Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
May 2025

Simulating Human Behavior with AI Agents

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie J. Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, Michael S. Bernstein

AI agents have been gaining widespread attention among the general public as AI systems that can pursue complex goals and directly take actions in both virtual and real-world environments. Today, people can use AI agents to make payments, reserve flights, and place grocery orders for them, and there is great excitement about the potential for AI agents to manage even more sophisticated tasks.

However, a different type of AI agent—a simulation of human behaviors and attitudes—is also on the rise. These simulation AI agents aim to be useful at asking "what if" questions about how people might respond to a range of social, political, or informational contexts. If these agents achieve high accuracy, they could enable researchers to test a broad set of interventions and theories, such as how people would react to new public health messages, product launches, or major economic or political shocks. Across economics, sociology, organizations, and political science, new ways of simulating individual behavior—and the behavior of groups of individuals—could help expand our understanding of social interactions, institutions, and networks. While work on these kinds of agents is progressing, current architectures must cover some distance before their use is reliable.

Key Takeaways

Simulating human attitudes and behaviors could enable researchers to test interventions and theories and gain real-world insights.

We built an AI agent architecture that can simulate real people in ways far more complex than traditional approaches. Using this architecture, we created generative agents that simulate 1,000 individuals, each using an LLM paired with an in-depth interview transcript of the individual.

To test these generative agents, we evaluated the agents' responses against the corresponding person's responses to major social science surveys and experiments. We found that the agents replicated real participants' responses 85% as accurately as the individuals replicated their own answers two weeks later on the General Social Survey.

Because these generative agents hold sensitive data and can mimic individual behavior, policymakers and researchers must work together to ensure that appropriate monitoring and consent mechanisms are used to help mitigate risks while also harnessing potential benefits.

In our paper, "Generative Agent Simulations of 1,000 People," we introduce an AI agent architecture that simulates more than 1,000 real people. The agent architecture—built by combining the transcripts of two-hour, qualitative interviews with a large language model (LLM) and scored against social science benchmarks—successfully replicated real individuals' responses to survey questions 85% as accurately as participants replicate their own answers across surveys staggered two weeks apart. The generative agents performed comparably in predicting people's personality traits and experiment outcomes and were less biased than previously used simulation tools.

This architecture underscores the benefits of using generative agents as a research tool to glean new insights into real-world individual behavior. However, researchers and policymakers must also mitigate the risks of using generative agents in such contexts, including harms related to over-reliance on agents, privacy, and reputation.

Introduction

Simulations in which agents are used to model the behaviors and interactions of individuals have been a popular tool for empirical social research for years, even before the emergence of AI agents. Traditional approaches to building agent architectures, such as agent-based models or game theory, rely on clear sets of rules and environments manually specified by the researchers. While these rules make it relatively easy to interpret results, they also limit the contexts in which traditional agents can act while oversimplifying the real-life complexity of human behavior. This, in turn, can limit the generalizability and accuracy of the simulation results.

Generative AI models offer the opportunity to build general purpose agents that can simulate human attitudes across a variety of contexts.

Generative AI models offer the opportunity to build general purpose agents that can simulate human attitudes across a variety of contexts. To create simulations that better reflect the myriad, often idiosyncratic factors that influence individuals' attitudes, beliefs, and behaviors, we built a novel generative agent architecture that combines LLMs with in-depth interviews with real individuals.

We recruited 1,052 individuals—representative of the U.S. population across age, gender, race, region, education, and political ideology—to participate in two-hour qualitative interviews. These in-depth interviews, which included both pre-specified questions and adaptive follow-up questions, are a foundational social science method that has been successfully used by researchers to predict life outcomes beyond what could be learned from traditional surveys and demographic instruments. We also developed an AI interviewer to ask participants the questions based on a semi-structured interview

protocol from the American Voices Project—which ranged from life stories to people's views on current social issues.

Then, we built the generative agents based on participants' full interview transcripts and an LLM. When a generative agent was queried, the full transcript was injected into the model prompt, which instructed the model to imitate the relevant individual when responding to questions, including forced-choice prompts, surveys, and multi-stage interactional settings.

Once the generative agents were in place, we evaluated them on their ability to predict participants' responses to common social science surveys and experiments, which the participants completed after their in-depth interviews. We tested on the core module of the General Social Survey (widely used to assess survey respondents' demographic backgrounds, behaviors, attitudes, and beliefs); the 44-item Big Five Inventory (designed to assess an individual's personality); five well-known behavioral economic games (the dictator game, first and second player trust games, public goods game, and prisoner's dilemma); and five social science experiments with control and treatment conditions. For the General Social Survey (which has categorical responses), we measured accuracy and correlation based on whether the agent selects the same survey response as the person. For the Big Five Inventory and the economic games (which have continuous responses), we assessed accuracy and correlation using mean absolute error.

The generative agents proved remarkably effective in simulating individuals' real-world personalities.

Research Outcomes

Overall, the generative agents proved remarkably effective in simulating individuals' real-world personalities. For example, the generative agents predicted participants' responses to the General Social Survey with an average normalized accuracy of 85%—meaning that, on average, generative agents replicated participant responses 85% as accurately as the participants themselves when they were asked to retake the surveys and experiments two weeks later. This result is 14 to 15 percentage points higher than the accuracy of traditional demographic-based and persona-based agents that use the same LLMs but do not have access to the interviews.

