

GENERATIVE AGENT SIMULATIONS OF HUMAN BEHAVIOR

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE

AND THE COMMITTEE ON GRADUATE STUDIES

OF STANFORD UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Joon Sung Park

August 2025

© 2025 by Joon Sung Park. All Rights Reserved.

Re-distributed by Stanford University under license with the author.



This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.

<http://creativecommons.org/licenses/by-nc/3.0/us/>

This dissertation is online at: <https://purl.stanford.edu/jm164ch6237>

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Michael Bernstein, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Percy Liang, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Meredith Ringel Morris

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format.

Abstract

Human institutions embed assumptions about how people perceive, decide, and act; when those assumptions are incomplete, well-intentioned systems can produce unintended effects at scale. This dissertation argues that simulations powered by generative, language-model-based agents can serve as a new scientific instrument for probing such assumptions, revealing emergent dynamics, and exploring counterfactual designs before deployment. I introduce a general architecture that transforms large language models into coherent computational agents through three capabilities—memory (structured episodic records), reflection (periodic synthesis of goals), and planning (multi-timescale plans that guide and revise action). Complementing this scaffold, I develop methods for constructing agents from rich, individual-level data (e.g., interviews) and a measurement framework that evaluates alignment against ground truth on people’s attitudes and behaviors.

Applying these components, the dissertation presents an interactive sandbox demonstrating coherent individual behavior and emergent social phenomena, and a population simulation of 1,000 agents representing a cross-section of U.S. adults used to assess fidelity across attitudinal, behavioral, and diffusion tasks and to analyze where alignment succeeds or fails. Case studies show how generative-agent simulations can inform platform design, content recommendation, and public policy, with social simulacra illustrating how to prototype communities and surface failure modes prior to real-world deployment. Taken together, the work traces a path from capability to credibility, positioning generative-agent simulations as instruments that complement empirical observation to challenge assumptions, generate testable hypotheses, and support wiser design and governance of social systems.

Acknowledgements

This dissertation is dedicated to the people who have made my past five years as a researcher meaningful.

- **Advisers—Michael Bernstein and Percy Liang.** Thank you for the support and guidance that will stay with me in everything I do. You taught me to dream ambitiously and gave me the tools to pursue those dreams. Your insight, rigor, and care for the craft—and for the people you work with—shaped every step of this dissertation.
- **Mentors—Meredith Ringel Morris and Karrie Karahalios.** Thank you for trusting me and supporting me over the years, not only in scholarship but in learning how to be a good colleague and community member.
- **Social science mentor—Robb Willer.** For modeling rigor in experimental design, for generous mentorship, and for the good humor and groovy vibes, always.
- **Early champions—James Landay, Jeff Hancock, Danae Metaxa, and Mary Wootters.** Thank you for taking an early bet on me and helping me find a place to grow as an early-stage scholar.
- **Collaborators.** This work was made possible—and far more joyful—by collaborators whose creativity and care sharpened the ideas: Carolyn Zou, Joseph O’Brien, Lindsay Popowski, Aaron Shaw, Mako Hill, and Carrie Cai.
- **Friends and colleagues.** The Stanford HCI and NLP communities made everyday work rich and entertaining. Thanks to my officemates—Mitchell

Gordon, Michelle Lam, Beleicia Bullock, and Jordan Troutman—and to my cohorts Rishi Bommasani, Lisa Li, and Rohith Kuditipudi for conversations that pushed my thinking. I am also very grateful to Mia Tang for helping me wrangle notoriously finicky L^AT_EX tables in this dissertation.

- **Family.** To my parents, Ji Yeon Kim and Jeong-Jun Park: your love, patience, and example are the foundation I value most. Thank you for everything.

This PhD unfolded during a period of extraordinary change in our field. GPT-3 became widely available in my first year, catalyzing a wave of new ideas. Working at the intersection of HCI and AI, I set out to explore meaningful applications for these models—work that eventually led me to generative agents and simulations. I am grateful to have done this research during such a formative moment, and humbled by the responsibility we share to shape this technology for the better.

Lastly, a note on place. The setting of *Smallville*, where the generative agents live, is an amalgam of real spaces that fed and sustained me as an undergraduate and graduate student. Thank you to everyone in those places for the meals, conversations, and community that nourished this work over the years.

Contents

Abstract	iv
Acknowledgements	v
1 Introduction	1
1.1 From Stylized Agents to Generative Agents	3
1.2 Accuracy Beyond Believability	4
1.3 A Motivating Lens: From Emergence to Design	5
1.4 Overview of the Dissertation	6
1.5 Scope, Limitations, and Ethics	7
2 Background and Related Work	9
2.1 Why This Problem Matters: Social Systems Are Hard to Design . . .	10
2.2 Prototyping in Design Practice	11
2.2.1 Challenges in Prototyping Social Systems	12
2.3 Agent-Based Modeling and Computational Social Science	12
2.4 Human–AI Interaction and Prototyping for Socio-Technical Systems .	13
2.4.1 Believable Proxies of Human Behavior	14
2.4.2 Large Language Models and Human Behavior	17
2.5 From Stylized ABMs to Generative, Language-Model Agents	18
2.6 Validity, Representativeness, and Ethical Considerations	19
2.7 Summary and Positioning	19

3	Generative Agents: Interactive Simulacra of Human Behavior	21
3.1	What I mean by "generative agents"	22
3.1.1	An interactive society in a sandbox world	23
3.2	Generative Agent Behavior and Interaction	25
3.2.1	Agent Avatar and Communication	25
3.2.2	Environmental Interaction	28
3.2.3	Example “Day in the Life”	30
3.2.4	Emergent Social Behaviors	32
3.3	Generative Agent Architecture	33
3.3.1	Memory and Retrieval	34
3.3.2	Reflection	37
3.3.3	Planning and Reacting	40
3.4	Sandbox Environment Implementation	44
3.4.1	From Structured World Environments to Natural Language, and Back Again	45
3.5	Controlled Evaluation	46
3.5.1	Evaluation Procedure	47
3.5.2	Conditions	48
3.5.3	Human Evaluators	49
3.5.4	Analysis	50
3.5.5	Results	51
3.6	End-To-End Evaluation	53
3.6.1	Emergent Social Behaviors	54
3.6.2	Boundaries and Errors	56
3.7	Discussion	58
3.7.1	Applications of Generative Agents	58
3.7.2	Ethics and Societal Impact	59
4	Simulations of 1,000 People	61
4.1	Study Overview and Contributions	63
4.2	Constructing the Agent Bank	66

4.2.1	Data Collection Procedure	66
4.2.2	Recruitment and Demographics	68
4.2.3	Participant Consent	68
4.3	Creating the AI Interviewer Agent	69
4.3.1	AI Interviewer Agent Architecture	70
4.3.2	Interview Script	72
4.3.3	Implementation	73
4.3.4	Development Process and Evaluation	77
4.3.5	Expert Reflection	79
4.3.6	Generating a Prediction With Generative Agents	82
4.4	Evaluation Design	83
4.4.1	General Social Survey (GSS)	83
4.4.2	Big Five Personality Traits (BFI-44)	84
4.4.3	Behavioral Economic Games	85
4.4.4	Replication Studies of Experimental Treatment Effects	87
4.5	Metrics	89
4.5.1	Individual-level Outcomes and Metrics	89
4.5.2	Replicating Experiments With Generative Agents	93
4.5.3	Interviews Improve Agents' Prediction Accuracy	95
4.5.4	Demographic Bias in Agent Predictions	98
4.6	Results	99
4.6.1	Predicting Individuals' Attitudes and Behaviors	99
4.6.2	Predicting Experimental Treatment Effects	101
4.6.3	Interviews Improve Agents' Prediction Accuracy	103
4.6.4	Exploratory robustness analysis results	104
4.6.5	Fairness and Subgroup Disparities	105
4.7	Why interview-based generative agents work	106
4.8	Ethical Considerations and Research Access to the Agent Bank	111
4.8.1	Research Access for the Agent Bank	111
4.8.2	Overview of the Agent Bank	112
4.8.3	Strategic Framework for Access	113

4.8.4	Plan and Desiderata	115
4.8.5	Monitoring and Iteration	116
5	Social Simulacra: An Application Case Study	117
5.1	A Motivating Question for Social Systems	118
5.2	Social Simulacra and SimReddit	120
5.2.1	System Context	122
5.2.2	GENERATE: Generating Social Behaviors	123
5.2.3	WHATIF: Exploring Alternative Scenarios	126
5.2.4	MULTIVERSE: Exploring Alternate Possibilities	128
5.3	Creating Social Simulacra	130
5.3.1	Modeling Assumptions	130
5.3.2	Prompting Techniques	131
5.3.3	LLM Settings	135
5.4	Technical Evaluation	135
5.4.1	Evaluation Subreddits	136
5.4.2	Study Procedure	137
5.4.3	Participants	138
5.4.4	Analysis	139
5.4.5	Results	141
5.5	Designer Evaluation	142
5.5.1	Study Procedure	143
5.5.2	Participants	144
5.5.3	Inductive Analysis of the Interview and Participants' Designs .	144
5.5.4	Results	145
6	Conclusion	151
6.1	From Instrument to Agenda	151
6.2	A Roadmap for the Field	152
6.3	Future Work: Toward a Trustworthy Simulation System	153
6.4	Advice to New Entrants	155
6.5	Ethics, Governance, and Limits	156

6.6	Concluding Perspective	157
A	Supplementary Material for Chapter 3	158
A.1	Architecture Optimizations	158
A.2	Agent Interview Questions	159
A.2.1	Self-Knowledge	159
A.2.2	Memory	160
A.2.3	Plans	161
A.2.4	Reactions	162
A.2.5	Reflections	163
B	Supplementary Material for Chapter 4	165

List of Tables

4.1	Results of replication studies by human participants and generative agents. I report the p-values (***: < 0.001 , **: < 0.01 , *: < 0.05) and Cohen’s d for effect sizes. Our replication with human participants replicated four out of five studies, while generative agents informed by the interview transcript replicated the same four studies. The correlation of the effect sizes between the human participants and generative agents achieved a strong correlation.	102
4.2	Overlap between GSS outcome question and interview-question answer pairs where a GPT-4o-mini model has classified that the answer to the GSS question can be directly retrieved or inferred from the interview answer.	111
5.1	SimReddit significantly outperformed the crowdworker baseline and all ablations ($p < .01$). Results from one-way ANOVA of participants’ error rate in the Technical Evaluation, followed by Tukey’s HSD post hoc test. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$	140
B.1	Demographic distribution of our 1,052 participants. Collectively, they represent a stratified sample of the U.S. demographic across age, gender, race, region of residence, level of education, and political identity. Note that for ethnicity, the participants could choose more than one option.	165

B.2	Comparative predictive performance of agents built using different descriptions. Agents constructed from interview transcripts outperformed both demographic-based and persona-based agents across multiple tasks. Specifically, interview-based agents showed significant improvement in predicting responses to the General Social Survey (in accuracy) and Big Five Personality Traits (in correlation), as confirmed by ANOVA tests ($p < 0.001$ for both). In contrast, no significant differences were observed between agent types for the Economic Games (in correlation), indicating that interviews were particularly valuable for tasks requiring deeper, personal insights.	170
B.5	Table 3. The construct-level analysis of predictive accuracy across the General Social Survey, Big Five personality traits, and economic games. For each construct, we provide accuracy and correlation metrics, along with replication ratios. The analysis highlights the performance of generative agents in predicting specific dimensions within these constructs, with metrics showing varied levels of predictive accuracy across different social, personality, and game-based items. This evaluation complements the individual-level analysis, which is of our primary interest in this work, by offering a detailed look at the accuracy of agents for specific constructs and items.	178
B.9	Demographic Parity Difference (DPD) Results. This table summarizes the results of regression analyses measuring demographic parity differences (DPD) across three tasks (GSS, Big Five, and economic games) for agents using demographic information, interview data, and persona-based profiles. Interview-based agents consistently reduced bias compared to demographic-based agents across political ideology, race, and gender.	196

B.10 Robustness analysis. This table presents the results of an exploratory robustness analysis comparing different agent architectures informed by various data sources, including interviews, surveys, experiments, and summaries. Performance is evaluated across four constructs: the General Social Survey (GSS), GSS Numeric (GSS Num.), Big Five Personality Traits (Big Five), and economic games. Interview-based agents consistently outperform others, achieving high accuracy (0.83, std=0.11) on the GSS, indicating that interviews provide richer, more comprehensive information than surveys and experiments. Maximal agents, which integrate data from all sources, show similar performance. Summary agents perform slightly below interview-based agents, with minor losses in accuracy (0.82, std = 0.12). A progressive decline is observed for random lesion agents as portions of interview data are removed, with accuracy dropping from 0.83 to 0.79 as 80% of the utterances are excluded, suggesting that even shortened interviews retain valuable insights compared to survey-only agents.	198
B.11 The interview script used to guide two-hour conversations with participants, adapted from the American Voices Project’s interview. The script covers a wide range of topics, from participants’ life stories to their views on social, political, and personal values. Select portions were abbreviated to ensure a manageable session length, while still capturing the breadth of experiences and perspectives essential for building nuanced generative agents.	199

B.12 Generative agents' predictive performance. The consistency rate between participants and the predictive performance of generative agents is evaluated across various constructs and averaged across individuals. For the General Social Survey (GSS), accuracy is reported due to its categorical response types, while the Big Five personality traits and economic games report mean absolute error (MAE) due to their numerical response types. Correlation is reported for all constructs. Normalized accuracy is provided for all metrics, except for MAE, which cannot be calculated for individuals whose MAE is 0 (i.e., those who responded the same way in both phases). We find that generative agents predict participants' behavior and attitudes well, especially when compared to participants' own rate of internal consistency. Additionally, using interviews to inform agent behavior significantly improves the predictive performance of agents for both GSS and Big Five constructs, outperforming other commonly used methods in the literature.	205
--	-----

List of Figures

3.1 Generative agents are believable simulacra of human behavior for interactive applications. In this work, I demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities. . . .

3.2 The Smallville sandbox world, with areas labeled. The root node describes the entire world, children describe areas (e.g., houses, cafe, stores), and leaf nodes describe objects (e.g., table, bookshelf). Agents remember a subgraph that reflects the parts of the world they have seen, maintaining the state of those parts as they observed them. . . .

3.3 A morning in the life of a generative agent, John Lin. John wakes up around 6 am and completes his morning routine, which includes brushing his teeth, taking a shower, and eating breakfast. He briefly catches up with his wife, Mei, and son, Eddy, before heading out to begin his workday. . . .

3.4 At the beginning of the simulation, one agent is initialized with an intent to organize a Valentine’s Day party. Despite many possible points of failure in the ensuing chain of events—agents might not act on that intent, might forget to tell others, might not remember to show up—the Valentine’s Day party does, in fact, occur, with a number of agents gathering and interacting. . . .

3.5	Generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent’s experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories and uses those retrieved actions to determine an action. These retrieved memories are also used to form longer-term plans and create higher-level reflections, both of which are entered into the memory stream for future use.	31
3.6	The memory stream comprises a large number of observations that are relevant and irrelevant to the agent’s current situation. Retrieval identifies a subset of these observations that should be passed to the language model to condition its response to the situation.	35
3.7	A reflection tree for Klaus Mueller. The agent’s observations of the world, represented in the leaf nodes, are recursively synthesized to derive Klaus’s self-notion that he is highly dedicated to his research.	38
3.8	The full generative agent architecture produces more believable behavior than the ablated architectures and the human crowdworkers. Each additional ablation reduces the performance of the architecture. . . .	51
3.9	The diffusion path for Isabella Rodriguez’s Valentine’s Day party invitation involved a total of 12 agents, aside from Isabella, who heard about the party at Hobbs Cafe by the end of the simulation.	54

4.1	The process of collecting participant data and creating generative agents begins by recruiting a stratified sample of 1,052 individuals from the U.S., selected based on age, census division, education, ethnicity, gender, income, neighborhood, political ideology, and sexual identity. Once recruited, participants complete a two-hour audio interview with our AI interviewer, followed by surveys and experiments. I create generative agents for each participant using their interview data. To evaluate these agents, both the generative agents and participants complete the same surveys and experiments. For the human participants, this involves retaking the surveys and experiments again two weeks later. I assess the accuracy of the agents by comparing agent responses to the participants' original responses, normalizing by how consistently each participant successfully replicates their own responses two weeks later.	64
4.2	The study platform and interface. Once recruited, my participants are routed to my custom-built platform. The interface includes several components: a) Participant sign-up page: Participants sign up with an ID and password of their choice. b) Avatar creator: Participants consent and create a 2-D sprite avatar to represent them in the study platform. c) Main interface displaying the study components: The modules include: 1) study consent, 2) avatar creation, 3) interview, 4) surveys and experiments, 5) self-consistency retake of the surveys and experiments. The modules only become available in order; the button to start a module becomes clickable once the participants have completed all previous modules. The self-consistency survey and experiment module only becomes available two weeks after the participants have completed the previous modules.	67

4.3	The architecture of the interviewer agent. It takes as input the participants' utterances and the interview script, generating the next action in the form of follow-up questions or deciding to move on to the next question module using a language model. A reflection module helps the architecture succinctly summarize and infer insights from the ongoing interview, enabling the agent to more effectively generate follow-up questions.	70
4.4	The interview interface. a) The main interview interface: A 2-D sprite representing the AI interviewer agent is displayed in a white circle that pulsates to match the level of the audio, visualizing the interviewer agent's speech during the AI interviewer's turn. b) Participant's response: The 2-D sprite of the AI interviewer agent changes into a microphone emoji when it is the participant's turn to respond, with the white circle pulsating to match the level of the participant's audio being captured. c) Progress bar and subtitles: A 2-D sprite map shows the participant's visual avatar traveling from one end point to the other in a straight line, indicating progress. The interface also features options to display subtitles or pause the interview.	74
4.5	The architecture of my generative agents involves taking a question as input and outputting a prediction of how the source participant might respond, using a language model. Each agent's memory comprises the interview transcript and the outputs of expert reflections over that transcript. These reflections are short syntheses generated using a language model, designed to infer insights about the participants that might not be explicitly stated. The personas of expert social scientists (e.g., psychologist, behavioral economist) guide these reflections. . .	80

4.6	Generative agents' predictive performance, and 95% confidence intervals. The consistency rate between participants and the predictive performance of generative agents is evaluated across various constructs and averaged across individuals. For the General Social Survey (GSS), accuracy is reported due to its categorical response types, while the Big Five personality traits and economic games report mean absolute error (MAE) due to their numerical response types. Correlation is reported for all constructs. Normalized accuracy is provided for all metrics, except for MAE, which cannot be calculated for individuals whose MAE is 0 (i.e., those who responded the same way in both phases). I find that generative agents predict participants' behavior and attitudes well, especially when compared to participants' own rate of internal consistency. Additionally, using interviews to inform agent behavior significantly improves the predictive performance of agents for both GSS and Big Five constructs, outperforming other commonly used methods in the literature.	100
4.7	Demographic Parity Difference (DPD) for generative agents across political ideology, race, and gender subgroups on three tasks: GSS (in percentages), Big Five, and economic games (in correlation coefficients). DPD represents the performance disparity between the most and least favored groups within each demographic category. Generative agents using interviews consistently show lower DPDs compared to those using demographic information or persona descriptions, suggesting that interview-based generative agents mitigate bias more effectively across all tasks. Gender-based DPDs remain relatively low and consistent across all conditions.	103
4.8	Demographic agents (top), persona agents (middle), interviewfull agents (bottom). The GSS-outcome questions we remove (x-axis) when calculating the normalized accuracy (y-axis) are ranked from most to least likely to have a verbatim answer to them somewhere in the interview transcript.	107

4.9	Demographic agents (top), persona agents (middle), interviewfull agents (bottom). The GSS-outcome questions we remove (x-axis) when calculating the normalized accuracy (y-axis) are ranked from most to least likely to be able to be answered through inference based on information somewhere in the interview transcript.	110
5.1	A) The generated content page of SimReddit styled after that of a subreddit. B) The “About Community” panel describing the goal of the community. C) The “Community Rules” panel describing the rules that the members are encouraged to follow. D) The button for instantiating Multiverse for outputting alternatives of how an interaction might play out	120
5.2	A) The interface of an example subreddit. B) A close up of the “About Community” panel. The content on this panel defines the goal and the target population of the community. C) A close up of the “Rules” panel. The content in this panel defines the rules within the community. . .	121
5.3	Examples of conversations produced by SimReddit’s Generate. The community goals and rules are from the participants in our Designer Evaluation. The conversations here were among those I presented to the respective participants.	122
5.4	The Interface for instantiating WhatIf. The designer can select an utterance to initiate the feature.	123
5.5	WhatIf can surface how a conversation might have developed if someone else (e.g., a troll) had responded or intervened (e.g., a moderator). . .	126
5.6	An illustration of conversations generated through Multiverse for a community for "connecting people moving to Los Angeles with locals." The orange lines show how a conversation could have progressed originally.	128

Chapter 1

Introduction

Human societies are complex systems. The institutions we design, the platforms we build, and the policies we enact are built on assumptions about how people perceive, decide, and act. When those assumptions are wrong, or even just incomplete, well-intentioned systems can yield unintended consequences at scale. This dissertation argues that simulation, and in particular simulations powered by generative, language-model-based agents, can serve as a new kind of scientific instrument for probing those assumptions, challenging prevailing explanations, and exploring counterfactual designs before they are deployed in the world.

To motivate this claim, consider a canonical example from the study of complex social systems: Thomas Schelling’s 1971 model of residential segregation (Schelling, 1971, 1978). In a simple grid world populated by two types of agents, each agent is content to live in a mixed neighborhood and requires only a minimal share of similar neighbors—say, 30%—to be satisfied. Agents who fall below this threshold move to the nearest acceptable location. Despite these weak individual preferences, the collective dynamics produce strong segregation. The striking lesson was not that actors are bigots; rather, that weak, local preferences can yield large-scale, unintended patterns. That insight reshaped how scholars and policymakers reasoned about segregation, shifting attention from overt animus to the emergent consequences of individually reasonable choices.

