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

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

Generation Next: Experimentation with AI
Gary Charness, Brian Jabarian, and John A. List
NBER Working Paper No. 31679
September 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 discuss the implementation of experiments using LLMs, focusing on enhancing causal inference by creating consistent experiences, improving comprehension of instructions, and monitoring participant engagement in real time. Third, we highlight how LLMs can help analyze experimental data, including pre-processing, data cleaning, and other analytical tasks while helping reviewers and replicators investigate studies. Each of these tasks improves the probability of reporting accurate findings.

Gary Charness
Department of Economics
University of California, Santa Barbara
2127 North Hall
Santa Barbara, CA 93106-9210
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 generative capacity for creating original content and their status as generative Artificial Intelligence (AI) (Bubeck et al., 2023). Even with their recent emergence and already-proven advancements, researchers believe that the full extent of the rapid effects of generative AI in science, policy, and society remains to be experienced (Bommasani et al., 2022), including in economics (Acemoglu and Johnson 2023, Korinek 2023). Hence, although machine-learning algorithms have improved analysis methods for causal inference in economics (Athey 2019), these new models may also radically improve all critical areas of scientific knowledge production. A natural venue in economics is to generate data for causal inference in experimental settings. And, 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 generating causal insights using online experiments will continue to increase (Fréchette et al., 2022).

However, one key feature of online experiments tempering the optimism of even the 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 identification assumption underlying 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 experimental instructions is often tricky in an online experiment, since people cannot usually ask questions and receive live responses. While one remedy might involve incorporating real-time human support, this would require satisfying either having a sizable skilled labor force to accommodate simultaneous questions from many participants or providing extended availability to cover the protracted timelines with online experiments.

LLMs can be fine-tuned as chat assistants to simulate sophisticated human interactions while reducing labor costs. Given the inherent scalability and versatility, such integration could become standard practice for future online experiments, revolutionizing the field and fostering unprecedented advancements across online experiments and surveys. In addition, this approach requires only minimal coding knowledge and is compatible with many online experimental 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 when minimizing the burden on subjects by maintaining participant focus and engagement.

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 (Chen et al., 2023). Using this technology, we can create more immersive online experiences, facilitate real-time monitoring of participant engagement, and improve the quality and replicability of experiments. Furthermore, its use can promote open science opportunities, fostering increased collaboration.

We often refers to LLMs and what they can do, having foundational and fine-tuned models in mind under the umbrella of generative AI (Bommasani et al., 2022). However, users need to do more than just copy our suggested directions and details of the experiment into ChatGPT to obtain satisfactory results. Generative AI is a fast-moving technology that susceptible to inputs and produces unpredictable outputs (Ganguli et al., 2022). And working out which inputs lead to the most desired outputs is a growing section 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 ex-post (Davies et al., 2021) or could launch A/B tests to determine which prompt is the most effective in delivering the expected result.

In their best light, we envision these language models as becoming the wise sage always available at the experimentalist's beck and call. Within this framework, we explore their implementation 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, pre-processing, and cleaning, while Section 4 considers data analysis. The final section discusses the broader risks and benefits of generative AI in behavioral and experimental economics, as well as implications for open science and for scaling a culture of experimentation in business and policy-making. We offer some speculative pointers on managing these.

2. Designing Experiments with LLMs

By enhancing researcher productivity, LLMs free resources during the design phase, broadening the scope of potential research questions and the focus dedicated to ensuring validity. LLMs can generate ideas for research, 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, LLMs can assist in selecting the ideal experimental setting – 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. In particular, they can ensure balanced and comparable groups using random assignment, stratification, or matching. Their extensive training data allow 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 could also guide ethical considerations, such as deception, and potential harm to participants, helping researchers design rigorous and ethically-

sound studies.

Critically, LLMs can write clear, concise instructions and comprehension checks. Researchers can obtain more meaningful data by ensuring that participants understand the experimental setup and tasks. Models such as GPT-4, trained on billions of data points, can tailor instructions' language, tone, and complexity to participants with varying language proficiency, education, or familiarity with economic concepts. They could even generate relevant examples, such as textual descriptions, hypothetical scenarios, or visual representations of economic concepts or analogies to illustrate key issues. By having LLMs generate multiple versions of instructions and comprehension checks and iteratively providing feedback from researchers or participants, 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., 2023). In addition, they can adapt the generated code to specific experimental platforms, ensuring seamless integration with the requirements of a platform.

