EPiC Series in Technology

Volume 3, 2026, Pages 77–84

Proceedings of AI for Accelerated Research Symposium

Synthetic AI Agents for Experimental Social Science

Thomas Henning[1] and Colin F. Camerer[1,2]

[1] Division of Humanities and Social Science, California Institute of Technology, Pasadena, CA, USA
[2] Computational and Neural Systems, California Institute of Technology, Pasadena, CA, USA
{thenning,camerer}@caltech.edu

Abstract

AI will transform social sciences. LLM agents can converse, act autonomously, and arguably reason. This position paper provides three AI-human complementary research uses. First, as interactive partners, AI agents let scholars learn more about human–AI social. Second, finely-tuned AI models can substitutes for human subjects, allowing rapid piloting and hypothesis generation, cross-cultural "synthetic cultural agents", and "out-of-this-world" (OOTW) speculations about experiments that can't be conducted with humans. Third, AI agents can co-science by coding and summarizing data at scales much faster than humans with minimal error (in some cases). This Position paper outline elements of a research agenda in which AI complements human subjects and scientists.

1 Introduction

We are in a golden age of social science [9]. This golden age is created by a torrent of cheaper and better data, new open science sharing norms, and computational power. AI tools will accelerate this boom.

While social science will be accelerated by AI, social science can repay the favor by economic analysis of the impact of AI on the workplace, inferring 'preferences' that implicitly motivate AI agents, and using causal inference to understand how AI changes make a difference [19]. Theory can also be used to understand mathematical limits to value creation in human-AI collaboration [27].

We focus on three applications: (1) Mixed human–AI interactions; (2) using AI agents as substitutes for human experiments; and (3) using AI agents to analyze qualitative social data at scale and generate new quantitative data. The last two are examples of "AI co-science"– using AI to improve the quality and pace of scientific discovery and knowledge accumulation.

Our goal is to educate readers about how AI agents can complement social science methods and accelerate progress. Anthis et al. [2] also argue for this position (and include a list of 53 studies in AI-Behavioral Science).

2 AI-human interactions

People already interact with AI in many ways: We watch movies because they were recommended by a Netflix algorithm, use LLMs to improve essay writing, talk to customer service chatbots, and

increasingly face "personalized pricing" (called "first-degree price discrimination" in economics). Evidence is accumulating about what people like and dislike about such AI interactions. People dislike bargaining with AI [12], cheat more often when reporting results to a computer [31], feel less guilty about exploiting AI counterparts [20], and underreport how much they use AI [22].

The study of human-computer interaction (HCI) is a powerful way for social scientists to probe these phenomena. AI agents can be deployed as highly controlled "confederates" in experiments, allowing researchers to systematically vary a counterpart's attributes (e.g., presenting it as a human or an AI, or giving it an approachable persona vs. an abstract one) limiting perceptual noise [23]. For example, one could test how people's willingness to trust a teammate changes based on its personality traits, while enabling them to actually "interact" with that teammate in a relatively unstructured way (helping to hide the fact that it isn't a human).

Recent reviews highlight several factors that can be manipulated in human–AI interaction studies, including emotional framing and cultural cues [18, 5, 21]. By leveraging AI agents with controlled behaviors, researchers can isolate human biases and preferences in mixed human–AI settings with greater precision than ever before. This approach will improve AI design (by creating agents that humans find more trustworthy or fair) and advance social science by revealing how we treat non-human actors in social and economic exchanges.

Another frontier for HCI is using AI in clinical practices such as diagnosis and mental health therapy. The FDA has currently approved 1247 AI devices; that number will grow rapidly. Meta-analysis of 83 studies found that AI diagnoses were as accurate as most physicians, but worse than the most expert [30].

3 AI agents as substitutes for human subjects

A dreamy use case for AI-Behavioral Science is to create artificial agents who, properly trained on human data, can act sufficiently human-like to predict accurately in never-seen OOD tasks. This would enable rapid piloting and hypothesizing, in which human subject experiments come at the end.