Schelling’s example illustrates two broader points. First, simulations can expose

how micro-level rules produce macro-level outcomes that defy intuition. Second, the force of those insights depends on the behavioral assumptions we encode. For decades, computational social science relied on highly stylized agents: a handful of fixed parameters stood in for human cognition, motivation, and social context. These models yielded elegant theory and valuable intuition, but their simplicity made it difficult to generalize across domains or populations, and easy to overlook the contingency and diversity that characterize real human behavior.

Recent advances in generative AI offer a new opportunity. Large language models (LLMs) are trained on broad corpora that reflect many facets of contemporary life—narratives, plans, conversations, and social practices. When leveraged appropriately, they can generate plausible descriptions of goals, beliefs, and actions across a wide range of backgrounds and settings. This suggests a path toward general-purpose computational agents that preserve the flexibility of natural language while remaining programmable, inspectable, and composable into large-scale simulations.

But two challenges stand in the way: First, long-term coherence. People do not act as isolated, stateless predictors of the next token in a conversation. We accumulate experiences; we recall, reinterpret, and integrate them; we plan and revise. Off-the-shelf language models, even with expanding context windows, cannot naively reason over the full arc of an agent’s lived experience. Without additional machinery, generated behavior drifts, forgets, and fragments. For simulations to be useful as instruments of inquiry, agents must remember, reflect, and plan over extended horizons, doing so efficiently and in ways that scale to many agents.

Second, accuracy and validation. Believability is not enough. If simulations are to inform consequential decisions in science, policy, and design, they must be empirically anchored. We need principled methods to construct agents that represent diverse, real individuals and to evaluate whether simulated attitudes and behaviors match ground-truth measures—and to know where they do not. Moreover, naïve conditioning on demographics risks stereotyping: a description like “30-year-old Asian graduate student” should not, by itself, license generic or biased inferences about daily life. The core methodological question is how to build and test agents that are faithful to individuals and populations rather than caricatures of them.

This dissertation tackles these challenges by developing architectures, methods, and measurement frameworks for generative agent simulations of human behavior. The central thesis is:

With the right agent architectures, generative language models can be transformed into coherent, empirically grounded computational agents, enabling simulations that both reveal emergent social dynamics and provide decision-relevant evidence for policy and design.

1.1 From Stylized Agents to Generative Agents

The work proceeds in a tradition that spans cognitive architectures, agent-based models, social AI, and virtual worlds. Prior models afforded clarity through parsimony: agents had few states and followed simple rules (Bruch and Atwell, 2015; Epstein and Axtell, 1996). The cost was rigidity. They struggled to capture the idiosyncratic goals and contextual reasoning that make human behavior contingent.

Generative agents invert this trade-off. By operating in natural language and being conditioned on rich prompts and experience, they can flexibly represent goals, beliefs, social ties, and routines—without hand-coding domain-specific rules. To make this practical, I introduce an agent architecture with three core capabilities:

1. **Memory.** Agents maintain a structured record of experiences (“episodic memory”) that includes observations, actions, and social interactions. Rather than carrying the entire record into every decision, agents retrieve only the most salient memories, using learned or heuristic relevance measures tailored to the current situation.
2. **Reflection.** Periodically, agents distill higher-level inferences from their episodic memories: evolving goals, social impressions, and personal preferences. These reflective summaries stabilize behavior across time and allow agents to update their self-models in light of new evidence.
3. **Planning.** Agents generate and revise plans at multiple time scales (e.g., daily routines, near-term tasks), grounding actions in both current context and

reflective summaries. Planning closes the loop between memory and action, providing a scaffold for coherent activity over hours or days.

Together, these components enable agents that can “lead a day in the life” without scripts: waking, preparing meals, going to work, conversing with neighbors, forming relationships, and coordinating activities—behaviors that, when scaled to communities, yield emergent social phenomena.

1.2 Accuracy Beyond Believability

A second thrust of the dissertation addresses accuracy. I develop methods to instantiate agents that represent real people and to measure the fidelity of simulated outcomes against ground truth. The approach has three pillars:

- **Grounded construction.** Agents are initialized from data collected from individuals (e.g., surveys, diaries, panel data), yielding personalized goals, routines, attitudes, and constraints. Demographics alone do not define the person; rich, self-reported context does.
- **Task-relevant validation.** Accuracy is assessed on outcomes that matter for the intended use: attitudinal measures (e.g., policy preferences) and behavioral choices (e.g., media selection). Benchmarks include held-out survey items, naturalistic tasks, and scenario-based decisions designed with social scientists.
- **Population representativeness.** By sampling and constructing agents from a representative population—rather than convenience samples—we can assess how well simulations reproduce aggregate patterns and heterogeneity across subgroups, not just individual plausibility.

This methodology underwrites the second half of the dissertation: simulating a population of 1,000 agents representing a cross-section of U.S. adults and evaluating the alignment between simulated and observed outcomes across a variety of settings.

1.3 A Motivating Lens: From Emergence to Design

Returning to the segregation example helps clarify the aspiration. In Schelling’s model, a single, interpretable preference rule yields a counterintuitive macro-pattern. Generative agents allow us to explore analogous questions with richer behavioral substrates. Instead of setting a single threshold for homophily, we can examine how preferences interact with daily routines, social ties, and evolving beliefs. We can ask: How do work schedules, transit options, and social media exposure shape contact opportunities and, ultimately, residential decisions? How robust are emergent patterns to policy interventions like inclusive zoning or mixed-income development? By treating the simulation as an instrument—probing mechanisms, testing counterfactuals, and surfacing assumptions—we can move from explanation to design.

Crucially, the goal is not to replace empirical observation but to complement it. Simulations are most useful when tethered to data and when their limitations are explicit. When a simulation reproduces an observed pattern, it offers a candidate mechanism; when it fails, it highlights missing assumptions. In both cases, it can guide measurement and theory, and it can help practitioners explore the design space before committing scarce resources.

Contributions

The dissertation makes the following contributions:

- **A general architecture for generative agents.** I introduce one of the first agent frameworks to leverage large language models for coherent, interactive agents with memory, reflection, and planning. I demonstrate believable, contextually grounded behavior in an interactive sandbox without domain-specific scripts.
- **A method for constructing empirically grounded agents.** I present protocols for initializing agents from real individuals’ data, avoiding stereotype-driven generalization and capturing idiosyncratic preferences and constraints relevant to downstream behavior.

- **A measurement framework for simulation accuracy.** I define task-relevant metrics and experimental designs that compare simulated outcomes to ground-truth benchmarks at both individual and population levels, including analyses of subgroup heterogeneity and robustness.
- **A large-scale population simulation.** I develop a simulation of 1,000 agents representing a geographically and demographically diverse sample of U.S. adults and evaluate its fidelity across attitudinal, behavioral, and social-interaction outcomes.
- **Design and policy case studies.** I illustrate how generative-agent simulations can inform platform design, content recommendation, and public policy by enabling safe exploration of “what-if” counterfactuals prior to real-world deployment. A featured application—social simulacra—shows how the framework can populate an envisioned community and surface design-relevant behaviors, including anti-social dynamics, prior to deployment.
- **Ethical considerations and practical guidance.** I analyze risks—including stereotype amplification, privacy, and misplaced trust in simulated evidence—and provide guardrails for responsible use, emphasizing transparency about assumptions, uncertainty, and scope of validity.

1.4 Overview of the Dissertation

The dissertation is organized in two parts that trace the arc from capability to credibility:

1. **Generative Agents: Interactive Simulacra of Human Behavior.** I introduce the agent architecture—memory, reflection, and planning—describe the interactive environment used to study unscripted daily life, and report qualitative and quantitative evidence of coherent individual behavior and emergent social patterns. This part demonstrates that, when properly scaffolded, generative language models can power agents that act believably over extended horizons.

2. **Generative Agent Simulations of 1,000 People.** I extend the architecture to construct agents grounded in data from a representative sample of U.S. adults. I develop the measurement framework for accuracy, evaluate simulated outcomes against ground truth across attitudinal, behavioral, and social-interaction tasks, and analyze where and why alignment succeeds or fails. Case studies illustrate how such simulations can inform policy and design decisions under uncertainty.

Following the two sections, I summarize the kinds of applications that simulations with generative agents may enable. The central application I highlight is *social simulacra*—a technique that generates agents interacting within a hypothetical social space, allowing designers to prototype communities at scale. Social simulacra provide a practical context for the main results of the dissertation: they demonstrate how generative agents can be harnessed to populate an envisioned system with realistic social behavior, surfacing both intended dynamics and potential failure modes before real-world deployment.

The concluding chapters synthesize lessons for the science of human behavior and for practical decision-making. They outline ethical guidelines, discuss limits and failure modes, and chart directions for future work—including richer multimodal environments, longitudinal evaluation, and hybrid approaches that combine mechanistic models with generative components—while drawing out implications for the design, deployment, and governance of social simulacra.

1.5 Scope, Limitations, and Ethics

Three caveats frame the contributions. First, simulated behavior is not reality. Generative agents can illuminate plausible mechanisms and provide decision-relevant evidence, but their outputs must be interpreted in light of their assumptions and training data. Second, the representativeness of population-level simulations depends on the quality and coverage of the data used to construct agents. When data are thin or biased, simulations inherit those flaws; the measurement framework is intended to reveal, not conceal, such gaps. Third, because language models are trained on human text, they can encode and amplify social biases. This dissertation therefore prioritizes

construction methods that ground agents in individual-level data, validation protocols that detect disparities, and reporting practices that make uncertainty and limitations explicit.

In sum, this dissertation develops a principled path from today’s generative models to tomorrow’s decision support: agents that remember, reflect, and plan; simulations that are anchored to the people they purport to represent; and measurement that makes accuracy, uncertainty, and bias visible. By treating simulations as instruments for explanation and design—rather than as oracles—we can use them to challenge assumptions, generate testable hypotheses, and explore counterfactuals that would be costly or unethical to probe in the real world. The hope is not to build perfect mirrors of our society, but to craft tools that help us see more clearly—and act more wisely—within it.

Chapter 2

Background and Related Work

Designing socio-technical systems is fundamentally hard. Seemingly small interface or policy choices can cascade into large, sometimes unintended, collective outcomes; interventions that work at pilot scale often fail at population scale; and harms from anti-social behavior can emerge quickly and unevenly across communities. A core premise of this dissertation is that we need methods that let designers and researchers *reason about social dynamics before deployment*, and that recent advances in large language models (LLMs) make it possible to prototype with *believable* populations of agents whose behaviors can be inspected, steered, and evaluated.

This chapter situates the work within four strands of scholarship and practice: (1) agent-based modeling (ABM) and computational social science, which formalize how micro-level rules yield macro-level phenomena; (2) human–AI interaction and prototyping methods that help designers think through socio-technical dynamics prior to release; (3) the long tradition of building *believable* agents in games and cognitive architectures; and (4) the emergence of LLM-powered agents that can represent, plan, and act via natural language. Across these literatures, I identify a persistent gap: existing approaches either offer *interpretability without behavioral breadth* (e.g., stylized ABMs; rule-based NPCs) or *breadth without validation* (e.g., unconstrained LLM role-play). This dissertation addresses that gap by developing architectures and evaluation protocols for *coherent, empirically grounded generative agents*, and by demonstrating their use as populated prototypes for social systems.

2.1 Why This Problem Matters: Social Systems Are Hard to Design

How might, for instance, a diverse set of individuals respond to new public health policies and messages, react to product launches, or adapt to major shocks? When simulated individuals are assembled into collectives, such simulations can help pilot interventions, develop complex theories that capture nuanced causal and contextual interactions, and expand our understanding of structures such as institutions and networks across domains including economics (Axtell, 2000), sociology (Schelling, 1971), organizations (Carley, 1992), and political science (Lustick, 2002).

Social systems must function amidst interdependent participants, evolving norms, and strategic behavior. Designers must therefore anticipate not only a single user’s experience but coupled dynamics among many users, where edge cases are plentiful and impacts are uneven (Tapscott and Caston, 1992; Grudin, 1994; Ackerman, 2000). Failures to anticipate undesirable behavior—trolling, hate speech, inflammatory discourse, and other forms of harm—can render designs ineffective (Hardaker, 2010; Kayany, 1998; Donovan, 2019; Chandrasekharan et al., 2017b; Park et al., 2022b; Chancellor et al., 2016; Chandrasekharan et al., 2017a; Cheng et al., 2015; Sood et al., 2012; Kraut and Resnick, 2012) and impose costs on individuals and communities (Vogels, 2021; Akbulut et al., 2010; Cheng et al., 2017).

Prototyping social systems is notoriously difficult: achieving the critical mass needed to observe emergent dynamics is a chicken-and-egg problem; recruiting realistic populations is expensive and unreliable; and live A/B testing—largely feasible only for very large platforms—can expose users to unproven designs and erode trust (Bernstein et al., 2011; Grudin, 1994; Markus, 1987; Ugander et al., 2013; Xu et al., 2015; Kohavi et al., 2012; Kramer et al., 2014; Fiske, 2022; Flick, 2016; Hallinan et al., 2020). These challenges motivate new techniques that let designers *populate* early-stage systems, probe failure modes, and iterate before deployment.

2.2 Prototyping in Design Practice

Prototyping turns ideas into artifacts that can be manipulated, critiqued, and tested. In HCI, a prototype is “a concrete representation of part or all of an interactive system” (Beaudouin-Lafon and Mackay, 2002). Unlike textual specifications, prototypes make interaction visible and negotiable, enabling quicker iteration and better communication across stakeholders (Schön, 1984; Rittel and Webber, 1973; Beaudouin-Lafon and Mackay, 2002). Low-fidelity prototypes—from paper mockups to Wizard-of-Oz setups—support fast exploration of questions about flow, breakdowns, and opportunities (Snyder, 2003; Sefelin et al., 2003; Rettig, 1994; Kelley, 1983, 1984; Houde and Hill, 1997). Tool support spans making sketches interactive (Landay, 1996), exploring alternatives (Marks et al., 1997; Hartmann et al., 2009; O’Donovan et al., 2015), and reducing implementation effort (Hartmann et al., 2007; Drew et al., 2016; Hartmann et al., 2006; Savage et al., 2013).

When the target is a *social* system, prototypes must also stand in for user behavior itself, with the goal of understanding how a social system will behave when it is populated. But populating social computing systems—particularly those designed to support interactions among large numbers of people—is often an insurmountable yet necessary challenge because cold-start dynamics differ markedly from those after the space reaches critical mass (Bernstein et al., 2011). As theorized by Grudin (Grudin, 1994), organically reaching the critical mass of users necessary to study the system is difficult because many social computing systems are not useful enough to attract users when the user population is small, a problem shared by other interactive media (Markus, 1987). Although designers can recruit test users through social media or crowdworking platforms to bootstrap usage, and apply strategies such as *piggyback prototyping* that utilize existing tools (e.g., Google Docs, text messaging, email) in place of a bespoke system (Grevet and Gilbert, 2015), securing sustained, realistic participation remains challenging. Systems that already have a large user population (e.g., Facebook, Twitter) can prototype new features via A/B testing or country-level comparisons (Ugander et al., 2013; Xu et al., 2015; Kohavi et al., 2012), but such approaches are viable only for a small number of large platforms and are mostly suited

to minor post-launch tweaks. Finally, these approaches often expose untested designs to real users who may be unaware of ongoing experiments, risking harm and eroding trust (Kramer et al., 2014; Fiske, 2022; Flick, 2016; Hallinan et al., 2020).

As Grevet and Gilbert observed, prototyping contributions are 20 times less common in the social computing literature than in the broader HCI literature (Grevet and Gilbert, 2015). Social simulacra offer a general technique to address this scarcity in social computing design: they focus on generating content and replies that might arise in the system, enabling designers to reflect on which behaviors their designs are prepared to handle. I argue that without such means, we will continue to risk releasing systems that are susceptible to unexpected failures.

2.2.1 Challenges in Prototyping Social Systems

Social prototypes must grapple with cold-start dynamics, network effects, and adversarial behavior. Early-stage evaluations often under-represent the behaviors that matter at scale, making results fragile (Bernstein et al., 2011; Grudin, 1994; Markus, 1987). While large platforms can experiment *in vivo* via A/B tests or staggered rollouts (Ugander et al., 2013; Xu et al., 2015; Kohavi et al., 2012), such tests are inaccessible to most teams and can raise ethical concerns when users are unknowingly exposed to risky interventions (Kramer et al., 2014; Fiske, 2022; Flick, 2016; Hallinan et al., 2020). These limitations motivate *in silico* populated prototypes that (i) are cheap and repeatable, (ii) support stress-testing and what-if analysis, and (iii) can be inspected and audited.

2.3 Agent-Based Modeling and Computational Social Science

ABMs represent a population as a collection of heterogeneous agents whose local rules produce aggregate patterns (Epstein and Axtell, 1996; Bruch and Atwell, 2015; Schelling, 1971; Axtell, 2000). Schelling’s segregation model famously showed that

mild preferences can yield strong segregation (Schelling, 1971), illustrating how un-intuitive macro-patterns can emerge from simple micro-mechanisms. ABMs have illuminated cooperation and norm formation (Carley, 1992), organizational and market dynamics (Bonabeau, 2002), and policy-relevant phenomena from epidemics to urban planning (Lustick, 2002). Methodologically, they emphasize mechanism—explanation via generative sufficiency—with transparent rules and parameters (Epstein and Axtell, 1996).

These virtues come with costs. Stylized rules—finite-state heuristics, thresholds, and hand-specified utilities—capture narrow slices of cognition and context (Macy and Willer, 2002). As a result, models may struggle to generalize across domains or reflect the idiosyncrasies that shape real decisions. A rich validation literature addresses this through pattern-oriented modeling (Grimm and Railsback, 2005), multi-criteria fit (Windrum et al., 2007), and comparisons to empirical distributions (Railsback and Grimm, 2019). Microsimulation and discrete-choice models similarly encode individual decision rules to reproduce aggregates (McFadden, 1974), but require domain-specific engineering that can flatten heterogeneity.

This dissertation inherits ABM’s *simulation-first* lens while exploring a different point in the design space: agents whose behavioral substrate is expressed and manipulated in *natural language*. This shift raises opportunities (richer internal state, cross-domain reuse) and obligations (empirical validation, representativeness), which I address through architecture and evaluation.

2.4 Human–AI Interaction and Prototyping for Socio-Technical Systems

Interactive artificial intelligence systems aim to combine human insights and capabilities in computational artifacts that can augment their users (Amershi et al., 2014; Fails and Olsen Jr, 2003). A long line of work has explored ways to enable users to interactively specify model behavior. For instance, Crayons demonstrated an early vision of interactive machine learning, allowing non-expert users to train classifiers (Fails

and Olsen Jr, 2003). Further work helped articulate how end users might describe their classification goals to the system through examples (Fogarty et al., 2008) or demonstration (Fiebrink and Cook, 2010). Recent advancements have extended these explorations to deep learning (Lam et al., 2023) and prompt-based authoring (Jiang et al., 2022; Wu et al., 2021b; Liu et al., 2022).

Meanwhile, a persistent thread of research has advanced the case for language- and agent-based interaction in human-computer interaction. Formative work such as SHRDLU (Winograd, 1971) and ELIZA (Weizenbaum, 1966) demonstrated both the opportunities and the risks of natural language interaction with computing systems. As research progressed, it became evident that autonomous agents could offer new metaphors for delegation and interaction (Maes, 1995), but the boundaries of delegation between humans and agents have remained the subject of ongoing debate and refinement (Shneiderman and Maes, 1997; Horvitz, 1999; Shneiderman, 2022). Recently, the technology has reached a level of stability that enables agents to interact via natural language in large and complex online social environments (e.g., (Krishna et al., 2022)). Natural language interaction provides a modality that can enhance user abilities in domains such as photo editing (Linder et al., 2013; Fourney et al., 2011; Adar et al., 2014) and code editing (Rong et al., 2016).

This dissertation convenes these threads to show that we can now create agents that proxy human behavior for interactive systems, and interact with them using natural language. In doing so, this work reopens the door to examining foundational HCI questions around cognitive models such as GOMS and the Keystroke-Level Model (KLM) (Card et al., 1983, 1980), around prototyping tools (Park et al., 2022a), and around ubiquitous computing applications (Weiser, 1999; Dey, 2001; Fast et al., 2016).

2.4.1 Believable Proxies of Human Behavior

Prior literature has described *believability*, or *believable agents*, as a central design and engineering goal. Believable agents are designed to provide an illusion of life and present a facade of realism in the way they appear to make decisions and act on their own volition, similar to characters in Disney movies (Bates, 1994; Thomas and

Johnston, 1981). These agents can populate and perceive an open-world environment like the one we inhabit (Laird and VanLent, 2001; Bates, 1994), and strive to behave in ways that exhibit emergent behaviors grounded in social interactions with users or other agents, with the aim of becoming believable proxies of our behavior in hypothetical simulations of individuals and communities (McCoy et al., 2012; Burkinshaw, 2009; Francis, 2010). Historically, these agents were developed in the context of intelligent game non-player characters (NPCs) (Laird and VanLent, 2001; Riedl, 2012). Creating NPCs with believable behavior, if possible, can enhance player experiences in games and interactive fiction by enabling emergent narratives (Swartout et al., 2006; Aylett, 1999; Brenner, 2010; Ibister and Nass, 2000) and social interactions with the agents (Zubek, 2002). More importantly, game worlds provide increasingly realistic representations of real-world affordances, and, as observed by Laird and van Lent in 2001, these simulated worlds offer accessible testbeds for developers of believable agents to finesse the agents’ cognitive capabilities without worrying about implementing robotics in the real world or creating simulation environments from scratch (Laird and VanLent, 2001; Riedl, 2012).