In the same vein, they can reason through experiments, detect coding errors in the design, and even simulate participants (Horton, 2023). By simulating demographic characteristics, backgrounds, and language proficiency, these virtual participants can 'interact' with the experiment materials and highlight confusion or misunderstanding related to the wording of the instructions or logical errors in the experimental code, which may be innocuous to researchers but crucial to the results (Charness et al., 2004). By investigating how simulated participants interact, researchers can examine these topics in a controlled experimental setting while maintaining a high degree of ecological and external validity. Researchers can generate multiple versions of instructions and comprehension checks by providing feedback from researchers or participants, with LLMs assisting in critiquing human- or AI-generated concise instructions against lengthy but precise benchmarks provided by the researcher (Saunders et al., 2022). This iterative approach enables researchers to refine the instructions, optimizing them for clarity and effectiveness.

Furthermore, LLMs can help researchers create replication packages, 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 help develop other documentation (e.g., IRB submissions) to more effectively explain the experimental design. If given access through plug-ins, 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, Sargent 2023). Similarly, LLMs can be crucial in standardizing approaches since the literature suggests that design flexibility and associated experimental decisions are inversely related to actual research findings (see Ionannides, 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, such as Qualtrics (Molnar 2019) and oTree (Chen et al. 2016).

Predictive modeling is social scientists' third area of interest (Hofman et al., 2021), where generative AI can make significant contributions. It can formulate and test policy hypotheses by analyzing vast amounts of data and determining the plausibility of policy assumptions. By developing and refining predictive models, LLMs enable governmental and nonprofit agencies to simulate the effects of policy experiments, anticipating the consequences of policy changes before full implementation and thus informing iterative policy interventions design and adjustments. Similar capabilities could encourage a more experimental, 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, this issue is addressed when LLMs supervise the experiment.

Finally, any bias LLMs introduce can be inexpensively and rigorously tested and before implementation, unlike with human experts.

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, monitor data quality in live time, and facilitate follow-up surveys.

Chatbots demonstrate significant potential to provide detailed, instant responses to inquiries. Recent evidence from 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 have questions or need clarification of the instructions. In its most scalable version, perhaps fully-autonomous chatbots could provide 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 enhance 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 to each participant's communication and comprehension style during the experiment. This helps ensure that participants understand the experimental instructions before performing tasks, a necessary step to ensure compliance and avoid construct validity concerns.

Additionally, providing all necessary guidance within the chat interface maintains participant focus and engagement, preventing participants from feeling overwhelmed or

confused and minimizing distractions 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 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 plug-ins, automatically implement different techniques already done through JavaScript: tracking participant browser activity by opening new tabs, switching between windows, or spending excessive time away from the experiment (Jabarian and Sartori, 2023). With systemic monitoring, 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' chance of forgetting critical information. Additionally, chat assistants could engage participants in dynamic and interactive conversations. This real-time interaction facilitates higher-quality data: 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 participation. 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 an experiment's design 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 subsequent tasks accordingly, allowing better identification of the type of participants. This maintains engagement and accurately measures the participant's abilities. Similarly, chatbots could prioritize 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 when appropriate. Chat assistants can simulate social interactions such as negotiations or group decision-making to study complex social phenomena. Thus, researchers can examine these topics in a controlled experimental setting while maintaining a high degree of ecological and external validity.

Finally, 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, 2012). 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 augment the analysis of experimental data in two ways: First, by automating data-analysis tasks such as sanitization, examining relationships within data, and streamlining data visualization using the Code Interpreter on ChatGPT. Second, and less obviously, by creating data unexplored so far in standard economic experiments. Leveraging natural-language processing (NLP) techniques with live-chat logs from experiments can yield insights into participant behavior, uncertainty, and cognitive processes. Such insights open a window to scrutinize new variables for statistical tests and identify factors that could influence results. Understanding these variables can illuminate the correlation between participant behavior and experimental results, offering a more nuanced comprehension 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 undertake content analysis — identifying and categorizing frequently-expressed concerns from participants, analyzing emotions conveyed, and gauging the efficacy of instructions, responses, and interactions. Models of this nature can also pinpoint areas of confusion. This would aid in enhancing experimental designs, instructions, and training protocols for chatbots. 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.

