate multiple versions of instructions and comprehension checks and iteratively provide feedback from researchers or pilot study participants to AI, interactions with participants can be optimized for clarity. LLMs can effectively assist in critiquing human- or AI-generated concise instructions against lengthy but precise benchmarks provided by the researcher (Saunders et al., 2022).

Language models offer versatile and efficient solutions for implementing experiments in various programming languages. LLMs can convert experimental setups, desired variable manipulations, and data collection, plain English, into complete code of Python, JavaScript, HTML, or R scripts (Chen et al., 2021). In addition, they can adapt the generated code to specific experimental platforms, ensuring seamless integration and adapting the code to the requirements of a platform. In the same vein, they can reason through experiments, detect design coding errors, and even simulate participants (Horton, 2023). By simulating demographic characteristics, backgrounds and language proficiency, these virtual participants can 'interact' with experiment materials and highlight confusion or misunderstanding with respect to the wording of the instructions, logical errors, or unexpected results in the experimental code, all of which may be harmless to researchers but crucial to the validity of the results (Charness et al., 2004). By investigating how simulated participants interact, researchers can examine these issues in a controlled experimental setting while maintaining a high degree of ecological and external validity. Researchers can also generate multiple versions of instructions and comprehension checks by providing feedback from other researchers or pilot study participants with LLM that helps evaluate concise human or AI-generated instructions against lengthy but precise benchmarks provided by the researcher (Saunders et al., 2022). This iterative approach enables researchers to refine instructions, optimizing them for clarity and effectiveness.

Furthermore, LLMs can help researchers create replication packages in the same or different settings, including the necessary materials to reproduce a study. Given the observed differences between online and offline samples (Snowberg and Yariv, 2021), such abilities are valuable for the field. They can also assist in developing other documentation (for example, IRB submissions) to more effectively explain the design of the experiment and its impact on the design of the experiment. If given access through plugins, they could automate the use of ML tools designed to make pre-registration less restrictive and calculate the lost power from such tools (Ludwig et al., 2019). Similarly, LLMs can be crucial in standardizing approaches since the literature suggests that design flexibility and associated experimental decisions are inversely related to ac-

tual research findings (Ioannidis (2005). Overall, such standardization of experimental design templates could go further and be made through a public library of standard experimental design templates, ready-to-use, scaling an approach developed manually in different languages, to name a few, in Qualtrics (Molnar, 2019) and oTree (Chen et al., 2016).

Besides, generative AI also presents opportunities to design policy experiments, RCT design, and predictive models. It helps to formulate and test policy hypotheses by analyzing vast amounts of data, which helps to determine the plausibility of policy assumptions and predict possible outcomes. Generative AI has already succeeded in guiding mathematicians by generating conjectures, using feedback from mathematicians to narrow down the space of significant relationships between variables to those that might be relevant (Davies et al., 2021). Similar capabilities could encourage a more experimental and evidence-based approach to policymaking. LLMs also have a role in the design of Randomized Controlled Trials (RCTs), a standard in policy experimentation. An essential critique of RCTs in economics is their frequent inability to meet the "double-blind" standards of medicine (Deaton and Cartwright, 2018). However, if LLMs supervise an experiment, neither human participants nor researchers interacting with the data need to know the specific treatment administered to each group.

Finally, any potential bias introduced by LLMs can be tested cheaply and rigorously and addressed before implementation, in contrast to human experts. These models can generate various experimental design options, reducing the amount of human labor required. Predictive modeling is the third area of interest of social scientists (Hofman et al., 2021), where generative AI can make significant contributions. By developing and refining predictive models, LLMs allow government and nonprofit agencies to simulate the effects of policy experiments, anticipating the consequences of policy changes before full implementation and thus informing the design and adjustments of iterative policy interventions.

3 Implementing Experiments with LLMs

Incorporating AI agents into online experiments can streamline aspects of the implementation of experiments while enhancing data quality collection. They could help with participant recruitment, provide real-time assistance, increase engagement, live monitor data quality, and facilitate follow-up surveys.