There are many successful examples. Hewitt et al. [17] showed that LLM (GPT-4) responses, to social science surveys of representative US human samples in 70 experiments, matched previous human data with correlation r=.91. Binz et al. [6] introduced a Centaur model that reproduced many findings about human cognition. Park et al. [26] created an LLM 1000-agent sample that matched the results of the General Social Survey with 85% precision.

Despite these apparent successes, there is intense debate about the current and future value of AI subjects. An example of this useful debate began with Dillion et al. [10] showing that LLMs could reproduce human moral judgments. Then Schröder et al. [29] showed that in perturbations of those scenarios, LLMs were more sensitive to small prompt variations than humans. They pessimistically concluded that "...LLMs [are] useful but fundamentally unreliable tools."

Our Position is that the pessimism expressed by [29] is a myopic mistake. The cycle of improvement after AI flaws are evident has been incredibly fast (months or years, not decades). AI images no longer show wonky fingers, debiasing for fairness is a hot topic, blackbox interpretability is being improved by xAI, and LLM hallucination and sycophancy are being conquered. The checklist of AI flaws which led Schröder et al. [29] to call methods "unreliable" is actually a to-do list for active researchers, who are eager to improve models quickly and will do so.

Describing a specific AI co-science experimental design will illustrate some of our ideas. Our group put humans in a simplified experimental financial market. In the market they trade assets that pay a financial dividend in each of T periods, or can invest experimental cash into a risk-free bond. The market is designed so there is a constant fundamental ("intrinsic") value. In these markets, humans often trade at accelerating prices in a "bubble" (about 3x the fundamental), which crashes toward a terminal period T.

Henning et al. [15] created AI agent-based trading agents in the same trading environment (with no steered prompting) and compared them with human subjects. Unlike human traders, AI agents priced assets close to their fundamental value (Figure 1). These findings suggest that these AI agents behave more like textbook rational agents than humans do.

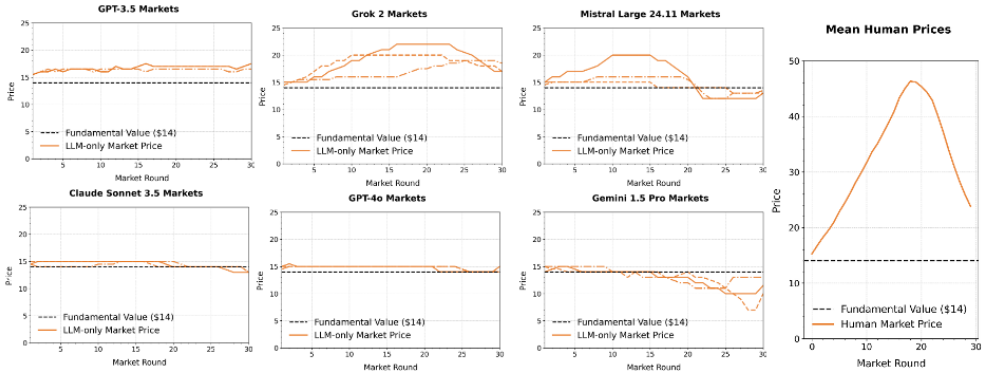

Figure 1: **AI agent vs. human market prices across 30 periods.** The six left panels (top and bottom rows) show price paths for six different LLMs prompted by human instructions, compared to the constant fundamental value (14, dotted line). The rightmost plot ("Mean human prices") shows data from human experiments averaged over 19 sessions. None of the LLMs reproduces the magnitude or trajectory of the human bubble–crash pattern, though several exhibit similar tendencies. This invites agent prompt optimization (APO) tuning of LLMs to better match human data (ongoing research).

The next step in this research is to steer the AI agents, through prompt optimization, to look more human-like. Xie et al. [33] showed the promise of this agenda. They trained AI agents across a wide range of canonical games, such as the prisoner's dilemma, dictator and public goods games. They were able to steer AI agents through language prompting to approximate overall human data. The language prompts represent "behavioral codes" which can be considered candidate mental states corresponding to what people prefer and choose.