Regarding automating data-analysis tasks, specific LLMs such as Code Interpreter can help at different stages of the knowledge-production chain by conducting statistical tests, developing econometric models, determining causal relationships, and performing robustness checks using state-of-the-art techniques. Automating these tasks has a two-fold benefit. On one hand, it saves researchers time, which can be allocated to other tasks in knowledge production. On the other hand, it maximizes research flexibility across all tasks, bolstering the likelihood that reported findings represent genuine associations (Ionannides, 2005). Besides, they can generate data visualizations in concert with other features. This aids researchers in both understanding their results and communicating their findings effectively.

Second, after the analysis is complete, LLMs can aid researchers and the broader scientific community in comparing the finished work and pre-registration plans. In particular, these tools could pinpoint and highlight significant divergences, including the unexpected introduction of new variables, the omission of pre-determined variables, or modifications in the specified data-acquisition methods. Furthermore, these models can be fine-tuned to distinguish between confirmatory and exploratory analysis. Confirmatory analysis aims to validate pre-determined hypotheses. In contrast, exploratory analysis, devoid of a rigid plan, allows a more

flexible approach to data interpretation. These models can meticulously scrutinize the manuscript for sections indicating a diversion from the pre-registered schema towards exploratory analysis, useful not only for the authors, but also for referees and editors.

In the same vein, while these models may still struggle to indentify AI versus human-generated text, they are becoming rapidly proficient in accurately detecting AI-generated code, anomalies, or red flags within code analysis. 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 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 misalignment between the quantitative findings and their qualitative exposition or significant findings that are overlooked could be flagged by LLMs for further examination.

Generative AI also offers significant opportunities for peer review, replication, and dissemination of research. Tasks, such as comparison with pre-registration plans, checking for tampering, and analysis of code and supplemental text, that would be highly time intensive for human researchers become far less so.

Generative AIs can cross-compare claims in the body of a paper with the code, ensuring 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 summarize appendencies, allowing reviewers and replicators to quickly see if their concerns are addressed. Given the experimental setup and treatment, it can check whether appropriate tests for main results and robustness have been carried out. In the limit, we even envisage simulated replications using existing code and information in the paper 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 LLMs are unlikely to internalize. However, this idea could hold promise with the ability of independent participant simulation and the capacity for many simulations. These various abilities can boost the speed of review and rate of replication, two common concerns in the field, and benefit research efficiency.

5. Discussion: Risks and Opportunities

Using LLMs in economic research may pose risks (Bommasani et al., 2021), including intellectual-property (IP) concerns, digital-privacy issues, user deception, scientific fraud by fabricating data or strategies to hide data manipulation, and challenging creativity by excessively-homogenizing the human-AI interface. For example, generative AI can effectively be unintentional plagiarism or copyright infringement; 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 a more transparent process. Beyond IP concerns, other potential issues remain.

First, the vast amounts of data these language models process can create privacy concerns (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. Citations to academic publications can look so natural and even more so with fakes of authoritative figures speaking false scientific claims on social media. Evidence shows that this shrinks as models grow (Brown et al., 2020). Specifically, it is lessened (Ouyang et al., 2022) by not only focusing on training AI to recognize errors but also by training/fine-tuning the model based on its thought process, not just its outcome (Lightman et al., 2023).

Given the rapid spread of misinformation on social media (Lazer et al., 2018; Pennycook et al., 2021), attention manipulation is a severe risk. The manipulation of human attention compounds this challenge. Regardless of its veracity, directing attention toward certain information can significantly influence decision-making, heightening the necessity for rigorous scrutiny of AI-generated content. Such manipulation is facilitated by the fact that directing attention to a particular item increases its selection likelihood, regardless of its quality (Gossner et al., 2023). Hence, this underlines the necessity for a relentless focus on information quality and credibility, particularly amidst the surge of AI-generated content that seems a pressing societal issue; the quality of the propaganda is increasing with "scientific troll farms", where agents strategically rely on sophisticated scientific fakes to serve specific manipulative goals. Research efforts have gravitated toward using AI to automate fact-checking (Guo et al., 2021). 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 low-cost propagation at high frequency poses clear challenges to standard slow and costly fact-checking methods (Goldstein et al., 2023). To re-imagine 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 emergence of this new technology raises new challenges in education sciences regarding which tools future economists should learn. Prompt engineering (how to use existing string-input generative AI) is a quickly-growing section of the industry, since the quality of AI's output is highly sensitive to the prompts fed into it; this makes discovering the best prompt technique an intensive task. The output quality is highly sensitive to the prompts fed into it, making discovering the best prompt techniques an intensive task. However, to our knowledge, this investigation has been unsystematic, an oversight solvable using the behavioral- and experimental-economist toolkit, mainly relying 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, interactions during the knowledge production between the researcher and the machine could be recorded while adhering to standards of replicability.