Chatbots demonstrate significant potential to provide detailed, instant responses to inquiries. Recent evidence from Eloundou et al. (2023), Noy and Zhang (2023) and

Brynjolfsson et al. (2023) show in different settings that granting humans access to AI-powered chat assistants can significantly increase their productivity. AI assistance allows human support to provide faster and higher quality responses to a more extensive customer base. This technique can be imported to experimental research, where participants might need clarification of the instructions or have other questions. In its most scalable version, we imagine having fully autonomous chatbots providing simultaneous support to hundreds of participants with few or no live support human agents. Consequently, these new generative AI models can represent a highly scalable solution with significant potential to improve the viability of the compliance assumption. This, in turn, may improve causal inference of data collected broadly in online experiments. As in the discussion of instruction comprehension in Section 2, chat assistants can answer questions and personalize their interaction according to the communication and comprehension style of each participant during the experiment. This is important to ensure that participants understand the experimental instructions before performing tasks, a necessary step to ensure compliance and avoid construct validity concerns.

Furthermore, providing all necessary guidance within the chat interface helps to maintain participant focus and engagement, thus preventing participants from feeling overwhelmed or confused, and minimizing distractions or multitasking that could introduce noise. Thus, observability, the third exclusion restriction, is more likely to hold by lessening the experimental burden on subjects. One can even learn why a participant failed a comprehension question so that researchers can decide how to proceed in real time. Finally, AI agents can automate the administration of follow-up surveys or debrief questionnaires, collecting additional data cost-effectively.

AI agents can also be fine-tuned to monitor (at scale) cheating in several ways. For example, they can, with relevant plugins, automatically implement different techniques done through JavaScript so far: tracking participant browser activity by opening new tabs, switching between windows, or spending excessive time away from the experiment, as was done by (Jabarian and Sartori, 2020). By systemically monitoring these actions, AI agents can detect potential cheating and remind participants to focus on the task. Second, they can analyze real-time participant responses for patterns suggesting cheating, repeating answers, or providing contradictory information. Researchers can then review these flagged cases in real time and determine whether further action is necessary. Currently, chat assistants can effectively handle unexpected scenarios and technical issues to ensure a smooth experimental process. They contribute to a more controlled environment by providing real-time reminders, reducing participants' chances of forgetting critical information. Additionally, chat assistants

could reduce attrition by engaging participants in dynamic and interactive conversations. This real-time interaction facilitates higher-quality data, as more engaged participants are less likely to make errors or random responses. AI agents and chat assistants create a more efficient and reliable experimental setup, balancing strict supervision of participant behavior with real-time assistance and participant engagement. Finally, they can select participants for eligibility criteria, ensuring a representative and appropriate sample. By automating this process, researchers can save time and reduce the risk of human error.

Chat assistants can dynamically tailor the design of an experiment to enhance data collection. For example, in a cognitive-ability experiment, if a task is too easy or difficult for a participant, a chatbot could adjust the difficulty of following tasks accordingly, allowing better identification of the type of participants. This maintains participation and accurately measures the participant's abilities. Similarly, the chatbot can prioritize or deprioritize questions based on a participant's earlier responses in a personality assessment. This personalized approach, made possible by chatbots, allows for more nuanced data collection, offering a more effective and customized experimental process.

In addition, depending on the research question and design, providing immediate feedback can improve motivation and understanding of tasks when appropriate. Chat assistants can simulate social interactions, such as negotiations, group decision making, or trust building exercises, in the study of complex social phenomena. Thus, researchers can examine these topics in a controlled experimental setting while maintaining a high degree of ecological and external validity. Automating the data collection process through chat assistants reduces the risk of experimenter bias or demand characteristics that influence participant behavior, resulting in a more reliable evaluation of research questions (Fréchette et al., 2022).

Finally, chat assistants offer scalability by handling multiple participants simultaneously, facilitating large-scale data collection cost-effectively and timelessly, and allowing greater control over time-of-day-caused session effects. This generates more robust and generalizable findings by accessing diverse and representative samples.