4 What AI-behavioral science lies ahead?

Experiments on human subjects are not easy. They must be recruited to a physical lab, or connected and filtered online. IRB rules means humans have to self-select to volunteer and can't be put into many lifelike situations. Humans get tired, and drop out of gold-standar longitudinal studies that require repeated contact. AI subjects could overcome *many* of these limits. Furthermore, social science vastly oversamples people from WEIRD (Western, educated, industrialized, rich, and democratic) societies [16]. Synthetic AI agents could be used to study non-WEIRD societies that are often hard to sample (such as small-scale societies in poor countries with limited internet.

Synthetic cultural agents (SCAs) can be fine-tuned using web scraping and retrieval-augmented generation (RAG). Gonzalez-Bonorino et al. [14] created SCAs for 6 small-scale societies for which there was some human experimental data about economic games. The SCAs displayed cross-cultural qualitative patterns resembling human behavior. SCA predictions can be used as pilot data to decide which cultures are worth spending money to travel to and study up close.

AI agents could be used to conduct what we call "out-of-this-world" (OOTW) experiments (an acronym extending OOD). Some human subjects and conditions cannot ethically or feasibly be done with humans. Can we learn about how ancient Athenians trade, bargain, and speak politically? What about how traders create and crash bubbles over a 1,000-period trading horizon? Do groups of 250 AI agents create threshold public goods (such as political revolutions) differently than groups of 10, on which we have experimental human data, or groups of 1M, on which we have no controlled data? These are experiments that range from impossible to impractical with humans. All of them could be done using AI agents [25].

Of course in many experimental domains LLMs can be bad predictors of human behavior if they are not well-prompted [13]. However, theory-guided prompting in the same games studied by Gao et al. [13] can produce very good predictions, including OOD in games in a large parametrized set of novel designs [24]. A feature of this debate is the speed of resolution. Gao et al. [13] was published on June 13, 2025. Manning & Horton [24] was on arXiv less than 3 months later.

5 AI tools for analyzing and generating social data

AI agents also show promise as tools for processing qualitative data, coding open-ended responses, and generating new textual data. In computational social science (CSS), Ziems et al. [34] found that zero-shot AI agents (without fine-tuning) can achieve near-human agreement on classification tasks, and excel at free-form coding.

Researchers studying social movements could use an AI agent to classify millions of tweets by protest topics and summarize long transcripts of public hearings or analyze the aggressiveness of wartime communications between adversaries. AI agents also enable rapid quantification of qualitative data by embedding text into high-dimensional representations, which can then be analyzed with conventional statistical techniques, or by simply providing numerical ratings in lieu of human labelers. However, the authors caution that AI agent output can be non-deterministic and highly sensitive to the phrasing of the prompt, necessitating careful documentation. Though, this is often true in humans as well.

It is crucial to recognize that AI agents' open-ended survey coding will not always be analogous to humans, von der Heyde et al. [32] compared different AI agents for classifying free-text survey responses, only a fine-tuned model achieved satisfactory performance. Off-the-shelf AI agents misclassified responses across categories, and the distribution of codes varied widely. While AI agents may reduce labor costs, researchers should be concerned about the validity of their performance and should use humans to validate. This echoes a broader theme: AI tools are best used to augment, rather than fully replace, human analysis.

6 Conclusions and future directions

AI agents will transform behavioral science. They allow scholars to create and explore AI-human interactions, create synthetic AI agents to pilot-test design substitute for human participants

OOD or over impossible OOTW historical and numerical scales, and use AI tools for analysis of text and do other co-science, such as detecting fraudulent science or predicting replicability [1].

There are certainly limits to using AI agents in social science right now. AI behavior is often different from that of humans (sometimes "hyper-rational"), and these differences may be quirkily domain-specific- e.g., LLMs evaluate risks like humans but discount future rewards hugely more than humans do [28]. AI agents are also trained on a corpus that reflects dominant cultures and may reflect harmful biases. Asking LLMs to compete for success (as many applications will do) creates inadvertent misalignment [11].