Vigilance in identifying biases during model training and data analysis is also essential (Luca et al., 2016. Kleinberg et al., 2018). The accuracy-fairness tradeoff of algorithms has been researched (Lang et al., 2023), leading some to claim that the optimal approach could not directly tamper with algorithmic bias but factor it in during later analysis (Rambachan et al., 2020). Such adjustments still require producers and consumers of research to understand the possible biases involved, motivating the need for detail and transparency in training, fine-tuning, and use of any 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 affect research by homogenizing thought, relying only on standardized prompts when interacting with AI. This new technology could potentially 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.

One essential role of LLMs is generating standardized documentation, which follows best practices and established guidelines for open-science norms. Consistent formats and content reduce barriers to replication by people or generative AIs. They 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 increases our knowledge creation immensely (see 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 publishing it rather than letting the replication as a hoped-for positive externality later performed by other researchers. Hence, in addition to working on better aligning professional incentives with transparent scientific behavior, a concrete and fully-operational institutional change through AI-engineering assistance could make a difference in a desired change in the scientific culture.

Such fine-tuned models can also facilitate collaboration by managing collaborative replication projects through 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. These opportunities aid in the “credibility revolution”, which has recently taken on a more critical role in the social sciences (see, e.g., Jennions and Moller (2002); Moonesinghe et al. (2007); 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 develop training materials, online courses, or educational workshops for conducting replication studies. Making these resources widely available demonstrates to researchers the importance of replication and transparency in scientific research. Additionally, this can facilitate communication between researchers, editors, and others by generating standardized correspondence templates and streamlining the review process.

These opportunities can trickle down beyond the academic world, helping to standardize a scientific culture of experimentation in technology, artificial-intelligence companies, and government agencies. It has already been argued that ML could help with pre-registration, creating a flexible compromise between the ideal open-science pre-registration requirements for experimental work and the current exploratory nature of some research by suggesting additional variables of interest (Ludwig et al., 2019).

Generative AI might foster a culture of systematic experimentation in technology firms that could 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 these individuals' pivotal roles in resolving multi-faceted business challenges (Athey et al., 2019). These economists navigate various issues, including pricing, auctions, matching, market design, consumer behavior, product design, and strategic decision-making. They tackle managerially-relevant issues by employing company-specific data. Illustrative of this trend are tech giants like Microsoft and Amazon. Microsoft's business-oriented chief economist leads a team recruiting Ph.D. economists to address issues ranging from cloud computing to search advertising. Similarly, Amazon employs economists to resolve business-specific challenges across its multiple divisions, including e-commerce platforms, digital content, and platforms designed to evaluate innovations.

The rising prominence of economists in technology firms underscores their crucial role in creating a culture of experimentation. They draw upon their expertise to conduct evaluations of changes and innovations. This process echoes the Bob Wilson’s pioneering work in auctions during the 1970s (Wilson, 1969 and 1977). His groundbreaking efforts blended novel theoretical insights with empirical work and experiments to address real-world problems. With the advent of foundation models, technology corporations could now instill a comprehensive culture of experimentation. This approach echoes the rigors and originality of academia, paving the way to ground business decisions more on scientific principles. Building such a culture of experimentation within governmental agencies involves a more systematic approach to policymaking. 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 of experiments, LLMs can promote transparency, build public trust and contribute to technology literacy for different stakeholders. This step is critical in fostering a culture that values and understands the importance of experimentation. This 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 these tools' ethical use and to strike a balance between automated insights and human expertise.

6. Conclusion

We have explored the diverse benefits of AI across all stages of experimental work: design, execution, and analysis. LLMs can profoundly impact experimental science if used carefully with appropriate scientific governance and recognition of potential negative externalities. By standardizing templates, guidelines, and other resources to help harmonize good research practices among scientists, LLMs will ultimately be a critical advance that enhances science by promoting greater scientific research transparency, rigor, and reproducibility. In the best light, knowledge creation will rapidly advance with LLMs.