4 Analyzing Experimental Data with LLMs

LLMs can substantially increase the analysis of economic experiments by assisting with data analysis in two primary ways. First, they could automate data analysis tasks such as sanitization, examining relationships within the data, and analyzing the data

using the Code Interpreter on ChatGPT. Second and less obvious, they could create and explore new data unexamined so far in standard economic experiments.

Regarding exploring new data, the use of natural language processing (NLP) techniques with live chat logs from experiments can yield insights into participant behavior, uncertainty, and cognitive processes. Such insights provide opportunities to observe and scrutinize new variables for statistical tests and identify factors that could influence conventional results. Variables of potential interest that emerge from chat logs encompass the frequency of questions posed, the degree of engagement in conversation, or sentiments expressed by participants. Understanding these variables can illuminate the correlation between participant behavior and experimental results, offering a more nuanced understanding of the factors shaping outcomes. This "under-the-hood" perspective can spark novel hypotheses and insights.

During data pre-processing, language models can distill pertinent details from chat logs, organize the data into an analytical-friendly format, and manage any incomplete or missing entries. Beyond these tasks, such models can perform content analysis - identifying and categorizing frequently expressed concerns of participants, analyzing sentiments and emotions conveyed, and gauging the efficacy of instructions, responses, and interactions. Models of this nature are also equipped to pinpoint areas of confusion. This ability can help improve experimental designs, instructions, and training protocols for chatbots covered in Section 2. It could also be used in the final data analysis for new types of behavioral heterogeneity analysis. Participant characteristics such as demographics and cognitive abilities, and their influence on chat interactions and experimental outcomes, can also be explored through these models.

With respect to automating data analysis tasks, specific LLMs, such as Code Interpreter, can help at different stages of the knowledge production chain. First, it can help authors conduct statistical tests, develop econometric models, determine causal relationships, and perform robustness checks by harnessing state-of-the-art techniques. Automating these tasks offers dual benefits. First, it saves researchers' time, which can be allocated to other tasks in knowledge production. Second, it minimizes research flexibility across all tasks and strengthens the likelihood that reported research findings represent genuine associations (Ioannidis, 2005). In addition, they can generate data visualizations in concert with other features. This helps researchers understand both their results and communicate their findings effectively.

Second, after authors have conducted their analysis, LLMs can aid them and the broader scientific community in comparing the finished work and pre-registration plans. In particular, these tools would be adept at pinpointing and highlighting signif-

icant divergences, including the unexpected introduction of new variables, the omission of predetermined variables, the application of alternative econometric paradigms, or modifications in the specified data acquisition methods. Furthermore, these language models can be fine-tuned to distinguish between confirmatory and exploratory analysis. Confirmatory analysis, anchored to a pre-registered plan, aims to validate predetermined hypotheses. In contrast, exploratory analysis allows for a more flexible approach to data interpretation. These models can scrutinize the manuscript for sections indicating a diversion from the pre-registered schema towards exploratory analysis, valid not only for the authors in their submission process but also for referees and editors in the review process.

In the same vein, although these models may still encounter difficulties discerning between AI versus human-generated text, they are becoming rapidly proficient in readily and accurately detecting AI-generated code, anomalies, or red flags within code analysis and using LLMs under the impulse of OpenAIs classifiers (Kirchner et al., 2023). In a specific application, we could imagine presenting models with a pair – the result and its corresponding interpretation – to determine the fidelity of the interpretation relative to the actual result. Far from being merely speculative, this task could rapidly be implemented since, when aiding a human expert, LLMs have been shown to improve significantly at coding challenges if repeated sampling is allowed (Chen et al., 2021). This capability could be pivotal in identifying instances of overclaiming, where interpretations may exceed the implications of the results or, conversely, underclaiming, where the interpretation fails to capture the full potential of the results. Anomalies such as 1) misalignment between the quantitative findings and their qualitative exposition or 2) significant findings that are overlooked due to suboptimal communication or missing facts could be flagged by LLMs for further examination.