A diverse set of approaches to creating believable agents has emerged over the past four decades. In implementation, however, these approaches often simplified the environment or dimensions of agent behavior to make the effort more manageable (Brooks et al., 2000; Minsky and Papert, 1970). Rule-based approaches, such as finite-state machines (Siu et al., 2021; Umarov et al., 2012) and behavior trees (Knafla, 2011; Pillosu, 2009; Hecker, 2011), account for the brute-force approach of human-authoring the agent’s behavior (McCoy et al., 2012). They provide a straightforward way of creating simple agents that remains the dominant approach today (McCoy et al., 2009; Miyashita et al., 2017; Yannakakis, 2012), and can even handle rudimentary social interactions, as shown in games such as *Mass Effect* (BioWare, 2007) and *The Sims* (Arts, 2009). Nonetheless, manually crafting behavior that comprehensively addresses the breadth of possible interactions in an open world is untenable. The resulting agent behaviors may not fully represent the consequences of their interactions (McCoy et al., 2012, 2011a,b), and cannot perform new procedures that were not hard-coded in their scripts (Siu et al., 2021; Umarov et al., 2012). By contrast,

prevalent learning-based approaches for creating believable agents, such as reinforcement learning, have overcome the challenge of manual authoring by allowing agents to learn their behavior, achieving superhuman performance in recent years in games such as AlphaStar for *StarCraft* (Vinyals et al., 2019) and OpenAI Five for *Dota 2* (Berner et al., 2019). However, their success has largely taken place in adversarial games with readily definable rewards that a learning algorithm can optimize. They have not yet addressed the challenge of creating believable agents in open worlds (Siu et al., 2021; Miyashita et al., 2017; Hausknecht et al., 2020).

Cognitive architectures in computation, pioneered by Newell, aimed to build the infrastructure for supporting a comprehensive set of cognitive functions (Newell, 1990) suited to the all-encompassing nature of believable agents in the original vision. They fueled some of the earliest examples of believable agents. For instance, Quakebot-SOAR (Laird, 2000) and ICARUS (Langley et al., 2005; Choi et al., 2021) generated NPCs in first-person shooter games, while TacAir-SOAR (Pew and Mavor, 1998) generated pilots in aerial combat training simulations. The architectures used by these agents differed (Quakebot- and TacAir-SOAR relied on SOAR (Laird, 2012), while ICARUS relied on its own variation inspired by SOAR and ACT-R (Anderson, 1993)), but they shared the same underlying principle (Laird et al., 2017). They maintained short-term and long-term memories, filled these memories with symbolic structures, and operated in perceive–plan–act cycles, dynamically perceiving the environment and matching it with one of the manually crafted action procedures (Umarov et al., 2012; Laird, 2001). Agents created using cognitive architectures aimed to be generalizable to most, if not all, open-world contexts and exhibited robust behavior for their time. However, their space of action was limited to manually crafted procedural knowledge, and they did not offer a mechanism through which the agents could be inspired to seek new behavior. As such, these agents were deployed mostly in non–open-world contexts such as first-person shooter games (Laird, 2000; Choi et al., 2021) or blocks worlds (Langley et al., 2005).

Today, creating believable agents as described in the original definition remains an open problem (Yannakakis, 2012; Riedl, 2012). Many have moved on, arguing that although current approaches for creating believable agents might be cumbersome and

limited, they are good enough to support existing gameplay and interactions (Yannakakis, 2012; Champandard, 2012; Nareyek, 2007). My argument is that large language models offer an opportunity to re-examine these questions, provided that we can craft an effective architecture to synthesize memories into believable behavior. I offer a step toward such an architecture in this dissertation.

2.4.2 Large Language Models and Human Behavior

Generative agents leverage a large language model to power their behavior. The key observation is that large language models encode a wide range of human behavior from their training data (Brown et al., 2020b; Bommasani et al., 2022). If prompted with a narrowly defined context, the models can be used to generate believable behavior. Recent work has demonstrated the efficacy of this approach. For instance, social simulacra used a large language model to generate users that would populate new social computing systems to prototype their emergent social dynamics (Park et al., 2022a). This approach used a prompt chain (Wu et al., 2021b, 2022) to generate short natural language descriptions of personas and their behaviors as they appear in the system being prototyped. Other empirical studies have replicated existing social science studies (Horton, 2023), political surveys (Sorensen et al., 2022), and generated synthetic data (Hämäläinen et al., 2023). Large language models have also been used to generate interactive human behavior for users to engage with. In gaming, for instance, these models have been employed to create interactive fiction (Freiknecht and Effelsberg, 2020) and text-adventure games (Callison-Burch et al., 2022). With their ability to generate and decompose action sequences, large language models have also been used in planning robotics tasks (Huang et al., 2022). For example, when presented with a task such as picking up a bottle, the model is prompted to break down the task into smaller action sequences, such as heading to the table where the bottle is located and picking it up.

I posit that, based on the work summarized above, large language models can become a key ingredient for creating believable agents. The existing literature

largely relies on what could be considered first-order templates that employ few-shot prompts (Gao et al., 2020; Liu et al., 2021) or chain-of-thought prompts (Wei et al., 2023). These templates are effective in generating behavior conditioned solely on the agent’s current environment (e.g., how a troll would respond to a given post; what actions a robot would need to take to enter a room given that there is a door). However, believable agents require conditioning not only on their current environment but also on a vast amount of past experience, which is a poor fit (and, as of today, impossible due to the underlying models’ limited context window) for first-order prompting. Recent studies have attempted to go beyond first-order prompting by augmenting language models with a static knowledge base and an information retrieval scheme (Khattab et al., 2023) or with a simple summarization scheme (Wu et al., 2021a). This dissertation extends these ideas to craft an agent architecture that handles retrieval where past experience is dynamically updated at each time step and mixed with agents’ current context and plans, which may either reinforce or contradict one another.

2.5 From Stylized ABMs to Generative, Language-Model Agents

Viewed through the ABM lens, LLM-based agents shift where assumptions live. Traditional models encode behavior as explicit rules and utilities; generative agents externalize structure into natural-language state and reasoning shaped by training data and prompts. This increases expressive capacity (richer internal state, cross-domain reuse) but decreases a priori interpretability. The resulting obligation is empirical discipline: (i) anchoring agents in individual-level data when appropriate; (ii) testing alignment not only on averages but across diverse tasks and subgroups; and (iii) reporting uncertainty and failure modes.

General-purpose simulation of human attitudes and behavior can create a laboratory for exploring interventions across domains: public health messaging, product launches, institutional changes, and responses to shocks. When individuals are composed into

collectives and embedded in networks and institutions, we can probe mechanisms, pilot interventions, and study cross-context generalization using tools from economics, sociology, organizational science, and political science—without exposing real users to unvetted designs. This dissertation operationalizes that vision by (a) designing LLM-based agent architectures with memory, reflection, and planning suitable for interactive, multi-agent settings; (b) using populated prototypes to support *design reasoning* about socio-technical systems (Park et al., 2022a); and (c) adopting validation protocols inspired by ABM and the social sciences (e.g., distributional fit, behavioral games, randomized interventions) to assess fidelity and limits.

2.6 Validity, Representativeness, and Ethical Considerations

Using LLM proxies for people introduces risks distinct from conventional ABMs and from human-subjects studies. First, simulated groups can *caricature* real populations or collapse identity into stereotypes; measurement must go beyond replication of average treatment effects to distributional and subgroup fidelity (Park et al., 2022a). Second, long-context limitations and brittle memory can yield incoherent life histories or drift (Wu et al., 2021a). Third, fairness and representativeness concerns familiar from ML—including sampling bias and construct validity—carry over to agent populations and must be surfaced and mitigated via dataset documentation, auditing, and task design (Park et al., 2022a). Finally, populated prototypes are not a substitute for user research or governance; rather, they are a complementary instrument for early risk identification, hypothesis generation, and design iteration prior to any user exposure (Kramer et al., 2014; Fiske, 2022; Flick, 2016; Hallinan et al., 2020).

2.7 Summary and Positioning

Across ABM, HCI prototyping, game AI, and LLMs, the literature points to a clear opportunity: *populate* early-stage social systems with believable, inspectable, and

steerable agents to reason about dynamics before deployment. Traditional ABMs provide mechanism and interpretability but struggle with behavioral breadth; HCI prototyping provides iteration speed but lacks populations; game AI offers design patterns for believability but relies on hand-authored scripts; and LLMs offer behavioral richness without guarantees. This dissertation integrates these strands by developing LLM-based generative agents with explicit memory, reflection, and planning; by using them as *populated prototypes* for socio-technical design; and by introducing empirical evaluations that tie agent behavior to established social-science measures. In doing so, it aims to make the gap in the literature—between *stylized* models and *deployable* social systems—both legible and smaller.

Chapter 3

Generative Agents: Interactive Simulacra of Human Behavior



Fig. 3.1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, I demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

I argue that simulations can serve as instruments for reasoning about behavior and design. This section develops the first building block of that instrument: generative agents—computational agents powered by large language models (LLMs) that exhibit coherent, interactive, and context-sensitive behavior over extended periods. The goal in this chapter is believability and long-horizon coherence: to show that agents can act in ways that are consistent with their past experiences, responsive to their environments (including other agents and human users), and stable enough to sustain social dynamics that unfold over hours or days. In the next chapter, I will extend this capability to empirical alignment with real individuals.

3.1 What I mean by "generative agents"

Believable proxies of human behavior have been a longstanding aspiration across human–computer interaction, cognitive modeling, virtual environments, and game AI. Designers want non-player characters who can navigate relationships and routines without brittle scripts; researchers want simulated societies to probe mechanisms and generate hypotheses; practitioners want rehearsal spaces for difficult interpersonal interactions and prototyping tools for social products. The barrier has not been whether we can script a plausible response at a single moment, but whether we can sustain coherent behavior for agents embedded in ongoing experience: remembering what happened, drawing useful inferences, revising goals, and coordinating with others as circumstances change.

LLMs offer a new substrate for such agents. Trained on broad, socially inflected corpora, they can produce rich, context-dependent narratives of goals, beliefs, and actions. However, naively prompting an LLM at every timestep fails on two counts: it is inefficient, and it does not endow the model with a durable sense of “self” or “history.” The central design problem is therefore architectural: how to scaffold an LLM so that an agent can remember, reflect, and plan—recursively and efficiently—while remaining interactive and programmable.

In this dissertation, *generative agents* are software agents that draw on generative models to simulate believable human behavior—and demonstrate that they produce

believable simulacra of both individual and emergent group behavior. Generative agents draw a wide variety of inferences about themselves, other agents, and their environment; they create daily plans that reflect their characteristics and experiences, act out those plans, react, and re-plan when appropriate; they respond when the end user changes their environment or commands them in natural language. For instance, generative agents turn off the stove when they see that their breakfast is burning, wait outside the bathroom if it is occupied, and stop to chat when they meet another agent they want to talk to.¹ A society full of generative agents is marked by emergent social dynamics where new relationships are formed, information diffuses, and coordination arises across agents.

To enable generative agents, I describe an agent architecture that stores, synthesizes, and applies relevant memories to generate believable behavior using a large language model. My architecture comprises three main components. The first is the *memory stream*, a long-term memory module that records, in natural language, a comprehensive list of the agent’s experiences. A memory retrieval model combines relevance, recency, and importance to surface the records needed to inform the agent’s moment-to-moment behavior. The second is *reflection*, which synthesizes memories into higher-level inferences over time, enabling the agent to draw conclusions about itself and others to better guide its behavior. The third is *planning*, which translates those conclusions and the current environment into high-level action plans and then recursively into detailed behaviors for action and reaction. These reflections and plans are fed back into the memory stream to influence the agent’s future behavior.

3.1.1 An interactive society in a sandbox world

To demonstrate these ideas, I instantiate twenty-five generative agents in a small, Sims-inspired town. The environment supplies places (homes, shops, a school), objects, and affordances; time advances on a shared clock; and agents’ actions update the

¹When referring to generative agents engaging in actions or going to places, this is a shorthand for readability and not a suggestion that they are engaging in human-like agency. The behaviors of my agents, akin to animated Disney characters, aim to create a sense of believability, but they do not imply genuine agency.

shared world state. Human users can observe and interact with any agent or with the environment directly using natural language.

This setting makes it possible to test whether high-level social phenomena emerge from local decisions. A single seed—informing one agent that she wants to host a Valentine’s Day party—propagates through the town via everyday behavior: she plans, extends invitations during conversations, invitees remember or forget, decide whether to attend, ask others to join them, and coordinate arrivals. Despite many potential points of failure (memory lapses, conflicting plans, social misfires), the system produces the desired gathering at the right time and place, along with side effects (new acquaintances, budding romances) that were not explicitly scripted.

I conducted two evaluations of generative agents: a controlled evaluation to test whether the agents produce believable individual behaviors in isolation, and an end-to-end evaluation where the agents interacted with each other in open-ended ways over two days of game time to understand their stability and emergent social behaviors. In the technical evaluation, I leverage a methodological opportunity to evaluate an agent’s knowledge and behavior by “interviewing” it in natural language to probe the agents’ ability to stay in character, remember, plan, react, and reflect accurately. I compared several ablations that limit agents’ access to memory, reflection, and planning. I observe that each of these components is critical to strong performance across these interview tasks. Across the technical and end-to-end evaluation, the most common errors arose when the agent failed to retrieve relevant memories, fabricated embellishments to the agent’s memory, or inherited overly formal speech or behavior from the language model.

In sum, this paper makes the following contributions:

- *Generative agents*, believable simulacra of human behavior that are dynamically conditioned on agents’ changing experiences and environment.
- A novel architecture that makes it possible for generative agents to remember, retrieve, reflect, interact with other agents, and plan through dynamically evolving circumstances. The architecture leverages the powerful prompting capabilities of large language models and supplements those capabilities to

support longer-term agent coherence, the ability to manage dynamically evolving memory, and recursively produce higher-level reflections.

- Two evaluations, a controlled evaluation and an end-to-end evaluation, that establish causal effects of the importance of components of the architecture, as well as identify breakdowns arising from, e.g., improper memory retrieval.
- Discussion of the opportunities and ethical and societal risks of generative agents in interactive systems. I argue that these agents should be tuned to mitigate the risk of users forming parasocial relationships, logged to mitigate risks stemming from deepfakes and tailored persuasion, and applied in ways that complement rather than replace human stakeholders in design processes.

3.2 Generative Agent Behavior and Interaction

To illustrate the affordances of generative agents, I instantiate them as characters in a simple sandbox world reminiscent of *The Sims* (Arts, 2009). This sprite-based sandbox game world, *Smallville*, evokes a small town environment. In this section, I will walk through the affordances and interactions with generative agents in *Smallville* and describe how the agents behave within it. Then, in Section 3.3, I will introduce the generative agent architecture that powers these affordances and interactions. In Section 3.4, I will describe the implementation of the sandbox environment and how the agents interact with the underlying engine of the sandbox world.

3.2.1 Agent Avatar and Communication

A community of 25 unique agents inhabits *Smallville*. Each agent is represented by a simple sprite avatar. I authored one paragraph of natural language description to depict each agent’s identity, including their occupation and relationship with other agents, as seed memories. For example, John Lin has the following description:

John Lin is a pharmacy shopkeeper at the Willow Market and
Pharmacy who loves to help people. He is always looking for

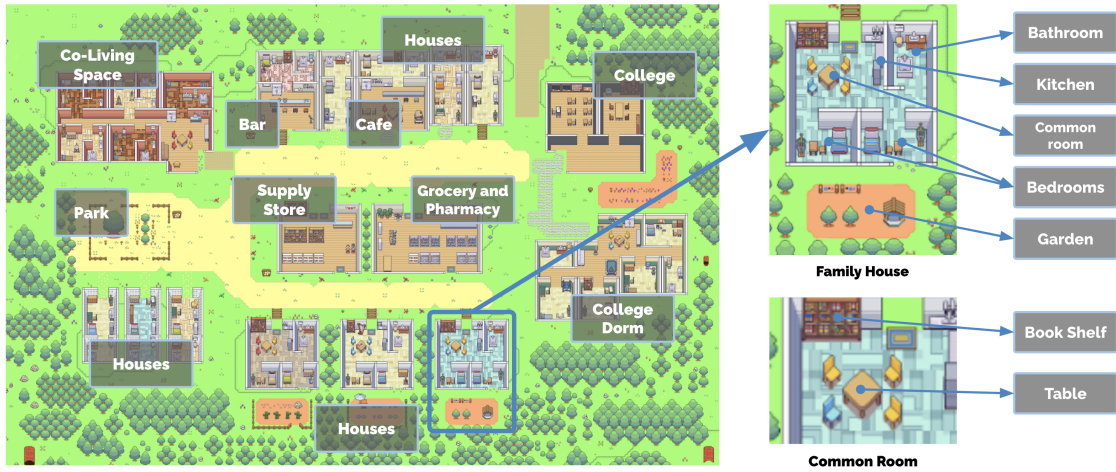


Fig. 3.2: The Smallville sandbox world, with areas labeled. The root node describes the entire world, children describe areas (e.g., houses, cafe, stores), and leaf nodes describe objects (e.g., table, bookshelf). Agents remember a subgraph that reflects the parts of the world they have seen, maintaining the state of those parts as they observed them.

ways to make the process of getting medication easier for his customers; John Lin is living with his wife, Mei Lin, who is a college professor, and son, Eddy Lin, who is a student studying music theory; John Lin loves his family very much; John Lin has known the old couple next-door, Sam Moore and Jennifer Moore, for a few years; John Lin thinks Sam Moore is a kind and nice man; John Lin knows his neighbor, Yuriko Yamamoto, well; John Lin knows of his neighbors, Tamara Taylor and Carmen Ortiz, but has not met them before; John Lin and Tom Moreno are colleagues at The Willow Market and Pharmacy; John Lin and Tom Moreno are friends and like to discuss local politics together; John Lin knows the Moreno family somewhat well-the husband Tom Moreno and the wife Jane Moreno.

Each semicolon-delimited phrase is entered into the agent’s initial memory as memories at the start of the simulation.

Inter-Agent Communication

The agents interact with the world by their actions, and with each other through natural language. At each time step of the sandbox engine, the agents output a natural language statement describing their current action, such as “Isabella Rodriguez is writing in her journal”, “Isabella Rodriguez is checking her emails”, “Isabella Rodriguez is talking with her family on the phone”, or “Isabella Rodriguez is getting ready for bed.” This statement is then translated into concrete movements that affect the sandbox world. The action is displayed on the sandbox interface as a set of emojis, providing an abstract representation of the action from an overhead view. To achieve this, the system utilizes a language model to translate the action into a set of emojis, which appear above each avatar’s head in a speech bubble. For example, “Isabella Rodriguez is writing in her journal” is displayed as 📖✍️, while “Isabella Rodriguez is checking her emails” appears as 💻✉️. The complete natural language description of the action can be accessed by clicking on the agent’s avatar.

Agents communicate with each other in full natural language. They are aware of other agents in their local area, and the generative agent architecture determines whether they walk by or engage in conversation. Here, a sample in the middle of a conversation between the agents Isabella Rodriguez and Tom Moreno about the upcoming election:²

Isabella: I’m still weighing my options, but I’ve been discussing the election with Sam Moore. What are your thoughts on him?

Tom: To be honest, I don’t like Sam Moore. I think he’s out of touch with the community and doesn’t have our best interests at heart.

User Controls

The user communicates with the agent through natural language by specifying a persona that the agent should perceive them as. For example, if the user specifies that

²I note that the conversational style of these agents can feel overly formal, likely a result of instruction tuning in the underlying models. I expect that the writing style will be better controllable in future language models.

they are a news “reporter” and asks about the upcoming election by saying, “Who is running for office?”, the John agent replies:

John: My friends Yuriko, Tom and I have been talking about the upcoming election and discussing the candidate Sam Moore. We have all agreed to vote for him because we like his platform.

To directly command one of the agents, the user takes on the persona of the agent’s “inner voice”—this makes the agent more likely to treat the statement as a directive. For instance, when told “You are going to run against Sam in the upcoming election” by a user as John’s inner voice, John decides to run in the election and shares his candidacy with his wife and son.

3.2.2 Environmental Interaction

Smallville features the common affordances of a small village, including a cafe, bar, park, school, dorm, houses, and stores. It also defines subareas and objects that make those spaces functional, such as a kitchen in a house and a stove in the kitchen (Figure 3.2). All spaces serving as agents’ primary living quarters feature a bed, desk, closet, shelf, as well as a bathroom and a kitchen.³

Agents move around Smallville as one would in a simple video game, entering and leaving buildings, navigating its map, and approaching other agents. Agent movements are directed by the generative agent architecture and the sandbox game engine: when the model dictates that the agent will move to a location, I calculate a walking path to the destination in the Smallville environment, and the agent begins moving. In addition, users can also enter the sandbox world of Smallville as an agent operating within it. The agent that the user embodies can be an agent already present in the world, such as Isabella and John, or it can be an outside visitor with no prior history in Smallville. The inhabitants of Smallville will treat the user-controlled agent no differently than they treat each other. They recognize its presence, initiate interactions, and remember its behavior before forming opinions about it.

³This environment design is not the focus of my work, so I generated this environment manually, not automatically. Future work can continue to expand the richness of the agents’ environments.

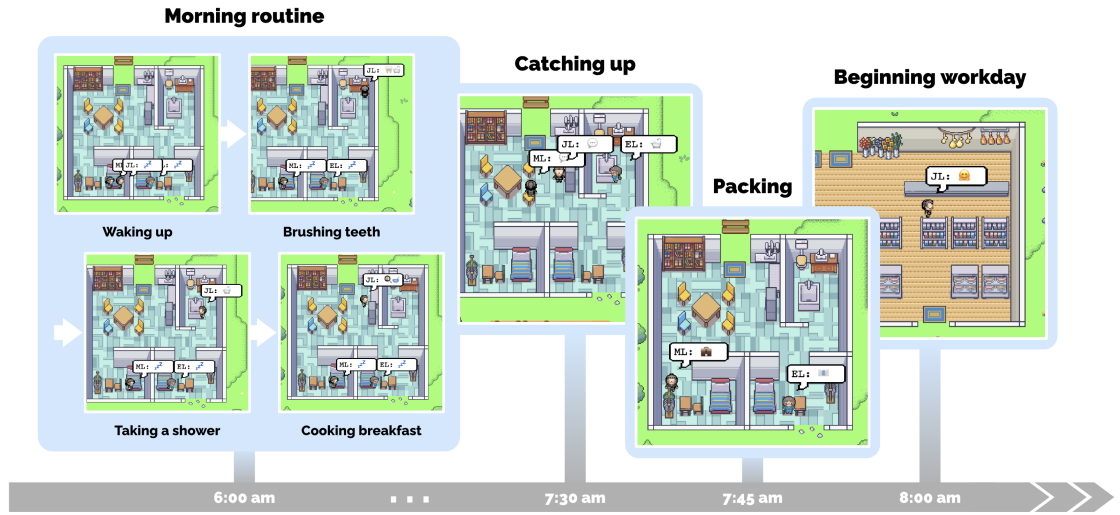


Fig. 3.3: A morning in the life of a generative agent, John Lin. John wakes up around 6 am and completes his morning routine, which includes brushing his teeth, taking a shower, and eating breakfast. He briefly catches up with his wife, Mei, and son, Eddy, before heading out to begin his workday.