We also explore these advancements' potential impacts on academia, policy, and industry. Generative AI's societal application is still confined to speculative or non-standard experimental work. The significance of the societal opportunities posed by generative AI warrants a more systematic approach. This is where behavioral theory and experimental economics can, in turn, contribute to refining AI by improving both research related to the societal implications of upcoming transformative technologies and research related to improving these technologies by relying on the culture of theory and experimentation laid out in our last section. This more global approach to technology shifts can provide better guidance on effectively combining AI and humans in diverse policy and industry sectors.

References

1. Acemoglu, Daron and Johnson, Simon. Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity. Hachette UK, 2023.
2. Athey, Susan. "Machine learning and causal inference for policy evaluation." Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining. 2015.
3. Athey, Susan, and Guido W. Imbens. "Machine learning methods that economists should know about." Annual Review of Economics 11 (2019): 685-725.
4. Athey, Susan, and Michael Luca. "Economists (and economics) in tech companies." Journal of Economic Perspectives 33.1 (2019): 209-30. DOI: 10.1257/jep.33.1.209
5. Berg, Justin M., Manav Raj, Robert Seamans "Capturing Value from Artificial Intelligence," Academy of Management Discoveries (2023)
6. Bettis, Richard A. "The search for asterisks: Compromised statistical tests and flawed theories." Strategic Management Journal 33, no. 1 (2012): 108-113.
7. Boiko, Daniil A., Robert MacKnight, and Gabe Gomes. "Emergent autonomous scientific research capabilities of large language models." arXiv preprint arXiv:2304.05332 (2023).
8. Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2022).

9. Brown, T. et al. Language Models are Few-Shot Learners. in Advances in Neural Information Processing Systems (eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) vol. 33 1877–1901 (Curran Associates, Inc., 2020).
10. Brynjolfsson, Erik, Avinash Collis, and Felix Eggers. "Using massive online choice experiments to measure changes in well-being." Proceedings of the National Academy of Sciences 116.15 (2019a): 7250-7255.
11. Brynjolfsson, Erik, et al. GDP-B: Accounting for the value of new and free goods in the digital economy. No. w25695. National Bureau of Economic Research, 2019b.
12. Snowberg, Erik, and Leeat Yariv. 2021. "Testing the Waters: Behavior across Participant Pools." *American Economic Review*, 111 (2): 687-719.
13. Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. Generative AI at work. No. w31161. National Bureau of Economic Research, 2023.
14. Bubeck, Sébastien, et al. "Sparks of artificial general intelligence: Early experiments with GPT-4." arXiv preprint arXiv:2303.12712 (2023).
15. Butera, Luigi, Philip J. Grossman, Daniel Houser, John A. List, and Marie-Claire Villeval. A new mechanism to alleviate the crises of confidence in science with an application to the public goods game. No. w26801. National Bureau of Economic Research, 2020.
16. Camerer, Colin F. "Artificial intelligence and behavioral economics." In Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, eds., The economics of artificial intelligence: An agenda, pp. 587-608. University of Chicago Press, 2018.
17. Charness, Gary, Guillaume R. Frechette, and John H. Kagel. "How robust is laboratory gift exchange?." Experimental Economics 7 (2004): 189-205.
18. Chen, Daniel L., Martin Schonger, and Chris Wickens. "oTree—An open-source platform for laboratory, online, and field experiments." *Journal of Behavioral and Experimental Finance* 9 (2016): 88-97.
19. Davies, Alex, et al. "Advancing mathematics by guiding human intuition with AI." *Nature* 600.7887 (2021): 70-74.
20. Deaton, Angus, and Nancy Cartwright. "Understanding and misunderstanding randomized controlled trials." *Social Science & Medicine* 210 (2018): 2-21.
21. Dreber, Anna, and Magnus Johannesson. "A Framework for Evaluating Reproducibility and Replicability in Economics." Available at SSRN 4458153 (2023).
22. Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. "Using prediction markets to estimate the reproducibility of scientific research." Proceedings of the National Academy of Sciences 112, no. 50 (2015): 15343-15347.
23. Fréchette, Guillaume R., Kim Sarnoff, and Leeat Yariv. "Experimental economics: Past and future." Annual Review of Economics 14 (2022): 777-794.