In co-science AI, a recent large-scale study compared human-only, AI-assisted, and AI-led teams reproducing quantitative social science results. AI-only approaches were not capable of reliably reproducing complex scientific analyses without substantial human oversight [8].

Our position is that research on AI-Behavioral Science benefits from the conversational cycle between AI-derived insights and criticisms that those insights provoke. An adversarial process is the best way to make progress, as long as it's based on new evidence. In political science, for example, evidence that LLMs can reproduce human election polling [3] was criticized [7], leading to a thoughtful assessment of LLM pros and cons one year later [4].

There is a lot more to learn about meta-science questions: How does the presence of AI agents affect human subjects' beliefs, trust, and cooperation? Can AI be used to pre-test experimental designs and predict outcomes? Ethicists and policymakers must also help ensure that AI benefits are equitably distributed and that potential harms are quickly identified and limited.

References

[1] Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. Predicting the replicability of social science lab experiments. *PloS one*, 14(12), 2019. doi: e0225826.

[2] Jacy R. Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. Llm social simulations are a promising research method. *arXiv preprint*, arXiv:2504.02234, 2025. URL https://arxiv.org/abs/2504.02234. Version v2, revised 5 June 2025.

[3] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

[4] Lisa P. Argyle, Ethan C. Busby, Joshua R. Gubler, Bryce Hepner, Alex Lyman, and David Wingate. Arti-"fickle" intelligence: Using llms as a tool for inference in the political and social sciences, 2025. URL https://arxiv.org/abs/2504.03822.

[5] A. Baskaran, R. Kowalewski, J. R. Sutherland, and M. McTear. A review of emotions in human–conversational agent interaction. *Proceedings of the AAAI Symposium Series*, 178: 103040, 2023. doi: https://doi.org/10.1609/aaaiss.v1i1.27477.

[6] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Eltetö, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath,

Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09215-4. URL https://doi.org/10.1038/s41586-025-09215-4.

[7] James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024. doi: 10.1017/pan.2024.5.

[8] Abel Brodeur, David Valenta, Alexandru Marcoci, et al. Comparing human-only, ai-assisted, and ai-led teams on assessing research reproducibility in quantitative social science. Technical Report DP 195, Institute for Replication, 2025. URL https://www.econstor.eu/bitstream/10419/308508/1/I4R-DP195.pdf.

[9] Anastasia Buyalskaya, Marcos Gallo, and Colin F. Camerer. The golden age of social science. *Proceedings of the National Academy of Sciences*, 118(5):e2002923118, 2021. doi: 10.1073/pnas.2002923118.

[10] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023. ISSN 1364-6613. doi: 10.1016/j.tics.2023.04.008.

[11] Batu El and James Zou. Moloch's bargain: Emergent misalignment when llms compete for audiences. *arXiv preprint*, arXiv:2510.06105, 2025. URL https://arxiv.org/abs/2510.06105. Preprint, v1, 7 Oct 2025.

[12] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with ai systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2022. doi: 10.1145/3491102.3517734.

[13] Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 2025. doi: 10.1073/pnas.2501660122. URL https://www.pnas.org/doi/10.1073/pnas.2501660122.

[14] Augusto Gonzalez-Bonorino, Emilio Pantoja, and C. Monica Capra. Llms model non-weird populations: Experiments with synthetic cultural agents. SSRN working paper, 2024.

[15] Thomas Henning, Siddhartha M. Ojha, Ross Spoon, Jiatong Han, and Colin F. Camerer. Llm trading: Analysis of llm agent behavior in experimental asset markets. *arXiv preprint arXiv:2502.15800*, 2025. URL https://arxiv.org/abs/2502.15800. Advances in Financial AI Workshop at ICLR 2025.

[16] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3):61–83, 2010. doi: 10.1017/S0140525X0999152X.

[17] Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting results of social science experiments using large language models. *Working Paper*, 2024. *equal contribution.

[18] John J. Horton. Large language models as simulated economic agents: What Can We Learn from Homo Silicus? *arXiv preprint arXiv:2301.07543*, 2023. doi: 10.48550/ARXIV.2301.07543. Version 1, January 18.