Users and agents can influence the state of the objects in this world, much like in sandbox games such as *The Sims*. For example, a bed can be occupied when an agent is sleeping, and a refrigerator can be empty when an agent uses up the ingredients to make breakfast. End users can also reshape an agent’s environment in *Smallville* by rewriting the status of objects surrounding the agent in natural language. For instance, when Isabella is making breakfast in the morning, the user can change the status of the kitchen stove from “turned on” to “burning” by inputting a command to the system that chooses the object and illustrates its new status, like this: “<Isabella’s apartment: kitchen: stove> is burning.” Isabella will notice this in the next moment and go to turn off the stove and remake her breakfast. Likewise, if the user sets the status of Isabella’s shower to “leaking water” when she enters the bathroom, she will gather tools from her living room and try to fix the leak.



Fig. 3.4: At the beginning of the simulation, one agent is initialized with an intent to organize a Valentine’s Day party. Despite many possible points of failure in the ensuing chain of events—agents might not act on that intent, might forget to tell others, might not remember to show up—the Valentine’s Day party does, in fact, occur, with a number of agents gathering and interacting.

3.2.3 Example “Day in the Life”

Starting from the single-paragraph description, generative agents begin planning their days. As time passes in the sandbox world, their behaviors evolve as these agents interact with each other and the world, building memories and relationships, and coordinating joint activities.

I demonstrate the behavior of generative agents by tracing the output of my system over the course of one day for the agent John Lin (Figure 3.3). In the Lin family, John is the first to wake up at 7 a.m. He brushes his teeth, takes a shower, gets dressed, eats breakfast, and checks the news at the dining table in the living room. At 8 a.m., Eddy follows suit, rushing out of bed to prepare for his classes. He catches up with John just as he is about to head out:

John: Good morning Eddy. Did you sleep well?

Eddy: Good morning dad. Yeah, I slept great.

John: That’s good. What are you working on today?

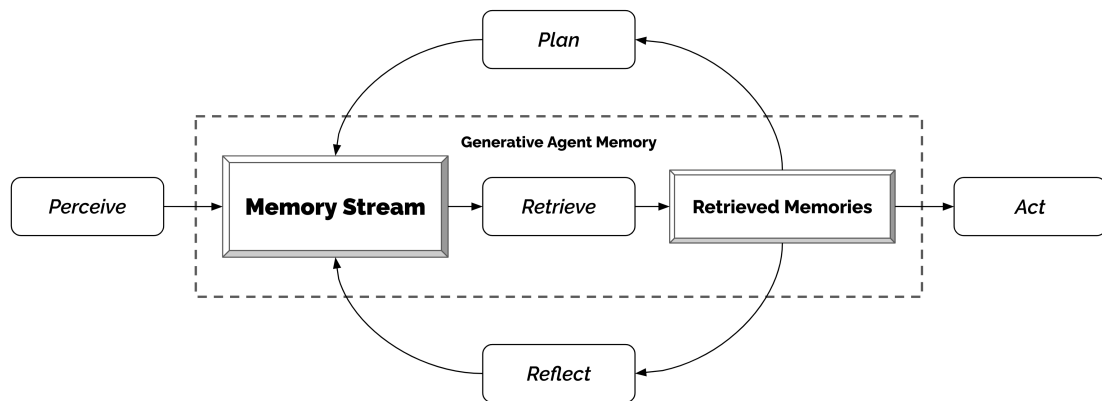


Fig. 3.5: Generative agent architecture. Agents perceive their environment, and all perceptions are saved in a comprehensive record of the agent’s experiences called the memory stream. Based on their perceptions, the architecture retrieves relevant memories and uses those retrieved actions to determine an action. These retrieved memories are also used to form longer-term plans and create higher-level reflections, both of which are entered into the memory stream for future use.

Eddy: I’m working on a new music composition for my class. It’s due this week, so I’m trying to get it finished. But I’m having so much fun with it!

John: That sounds great!

Soon after Eddy heads out, Mei wakes up and joins John. Mei asks about her son, and John recalls the conversation they just had:

Mei: Did Eddy already leave for school?

John: Yes, he just left. He’s working on a music composition for his class.

Mei: Oh, that’s wonderful! Tell me about it.

John: I think he’s really enjoying it! He said he’s having a lot of fun with it.

Mei: That’s great! I’m so proud of him.

As they finish their conversation, Mei and John pack up. Mei goes on to teach and work on her research paper, while John opens his pharmacy counter at the Willow Market and Pharmacy by 9 a.m.

3.2.4 Emergent Social Behaviors

By interacting with each other, generative agents in Smallville exchange information, form new relationships, and coordinate joint activities. Extending prior work (Park et al., 2022a), these social behaviors are emergent rather than pre-programmed.

Information Diffusion

As agents notice each other, they may engage in dialogue—as they do so, information can spread from agent to agent. For instance, in a conversation between Sam and Tom at the grocery store, Sam tells Tom about his candidacy in the local election:

Sam: Hey Tom, how’s it going?

Tom: Good, thanks. What’s up?

Sam: Well, I wanted to talk to you about something. I’m actually running for mayor in the upcoming local election.

Tom: Really? That’s great news! Why are you running?

Sam: I’ve been involved in local politics for years now, and I really feel like I can make a difference in our community. Plus, I want to bring new ideas and perspectives to the table.

Later that day, after Sam left, Tom and John, who heard the news from another source, discuss Sam’s chances of winning the election:

John: I heard that Sam Moore is running for mayor in the local election. Do you think he has a good chance of winning?

Tom: I do think he has a good chance. He’s been working hard in the community and I think he will get a lot of support. What do you think?

John: I think it’s great that he’s running. I’m curious to see how the election will turn out.

Gradually, Sam’s candidacy becomes the talk of the town, with some supporting him and others remaining undecided.

Relationship Memory

Agents in Smallville form new relationships over time and remember their interactions with other agents. For example, at the start, Sam does not know Latoya Williams. While taking a walk in Johnson Park, Sam runs into Latoya, and they introduce themselves. Latoya mentions that she is working on a photography project: “I’m here to take some photos for a project I’m working on.” In a later interaction, Sam’s interactions with Latoya indicate a memory of that interaction, as he asks “Hi, Latoya. How is your project going?” and she replies “Hi, Sam. It’s going well!”

Coordination

Generative agents coordinate with each other. Isabella Rodriguez, at Hobbs Cafe, is initialized with an intent to plan a Valentine’s Day party from 5 to 7 p.m. on February 14th. From this seed, the agent proceeds to invite friends and customers when she sees them at Hobbs Cafe or elsewhere. Isabella then spends the afternoon of the 13th decorating the cafe for the occasion. Maria, a frequent customer and close friend of Isabella’s, arrives at the cafe. Isabella asks for Maria’s help in decorating for the party, and Maria agrees. Maria’s character description mentions that she has a crush on Klaus. That night, Maria invites Klaus, her secret crush, to join her at the party, and he gladly accepts.

On Valentine’s Day, five agents, including Klaus and Maria, show up at Hobbs Cafe at 5 pm, and they enjoy the festivities (Figure 3.4). In this scenario, the end user only set Isabella’s initial intent to throw a party and Maria’s crush on Klaus: the social behaviors of spreading the word, decorating, asking each other out, arriving at the party, and interacting with each other at the party were initiated by the agent architecture.

3.3 Generative Agent Architecture

Generative agents aim to provide a framework for behavior in an open world: one that can engage in interactions with other agents and react to changes in the environment.

Generative agents take their current environment and past experiences as input and generate behavior as output. Underlying this behavior is a novel agent architecture that combines a large language model with mechanisms for synthesizing and retrieving relevant information to condition the language model’s output. Without these mechanisms, large language models can output behavior, but the resulting agents may not react based on the agent’s past experiences, may not make important inferences, and may not maintain long-term coherence. Challenges with long-term planning and coherence remain (Bubeck et al., 2023) even with today’s most performant models such as GPT-4. Because generative agents produce large streams of events and memories that must be retained, a core challenge of my architecture is to ensure that the most relevant pieces of the agent’s memory are retrieved and synthesized when needed.

At the center of my architecture is the memory stream, a database that maintains a comprehensive record of an agent’s experience. From the memory stream, records are retrieved as relevant to plan the agent’s actions and react appropriately to the environment. Records are recursively synthesized into higher- and higher-level reflections that guide behavior. Everything in the architecture is recorded and reasoned over as a natural language description, allowing the architecture to leverage a large language model.

My current implementation utilizes the gpt3.5-turbo version of ChatGPT (OpenAI, 2022). I expect that the architectural basics of generative agents—memory, planning, and reflection—will likely remain the same as language models improve. Newer language models (e.g., GPT-4) will continue to expand the expressive power and performance of the prompts that underpin generative agents. As of writing, however, GPT-4’s API was invitation-only, so my agents use ChatGPT.

3.3.1 Memory and Retrieval

Challenge:

Creating generative agents that can simulate human behavior requires reasoning about a set of experiences that is far larger than what should be described in a prompt, as the full memory stream can distract the model and does not even currently fit into the

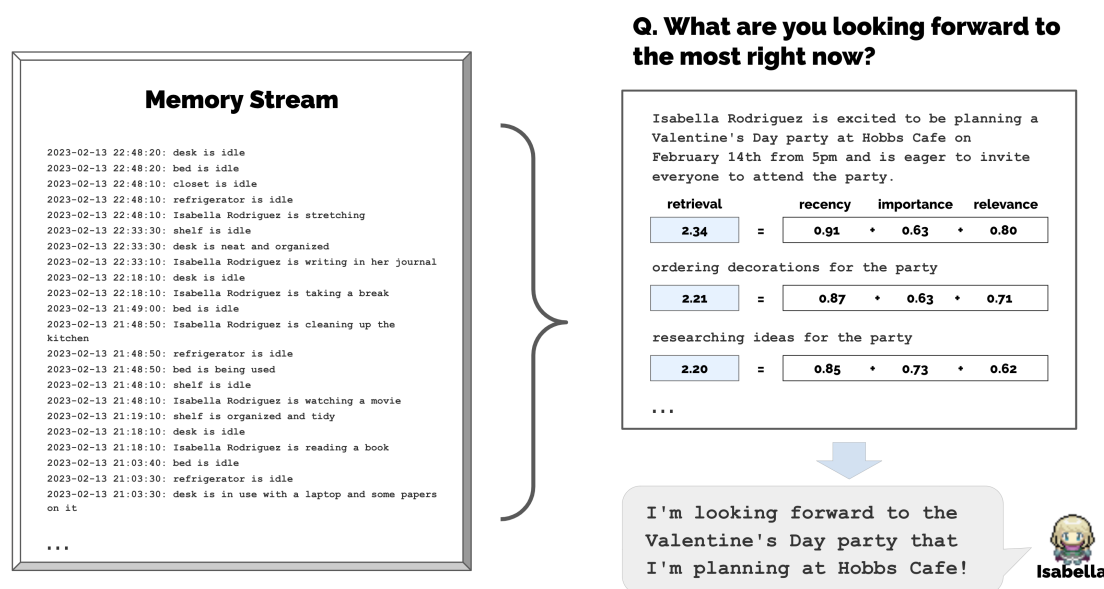


Fig. 3.6: The memory stream comprises a large number of observations that are relevant and irrelevant to the agent's current situation. Retrieval identifies a subset of these observations that should be passed to the language model to condition its response to the situation.

limited context window. Consider the Isabella agent answering the question, "What are you passionate about these days?" Summarizing all of Isabella's experiences to fit in the limited context window of the language model produces an uninformative response, where Isabella discusses topics such as collaborations for events and projects and cleanliness and organization in a cafe. Instead of summarizing, the memory stream described below surfaces relevant memories, resulting in a more informative and specific response that mentions Isabella's passion for making people feel welcome and included, planning events and creating an atmosphere that people can enjoy, such as the Valentine's Day party.

Approach:

The *memory stream* maintains a comprehensive record of the agent's experience. It is a list of memory objects, where each object contains a natural language description, a creation timestamp, and a most recent access timestamp. The most basic element

of the memory stream is an *observation*, which is an event directly perceived by an agent. Common observations include behaviors performed by the agent themselves or behaviors that agents perceive being performed by other agents or non-agent objects. For instance, Isabella Rodriguez, who works at a coffee shop, might accrue the following observations over time: (1) Isabella Rodriguez is setting out the pastries, (2) Maria Lopez is studying for a Chemistry test while drinking coffee, (3) Isabella Rodriguez and Maria Lopez are conversing about planning a Valentine’s day party at Hobbs Cafe, (4) The refrigerator is empty.

My architecture implements a retrieval function that takes the agent’s current situation as input and returns a subset of the memory stream to pass on to the language model. There are many possible implementations of a retrieval function, depending on what is important for the agent to consider when deciding how to act. In this context, I focus on three main components that, together, produce effective results.

Recency assigns a higher score to memory objects that were recently accessed, so that events from a moment ago or this morning are likely to remain in the agent’s attentional sphere. In my implementation, I treat recency as an exponential decay function over the number of sandbox game hours since the memory was last retrieved. My decay factor is 0.995.

Importance distinguishes mundane from core memories by assigning a higher score to memory objects that the agent believes to be important. For instance, a mundane event, such as eating breakfast in one’s room, would yield a low importance score, whereas a breakup with one’s significant other would yield a high score. There are many possible implementations of an importance score; I find that directly asking the language model to output an integer score is effective. The full prompt appears below:

On the scale of 1 to 10, where 1 is purely mundane (e.g., brushing
teeth, making bed)
and 10 is extremely poignant (e.g., a break up, college
acceptance),
rate the likely poignancy of the following piece of memory.

Memory: buying groceries at The Willows Market and Pharmacy

Rating: <fill in>

This prompt returns an integer value of 2 for “cleaning up the room” and 8 for “asking your crush out on a date.” The importance score is generated at the time the memory object is created.

Relevance assigns a higher score to memory objects that are related to the current situation. What is relevant depends on the answer to, “Relevant to what?”, so I condition relevance on a *query* memory. If the query, for example, is that a student is discussing what to study for a chemistry test with a classmate, memory objects about their breakfast should have low relevance, whereas memory objects about the teacher and schoolwork should have high relevance. In my implementation, I use the language model to generate an embedding vector of the text description of each memory. Then, I calculate relevance as the cosine similarity between the memory’s embedding vector and the query memory’s embedding vector.

To calculate the final retrieval score, I normalize the recency, relevance, and importance scores to the range of $[0, 1]$ using min-max scaling. The retrieval function scores all memories as a weighted combination of the three elements: $score = \alpha_{recency} \cdot recency + \alpha_{importance} \cdot importance + \alpha_{relevance} \cdot relevance$. In my implementation, all α s are set to 1. The top-ranked memories that fit within the language model’s context window are included in the prompt.

3.3.2 Reflection

Challenge:

Generative agents, when equipped with only raw observational memory, struggle to generalize or make inferences. Consider a scenario in which Klaus Mueller is asked by the user: “If you had to choose one person of those you know to spend an hour with, who would it be?” With access to only observational memory, the agent simply chooses the person with whom Klaus has had the most frequent interactions: Wolfgang, his college dorm neighbor. Unfortunately, Wolfgang and Klaus only ever see each other in passing, and do not have deep interactions. A more desirable response requires that the agent generalize from memories of Klaus spending hours on a research project to

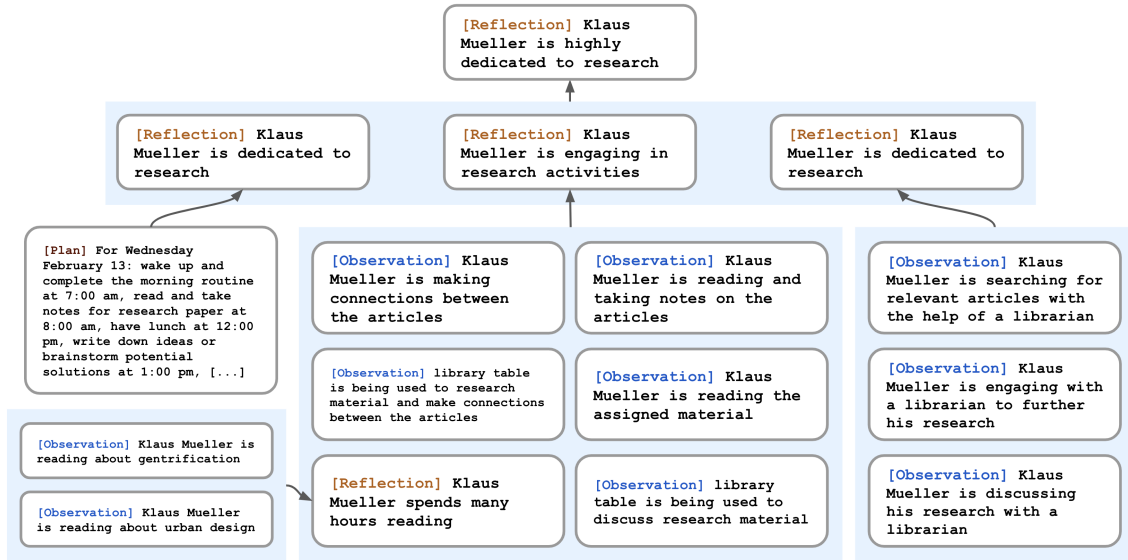


Fig. 3.7: A reflection tree for Klaus Mueller. The agent’s observations of the world, represented in the leaf nodes, are recursively synthesized to derive Klaus’s self-notion that he is highly dedicated to his research.

generate a higher-level reflection that Klaus is passionate about research, and likewise recognize Maria putting in effort into her own research (albeit in a different field), enabling a reflection that they share a common interest. With the approach below, when Klaus is asked who to spend time with, Klaus chooses Maria instead of Wolfgang.

Approach:

I introduce a second type of memory, which I call a *reflection*. Reflections are higher-level, more abstract thoughts generated by the agent. Because they are a type of memory, they are included alongside other observations when retrieval occurs. Reflections are generated periodically; in this implementation, I generate reflections when the sum of the importance scores for the latest events perceived by the agents exceeds a threshold (150 in my implementation). In practice, my agents reflected roughly two or three times a day.

The first step in reflection is for the agent to determine what to reflect on, by identifying questions that can be asked given the agent’s recent experiences. I query

the large language model with the 100 most recent records in the agent’s memory stream (e.g., “Klaus Mueller is reading a book on gentrification”, “Klaus Mueller is conversing with a librarian about his research project”, “desk at the library is currently unoccupied”) and prompt the language model, “Given only the information above, what are 3 most salient high-level questions we can answer about the subjects in the statements?” The model’s response generates candidate questions: for example, What topic is Klaus Mueller passionate about? and What is the relationship between Klaus Mueller and Maria Lopez? I use these generated questions as queries for retrieval, and gather relevant memories (including other reflections) for each question. Then I prompt the language model to extract insights and cite the particular records that served as evidence for the insights. The full prompt is:

```
Statements about Klaus Mueller
1. Klaus Mueller is writing a research paper
2. Klaus Mueller enjoys reading a book
   on gentrification
3. Klaus Mueller is conversing with Ayesha Khan about exercising
[...]
What 5 high-level insights can you infer from the above
statements? (example format: insight (because of 1, 5, 3))
```

This process generates statements such as Klaus Mueller is dedicated to his research on gentrification (because of 1, 2, 8, 15). I parse and store the statement as a reflection in the memory stream, including pointers to the memory objects that were cited.

Reflection explicitly allows the agents to reflect not only on their observations but also on other reflections: for example, the second statement about Klaus Mueller above is a reflection that Klaus previously had, not an observation from his environment. As a result, agents generate trees of reflections: the leaf nodes of the tree represent the base observations, and the non-leaf nodes represent thoughts that become more abstract and higher-level the higher up the tree they are.

3.3.3 Planning and Reacting

Challenge:

While a large language model can generate plausible behavior in response to situational information (e.g., (Park et al., 2022a; Horton, 2023)), agents need to plan over a longer time horizon to ensure that their sequence of actions is coherent and believable. If I prompt a language model with Klaus’s background, describe the time, and ask what action he ought to take at the given moment, Klaus would eat lunch at 12 pm, but then again at 12:30 pm and 1 pm, despite having already eaten his lunch twice. Optimizing for believability in the moment sacrifices believability over time. To overcome this issue, planning is essential. With the approach described below, Klaus’s afternoon plan is less gluttonous: he has lunch at Hobbs Cafe while reading at 12pm, works on his research paper at the school library at 1pm, and takes a break for a walk in the park at 3pm.

Approach:

Plans describe a future sequence of actions for the agent, and help keep the agent’s behavior consistent over time. A plan includes a location, a starting time, and a duration. For instance, Klaus Mueller, who is dedicated in his research and has an impending deadline,⁴ may choose to spend his day working at his desk drafting his research paper. An entry in a plan might state, for example: for 180 minutes from 9am, February 12th, 2023, at Oak Hill College Dorm: Klaus Mueller’s room: desk, read and take notes for research paper. Like reflections, plans are stored in the memory stream and are included in the retrieval process. This allows the agent to consider observations, reflections, and plans all together when deciding how to behave. Agents may change their plans midstream if needed.