Generative AI also offers significant opportunities to help facilitate peer review, replication, and dissemination of research. Previous tasks, such as comparison with pre-registration plans, analysis of data for tampering, analysis of code and supplemental text, and new tasks, such as separating human and AI work, that would be highly time intensive for human researchers are now far less so. These models can cross-compare claims in the body of a paper with the code, ensuring that the implementation matches the theory. It can examine datasets and highlight irregularities like outliers driving results, text, ordering, or meta-data that do not fit the implicit patterns it can identify. Furthermore, it can write summaries of the appendices, allowing reviewers and replicators to see if their concerns are addressed quickly. Given the experimental setup and treatment, it can be checked whether appropriate tests for the

main results and robustness have been carried out in the main text or the supplementary material. At the limit, we even envisage simulated replications using existing code and information in the paper and appendix that could help highlight coding errors or irregular results. Any such endeavor would be fraught with difficulties; especially with more novel results relaying on behaviors that it is less likely that LLMs will have internalized. However, this idea could hold promise with the possibility of simulating independent participants and the capacity for many simulations. These various abilities can help boost the speed of review and the rate of replication (two common concerns in the field) and benefit research efficiency.

5 Discussion: Risks and Opportunities

Using LLMs in economic research can pose several risks (Bommasani et al., 2022), including concerns about intellectual property (IP), digital privacy issues, user deception, scientific fraud by fabricating data or strategies to hide data manipulation, hallucinations, and challenging creativity by homogenizing the human-AI interface too much. For example, generative AI not citing its sources can be unintentional plagiarism or copyright infringement, and relying on technologies explicitly citing their sources, such as PerplexityAI or Elicit, seems desirable. Such possible drawbacks call for increased scrutiny from the scientific community and more transparent scientific practice, from collecting data to publishing the papers. Beyond IP concerns, several other potential issues remain. First, the vast amounts of data these language models process can create privacy concerns, especially since these data may contain sensitive participant information. Researchers fine-tuning such models should follow best practices such as anonymizing data, obtaining informed consent, and implementing secure data access controls and storage methods to protect data.

Second, deception may occur since, as mentioned in the previous section, it may require help distinguishing AI-generated content from human-generated content, particularly in high-frequency information settings such as social media. Such deception can occur in two main ways: hallucination and manipulation. For example, citations to academic publications can look so natural, even with deep voice and face fakes of authoritative figures speaking false scientific claims on social media. Evidence shows that hallucination decreases as models grow (Brown et al. (2020)). Specifically, it is reduced by fine-tuning (Ouyang et al. (2022)) not only by focusing on training AI to recognize erroneous outcomes but also, and even more promising, by training/fine-tuning the model based on its thought process, not only its outcome (Lightman et al.

(2023)).

Furthermore, attention manipulation is a serious global social risk, given the rapid spread of misinformation on social networks (Lazer et al., 2018; Pennycook et al., 2021). Manipulation of human attention compounds this challenge. Regardless of its veracity, directing attention to specific information can significantly influence decision-making, increasing the need for rigorous scrutiny of AI-generated content. Such manipulation is made more accessible by directing attention to a particular item, regardless of its quality, and increases its likelihood of selection, as shown by Gossner et al. (2021). Hence, focusing finding underlines the necessity for a relentless focus on information quality and credibility, particularly amidst the surge of AI-generated content seems a pressing societal issue, in particular when the quality of the propaganda is about increasing with what we name scientific troll farms, where agents strategically rely on sophisticated scientific hallucinations to serve specific manipulative goals. Research efforts have gravitated toward the use of artificial intelligence to automate fact checking (Guo et al., 2022). However, detecting and mitigating AI-generated misinformation remains a daunting task not only due to the ease of its creation (Gupta et al., 2022) but also because its propagation at high frequency and low cost pose clear challenges to standard slow and costly fact-checking methods (Goldstein et al., 2023). To reimagine fact- or authenticity-checking methods, we could rely on experimentation in the economics of networks (Jackson, 2009) to strategically allocate GPU-limited energy resources of generative AI to accurately predict and anticipate sources of misinformation given different treatments, for instance, the reputation of the information emitter, the process used to produce information, which could all be encapsulated in a blockchain, serving as an authentificator of expertise authenticity.