[19] Matthew O. Jackson, Qiaozhu Mei, Stephanie Wang, Yutong Xie, Walter Yuan, Seth G. Benzell, Erik Brynjolfsson, Colin F. Camerer, James A. Evans, Brian Jabarian, Jon Kleinberg, Juanjuan Meng, Sendhil Mullainathan, Asuman E. Ozdaglar, Thomas Pfeiffer, Moshe Tennenholtz, Robb Willer, Diyi Yang, and Teng Ye. Ai behavioral science. Technical report, SSRN, August 2025. URL https://ssrn.com/abstract=5395006. Available at SSRN: https://ssrn.com/abstract=5395006 or http://dx.doi.org/10.2139/ssrn.5395006.

[20] Jurgis Karpus, Armin Krüger, Julia T. Verba, Bahador Bahrami, and Ophelia Deroy. Algorithm exploitation: Humans are keen to exploit benevolent ai. *iScience*, 24:102679, 2021. doi: 10.1016/j.isci.2021.102679.

[21] Sung Park Kim, Jin Lee, and S. Y. Park. Anthropomorphic response: Understanding interactions with ai agents. *Computers in Human Behavior*, 139:107512, 2023.

[22] Yier Ling and Alex Imas. Underreporting of ai use: The role of social desirability bias. Technical Report SSRN 5232910, SSRN, May 2025. URL https://ssrn.com/abstract=5232910. Revised May 18, 2025.

[23] Akihiro Maehigashi, Takahiro Tsumura, and Seiji Yamada. Experimental investigation of trust in anthropomorphic agents as task partners. In *Proceedings of the 10th International Conference on Human–Agent Interaction (HAI 2022)*, pp. 302–305. ACM, 2022. doi: 10.1145/3527188.3563921. URL https://dl.acm.org/doi/10.1145/3527188.3563921.

[24] Benjamin Manning and John J. Horton. General social agents. *Available at SSRN*, 2025. doi: 10.2139/ssrn.5521881. URL https://ssrn.com/abstract=5521881.

[25] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. A turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, Feb 2024. doi: 10.1073/pnas.2313925121. URL https://www.pnas.org/doi/10.1073/pnas.2313925121.

[26] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024. URL https://arxiv.org/abs/2411.10109.

[27] Kenny Peng, Nikhil Garg, and Jon Kleinberg. A no free lunch theorem for human–ai collaboration. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 2025.

[28] Jillian Ross, Yoon Kim, and Andrew W. Lo. Llm economicus? mapping the behavioral biases of llms via utility theory. (arXiv:2408.02784), August 2024. doi: 10.48550/arXiv.2408.02784. URL http://arxiv.org/abs/2408.02784. arXiv:2408.02784 [cs].

[29] Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. Large language models do not simulate human psychology. Technical report, 2025. URL https://arxiv.org/abs/2508.06950.

[30] H. Takita et al. A systematic review and meta-analysis of diagnostic performance comparison between generative ai and physicians. *npj Digital Medicine*, 2025. doi: 10.1038/s41746-025-01543-z. URL https://www.nature.com/articles/s41746-025-01543-z.

[31] Lauren S. Treiman, Chien-Ju Ho, and Wouter Kool. Humans forgo reward to instill fairness into ai. In *Proceedings of the 11th AAAI Conference on Human Computation and Crowdsourcing (HCOMP '23)*, pp. 152–162, 2023. doi: 10.1609/hcomp.v11i1.27556.

[32] Leah von der Heyde, Anna-Carolina Haensch, Alexander Wenz, and Bolei Ma. United in diversity? contextual biases in llm-based predictions of the 2024 european parliament elections. *arXiv preprint arXiv:2409.09045*, 2025.

[33] Yutong Xie, Qiaozhu Mei, Walter Yuan, and Matthew Jackson. Using large language models to categorize strategic situations and decipher motivations behind human behaviors. *Proceedings of the National Academy of Sciences*, 2025.

[34] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.