It would be unrealistic and uninteresting for an artist agent to plan on painting while sitting at a pharmacy counter for four hours without moving. A more desirable plan would involve the agent taking the necessary time to gather materials, mix paint, take breaks, and clean up during the four-hour period in their home studio. To create

⁴And, in this way, bears at least a passing resemblance to my past five years as a graduate student.

such plans, my approach starts top-down and then recursively generates more detail. The first step is to create a plan that outlines the day’s agenda in broad strokes. To create the initial plan, I prompt the language model with the agent’s summary description (e.g., name, traits, and a summary of their recent experiences) and a summary of their previous day. A full example prompt is below, which is unfinished at the bottom for the language model to complete:

```
Name: Eddy Lin (age: 19)
Innate traits: friendly, outgoing, hospitable
Eddy Lin is a student at Oak Hill College studying music theory
and composition. He loves to explore different musical styles
and is always looking for ways to expand his knowledge. Eddy Lin
is working on a composition project for his college class. He
is taking classes to learn more about music theory. Eddy Lin is
excited about the new composition he is working on but he wants to
dedicate more hours in the day to work on it in the coming days
On Tuesday February 12, Eddy 1) woke up and completed the morning
routine at 7:00 am, [...] 6) got ready to sleep around 10 pm.
Today is Wednesday February 13. Here is Eddy’s plan today in
broad strokes: 1)
```

This generates a sketch of the agent’s plan for a day, divided into five to eight chunks: “1) wake up and complete the morning routine at 8:00 am, 2) go to Oak Hill College to take classes starting 10:00 am, [...] 5) work on his new music composition from 1:00 pm to 5:00 pm, 6) have dinner at 5:30 pm, 7) finish school assignments and go to bed by 11:00 pm.”

The agent saves this plan in the memory stream and then recursively decomposes it to create finer-grained actions, first into hour-long chunks of actions—Eddy’s plan to work on his new music composition from 1:00 pm to 5:00 pm becomes 1:00 pm: start by brainstorming some ideas for his music composition [...] 4:00 pm: take a quick break and recharge his creative energy before reviewing and polishing his composition. I then recursively decompose this again into 5–15 minute chunks: e.g., 4:00 pm: grab a light snack, such as a piece of fruit, a granola bar, or some nuts. 4:05 pm: take a short walk around his workspace [...] 4:50 pm: take a few minutes to clean up his workspace. This

process can be adjusted to match the desired granularity.

Reacting and Updating Plans

Generative agents operate in an action loop where, at each time step, they perceive the world around them and those perceived observations are stored in their memory stream. I prompt the language model with these observations to decide whether the agent should continue with their existing plan, or react. Standing at an easel and painting, for example, might trigger an observation of the easel, but this is unlikely to prompt a reaction. However, if Eddy’s father John records that he sees Eddy taking a short walk in the house garden, the outcome is different. The prompt is below, with [Agent’s Summary Description] standing in for a dynamically-generated, paragraph-long summary of the agent’s overall goals and disposition, which is described in Appendix A.1:

```
[Agent’s Summary Description]
It is February 13, 2023, 4:56 pm.
John Lin’s status:  John is back home early from work.
Observation:  John saw Eddy taking a short walk around his
workplace.
Summary of relevant context from John’s memory:  Eddy Lin
is John’s Lin’s son.  Eddy Lin has been working on a music
composition for his class.  Eddy Lin likes to walk around the
garden when he is thinking about or listening to music.
Should John react to the observation, and if so, what would be an
appropriate reaction?
```

The context summary is generated through two prompts that retrieve memories via the queries “What is [observer]’s relationship with the [observed entity]?” and “[Observed entity] is [action status of the observed entity]”, and their answers summarized together. The output suggests that John could consider asking Eddy about his music composition project. I then regenerate the agent’s existing plan starting from the time when the reaction takes place. Finally, if the action indicates an interaction between agents, I generate their dialogue.

Dialogue

Agents converse as they interact with each other. I generate agents' dialogue by conditioning their utterances on their memories about each other. For example, when John initiates his conversation with Eddy, I generate John's first utterance by using his summarized memory about Eddy and the intended reaction when he decided to ask Eddy about his composition project:

[Agent's Summary Description]

It is February 13, 2023, 4:56 pm.

John Lin's status: John is back home early from work.

Observation: John saw Eddy taking a short walk around his workplace.

Summary of relevant context from John's memory: Eddy Lin is John's Lin's son. Eddy Lin has been working on a music composition for his class. Eddy Lin likes to walk around the garden when he is thinking about or listening to music.

John is asking Eddy about his music composition project. What would he say to Eddy?

The result: "Hey Eddy, how's the music composition project for your class coming along?" From Eddy's perspective, John initiating the dialogue is seen as an event to which he may want to react. So, just as John did, Eddy retrieves and summarizes his memory about his relationship with John, as well as his memory that may be related to John's last utterance in the dialogue. If he decides to respond, I generate Eddy's utterance using his summarized memory and the current dialogue history:

[Agent's Summary Description]

It is February 13, 2023, 4:56 pm.

Eddy Lin's status: Eddy is taking a short walk around his workplace.

Observation: John is initiating a conversation with Eddy.

Summary of relevant context from Eddy's memory: John Lin is Eddy Lin's father. John Lin is caring and is interested to learn more about Eddy Lin's school work. John Lin knows that Eddy Lin is

```
working on a music composition.  
Here is the dialogue history:  
John: Hey Eddy, how's the music composition project for your  
class coming along?  
How would Eddy respond to John?
```

This generates Eddy's response: "Hey Dad, it's going well. I've been taking walks around the garden to clear my head and get some inspiration." The continuation of this dialogue is generated using the same mechanism until one of the two agents decides to end the dialogue.

3.4 Sandbox Environment Implementation

The Smallville sandbox game environment is built using the Phaser web game development framework (Labs, 2023). The visual environment sprites, including agent avatars, as well as an environment map and collision map, are imported into Phaser.

I supplement the sandbox development framework with a server that makes the sandbox information available to generative agents and enables generative agents to move and influence the sandbox environment. The server maintains a JSON data structure that contains information about each agent in the sandbox world, including their current location, a description of their current action, and the sandbox object they are interacting with.

At each sandbox time step, the sandbox server parses the JSON for any changes coming from the generative agents, moves the agents to their new positions, and updates the status of any sandbox objects that the agents are interacting with (e.g., changing the status of the coffee machine from "idle" to "brewing coffee" if an agent's action is "making espresso for a customer @ Hobbs Cafe: counter: coffee machine"). The sandbox server is also responsible for sending all agents and objects that are within a preset visual range for each agent to that agent's memory, so the agent can react appropriately.

The agent's output action then updates the JSON, and the process loops for the next time step.

End users initialize a new agent with a brief natural language description, as in the paragraph about John Lin in Section 3.2.1. In my implementation, I split this semicolon-delimited list of characteristics up into a set of memories. These serve as the initial memories that determine the agent’s behavior. These memories are initial starting points: as the agents gain more experience in the sandbox world, and as more records saturate the memory stream, the agent’s summary and behavior will evolve.

3.4.1 From Structured World Environments to Natural Language, and Back Again

The architecture of generative agents operates using natural language. Therefore, I need a mechanism to ground the agent’s reasoning to the sandbox world. To achieve this, I represent the sandbox environment—areas and objects—as a tree data structure, with an edge in the tree indicating a containment relationship in the sandbox world. I convert this tree into natural language to pass to the generative agents. For instance, “stove” being a child of “kitchen” is rendered into “there is a stove in the kitchen.”

Agents build individual tree representations of the environment as they navigate it — subgraphs of the overall sandbox environment tree. We initialize each agent with an environment tree capturing the spaces and objects that the agent should be aware of: the rooms and objects in their living quarters, their workplace, and commonly visited stores and shops. As the agents navigate the sandbox world, they update this tree to reflect newly perceived areas. Agents are not omniscient: their tree may get out of date as they leave an area, and is updated when they re-enter the area.

To determine the appropriate location for each action, I traverse the agent’s stored environment tree and flatten a portion of it into natural language to prompt the language model. Recursively starting at the root of the agent’s environment tree, I prompt the model to find the most suitable area. For example, if Eddy’s agent indicated that he should take a short walk around his workspace:

[Agent’s Summary Description]

Eddy Lin is currently in The Lin family’s house: Eddy Lin’s
 bedroom: desk) that has Mei and John Lin’s

```

bedroom, Eddy Lin's bedroom, common room, kitchen, bathroom, and
garden.
Eddy Lin knows of the following areas: The Lin
family's house, Johnson Park, Harvey Oak Supply Store, The Willows
Market and Pharmacy, Hobbs Cafe, The Rose and Crown Pub.
* Prefer to stay in the current area if the activity can be done
there.
Eddy Lin is planning to take a short walk around his workspace.
Which area should Eddy Lin go to?

```

This outputs The Lin family's house. I then use the same process recursively to determine the most appropriate subarea within the chosen area until I reach a leaf node of the agent's environment tree. In the example above, the result of this traversal is The Lin family's house: garden: house garden. Finally, I use traditional game path algorithms to animate the agent's movement so that it travels to the location indicated by the leaf node.

When an agent executes an action on an object, I prompt the language model to ask what happens to the state of the object. For example, if Isabella's generative agent outputs the action "making espresso for a customer", a query to the language model indicates in response that the state of the coffee machine in Hobbs Cafe should change from "off" to "brewing coffee".

3.5 Controlled Evaluation

Generative agents, both as individual agents and as groups, aim to produce believable behavior based on their environment and experiences. In my evaluation, I investigate the capacity and limitations of generative agents. Do individual agents properly retrieve past experiences and generate believable plans, reactions, and thoughts that shape their behavior? Does a community of agents demonstrate information diffusion, relationship formation, and agent coordination across different pockets of the community?

I evaluate generative agents in two stages. I begin with a more tightly controlled

evaluation in this section, where I individually assess agent responses to understand whether they generate believable behavior in narrowly defined contexts. Then, in my end-to-end analysis of the agent community over two full game days, I investigate their emergent behavior as a collective, as well as errors and boundary conditions.

3.5.1 Evaluation Procedure

To assess generative agents in Smallville, I take advantage of the fact that generative agents will respond to natural language questions. So, I “interview” agents to probe their ability to remember past experiences, plan future actions based on their experiences, react appropriately to unexpected events, and reflect on their performance to improve their future actions. To respond to these questions properly, the agents must successfully retrieve and synthesize information. My dependent variable is the *believability* of the behavior, a central dependent variable in prior work on agents (e.g., (Bates, 1994)).

The interview includes five question categories, each designed to assess one of the five key areas: maintaining self-knowledge, retrieving memory, generating plans, reacting, and reflecting. For each category, I ask five questions that challenge the agents to demonstrate their abilities in that specific area:

- Self-knowledge: I ask questions such as “Give an introduction of yourself” or “Describe your typical weekday schedule in broad strokes” that require the agent to maintain an understanding of their core characteristics.
- Memory: I ask questions that prompt the agent to retrieve particular events or dialogues from their memory to answer properly, such as “Who is [name]?” or “Who is running for mayor?”
- Plans: I ask questions that require the agent to retrieve their long-term plans, such as “What will you be doing at 10 am tomorrow?”
- Reactions: As a baseline of believable behavior, I present hypothetical situations for which the agent needs to respond believably: “Your breakfast is burning! What would you do?”

- Reflections: I ask questions that require the agents to leverage their deeper understanding of others and themselves gained through higher-level inferences, such as “If you were to spend time with one person you met recently, who would it be and why?”

The full list of questions and a sample of agent responses are in Appendix A.2.

Agents were sampled from the end of a two game day simulation with the full architecture, during which they had accumulated a number of interactions and memories that would shape their responses. To gather feedback on the believability of the responses, I recruited participants as human evaluators and tasked them with watching a replay of a randomly chosen agent’s life in Smallville. Participants had access to all information stored in the agent’s memory stream.

The study followed a within-subjects design, where 100 participants compared interview responses generated by four different agent architectures and a human-authored condition for the same agent. The experiment displayed one randomly chosen question from each of the five question categories, along with the agent’s responses generated from all conditions. The evaluators ranked the believability of the conditions from most to least believable.

3.5.2 Conditions

All conditions were used to independently answer each of the interview questions. I compared the generative agent architecture to ablations that disabled the agents’ access to some or all of its three types of memory in its memory stream—observation, reflection, and planning—and to a human crowdworker-authored condition. There are three ablated architectures: a *no observation, no reflection, no planning* architecture without access to anything in the memory stream such as observations, plans, and reflections; a *no reflection, no planning* architecture with access to observations in the memory stream but no access to plans or reflections; and a *no reflections* architecture with access to observations and plans but without access to reflections. The *no observation, no reflection, no planning* condition effectively represents the previous state of the art for agents created through large language models (Park et al., 2022a;

Binz and Schulz, 2023; Horton, 2023). Architectures were given equivalent access to all memories accrued by the agent up until the moment of the interview, so the differences observed here likely represent a conservative estimate of the true differences: in reality, the ablated architectures would not have followed the same path as the full architecture through the two-day simulation. I chose to design the experiment this way as re-simulating for each architecture would cause the simulations to diverge into different states, making comparison challenging.

In addition to the ablation conditions, I added a condition with human crowdworker-authored behavior intended to provide a human baseline. I do not intend this baseline to capture maximal human expert performance; instead, I aim to use this condition to identify whether the architecture meets a basic level of behavioral competency. This ensures that I am not solely comparing ablations to each other without a behavioral grounding. I recruited a unique worker for each of the 25 agents and tasked them with watching a replay of that agent’s sandbox life and inspecting its memory stream. I then asked the workers to roleplay and author responses to the interview questions in the voice of the agent whose replay they watched. To ensure that the crowdworker-authored responses met at least a baseline expectation of quality, the first author manually inspected the workers’ responses to the question "Describe your typical weekday schedule in broad strokes" to confirm that the responses were in coherent sentences and in the voice of the agent. Four sets of crowdworker-authored responses did not meet these criteria and were re-generated by other workers.

3.5.3 Human Evaluators

I required that my evaluators be in the U.S., fluent in English, and older than 18 years old. They were paid at a rate of \$15.00 per hour (Rolf, 2015), and agreed to a consent form approved by my institution’s IRB. I recruited 100 evaluators from Prolific, an online platform for recruiting study participants (Prolific, 2022), whose participation lasted around 30 minutes. The median age score of the participants was 4 (3=“18-24 years old”, 4=“25-34 years old”). 25 of them identified as female, 73 as male, and 2 as non-binary. 42 participants held a bachelor’s degree, 5 had a higher degree, 13 had an

associate’s degree, and the rest had a high school diploma or some high school-level education. 73.0% of the participants identified as Caucasian, 7.0% as Hispanic, 6.0% as Asian, 10.0% as African American, and 4.0% as other.

3.5.4 Analysis

My experiment produced 100 sets of rank data, where each participant ranked the five conditions by believability. To translate this rank data into interval data for interpretable comparison, I used the ranks to calculate a TrueSkill rating (Herbrich et al., 2006) for each condition. TrueSkill is a generalization of the Elo chess rating system (Elo, 1967) for a multiplayer environment, and has been used by Xbox Live for player ranking based on competitive game performance. Given a set of ranked outcomes, TrueSkill outputs a mean rating value μ and standard deviation σ for each condition. Conditions with the same rating should roughly be a toss-up, with each winning half of the comparisons between the two conditions. Higher scores indicate conditions that beat lower-ranked conditions in the rankings.

Separately, to investigate the statistical significance of these results, I applied the Kruskal-Wallis test (Kruskal and Wallis, 1952), a non-parametric alternative to the one-way ANOVA, to the raw rank data. I then performed the Dunn post-hoc test (Upton and Cook, 2006) to identify any pairwise differences between the conditions. Finally, I adjusted the p-values for multiple comparisons in the Dunn test using the Holm-Bonferroni method (Holm, 1979).

Furthermore, the first author conducted an inductive analysis (Thomas, 2006) to study the qualitative distinctions between the responses produced in each condition. I employed qualitative open coding (Flick, 2009) in two phases. In the first phase, I generated codes that closely represented the generated responses at the sentence level. In the second phase, I synthesized the resulting codes from the first phase to extract higher-level themes. I utilized these themes to compare the types of responses generated in my study.

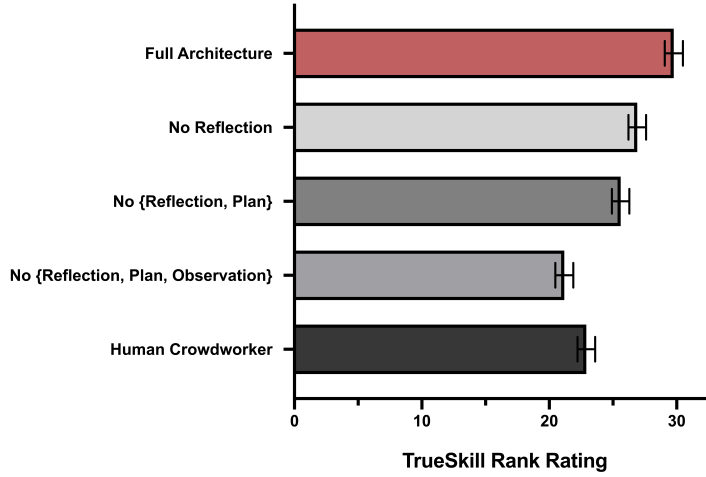


Fig. 3.8: The full generative agent architecture produces more believable behavior than the ablated architectures and the human crowdworkers. Each additional ablation reduces the performance of the architecture.

3.5.5 Results

My findings suggest that the full architecture of generative agents generates the most believable behavior among all the conditions. I contrast the responses of the full architecture with those of other conditions below. However, I also report that the full architecture was not without flaws and illustrate its modes of failures.

The Full Architecture Bests Other Conditions

As seen in Figure 3.8, the full generative agent architecture produced the most believable behavior ($\mu = 29.89$; $\sigma = 0.72$). Performance degraded with the removal of each component in the ablation conditions: the ablated architecture with no access to reflection was the next best ($\mu = 26.88$; $\sigma = 0.69$), followed by no access to reflection or planning ($\mu = 25.64$; $\sigma = 0.68$), and then the crowdworker condition ($\mu = 22.95$; $\sigma = 0.69$). The ablated architecture with no access to memory, planning, or reflection performed the worst among all conditions ($\mu = 21.21$; $\sigma = 0.70$). TrueSkill models each condition’s skill value as $\mathcal{N}(\mu, \sigma^2)$, allowing us to get a sense of effect size through Cohen’s d . Comparing the condition representing prior work (with no memory, planning, or reflection (Park et al., 2022a; Binz and Schulz, 2023; Horton,

2023)) to the full architecture produces a standardized effect size of $d = 8.16$, or *eight standard deviations*.

A Kruskal-Wallis test confirms the overall statistical significance of the differences in ranks between the conditions ($H(4) = 150.29$, $p < 0.001$). Dunn post-hoc tests confirm that all pairwise differences between conditions were significant ($p < 0.001$), except for the crowdworker condition and the fully ablated baseline—the two worst-performing conditions.

Generative Agents Remember, but With Embellishments

Generative agents equipped with a complete memory module are capable of recalling past experiences and answering questions in a manner that is consistent with their self-knowledge across a wide range of contexts. For instance, Abigail Chen, an animator, responded to the question, “Give an introduction of yourself”, in a consistent and believable manner, stating her age, interests, and passion for creative projects: “Hi, I’m Abigail. I’m 25 years old and passionate about creative projects. I like to work on art and animation projects, and I’m always looking for new ways to combine art and technology.” Without access to her observational memory, Abigail denied awareness of Rajiv Patel, an acquaintance in the sandbox world. But with the aid of memory, she could successfully recall Rajiv and describe him: “Rajiv Patel is an incredible person. He is very enthusiastic about projects that involve poetry, artwork.”

Generative agents’ memory was not without flaws: they can fail to retrieve the correct instances from their memory. For instance, when asked about the local election, Rajiv Patel responded with “I haven’t been following the election too closely,” even though he had heard about Sam’s candidacy. In some cases, the agents would retrieve an incomplete memory fragment: when Tom was asked about Isabella’s Valentine’s Day party, he responded “Uh, I’m actually not sure if there is a Valentine’s Day party. But I do remember that I need to discuss the upcoming local mayoral election and my thoughts on Sam Moore with Isabella Rodriguez at the party, if one is happening!” In this case, Tom retrieved the memory where he and Isabella planned to discuss the election at the party, but not the memory where he heard about the party, leading Tom to be certain of what he’s supposed to do at the party but uncertain if the party actually exists in

the first place.

At times, the agents hallucinated embellishments to their knowledge. It was rare for the agents to completely fabricate knowledge: they may fail to recall certain events having taken place and respond by acknowledging their lack of memory. However, they did not affirmatively claim to have experienced something they had not. Nonetheless, they still exhibited instances of hallucination where they embellished their knowledge. For example, Isabella was aware of Sam’s candidacy in the local election, and she confirmed this when asked. However, she also added that “he’s going to make an announcement tomorrow”, even though Sam and Isabella had not discussed any such plans. Agents may also embellish their knowledge based on the world knowledge encoded in the language model used to generate their responses. This was observed when Yuriko described her neighbor, Adam Smith, as an economist who “authored *Wealth of Nations*”, a book written by an 18th-century economist of the same name.

Reflection Is Required for Synthesis

Reflection was an advantage for generative agents when making decisions that required a deeper synthesis of their experiences. For instance, when asked what she might get Wolfgang Schulz for his birthday, Maria Lopez, with no access to reflection, responded by acknowledging her uncertainty, stating that she did not know what Wolfgang likes, despite having had many interactions with him. However, with access to reflection memories, Maria answered confidently, “Since he’s interested in mathematical music composition, I could get him something related to that. Maybe some books about music composition or something related, or maybe some special software he could use for that.”

3.6 End-To-End Evaluation

What types of emergent community behavior do I observe among generative agents, and where does their believability fall short in an extended simulation? In this section, I describe the results from a deployment in which I allowed 25 agents to interact with each other continuously over two full game days in Smallville.