24. Ganguli, Deep, et al. "Predictability and surprise in large generative models." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.
25. Gillen, Ben, Erik Snowberg, and Leeat Yariv. "Experimenting with measurement error: Techniques with applications to the Caltech cohort study." Journal of Political Economy 127, no. 4 (2019): 1826-1863.
26. Goldstein, Josh A., et al. "Generative language models and automated influence operations: Emerging threats and potential mitigations." *arXiv preprint arXiv:2301.04246* (2023).
27. Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos. "A survey on automated fact-checking." *Transactions of the Association for Computational Linguistics* 10 (2022): 178-206.
28. Gupta, Ankur, Neeraj Kumar, Purnendu Prabhat, Rajesh Gupta, Sudeep Tanwar, Gulshan Sharma, Pitshou N. Bokoro, and Ravi Sharma. "Combating fake news: Stakeholder interventions and potential solutions." *Ieee Access* 10 (2022): 78268-78289.
29. Hofman, Jake M., et al. "Integrating explanation and prediction in computational social science." Nature 595.7866 (2021): 181-188. https://doi.org/10.1038/s41586-021-03659-0
30. Horton, John J. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?." arXiv preprint arXiv:2301.07543 (2023).
31. Ioannidis, John PA. "Why most published research findings are false." PLoS Medicine 2, no. 8 (2005): e124.
32. Jabarian, Brian, and Elia Sartori. "Critical Thinking and Storytelling:." arXiv preprint arXiv:2303.16422 (2023).
33. Kirchner, Jan Hendrik, Lama Ahmad, Scott Aaronson, and Jan Leike. "New AI classifier for indicating AI-written text." *OpenAI* (2023).
34. Jackson, Matthew O. "Networks and economic behavior." Annu. Rev. Econ. 1.1 (2009): 489-511.
35. Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. "Algorithmic Fairness." AEA Papers and Proceedings, 108: 22-27.
36. Korinek, Anton. Language models and cognitive automation for economic research. No. w30957. National Bureau of Economic Research, 2023.
37. Lazer, David MJ, et al. "The science of fake news." Science 359.6380 (2018): 1094-1096.
38. Lightman, Hunter, et al. "Let's Verify Step by Step." arXiv preprint arXiv:2305.20050 (2023).
39. List, JA (2023), A course in experimental economics, University of Chicago Press, in press.
40. Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess. 2019. "Augmenting Pre-Analysis Plans with Machine Learning." AEA Papers and Proceedings, 109: 71-76.
41. Luca, Michael, Jon Kleinberg, and Sendhil Mullainathan. "Algorithms Need Managers, Too." Harvard Business Review 94, nos. 1/2 (January–February 2016): 96–101.

42. 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 (1). pp. 277-290. ISSN 0002-8282."
43. Moonesinghe, Ramal, Muin J. Khoury, and A. Cecile J. W. Janssens. "Most published research findings are false—but a little replication goes a long way." PLoS medicine 4, no. 2 (2007): e28.
44. Nosek, Brian A., Jeffrey R. Spies, and Matt Motyl. "Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability." Perspectives on Psychological Science 7, no. 6 (2012): 615-631.
45. Noy, Shakked and Zhang Whitney, Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187-192 (2023). DOI:10.1126/science.adh2586
46. Ouyang, Long, et al. "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems 35 (2022): 27730-27744.
47. Pennycook, Gordon, et al. "Shifting attention to accuracy can reduce misinformation online." Nature 592.7855 (2021): 590-595.
48. Saunders, William, et al. "Self-critiquing models for assisting human evaluators." *arXiv preprint arXiv:2206.05802* (2022).
49. Small, Christopher T., et al. "Opportunities and Risks of LLMs for Scalable Deliberation with Polis." *arXiv preprint arXiv:2306.11932* (2023).
50. Thaler, Richard H. "Nudge, not sludge." Science 361.6401 (2018): 431-431.
51. Wilson, Robert B. “Competitive bidding with disparate information”. Management Science, 15:446–448. (1969)
52. Wilson, Robert B. “A bidding model of perfect competition”. The Review of Economic Studies, 44:511–518. (1977)
53. Sargent, Thomas J. "Sources of Artificial Intelligence." (2023).