The emergence of this new technology raises new challenges in the education sciences regarding which tools future economists should learn. Prompt engineering, which is finding the optimal way to input commands into generative AI, is a rapidly growing industry sector, as the quality of an AI's output is susceptible to its prompts. which makes discovering the best prompt technique an intensive task. The quality of AI's output is highly sensitive to the prompts fed into it, which makes discovering the best prompt techniques an intensive task. However, to our knowledge, this investigation has been unsystematic, an oversight that could be solved with the behavioral and experimentalist economist toolkit, based mainly on the roles of nudges (Thaler, 2018) to build an effective human-AI interface or "UAI" (User-AI interface) and "UAIX" (User-AI experience). Furthermore, to maintain trust and standards of replicability, interactions during the knowledge production between the researcher and the machine could

be recorded and attached to submissions in the appendix.

Vigilance in identifying biases during model training and data analysis is also essential (Luca et al., 2016; Kleinberg et al., 2018). Research on the accuracy-fairness trade-off of algorithms has led some to claim that the optimal approach could not directly tamper with algorithmic bias (Xian et al., 2023), but factor it in during later analysis (Kleinberg et al., 2018). Such adjustments still require researchers and consumers to understand the possible biases involved, motivating the need for detail and transparency in the training, fine-tuning, and use of models. Researchers may consider using these models as supportive tools rather than a complete replacement for human expertise.

One final negative externality is that the broader use of generative AI could challenge our conception of creativity and homogenize too much thought by relying only on standardized prompts when interacting with AI. In its worst light, this new technology could, in principle, create research drones by taking the art and creativity out of the research and thought process, leading to decreased research quality. This would undoubtedly lead to lost opportunities for new wisdom, thought, hypotheses, and scholarship needed in the face of every new societal challenge. We should recognize this trade-off and continue to reward such creativity in the marketplace for ideas; without incentives, significant contributions that come about via critical thinking, creativity, and out-of-the-box ideas might be sacrificial lambs to this sophisticated standardizing of knowledge production.

An essential role for LLMs is to generate standardized documentation that follows best practices and established guidelines for open science norms. Consistent formatting and content reduce barriers to replication by human agents or generative AIs. LLMs can analyze the scientific literature, helping researchers identify relevant studies for replication. Researchers can replicate essential and influential studies by prioritizing novelty, impact, or methodological rigor, which greatly increases our knowledge creation (Maniadis et al. (2014) for the inferential power of replications). Trained on specific and small datasets, we could imagine LLMs predicting whether a submitted paper is likely replicable or even helping to replicate it before accepting it as publication rather than letting the replication as a hoped-for positive externality performed by researchers after publication, which could contribute to the emerging forecasting markets in social sciences (DellaVigna et al., 2020). Therefore, in addition to working to better align professional incentives with transparent scientific behavior, a concrete and fully operational institutional change through AI engineering assistance at the journal level can make a difference in the desired change in scientific culture.

Such fine-tuned models can also facilitate collaboration by managing collaborative replication projects by generating project management tools, coordinating communication, and maintaining version control for shared documentation. They have successfully coordinated large groups for communication (Small et al., 2023) and could present an opportunity for more extensive collaborations between academics. These opportunities help to advance the credibility revolution, which has recently taken on a more critical role in the social sciences (Jennions and Møller, 2002; Nosek et al., 2012; Bettis, 2012; Dreber et al., 2015; Butera et al., 2020; Dreber and Johannesson, 2023). By supporting the peer review process with standardized guidelines, these models can ensure that published studies adhere to the highest standards of scientific integrity. They can also develop training materials, online courses, or educational workshops for conducting and reporting replication studies. Making these resources widely available demonstrates to researchers the importance of replication and transparency in scientific research. Additionally, they can facilitate communication between researchers, editors, and other stakeholders by generating standardized correspondence templates and streamlining the submission and review process.