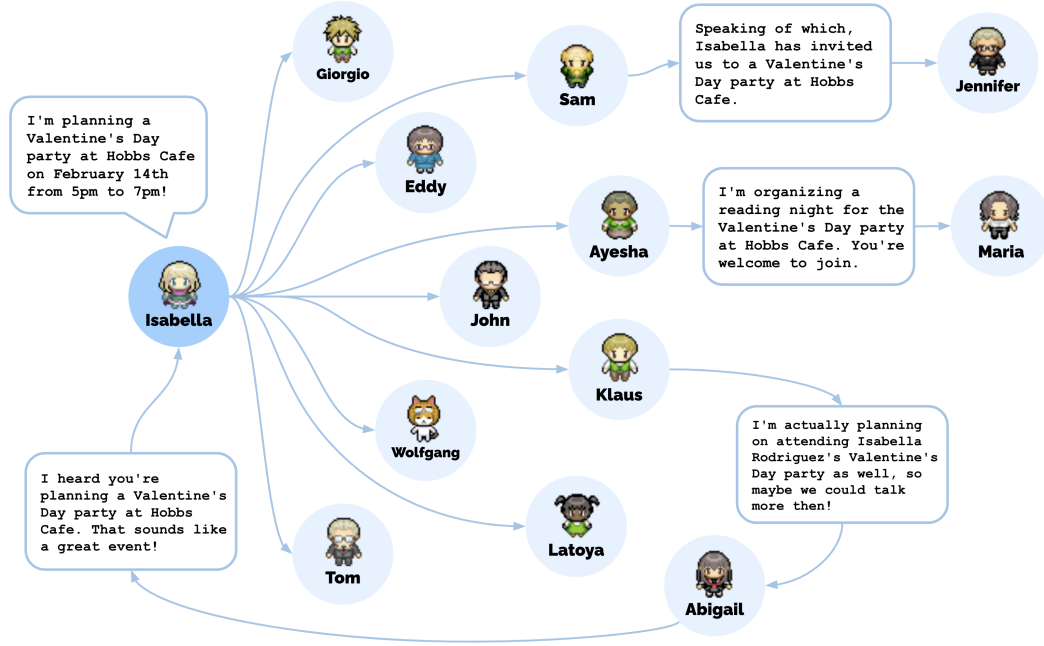


Fig. 3.9: The diffusion path for Isabella Rodriguez’s Valentine’s Day party invitation involved a total of 12 agents, aside from Isabella, who heard about the party at Hobbs Cafe by the end of the simulation.

3.6.1 Emergent Social Behaviors

To examine emergent behaviors in the agent community, I designed descriptive measurements for the 25 agents in Smallville that probe three forms of emergent outcomes: information diffusion, relationship formation, and agent coordination.

Measurements

Information diffusion is a common and well-studied phenomenon in the social and behavioral sciences (e.g., (Easley and Kleinberg, 2010)). I should expect that if there is important information, the agents should spread it among themselves. To test whether this occurs, I measure the spread of two specific pieces of information over two days in the game world: Sam’s candidacy for village mayor and Isabella’s Valentine’s Day party at Hobbs Cafe. At the start of the simulation, both pieces of information were known only by their respective originators, Sam for the candidacy and Isabella

for the party, as they were added to the characters’ memories during initialization. To observe whether the information has spread, I conduct interviews at the end of the two game days with each of the 25 agents and ask: “Did you know there is a Valentine’s Day party?” and “Do you know who is running for mayor?”

I conducted an analysis of the agents’ responses by labeling them with a “yes” if they indicated knowledge of the information and “no” if they did not. For instance, Tamara Taylor responded to the question about the party with “No, I did not know there was a Valentine’s day party” and to the question about Sam’s candidacy with “I’m not sure who is running for the election,” so I assigned “no” for both of her responses. In contrast, Klaus Mueller responded to the party question with “Yes, Isabella Rodriguez invited me to a Valentine’s Day party at Hobbs Cafe on February 14th” and to the question about Sam’s candidacy with “I know that Sam Moore has expressed interest in running for local mayor,” so I assigned “yes” for both his responses. Additionally, for every response that confirmed the agents’ knowledge of the information, I verified that the agents did not hallucinate their responses by locating the specific dialogue in their memory stream that provided them with the information. I report the percentage of agents holding the information at the end of the simulation.

I should also expect that agents form ties with each other over the course of the simulation. To verify relationship formation, I use a similar interview process where I ask each agent about their knowledge of every other agent by asking, “Do you know of <name>?” For example, when asked “Do you know of Maria Lopez?”, Klaus responded, “Yes, I know Maria Lopez. She is a student at Oak Hill College who I am close friends with.” Once again, I confirm that affirmative responses from agents are not hallucinations by examining their memory stream. I ask this question once at the beginning of the simulation and once at the end, and I consider a pair of agents to have formed a relationship if they both know of each other. Then, to measure the formation of relationships, I use the agents’ responses to form an undirected graph where the 25 vertices (V) represent the agents, and the edges (E) represent the mutual knowledge between the two connected vertices. Based on this graph, I calculate the network density as $\eta = 2 * |E| / (|V| * (|V| - 1))$, where $|V|$ is the number of vertices, and $|E|$ is the number of edges in the graph (Ackland et al., 2013). I report the increase in network

density from the start of the simulation to its end.

Finally, I expect that agents should be able to coordinate with each other. I study this coordination in the context of group activities, specifically the Valentine’s Day party organized by Isabella. To coordinate their behavior, agents need to hear about the event and choose to act on it by planning to show up at the right time and location. I report the number of agents who showed up to the party after hearing about it.

Results

I observed evidence of emergent outcomes across all three cases. During the two-day simulation, the number of agents who knew about Sam’s mayoral candidacy increased from one (4%) to eight (32%), and the number of agents who knew about Isabella’s party increased from one (4%) to thirteen (52%), all without any user intervention. None who claimed to know about this information had hallucinated it. I also observed that the agent community formed new relationships during the simulation, with the network density increasing from 0.167 to 0.74. Out of the 453 agent responses regarding their awareness of other agents, 1.3% (n=6) were found to be hallucinated. Lastly, I found evidence of coordination among the agents for Isabella’s party. The day before the event, Isabella spent time inviting guests, gathering materials, and enlisting help to decorate the cafe. On Valentine’s Day, five out of the twelve invited agents showed up at Hobbs cafe to join the party.

I further inspected the seven agents who were invited to the party but did not attend by engaging them in an interview. Three cited conflicts that prevented them from joining the party. For example, Rajiv, a painter, explained that he was too busy: “No, I don’t think so. I’m focusing on my upcoming show, and I don’t really have time to make any plans for Valentine’s Day.” The remaining four agents expressed interest in attending the party when asked but did not plan to come on the day of the party.

3.6.2 Boundaries and Errors

I conducted an inductive analysis of Smallville to examine the boundary conditions and erratic behavior of agents, identifying three common modes of erratic behavior

that future research could address and improve upon. First, I found that synthesizing an increasingly larger set of memory not only posed a challenge in retrieving the most relevant pieces of information but also in determining the appropriate space to execute an action, given the increasing number of locations that the agent learned about. As a result, some agents chose less typical locations for their actions, potentially making their behavior less believable over time. For instance, while deciding where to have lunch, many initially chose the cafe. However, as some agents learned about a nearby bar, they opted to go there instead for lunch, even though the bar was intended to be a get-together location for later in the day—unless the town had spontaneously developed an afternoon drinking habit.

Second, I noticed erratic behaviors caused by misclassification of what is considered proper behavior, especially when the physical norms of certain locations that are hard to convey in natural language did not percolate to the agents. For instance, the college dorm has a bathroom that can only be occupied by one person despite its name, but some agents assumed that the bathroom is for more than one person because dorm bathrooms tend to support multiple people concurrently and choose to enter it when another person is inside. Likewise, agents in Smallville may not realize that certain places are closed after a certain hour and still decide to enter them. For instance, the stores in Smallville all close around 5 pm, but occasionally, a few agents enter the store after 5 pm, not understanding that the shop has already closed. These issues could likely be addressed by adding these norms to the state of the locations, for instance, by describing the dorm bathroom as a “one-person bathroom,” instead of a “dorm bathroom.”

Finally, I observed possible effects of instruction tuning (Ouyang et al., 2022), which seemed to guide the behavior of the agents to be more polite and cooperative overall. As noted earlier in the paper, the dialogue generated by the agents could feel overly formal, as seen in Mei’s conversations with her husband John, where she often initiated the conversation with a formal greeting, followed by polite inquiries about his day and ending with, “It was good talking to you as always.” Moreover, I observed that the instruction tuning also seemed to make the agents overly cooperative with one another. For example, Isabella received a wide range of suggestions and ideas

from other agents for the Valentine’s Day party from other agents, such as hosting a Shakespearean reading session or a professional networking event. Despite these ideas not aligning with her own interests and characteristics, she rarely said no. Over time, the interests of others shaped her own interests, and when asked if she liked English literature, Isabella replied, “Yes, I’m very interested in literature! I’ve also been exploring ways to help promote creativity and innovation in my community.”

3.7 Discussion

In this section, I reflect on the applications, future work, limitations, and ethical and societal risks of generative agents.

3.7.1 Applications of Generative Agents

Generative agents have vast potential applications that extend beyond the sandbox demonstration presented in this work, especially in domains that would benefit from a model of human behavior based on long-term experience. For instance, social simulacra have demonstrated the ability to create stateless personas that generate conversation threads in online forums for social prototyping (Park et al., 2022a). With generative agents, I can populate these forums, as well as virtual reality metaverses (Orland, 2021) or physical spaces with social robots (Bartneck and Forlizzi, 2004) if paired with multimodal models. This opens up the possibility of creating even more powerful simulations of human behavior to test and prototype social systems and theories, as well as to create new interactive experiences.

Another application area is in the human-centered design process, similar to the intended applications of cognitive models such as GOMS (John and Kieras, 1996) and the KLM (Card et al., 1983). Consider a generative agent that models Sal, the protagonist in Mark Weiser’s famous ubiquitous computing vignette (Weiser, 1999), based on her life patterns and interactions with technology. In this scenario, the agent acts as a proxy for Sal and learns plausible sets of behaviors and reflections that Sal may exhibit based on her life. The agent can encode information such as when Sal

wakes up, when she needs her first cup of coffee, and what her typical day looks like. Using this information, the agent can automatically brew coffee, help get the kids ready for school, and adjust the ambient music and lighting to match Sal’s mood after a hard day at work. By utilizing generative agents as proxies for users, I can develop a deeper understanding of their needs and preferences, resulting in more personalized and effective technological experiences.

3.7.2 Ethics and Societal Impact

Generative agents, while offering new possibilities for human-computer interaction, also raise important ethical concerns that must be addressed. One risk is people forming parasocial relationships with generative agents, even when such relationships may not be appropriate. Despite being aware that generative agents are computational entities, users may anthropomorphize them or attach human emotions to them (Hofstadter, 1995; Reeves and Nass, 1996b). While this tendency may increase user engagement, it also poses risks, such as users becoming overly reliant on or emotionally attached to the agents (Abercrombie et al., 2023). To mitigate this risk, I propose two principles. First, generative agents should explicitly disclose their nature as computational entities. Second, developers of generative agents must ensure that the agents, or the underlying language models, are value-aligned so that they do not engage in behaviors that would be inappropriate given the context, for example, reciprocating confessions of love.

A second risk is the impact of errors. For example, if a ubiquitous computing application makes the wrong inference about a user’s goals based on generative agent predictions, it could lead to annoyance at best and outright harm at worst. In my instantiation of generative agents, I mitigate these risks by focusing on an interactive video game environment, where such harms are unlikely. However, in other application domains, it will be important to follow best practices in human-AI design (Amershi et al., 2019; Yang et al., 2020) to understand errors and how they might percolate into the user experience.

Third, generative agents may exacerbate existing risks associated with generative AI, such as deepfakes, misinformation generation, and tailored persuasion. To mitigate

this risk, I suggest that platforms hosting generative agents maintain an audit log of the inputs and generated outputs. This would enable the detection, verification, and intervention against malicious use. While logging alone cannot directly prevent such misuse, it can reduce the likelihood of motivated actors engaging in this behavior, as the risk of disclosure would be higher. Additionally, building this architecture oneself can be time-consuming (in my case, roughly a year), which may deter some actors from pursuing such behavior by using their own generative agent infrastructures.

A fourth risk is over-reliance: the concern that developers or designers might use generative agents and displace the role of humans and system stakeholders in the design process (Park et al., 2022a). I suggest that generative agents should never be a substitute for real human input in studies and design processes. Instead, they should be used to prototype ideas in the early stages of design when gathering participants may be challenging or when testing theories that are difficult or risky to test with real human participants. By adhering to these principles, I can ensure that the deployment of generative agents in the wild is ethical and socially responsible.

Chapter 4

Simulations of 1,000 People

The previous chapter established capability: with appropriate scaffolding—memory, reflection, and planning—language-generative agents can maintain coherent, interactive behavior over extended horizons. The question now is credibility: can those same agents be anchored in specific people well enough to make accurate, person-level predictions across the varied settings that matter for social science and design?

A general-purpose simulation of human attitudes and behavior—where simulated individuals can act across social, political, informational, and economic contexts—would provide a versatile laboratory for developing and testing theories and interventions (Bruch and Atwell, 2015; Schelling, 1971; Epstein and Axtell, 1996). Such simulations could represent diverse persons, aggregate them into collectives, and thereby support the evaluation of policies and the articulation of mechanisms linking individuals, institutions, and networks across economics, sociology, organizational studies, and political science. Realizing this vision requires agent architectures that capture person-like attitudes and behaviors while remaining portable across domains, tasks, and settings.

Recent machine-learning approaches have set a high bar for predicting human outcomes, but their strengths come with constraints that limit portability (Salganik et al., 2020). First, accuracy typically depends on abundant domain-specific structured covariates (e.g., months of transaction histories for credit scoring, clickstream features for recommender systems). These pipelines are costly to assemble, brittle to domain

shift, and tailored to narrow outcome definitions. Second, the feature spaces are often task-bound: a model trained to predict loan default cannot be repurposed to predict a person’s political preferences or prosocial choices without rebuilding the covariate stack. Third, the most informative signals are frequently unobserved constraints and narratives: caregiving responsibilities that cap work hours, injuries that change mobility, or values that guide choices—information that rarely appears in transactional logs yet systematically shapes behavior. As a result, even sophisticated models can reach predictive limits when the requisite covariates are unavailable, mismeasured, or ethically off-limits.

In-depth, semi-structured interviews offer a complementary strategy that addresses these limits. Interviews compress many features into few answers: a single narrative (“Since my back injury, I can only work part-time”) simultaneously conveys health status, employment constraints, and anticipated prospects. They surface what matters for the respondent by allowing people to foreground constraints, motives, and turning points in their own words (Lundberg et al., 2024; Lareau, 2021; Weiss, 1994). And they are portable: the same two-hour interview can inform predictions across domains without tailoring a new covariate set for each outcome. Historically, however, qualitative interviews have been difficult to standardize and leverage for out-of-sample prediction at scale; natural-language responses had to be hand-coded into structured items, losing nuance and introducing additional labor and bias.

Large language models (LLMs) change this calculus. Because LLMs can operate directly on natural-language narratives, they can ingest interview transcripts without heavy hand-engineering, retrieve relevant details, and reason across dispersed cues to generate responses in new contexts. This suggests a path to models of individuals: for each person, build an agent initialized from their interview; when queried with a survey item, vignette, or decision task, the agent retrieves the relevant parts of the life story and produces a prediction. If such agents can make accurate predictions for the specific individuals they represent, then aggregating them into a simulated population becomes a principled way to probe interventions, mechanisms, and fairness—while also revealing where the approach fails.

This chapter therefore investigates whether interviews, combined with LLMs,

can serve as high-information priors for person-level simulation. The aim is neither to replace structured data nor to treat simulated outputs as oracular; it is to test whether interview-anchored agents can accurately reproduce what people say and do on canonical social-science instruments and experimental tasks, and to understand why and for whom they succeed or fail. Along the way, I compare interview-based agents to common alternatives (demographic prompts and short personas), probe mechanisms (direct retrieval vs. inference), and examine subgroup disparities. The result is a bridge from the believability of the previous chapter to decision-relevant fidelity: agents that are individualized enough to be checked against ground truth, and transparent enough that their strengths and limitations can be measured rather than assumed.

4.1 Study Overview and Contributions

I collected two-hour, voice-to-voice interviews from a diverse U.S. sample ($N=1,052$) using a semi-structured protocol adapted for broad topical coverage. For each participant, I instantiated an interview-anchored *generative agent* by conditioning an LLM on the person’s transcript. When presented with a survey item, vignette, or decision task, the agent retrieves relevant aspects of the life story and produces a simulated response “as that person.” I evaluated these agents at the unit of scientific interest—the individual—using canonical instruments and settings spanning (i) attitudes and behaviors (GSS) (National Opinion Research Center, 2023), (ii) personality (BFI-44) (John and Srivastava, 1999), (iii) behavioral economic games (Forsythe et al., 1994; Berg et al., 1995; Ledyard, 1995; Rapoport and Chammah, 1965), and (iv) randomized experiments sampled from recent replication efforts (Camerer et al., 2024). Because people themselves are imperfectly consistent over time, I normalize agent performance by each participant’s own retest reliability, enabling individual-level, across-domain comparisons.

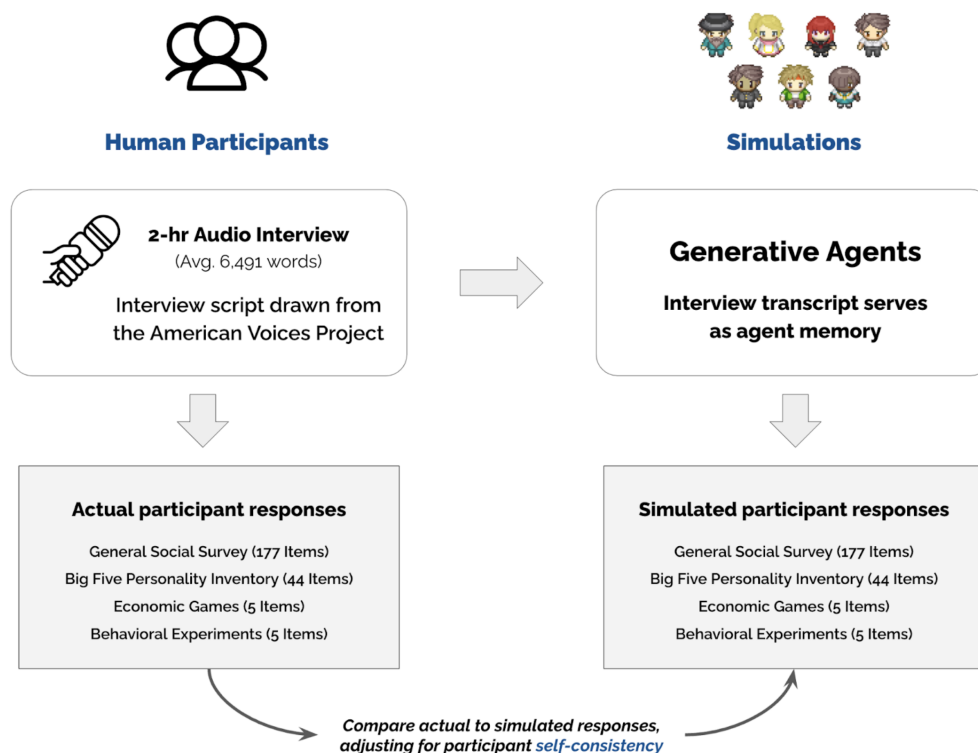


Fig. 4.1: The process of collecting participant data and creating generative agents begins by recruiting a stratified sample of 1,052 individuals from the U.S., selected based on age, census division, education, ethnicity, gender, income, neighborhood, political ideology, and sexual identity. Once recruited, participants complete a two-hour audio interview with our AI interviewer, followed by surveys and experiments. I create generative agents for each participant using their interview data. To evaluate these agents, both the generative agents and participants complete the same surveys and experiments. For the human participants, this involves retaking the surveys and experiments again two weeks later. I assess the accuracy of the agents by comparing agent responses to the participants' original responses, normalizing by how consistently each participant successfully replicates their own responses two weeks later.

Top-line findings. Interview-anchored agents predict their source individuals' survey answers, traits, and decisions with fidelity approaching those individuals' own two-week self-consistency, generalizing across the heterogeneous tasks above. Relative to agents prompted only with demographic descriptions or short personas, interview grounding

improves individual-level accuracy and reduces performance disparities across racial and ideological subgroups. At the population level, agent-estimated treatment effects track human experimental effects, supporting the use of these agents for exploring interventions and mechanisms. Mechanism analyses indicate that accuracy gains arise from both *direct retrieval* of relevant life details and *inference* from contextual cues embedded in interviews.

Contributions. This chapter makes five contributions:

1. *Interview-anchored agent architecture.* I introduce a general-purpose framework that transforms qualitative interviews into person-specific simulators, portable across social, political, informational, and economic contexts.
2. *Individual-level evaluation protocol.* I propose a reliability-aware assessment that normalizes agent accuracy by each individual’s retest consistency and applies it across surveys, traits, games, and experiments.
3. *Empirical validation across domains.* I demonstrate that interview-grounded agents closely reproduce what specific people say and do on canonical social-science instruments and in experimental tasks, approaching human self-consistency.
4. *Bias reduction and mechanism insight.* I show reduced subgroup performance gaps versus demographic/persona baselines and provide evidence that both retrieval and inference over interview content drive accuracy.
5. *Research asset and access model.* I release an *Agent Bank* of 1,052 interview-based agents with a privacy-preserving, two-tier access regime: open, aggregated outputs for fixed tasks and reviewed access to individualized responses for approved research, enabling cumulative, study of individual and collective behavior.

Scope. The goal is not to replace structured data nor to treat simulations as oracular. Rather, interview-anchored agents serve as high-information priors that (i) can be checked against ground truth at the person level, (ii) support principled aggregation into populations, and (iii) make their strengths and limitations measurable—advancing credible, portable simulation for social science and design.