The generative agents also outperformed demographic and persona-based agents on the Big Five personality test, achieving a normalized correlation of 80% when replicating real individuals' openness, conscientiousness, extraversion, agreeableness, and neuroticism. But they performed similarly as demographic and persona-based agents for the economic games, with a normalized correlation of 66% (i.e., 66% as high as the participants' own correlation with themselves two weeks later) across an

...the interview-based generative agents consistently reduced biases across tasks compared to demographic-based agents.

aggregate of the dictator game, the first and second player trust games, the public goods game, and the prisoner's dilemma.

Beyond those tests, we evaluated the generative agents' behavior in a set of social science experiments. These included investigations of how perceived intent affects blame assignment and how fairness influences emotional responses. Real-world participants and the generative agents agreed on the replication results of all five studies we tested.

The generative agents also lessened bias in predictive accuracy across social groups. Given rightful concerns about AI systems disadvantaging or misrepresenting underrepresented populations, we conducted a subgroup analysis focused on political ideology, race, and gender. These are dimensions of particular interest in the literature. We used the Demographic Parity Difference, which measures the performance difference between the best- and worst-performing groups, to quantify bias. Notably, we found that the interview-based generative agents consistently reduced biases across tasks compared to demographic-based agents. Drops in political ideology bias and racial bias vary depending on the survey, while gender-based Demographic Parity Difference remained fairly consistent across tasks (likely due to already low levels of discrepancy).

Policy Discussion

Generative agents could become useful tools for estimating attitudes and survey-based experimental treatment effects. For example, if you were considering the sorts of survey questions you might ask in a national survey, generative agents could help to estimate average responses the population might give. However, many open questions remain: How accurate are generative agents when simulating behavior, in addition to attitudes? What innovations are needed for generative agent simulations to accurately estimate the impacts of policy changes? While we will continue to build the empirical and technical research to expand the horizon of generative agents, we urge policymakers to critically examine analyses that overclaim what generative agents can actually achieve today.

One important risk for policymakers, practitioners, researchers, and others using generative agents is overreliance on generative agents when simulation accuracy is low. To ensure that policymakers don't rely on an inaccurate simulation, we must develop tools and methodologies so they know when they can, and can't, trust these simulations. Additionally, policymakers should not apply generative agents beyond the range of applications that have been validated. A second major risk relates to privacy: The interview

data used to build the generative agents is often sensitive, and data leaks could cause considerable harm to interviewees. Other concerns include the co-option of individuals' likenesses, as these agents can believably replicate a person's answers in a survey response or experiment. Significant reputational harm could also result from someone manipulating agent responses to falsely attribute defamatory statements to individuals whose data is used in the agent bank.

A range of other ethical and legal questions must also be considered. For example, what are the ethical implications of using AI agents that simulate a deceased person? How should human consent be managed? And what are the risks of agents being misused for fraudulent purposes? Given the inherent uncertainty of future advancements in generative AI, such as AI models' future reasoning abilities, managing these risks early on is crucial. Policymakers should consider establishing bright-line rules that determine how AI agents may or may not be used for human simulation purposes.

We made the decision not to release our generative agents for public use. Instead, to support further research while protecting participant privacy, we have chosen to provide controlled, research-only API access to our agent bank. We grant open access to aggregated responses on fixed tasks for general research use and restrict access to individual responses on open tasks for researchers following a review process, ensuring the agents are accessible while minimizing risks associated with the source interviews. Other researchers building similar systems should replicate our safeguards, and policymakers weighing how generative agents could be used in research settings should explore requirements for individual data rights, access, and deletion.

One important risk for policymakers, practitioners, researchers, and others using generative agents is overreliance on generative agents when simulation accuracy is low.

Policymakers and researchers should work together to ensure that appropriate monitoring and consent mechanisms are used to enhance trust, protect individual rights, and mitigate the risks of generative agent use. For example, our team proposed the possibility of an audit log for the use of every agent in our agent bank. That way, individuals who participated in a survey and had their preferences captured by a generative agent could see what the agent is doing and exert control over it over time. Permission could be granted one day and withdrawn a month later, reflecting individual consent. Translating such protections into policy—such as making them part of grant terms and conditions—would help researchers to detect and mitigate malicious use of generative agents built using people's personal data shared via in-depth interviews.

Looking forward, generative agents hold serious promise for enhancing human behavioral research and developing new insights into personal preferences and decision making. However, mitigating the risks of these innovations, through research and policy controls on agent access and auditing, will be crucial to harnessing their opportunities in economics, political science, and beyond.

Reference: The original article is accessible at Joon Sung Park, Carolyn Q. Zou, et al., **"Generative Agent Simulations of 1,000 People,"** arxiv.org, November 15, 2024, https://arxiv.org/abs/2411.10109.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Joon Sung Park is a PhD student in computer science in the Human-Computer Interaction and Natural Language Processing groups at Stanford University.

Meredith Ringel Morris is director and principal scientist for Human-AI Interaction at Google DeepMind.

Carolyn Q. Zou is a PhD student in computer science at Stanford University.

Robb Willer is a professor of sociology and, by courtesy, psychology and business at Stanford University.

Aaron Shaw is an associate professor in the Department of Communication Studies at Northwestern University.

Percy Liang is an associate professor of computer science at Stanford University and a senior fellow at Stanford HAI.

Benjamin Mako Hill is an associate professor in the Department of Communication at the University of Washington.

Michael S. Bernstein is an associate professor of computer science at Stanford University and a senior fellow at Stanford HAI.

Carrie J. Cai is a senior staff research scientist at Google DeepMind and manager/area lead of Human-AI Interaction in Google's People+AI Research group.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**