These opportunities extend beyond academia, fostering a standardized scientific culture of experimentation in technology, artificial intelligence firms, and government agencies. It has already been argued that ML could help with pre-registration, creating a flexible compromise between the ideal open science preregistration requirements (such as the AEAs RCT registry) for applied experimental microeconomic work and the current exploratory nature of the work by suggesting additional variables of interest (Ludwig et al., 2019).

The potential of generative AI to foster a culture of systematic experimentation in technology companies can significantly mitigate associated labor expenses related to human expertise (Berg et al., 2023). A rising trend of technology corporations actively recruiting Ph.D. economists demonstrates their pivotal role in resolving multifaceted business challenges (Athey and Luca, 2019). These economists engaged in strategic decision making and design choices navigate various issues, including pricing, auctions, matching, market design, consumer behavior, product design, and strategic decision making. They tackle issues relevant to management by employing company-specific data, often working in business-centric roles. Illustrative of this trend are tech giants like Microsoft and Amazon. Microsoft's business-oriented chief economist leads a team actively recruiting Ph.D. economists to address diverse issues ranging from cloud computing to search advertising. Similarly, Amazon employs economists to solve business-specific challenges in its multiple divisions, including e-commerce

platforms, digital content, and platforms designed to evaluate changes and innovations.

The increasing prominence of economists in technology companies underscores their crucial role in creating a culture of experimentation. They draw on their expertise to conduct changes and innovations evaluations. This process echoes the pioneering work of Paul Milgrom, Al Roth and Robert Wilson and in auctions (Wilson, 2020). His groundbreaking efforts blended novel theoretical insights with empirical work and experiments to address real-world problems. With the advent of foundation models, it is now possible for technology corporations to instill a comprehensive culture of experimentation across departments. This approach echoes the rigor and originality of academia, paving the way for even more business decisions to be grounded in scientific principles. Building such a culture of experimentation within government agencies requires a more systematic approach to policy making. This approach relies on a continuous low-cost cycle of tests, trials, and pilots to explore policy options, evaluate their impacts, and make informed, data-driven decisions.

Finally, by generating standardized documentation for experiments, LLMs can promote transparency, build public trust, and contribute to technology literacy for different stakeholders. For example, LLMs can help create educational materials and tools that instruct government personnel about experimental methods, data analysis, and evidence-based policymaking. This step is critical in fostering a culture that values and understands the importance of experimentation. Effective communication is crucial for accepting and institutionalizing an experimental culture. Incorporating LLMs into policy development can help governmental agencies promote systematic experimentation, fostering a culture of evidence-based policymaking. However, it remains vital to ensure their ethical use and strike a balance between automated insights and human expertise.

References

Acemoglu, Daron and Simon Johnson (2023), *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. Hachette UK.

Athey, Susan (2015), "Machine learning and causal inference for policy evaluation." In *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining*.

Athey, Susan and Guido W. Imbens (2019), "Machine learning methods that economists should know about." *Annual Review of Economics*, 11, 685–725.

Athey, Susan and Michael Luca (2019), "Economists (and economics) in tech companies." *Journal of Economic Perspectives*, 33, 209–30.

Berg, Justin M., Manav Raj, and Robert Seamans (2023), "Capturing value from artificial intelligence." *Academy of Management Discoveries*.

Bettis, Richard A. (2012), "The search for asterisks: Compromised statistical tests and flawed theories." *Strategic Management Journal*, 33, 108–113.

Boiko, Daniil A., Robert MacKnight, and Gabe Gomes (2023), "Emergent autonomous scientific research capabilities of large language models." *arXiv preprint arXiv:2304.05332*.

Bommasani, Rishi et al. (2022), "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258*.

Brown, T. et al. (2020), "Language models are few-shot learners." In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), volume 33, 1877–1901, Curran Associates, Inc.