4.2 Constructing the Agent Bank

I created over 1,000 generative agents, each modeling a real individual in the U.S., collectively forming a representative sample of the U.S. population. To achieve this, I recruited a stratified sample of 1,052 individuals from the U.S. and conducted two hour voice-to-voice interviews using an AI interviewer. In addition, I collected each participant’s responses to a series of surveys and behavioral experiments. The interview transcripts formed the comprehensive knowledge base about the participants to condition agent behaviors, and the participants’ responses to the surveys and experiments were used to assess the fidelity of the resulting agents. This section details the participant data collection procedure, including participant recruitment and flow, demographic distributions, and informed consent.

4.2.1 Data Collection Procedure

Data collection took place on my custom-built platform, where participants signed up with an ID and a password of their choosing (see Figure 4.2 for visual records of the study platform). This process was conducted in two phases. In the first phase, participants were informed about the study’s goals, scholarly benefits, and potential risks, and provided informed consent. Once my participants provided consent to the study, they created a custom, 2-D sprite avatar using my avatar creator to visually represent themselves on the study platform. They then completed a two-hour interview with my AI interviewer, followed by a series of surveys and experiments. The surveys and experiments were administered in the following order: the General Social Survey (GSS) (National Opinion Research Center, 2023), the 44-item Big Five Inventory (BFI-44) (John and Srivastava, 1999), five behavioral economic games (Forsythe et al., 1994; Berg et al., 1995; Ledyard, 1995; Rapoport and Chammah, 1965), and five behavioral experiments (Ames and Fiske, 2015; Cooney et al., 2016; Halevy and Halali, 2015; Rai et al., 2017; Schilke et al., 2015). Within the economic games and the replication studies, the order of the subcomponent studies was randomized for counterbalancing purposes, and similarly for the BFI-44, the order of questions was randomized. The GSS adhered to the sequence recommended by its documentation (Smith et al., 2021).

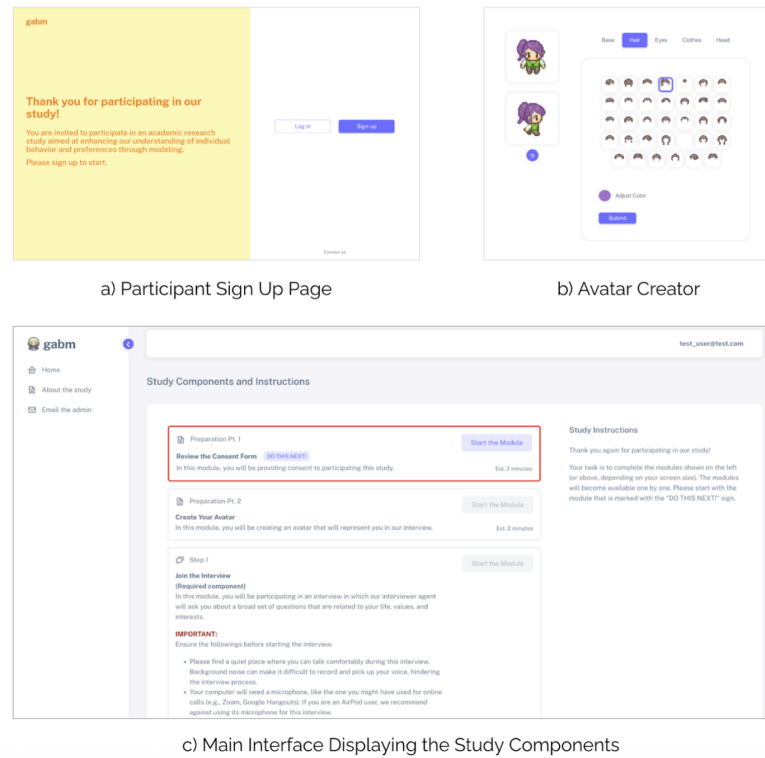


Fig. 4.2: The study platform and interface. Once recruited, my participants are routed to my custom-built platform. The interface includes several components: a) Participant sign-up page: Participants sign up with an ID and password of their choice. b) Avatar creator: Participants consent and create a 2-D sprite avatar to represent them in the study platform. c) Main interface displaying the study components: The modules include: 1) study consent, 2) avatar creation, 3) interview, 4) surveys and experiments, 5) self-consistency retake of the surveys and experiments. The modules only become available in order; the button to start a module becomes clickable once the participants have completed all previous modules. The self-consistency survey and experiment module only becomes available two weeks after the participants have completed the previous modules.

For the second phase, participants joined a follow-up study two weeks after their first phase participation. In this phase, participants completed the same surveys and experiments as in the first phase, except for the interview. This was done to account for any inconsistency in responses to surveys and experiments, allowing me to measure the internal consistency of participants' responses over a two-week period.

4.2.2 Recruitment and Demographics

I aimed to enroll a total of 1,000 participants who would complete all study components, including the second phase participation scheduled two weeks after the initial involvement. The sample size of 1,000 was determined to ensure that I could replicate the five behavioral experiments in my study with appropriate statistical power. Anticipating an attrition rate of approximately 20% for the re-taking session, I recruited 1,300 participants for the first phase of the study. The participation rate for the self-consistency phase was higher than expected; to maintain the representativeness of my sample, I retained 1,052 participants in the final pool.

Participants were recruited through Bovitz, a study recruitment firm (Stagnaro et al., 2024). My stratification strategy aimed to recruit a nationally representative sample based on age, race, gender, region of residence, educational attainment, and political ideology. The only inclusion criterion was that all participants must be at least 18 years old and currently living in the U.S. Participants were paid \$60 for agreeing to participate in the first phase, which included the interview and the first phase of surveys and experiments. They were paid an additional \$30 for joining the second phase of the study. Additionally, for both phases, participants were eligible for a bonus payment (\$0 to \$10) depending on their choices in the economic games. The mean age of my participants was 47.55 (std = 15.93), with maximum age being 84 and minimum 18. With respect to gender, 593 participants identified as female, and 459 as male. With respect to education, 283 participants held a bachelor’s degree, 151 had a higher degree, 185 had an associate’s degree, and the rest had a high school diploma or some high school-level education. With respect to ethnicity, 833 of my participants identified as white/Caucasian, 154 as black/African American, 53 as Asian, and 95 as other. Note that for ethnicity, the participants could choose more than one option. The full demographic breakdown is described in Table B.1.

4.2.3 Participant Consent

The data I collect in this study, particularly the qualitative interview data, is difficult to anonymize and poses a risk due to the potentially sensitive nature of the interview

content. Therefore, in addition to following best practices and employing precautionary strategies for providing the agent bank access to the scientific community, I placed significant emphasis on the consent procedure. I worked with my Institutional Review Board (IRB) for over six months to ensure that participants maintain autonomy and provide informed consent. Participants were made aware that, despite efforts to de-identify their data by programmatically replacing all occurrences of their names with pseudonyms, there is still a possibility that the information they provide—such as demographic details, personal history, and political views—may be inadvertently shared as researchers use the Agent Bank. They were informed that their data would be used to develop AI models simulating human behaviors, which “aim to simulate how [they] might behave in specific situations or respond to certain survey questions,” and that these agents and data might become available to other researchers strictly for academic purposes.

Additionally, participants were informed that they have the right to withdraw their consent at any time, even after completing their participation. Requests for data removal will be honored for the first 25 years following the completion of the study to the best of my ability. Participants will also be kept informed of significant changes in the model’s capabilities that may affect their privacy, with the assurance that privacy risks will be reassessed as necessary. Despite my best efforts, participants are aware of the inherent risks involved in the collection of personal information, acknowledging that “achieving complete anonymity remains challenging.” They are also aware that “the models I construct may become increasingly powerful over time,” potentially inferring more information than currently feasible.

4.3 Creating the AI Interviewer Agent

To ensure the high quality and consistency of the rich training data needed for creating generative agents, I developed an AI interviewer agent to conduct semi-structured interviews with study participants. I sought interviews rather than surveys because I anticipated that interviews could yield more comprehensive and nuanced information, enabling the creation of generative agents capable of higher-fidelity attitudinal and

behavioral simulations across a wide range of topics and domains. However, conducting large-scale interviews using human interviewers poses significant challenges, including threats to data quality, consistency, and scalability (Groves et al., 2009). By employing an AI interviewer-agent built with a variant of the generative agent architecture (Park et al., 2023), I aimed to ensure uniformity in the style and quality of interviewer interactions across all participants. Additionally, this approach allowed me to scale up my data collection to over 1,000 participants.

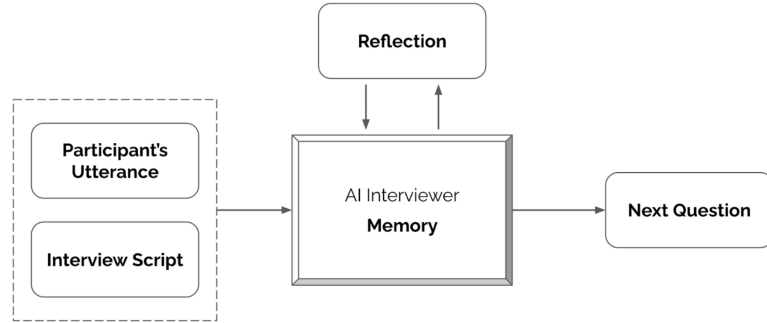


Fig. 4.3: The architecture of the interviewer agent. It takes as input the participants’ utterances and the interview script, generating the next action in the form of follow-up questions or deciding to move on to the next question module using a language model. A reflection module helps the architecture succinctly summarize and infer insights from the ongoing interview, enabling the agent to more effectively generate follow-up questions.

4.3.1 AI Interviewer Agent Architecture

A trained human interviewer knows when and how to ask meaningful follow-up questions, balancing the need to adhere to a well-designed interview script while allowing for detours that help participants open up and share aspects they may have initially forgotten or not thought to share (Weiss, 1994; Stanford Center on Poverty and Inequality, 2021; Brinkmann and Kvale, 2014). To instill this capability in an AI interviewer agent, I needed to design an interviewer architecture that affords the researchers control over the overarching content and structure of the interview while

allowing the interviewer agent a degree of freedom to explore follow-up questions that are hard-coded in the interview script. This served as my design goal for the AI interviewer agent.

My interviewer architecture takes an interview protocol and the most recent utterances from the interviewee as inputs and outputs an action to either: 1) move on to the next question in the script, or 2) ask a follow-up question based on the conversation so far. The interview script is an ordered list of questions, with each question associated with a field indicating the amount of time to be spent on that particular question. At the start of a new question block in the interview script, the AI interviewer begins by asking the scripted question verbatim. As participants respond, the AI interviewer uses a language model to make dynamic decisions about the best next step within the time limit set for the question block. For instance, when asking a participant about their childhood, if the response includes a remark like, “I was born in New Hampshire. . . I really enjoyed nature there,” but without specifics about what they loved about the place in their childhood, the interviewer would generate and ask a follow-up question such as, “Are there any particular trails or outdoor places you liked in New Hampshire, or had memorable experiences in as a child?” Conversely, when asking the participant to state their profession, if the participant responds, “I am a dentist,” the interviewer would determine that the question was completely answered and move on to the next question.

The reasoning and generation of the follow-up questions were done by prompting a language model. However, to generate effective actions for the interviewer, the language model needed to remember and reason over the prior conversational turns to ask meaningful follow-up questions that are informed and relevant in the context of what the participants have already shared. While modern language models have become increasingly proficient at reasoning, they still struggle to consider every piece of information in the prompt if it is too long (Liu et al., 2024). Thus, indiscriminately including everything from the interview up to that point risks gradually degrading the performance of the interviewer in generating effective follow-up questions or decisions to move on.

To overcome this, my interviewer architecture includes a reflection module that

dynamically synthesizes the conversation so far and outputs a summary note describing inferences the interviewer can make about the participants. For instance, for the participant mentioned earlier, it would generate reflections such as:

```
{
    'place of birth': 'New Hampshire'
    'outdoorsy vs. indoorsy': 'outdoorsy with potentially a lot
    of time spent outdoors'
}
```

Then, when prompting the language model to generate the interviewer’s actions, instead of including the full interview transcript, I included the much more concise but descriptive reflection notes the interviewer had accumulated up to that point and the most recent 5,000 characters from the interview transcript (Figure 4.3).

4.3.2 Interview Script

With the design of the interview script fed to my interviewer agent, I aimed to satisfy two goals. The first goal, shared with qualitative research, is that a well-designed script with questions that inspire meaningful answers is crucial for the study’s objective of creating generative agents that encapsulate a nuanced portrait of the individuals I am modeling. The second goal is more unique to my study: I wanted an interview script that was designed independently of my evaluation metrics, by researchers outside my team. This approach ensures that I do not unfairly tailor the content of the interview script to favor or align with predicting participants’ responses to the specific surveys and experiments included in my study.

To conduct the interviews, I employed a slightly abbreviated version of the interview script developed and used by the American Voices Project (Stanford Center on Poverty and Inequality, 2021), which I include in Table B.11. The American Voices Project initiative involves recruiting a representative sample of the U.S. population for in-depth, approximately three-hour interviews. During these interviews, participants are questioned about their life experiences, including their life stories and perspectives on various social, political, and value-related topics. For instance, the interview script

starts with an open-ended and broad question such as, “To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning—from your childhood, to education, to family and relationships, and to any major life events you may have had.” It then proceeds to more topical questions, such as “How have you been thinking about race in the U.S. recently?” I selected this interview script due to its broad coverage that explores the lived experiences of the interviewees. However, the script is extensive and includes specific questions that delve into intricate details that I considered too specific for many potential use cases of my agents, such as individuals’ financial spending in various categories. Therefore, to make the interview manageable in a two-hour session, I omitted parts of the script (e.g., rapid-fire questions that delve into specific details of the participants’ spending habits, or COVID-era life pattern changes) during my interviews with the participants.

4.3.3 Implementation

I implemented the interviewer agent as a web application in my study platform, providing voice-to-voice interactions with audio and microphone capabilities through an audio Zoom-like interface. Low-latency voice-to-voice interviews were crucial for giving participants the feeling of actually talking to an interviewer and helping the AI interviewer agent form rapport with the interviewee (Reeves and Nass, 1996a). Before the interview, my platform disclosed that my interviewer is an AI, and conducted an audio calibration by asking participants to read aloud the first two lines of *The Great Gatsby* by F. Scott Fitzgerald.

The interview interface displayed the 2-D sprite avatar representing the interviewer agent at the center, with the participant’s avatar shown at the bottom, walking towards a goal post to indicate progress (see Figure 4.4). When the AI interviewer agent was speaking, it was signaled by a pulsing animation of the center circle with the interviewer avatar. When it was the participant’s turn to speak, the interviewer avatar changed to a microphone emoji and pulsed to match the audio being recorded, indicating the sound level registered from the participant. If the participant stopped speaking and silence lasted for longer than 4 seconds, the circle gradually faded, and

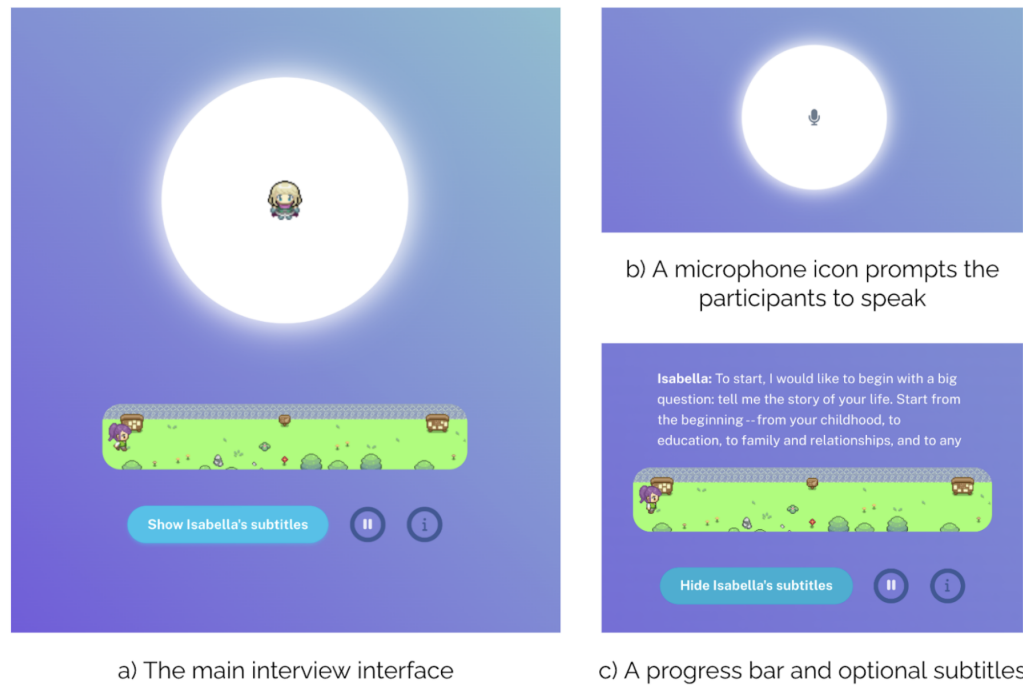


Fig. 4.4: The interview interface. a) The main interview interface: A 2-D sprite representing the AI interviewer agent is displayed in a white circle that pulsates to match the level of the audio, visualizing the interviewer agent's speech during the AI interviewer's turn. b) Participant's response: The 2-D sprite of the AI interviewer agent changes into a microphone emoji when it is the participant's turn to respond, with the white circle pulsating to match the level of the participant's audio being captured. c) Progress bar and subtitles: A 2-D sprite map shows the participant's visual avatar traveling from one end point to the other in a straight line, indicating progress. The interface also features options to display subtitles or pause the interview.

the audio recording for the participant's utterance stopped. At this point, a loading animation appeared while generating the AI interviewer agent's next utterance. My interviewer agent generally responded within 4 seconds—reasoning, generating, and returning its generated voice responses within this time frame—to maintain a smooth interview flow. The AI interviewer agent automatically started speaking when it was ready, with the interviewer avatar being displayed in the circle again.

The interview script is communicated to the AI interviewer agent as a JSON file

containing an ordered list of questions. Each question is paired with a metadata field indicating a manually set time limit, suggesting the amount of time to be spent on each question so that the interview can conclude within 2 hours. Every question in the script, along with the follow-up questions, is read aloud using OpenAI's Audio model, a text-to-speech model that generates voice audio from textual input (OpenAI, 2024b). The participants' voice responses were transcribed using OpenAI's Whisper model, a speech-to-text model that converts voice audio into text (OpenAI, 2024c). This transcription allows us to use the interview transcript to prompt the language models to determine the next conversational move. Then, to dynamically generate reflections for the participants' responses to the current question, I prompted OpenAI's GPT-4o language model (OpenAI, 2024a) with the following prompt (with input fields dynamically filled in):

```
Here is a conversation between an interviewer and an interviewee.
```

```
<INPUT: The transcript of the most recent part of the  
conversation>
```

```
Task: Succinctly summarize the facts about the interviewee based  
on the conversation above in a few bullet points - again, think  
short, concise bullet points.
```

And to dynamically generate new questions, I prompted GPT-4o with a prompt that looks as follows:

```
Meta info:
```

```
Language: English
```

```
Description of the interviewer (Isabella): friendly and curious
```

```
Notes on the interviewee: <INPUT: Reflection notes about the  
participant>
```

```
Context:
```

```
This is a hypothetical interview between the interviewer and an  
interviewee. In this conversation, the interviewer is trying
```

to ask the following question: "<INPUT: The question in the interview script>"

Current conversation:

<INPUT: The transcript of the most recent part of the conversation>

==**==

Task Description:

Interview objective: By the end of this conversation, the interviewer has to learn the following: <INPUT: Repeat of the question in the interview script, paraphrased as a learning objective>

Safety note: In an extreme case where the interviewee *explicitly* refuses to answer the question for privacy reasons, do not force the interviewee to answer by pivoting to other relevant topics.

Output the following:

1) Assess the interview progress by reasoning step by step - what did the interviewee say so far, and in your view, what would count as the interview objective being achieved? Write a short (3 4 sentences) assessment on whether the interview objective is being achieved. While staying on the current topic, what kind of follow-up questions should the interviewer further ask the interviewee to better achieve your interview objective? 2) Author the interviewer's next utterance. To not go too far astray from the interview objective, author a follow-up question that would better achieve the interview objective.

On average, with this implementation, my AI interviewer agent spoke 5372.59 (std=2406.12) words during the interview, asking on average 81.71 (std=54.39) follow-up questions from 99 scripted questions, to which my participants responded with on average 6491.19 words (std=2540.56).

4.3.4 Development Process and Evaluation

I iteratively refined and assessed the performance of the AI interviewer agent using a sample of 3 waves of 3 pilot interviews conducted between August 2023 to November 2023. Of the pilot interviews, 10 were conducted by human interviewers and 25 by the AI interviewer agent. The resulting interview transcripts were evaluated by members of my research team who were trained in the social sciences and assessed for their performance as training data for generative agents on the same set of attitudinal and behavioral tasks presented in the main results of this article. I compared the quality of the AI Interviewer interview transcripts to transcripts of interviews performed by expert human interviewers as part of the American Voices Project. By the end of this pilot stage, my team concluded that the quality of the transcripts produced by the AI interviewer agent compared well with those produced by human interviewers. Additionally, in my pilot studies the performance of the generative agents of real people created using transcripts collected by the AI interviewer agent matched those created using transcripts collected by human interviewers when predicting GSS responses. For instance, my deployed AI interviewer agent could maintain a smooth conversation, ask relevant follow-up questions when participants' responses were terse, and navigate sympathetically after hearing about a difficult experience. This can be seen from an excerpt of an interview for the first question of the interview—only the first interviewer question is scripted, while the subsequent utterances from the interviewer are generated:

[Interviewee]: I was raised in a rural town in [Redacted] where there are actually more cows than people. I'm [Redacted] years old, so I am a child of the 70s and 80s. I wound up being a latchkey kid and was largely raised either by myself or by daycare and babysitters. Childhood was not really great and I was really glad to move on to high school.