Brynjolfsson, Erik, Avinash Collis, and Felix Eggers (2019a), "Using massive online choice experiments to measure changes in well-being." *Proceedings of the National Academy of Sciences*, 116, 7250–7255.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond (2023), "Generative ai at work." Technical Report w31161, National Bureau of Economic Research.

Brynjolfsson, Erik and ANDREW Mcafee (2017), "Artificial intelligence, for real." *Harvard business review*, 1, 1–31.

Brynjolfsson, Erik et al. (2019b), "Gdp-b: Accounting for the value of new and free goods in the digital economy." Technical Report w25695, National Bureau of Economic Research.

Bubeck, Sébastien et al. (2023), "Sparks of artificial general intelligence: Early experiments with gpt-4." *arXiv preprint arXiv:2303.12712*.

Butera, Luigi, Philip J. Grossman, Daniel Houser, John A. List, and Marie-Claire Villeval (2020), "A new mechanism to alleviate the crises of confidence in science with an application to the public goods game." Technical Report w26801, National Bureau of Economic Research.

Camerer, Colin F. (2018), *Artificial intelligence and behavioral economics*, 587–608. University of Chicago Press.

Charness, Gary, Guillaume R. Frechette, and John H. Kagel (2004), "How robust is laboratory gift exchange?" *Experimental Economics*, 7, 189–205.

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016), "otree—an open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Chen, Mark et al. (2021), "Evaluating large language models trained on code."

Davies, Alex et al. (2021), "Advancing mathematics by guiding human intuition with ai." *Nature*, 600, 70–74.

Deaton, Angus and Nancy Cartwright (2018), "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine*, 210, 2–21.

DellaVigna, Stefano, Nicholas Otis, and Eva Vivalt (2020), "Forecasting the results of experiments: Piloting an elicitation strategy." In *AEA Papers and Proceedings*, volume 110, 75–79, American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203.

Dreber, Anna and Magnus Johannesson (2023), "A framework for evaluating reproducibility and replicability in economics." *Available at SSRN 4458153*.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson (2015), "Using prediction markets to estimate the reproducibility of scientific research." *Proceedings of the National Academy of Sciences*, 112, 15343–15347.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock (2023), "Gpts are gpts: An early look at the labor market impact potential of large language models."

Frank, Morgan R, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, et al.

(2019), "Toward understanding the impact of artificial intelligence on labor." *Proceedings of the National Academy of Sciences*, 116, 6531–6539.

Fréchette, Guillaume R., Kim Sarnoff, and Leeat Yariv (2022), "Experimental economics: Past and future." *Annual Review of Economics*, 14, 777–794.

Ganguli, Deep et al. (2022), "Predictability and surprise in large generative models." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Gillen, Ben, Erik Snowberg, and Leeat Yariv (2019), "Experimenting with measurement error: Techniques with applications to the caltech cohort study." *Journal of Political Economy*, 127, 1826–1863.

Goldstein, Josh A. et al. (2023), "Generative language models and automated influence operations: Emerging threats and potential mitigations." *arXiv preprint arXiv:2301.04246*.

Gossner, Olivier, Jakub Steiner, and Colin Stewart (2021), "Attention please!" *Econometrica*, 89, 1717–1751, URL https://doi.org/10.3982/ECTA17173.

Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos (2022), "A survey on automated fact-checking." *Transactions of the Association for Computational Linguistics*, 10, 178–206.

Gupta, Ankur, Neeraj Kumar, Purnendu Prabhat, Rajesh Gupta, Sudeep Tanwar, Gulshan Sharma, Pitshou N. Bokoro, and Ravi Sharma (2022), "Combating fake news: Stakeholder interventions and potential solutions." *Ieee Access*, 10, 78268–78289.

Hofman, Jake M. et al. (2021), "Integrating explanation and prediction in computational social science." *Nature*, 595, 181–188.

Horton, John J. (2023), "Large language models as simulated economic agents: What can we learn from homo silicus?" *arXiv preprint arXiv:2301.07543*.

Ioannidis, John PA (2005), "Why most published research findings are false." *PLoS Medicine*, 2, e124.

Jabarian, Brian and Elia Sartori (2020), "Critical thinking and storytelling."

Jackson, Matthew O. (2009), "Networks and economic behavior." *Annu. Rev. Econ.*, 1, 489–511.

Jennions, Michael D and Anders P Møller (2002), "Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution." *Proceedings. Biological sciences*, 269, 43–48.

Kirchner, Jan Hendrik, Lama Ahmad, Scott Aaronson, and Jan Leike (2023), "New ai classifier for indicating ai-written text." *OpenAI*.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan (2018), "Algorithmic fairness." *AEA Papers and Proceedings*, 108, 22–27.

Korinek, Anton (2023), "Language models and cognitive automation for economic research." *National Bureau of Economic Research*. No. w30957.

Lazer, David MJ et al. (2018), "The science of fake news." *Science*, 359, 1094–1096.

Lightman, Hunter et al. (2023), "Let's verify step by step." *arXiv preprint arXiv:2305.20050*.

List, JA (2023), *A course in experimental economics*. University of Chicago Press.

Luca, Michael, Jon Kleinberg, and Sendhil Mullainathan (2016), "Algorithms need managers, too." *Harvard Business Review*, 94, 96–101.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess (2019), "Augmenting preanalysis plans with machine learning." *AEA Papers and Proceedings*, 109, 71–76.

Maniadis, Zacharias, Fabio Tufano, and John A. List (2014), "One swallow doesn't make a summer: new evidence on anchoring effects." *American Economic Review*, 104, 277–290.

Manning, Sam, Pamela Mishkin, Gillian Hadfield, Tyna Eloundou, and Emily Eisner (2022), "A research agenda for assessing the economic impacts of code generation models."

Molnar, Andras (2019), "Smartriqs: A simple method allowing real-time respondent interaction in qualtrics surveys." *Journal of Behavioral and Experimental Finance*, 22, 161–169, URL https://doi.org/10.1016/j.jbef.2019.03.005.

Moonesinghe, Ramal, Muin J. Khoury, and A. Cecile J. W. Janssens (2007), "Most published research findings are false—but a little replication goes a long way." *PLoS medicine*, 4, e28.

Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl (2012), "Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability." *Perspectives on Psychological Science*, 7, 615–631.

Noy, Shakked and Whitney Zhang (2023), "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381, 187–192.

Ouyang, Long et al. (2022), "Training language models to follow instructions with human feedback." In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744.

Pennycook, Gordon et al. (2021), "Shifting attention to accuracy can reduce misinformation online." *Nature*, 592, 590–595.

Roth, Alvin E and Robert B Wilson (2019), "How market design emerged from game theory: A mutual interview." *Journal of Economic Perspectives*, 33, 118–143.

Saunders, William et al. (2022), "Self-critiquing models for assisting human evaluators." *arXiv preprint arXiv:2206.05802*.

Small, Christopher T. et al. (2023), "Opportunities and risks of llms for scalable deliberation with polis." *arXiv preprint arXiv:2306.11932*.

Snowberg, Erik and Leeat Yariv (2021), "Testing the waters: Behavior across participant pools." *American Economic Review*, 111, 687–719.

Thaler, Richard H. (2018), "Nudge, not sludge." *Science*, 361, 431–431.

Wang, X., Z. Wu, W. Huang, Y. Wei, Z. Huang, M. Xu, and W. Chen (2023), "Vis+ai: Integrating visualization with artificial intelligence for efficient data analysis." *Frontiers of Computer Science*, 17, 176709.

Wilson, Robert (1992), "Strategic analysis of auctions." *Handbook of game theory with economic applications*, 1, 227–279.

Wilson, Robert (2020), "Strategic analysis of auction markets." Technical report, Nobel Prize Committee.

Xian, Ruicheng, Lang Yin, and Han Zhao (2023), "Fair and optimal classification via post-processing."

Zhang, Daniel, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. (2021), "The ai index 2021 annual report." *arXiv preprint arXiv:2103.06312*.