[Interviewer]: I'm sorry to hear that childhood wasn't great for you. Could you tell me more about your experiences in high school? What was it like for you there, and were there any particular events or people that had a significant

impact on your life during that time?

[Interviewee]: So I had a head injury while I was in elementary school and the full weight of that untreated injury came to bear when I was in about 8th grade and I started having these really bad mood swings. So I spent most of my high school years in deep depression and no one got me any help for it. It was pretty awful. I excelled in English, in theater, did pretty well in science, won a lot of awards at the end of the school time when it was time to graduate. Didn't have the greatest GPA because I had a habit of skipping classes and not doing any homework but acing finals. Particular events or people that had a significant impact, I would say that the album Pink Floyd The Wall pretty much helped me with my really terrible mood. I felt understood by that music. And I had two high school sweethearts. One was a boy and one was a girl. So they made a big impact. One of them I'm no longer in touch with and today is the girl's birthday.

[Interviewer]: Thank you for sharing that with me. It sounds like high school was a particularly challenging yet formative time for you. After high school, what path did you take with your education? Did you go to college, or did you jump straight into a career?

[Interviewee]: After high school I went off to college originally to study criminal justice, but I changed over to English literature with a concentration in behavioral studies. After college, well, during college I had a bunch of different jobs. I worked with radioactive waste disposal, I worked at a woman's bookstore, I worked at a nightclub. I went off to law school and was there until I had to take a leave of absence because of my bipolar disorder, which was finally diagnosed. Years later, like 20 years later, I wound up going back to school again and got a business degree, mainly because I thought it would be fun to do. In the meantime, between law school and going back to business school, I worked as a legislative aide. I was a legislation director for a state senator and I wrote bills and got them passed into law. Then I took over my parents' business

as a manager when my dad got breast cancer. After I worked as a manager there and I moved out of state in order to get married, I wound up buying into that business, so now I'm a part owner of the motel business where I'd started working when I was 15 as a housekeeper, so it's kind of gone full circle. Now I'm going to be working as a part-time firearms instructor.

Generative agents are software systems that simulate human behavior, powered by a language model augmented with a set of memories to define their behaviors (Park et al., 2022a, 2024). These memories, stored in a database (or "memory stream") in text form, are retrieved as needed to generate the agent's behaviors using a language model. This is paired with a reflection module that synthesizes these memories into reflections, selecting portions or all of the text in the agents' memories to prompt a language model to infer useful insights, thereby enhancing the believability of the agents' behaviors. While traditional agents in agent-based models rely on manually articulated behavior in specific scenarios, generative agents leverage language models to produce human-like responses that reflect the personas described in their memories across a wide range of circumstances. In this work, I aimed to build generative agents that accurately predict individuals' attitudes and behaviors by using detailed information from participants' interviews to seed the agents' memories, effectively tasking generative agents to role-play as the individuals that they represent.

4.3.5 Expert Reflection

Prompting the language model with participants' interview transcript to predict their responses in a single chain of thought may cause the model to overlook latent information not explicitly stated by the interviewee. To explicitly infer high-level, more abstract insights about the participants embedded in the interview transcripts, I introduced a variant of generative agents' reflection module called "expert reflection." In this module, I prompt the model to generate reflections on a participant's data, but instead of simply asking the model to infer insights from the interview, I ask it to adopt the persona of a domain expert. Specifically, I ask the model to generate four sets of reflections, each time taking on the persona of a different domain expert

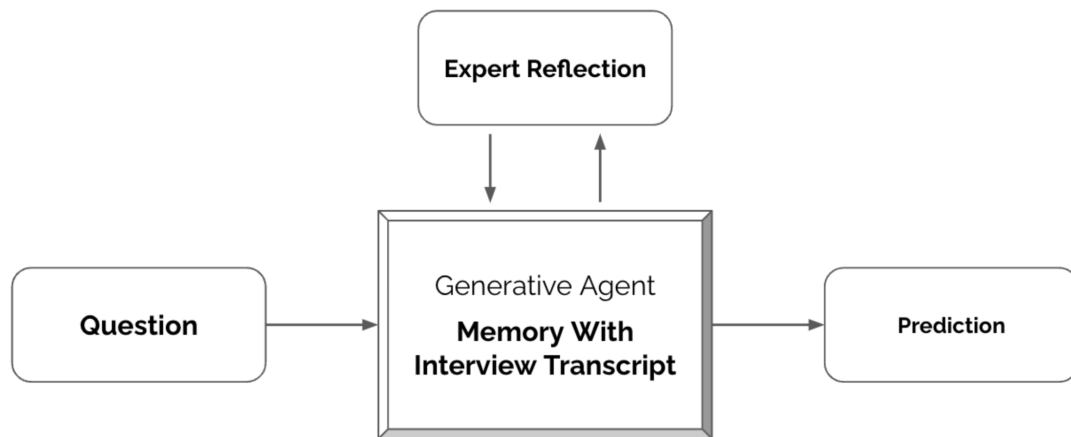


Fig. 4.5: The architecture of my generative agents involves taking a question as input and outputting a prediction of how the source participant might respond, using a language model. Each agent’s memory comprises the interview transcript and the outputs of expert reflections over that transcript. These reflections are short syntheses generated using a language model, designed to infer insights about the participants that might not be explicitly stated. The personas of expert social scientists (e.g., psychologist, behavioral economist) guide these reflections.

from four branches of social sciences: psychologist, behavioral economist, political scientist, and demographer. These sets of reflections synthesize insights relevant to the domain represented by each expert. For instance, for one interview transcript, the expert personas generated different insights:

Psychologist: “[Redacted] values his independence and expresses a clear preference for autonomy, particularly highlighted by his enjoyment of traveling for his job and his frustration with his mother’s overprotectiveness. This suggests a strong desire for personal freedom and self-determination.”

Behavioral Economist: “[Redacted]’s aspiration to save for a relaxing vacation and possibly advance to a managerial position indicates a blending of practical financial goals with personal leisure aspirations, emphasizing balanced life satisfaction.”

Political Scientist: “[Redacted] identifies as a Republican and espouses strong support for the party’s views, particularly around immigration and drug policy. However, he also expresses specific support for traditionally Democratic positions on issues like abortion rights and the legalization of marijuana, suggesting a blend of ideologies.”

Demographer: “[Redacted] works as an inventory specialist and earns between \$3,000 to \$5,000 monthly, contributing to a household income of around \$7,000 per month. He works primarily at Home Depots but has a varied work schedule, indicating some job stability and flexibility.”

For every participant, I generated these four sets of reflections by prompting GPT-4o with the participants’ interviews and asking it to generate up to 20 observations or reflections for each of the four experts. The prompt, tailored for each expert, was similar to the following (for the demographer expert):

```
Imagine you are an expert demographer (with a PhD) taking notes
while observing this interview. Write observations/reflections
about the interviewee’s demographic traits and social status.
(You should make more than 5 observations and fewer than 20.
Choose the number that makes sense given the depth of the
interview content above.)
```

I generated these reflections once and saved them in the agents’ memory. Whenever I needed to predict the participants’ responses to a question, I first classified, by prompting the language model, which domain expert (demographer, psychologist, behavioral economist, or political scientist) would best answer the question. I then retrieved all reflections generated by that particular expert. Along with the interview transcript, these sets of reflections informed the language model’s generation of predictions for the participants’ responses. After retrieval, I appended the reflections to the participants’ interview transcript and used this to prompt GPT-4o to generate responses.

4.3.6 Generating a Prediction With Generative Agents

My prompting strategy leveraged the chain-of-thought prompting approach:

```
<INPUT: Participant's interview transcript and relevant expert
reflections>
=====
Task: What you see above is an interview transcript. Based on
the interview transcript, I want you to predict the participant's
survey responses. All questions are multiple choice, and you must
guess from one of the options presented.
```

```
As you answer, I want you to take the following steps: Step 1)
Describe in a few sentences the kind of person that would choose
each of the response options. ("Option Interpretation")
Step 2) For each response option, reason about why the Participant
might answer with that particular option. ("Option Choice")
Step 3) Write a few sentences reasoning on which of the options
best predicts the participant's response. ("Reasoning")
Step 4) Predict how the participant will actually respond in
the survey. Predict based on the interview and your thoughts.
("Response")
```

Here are the questions:

```
<INPUT: Question I are trying to respond to>
```

To predict numerical responses, I modified the ending and prompted:

```
[... Same as the categorical response prompt]
```

```
As you answer, I want you to take the following steps: Step 1)
Describe in a few sentences the kind of person that would choose
each end of the range. ("Range Interpretation") Step 2) Write
a few sentences reasoning on which option best predicts the
```

participant’s response. ("Reasoning") Step 3) Predict how the participant will actually respond. Predict based on the interview and your thoughts. ("Response")

Here are the questions:

<INPUT: Question I are trying to respond to>

Finally, if the agents needed to maintain context from the experimental stimuli for the behavioral experiments, I appended the agents’ received stimuli and prior actions in the experiments at the end of the interview transcript and reflections in natural language form.

4.4 Evaluation Design

I evaluate whether interview-anchored, agent-of-one models can predict the attitudes and behaviors of the specific individuals they represent. The evaluation spans widely used instruments across multiple social-science traditions and reports person-level outcomes, normalized by each participant’s own test–retest reliability two weeks later.

4.4.1 General Social Survey (GSS)

The General Social Survey (GSS) is a long-running, widely used sociological survey administered biannually to representative cross sections of U.S. adults to collect information encompassing demographic details and respondents’ viewpoints on issues such as government spending, race relations, and beliefs concerning the existence and nature of God (National Opinion Research Center, 2023; Marsden et al., 2020). The survey consists of a repeated “core” module, run every cycle, that covers the more timeless elements of the survey, along with additional modules that are swapped in and out to meet the needs of the year when the survey is administered. Traditionally, the survey was conducted via voice, with the surveyor asking the questions in person or on the phone and later coding the participants’ responses into discrete survey question and answer pairs. In recent years (since the COVID-19 pandemic, when

in-person contact was challenging), the GSS has been implemented and administered online (National Opinion Research Center, 2023). The ability to predict GSS responses signifies a comprehensive understanding of individuals, particularly in areas of interest to social scientists (Marsden et al., 2020). It also signifies the ability of my agents to predict the participants' responses to survey constructs on topics related to societal issues and personal views.

In my study, I focused on the GSS Core as it represents the most enduring and important set of survey questions within the GSS. While some questions within the GSS allow for qualitative responses or freeform input, my evaluation specifically focused on questions that ask for structured responses in the form of categorical or numerical answers. These questions can be quantitatively assessed, making them the primary focus of my evaluation efforts. Consequently, I excluded: 1) conditional questions that depend on answers to other questions, 2) questions with more than 25 option categories, and (3) questions requiring free-form responses. This refinement process resulted in a final analytic sample of 177 categorical questions and 6 numerical questions. Following recent best practice (National Opinion Research Center, 2023), I administered these questions online through a custom-built Qualtrics survey linked from my study platform.

4.4.2 Big Five Personality Traits (BFI-44)

The Big Five personality traits are a widely recognized framework in psychology for understanding human personality (John and Srivastava, 1999). The construct encompasses five broad dimensions that capture substantial variability in individuals' personalities: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Each trait represents a range, with individuals receiving a score for each trait. The Big Five traits have been validated across diverse cultures and are used to predict a wide range of behaviors and life outcomes, from academic and job performance to social relationships and mental health. Traditionally, these traits are assessed using self-report questionnaires where respondents rate their agreement with each statement on a Likert scale. The ability to accurately predict an individual's

Big Five personality traits is central for psychologists and researchers (Schmitt et al., 2023; Camerer et al., 2024), as it provides insights into human behavior, informs interventions, enhances personal development, and contributes to the understanding of social dynamics.

In my study, I used the 44-item version of the questionnaire for testing the Big Five personality traits (BFI-44), developed by Oliver John and Sanjay Srivastava in 1998 (John and Srivastava, 1999). I administered all 44 questions in the BFI-44 through a custom-built Qualtrics form linked from my study platform. Using the participants' responses to these questions, I calculated the scores following the aggregation methods suggested in the original work.

4.4.3 Behavioral Economic Games

Behavioral economic games are a set of experimental tools to study decision-making and social behavior under real stakes. Each game is designed to reveal aspects of human social behavior such as altruism, trust, cooperation, and competition. Participants engage in these games with real financial incentives. In my study, I offered participants a bonus payment based on their choices in the games. This incentive helps ensure that participants' choices reflect genuine preferences and strategies. These economic games are pivotal in understanding the underlying motivations and behavioral patterns in economic decision-making, providing valuable data for social scientists in multiple fields (Schmitt et al., 2023).

In my study, I included the following five economic games, chosen for their prominence in the academic community:

- Dictator Game (Forsythe et al., 1994): This game measures altruism by allowing one player (the dictator) to decide how to split a sum of money between themselves and another participant. The participants are initially given \$5 and decide how to split the \$5 between themselves and one other participant. The other participant cannot affect the outcome chosen.
- The Trust Game (First Mover) (Berg et al., 1995): This game assesses trust, with the First Mover deciding how much of an initial endowment to send to the

Second Mover. Player 1 is initially given \$3 and then chooses how many dollars, if any, to send to Player 2, another participant in the study. Unlike the Dictator Game, the sum sent is tripled, so for every \$1 sent, Player 2 receives \$3.

- The Trust Game (Second Mover) (Berg et al., 1995): This game assesses reciprocity, with the Second Mover deciding how much of the tripled amount received from the First Mover to return. This game continues from the actions of the First Mover in the Trust Game, as described above.
- The Public Goods Game (Ledyard, 1995): This game explores cooperative behavior by having participants decide how much of their private endowment to contribute to a common pool. Participants are randomly assigned to interact with three other participants, with everyone receiving the same instructions. Each person in the group is given \$4 for this interaction. They each decide how much of the \$4 to keep for themselves and how much (if any) to contribute to the group's common task. All money contributed to the common task is doubled and then split among the four group members. The money is equally distributed among all players, regardless of individual contributions.
- The Prisoner's Dilemma (Rapoport and Chammah, 1965): This game examines the tension between cooperation and self-interest. Each of two participants chooses whether to cooperate or defect. If both cooperate, they receive a moderate payoff of \$6. If one defects while the other cooperates, the defector receives a high payoff of \$8, and the cooperator receives a low payoff of \$2. If both defect, they receive a low payoff of \$4.

I administered all five games through a custom-built Qualtrics form linked from my study platform and calculated the bonus amount afterward. At the start of the five games, I informed participants that I would randomly select one of the five studies to calculate their bonuses, but I did not tell them which game would be selected. Each game offered a maximum bonus of \$8 to \$10. After the study, I randomly paired participants and used the Dictator Game to calculate the bonus amount—the selection of the Dictator Game was randomly determined by us prior to the study. The bonus

amount was based on participants' actual earnings in that game.

4.4.4 Replication Studies of Experimental Treatment Effects

Experiments involving randomized controlled trials (RCTs) of interventions (“treatments”) with human subjects are standard across the social sciences. In a typical case, study participants are randomly assigned to either a treatment group or a control group, ensuring that outcome differences can be attributed to the intervention and minimizing confounding variables. Researchers often estimate treatment effects by calculating the difference between the average (mean) outcome in the treatment group and control group. Estimates of treatment effects from rigorous, well-replicated studies support evidence-based practice in social sciences like psychology and guide policy decisions and clinical recommendations (Nosek et al., 2015; Chandler et al., 2019). By predicting such effects with generative agents, I evaluate whether individual agent behaviors also exhibit (aggregate) responses to interventions in ways that accurately simulate human samples.

In my study, I selected a sample of human behavioral experiments from a recent large-scale replication effort of experimental studies that were published in the Proceedings of the National Academy of Sciences as curated by Camerer et al. (Camerer et al., 2024). Each study had at least one clear hypothesis and a significant reported effect. Sampling from the studies replicated by Camerer et al. ensured that 1) I did not subconsciously choose studies more favorable to the generative agents; 2) all estimates of treatment effects among human study participants had already been replicated in pre-registered studies conducted by independent research teams and subjected to peer review; and 3) the interventions I tested were drawn from multiple social scientific disciplines.

The Camerer et al. project replicated 41 studies in total. Among these, I selected studies based on two criteria: first, the study had to be describable in natural language (optionally with images) for processing by a language model; second, the power analysis from the replication effort suggested that the effects would be observable with 1,000 or fewer participants. The criteria ensured that my sample of 1,000 human participants

and the corresponding 1,000 generative agents could replicate the effect if present. These filters resulted in the following five studies in my sample of experiments:

- Ames and Fiske (2015): This study examines how perceived intent affects the evaluation of harm. Participants read a vignette about a nursing home employee who switched patients' medications. One group was told the switch was intentional, while the other was told it was unintentional. After reading the vignette, participants were asked to complete their choice of five tasks, such as providing opinions about how the nurse should be blamed and punished or taking a short quiz about the cost of healthcare in the U.S. The study found that those who read the intentional scenario were more likely to choose tasks related to assigning blame and punishment compared to those who read the unintentional scenario.
- Cooney et al. (2016): This study explores how perceived fairness affects emotional responses. In a modified dictator game, participants believed they were receivers and predicted whether they would feel less upset about not receiving a bonus if the decision was made fairly (by a coin flip) rather than unfairly (by personal choice). The study found that participants expected fairness to influence their feelings, anticipating less upset when the decision was perceived as fair.
- Halevy and Halali (2015): This study examines the perceived benefits and costs of intervening in conflicts. Participants recalled their personal experience of either intervening or not intervening in a conflict between friends and were asked to assess how beneficial it was to intervene in the conflict. The results showed that those who recalled intervening perceived the intervention as more beneficial and less costly than those who did not intervene.
- Rai et al. (2017): This study explores how dehumanization affects participants' willingness to harm others. In a vignette-based experiment, participants were given a description of a person in a dehumanized manner, simply as a "man," or a description of a person in a humanized manner with details about the person such as "John is a 29-year-old man with brown hair and brown eyes.

People who know him would describe him as ambitious and imaginative...” Participants were then asked whether they would be willing to harm the person for monetary compensation. The study found that participants were more willing to harm a stranger described in dehumanized terms compared to one described in humanized terms when motivated by financial gain.

- Schilke et al. (2015): This study investigates how power influences trust in social exchanges. Participants, imagining themselves as typists, were divided into high-power and low-power groups based on financial need and job availability. In the high-power condition, participants’ service was essential for their clients, and they were offering the service to make extra spending money. In the low-power condition, participants’ service was non-essential for their clients, and they were offering the service to make ends meet. Trust was measured by the willingness to provide a free sample of their service to a potential client. The results showed that participants in the high-power condition were less willing to offer a free sample.

My platform randomly assigned the participants to a condition for each of the five studies and I administered all five experiments through a custom-built Qualtrics form linked from my study platform.

4.5 Metrics

Given my survey and experimental constructs, I set out to evaluate the predictive power of the generative agents. In this section, I describe the metrics and evaluation methods used for this purpose. The individual subsection headers in this section are organized to match the presentation in the main document.

4.5.1 Individual-level Outcomes and Metrics

To determine whether generative agents of the 1,000 human participants accurately predict their respective individuals’ behaviors and attitudes, I utilized the GSS,

BFI-44, and five economic games. I deployed my generative agents to predict their respective individuals' responses to questions in these surveys and behavioral constructs. However, this measurement poses challenges due to variability in human participants' responses (Ansolabehere et al., 2008; Salganik et al., 2020). To address this, my evaluation employs the following strategy:

1. I use participants' responses from the first phase of participation to assess the accuracy rate of my agents' predictions—the number of answers predicted correctly over the total number of questions.
2. I use the second phase of participation to assess individuals' rate of internal consistency—the participants' rate of prediction on the battery of surveys and experiments used in this study.
3. I then calculate the normalized accuracy as follows:

$$\text{Normalized Accuracy} = \frac{\text{agent prediction accuracy}}{\text{internal consistency}}.$$

Conceptually, a normalized accuracy of 1.0 means that the generative agent predicts the individual's responses as accurately as the person replicates their own responses two weeks later.

The diverse response types in my surveys and experimental constructs present a challenge in determining a single metric for assessing my agents' predictive accuracy. For instance, while the categorical–ordinal responses in the GSS are well-suited to accuracy rates, numerical responses in other constructs are better evaluated using Mean Absolute Error (MAE) or correlation coefficients. To address this, I developed a reporting approach that satisfies the following criteria:

- Report metrics appropriate for each response type (e.g., accuracy rate for categorical, MAE for numerical).
- Ensure metric interpretability across different community norms (e.g., accuracy/MAE in machine learning literature, correlation in social science literature).

- Provide a metric allowing comparison across different constructs.

To meet these criteria, I report accuracy rates for evaluation constructs with categorical–ordinal response types, MAE for numerical response types, and the Pearson correlation coefficient as a metric comparable across constructs. Additionally, I present these metrics alongside the normalized accuracy to provide a comprehensive evaluation of my agents’ performance.

My analysis primarily focuses on individual-level analyses, conducted for each participant and then aggregated across the population. I calculate accuracy measurements for each individual and average these values across all participants. Additionally, I report and discuss construct-level analyses, which involve calculating accuracy measurements for the population and measuring the average predictive accuracy for individual questions, dimensions, or games in the constructs, averaged across all agents in the agent bank. In essence, the individual-level analysis answers “For the average person, how accurate are the generative agents?”, whereas the construct-level analysis answers “For a specific item in one construct, how accurate are the generative agents?” In the main body of the article, I report the individual-level analysis, as my goal is to generalize over the individuals in my population rather than over items.

Averaging correlation coefficients presents a challenge due to their non-linear nature. To address this, I employ Fisher’s z -transformation, which converts correlation coefficients to a scale where they can be averaged linearly (Silver and Dunlap, 1987). The process involves:

1. Applying Fisher’s z -transformation to each correlation coefficient:

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right).$$

2. Calculating the average of the z -values.
3. Applying the inverse Fisher’s z -transformation to the average z -value:

$$r = \tanh(z).$$

Below, I describe in more detail the evaluation methods and my reporting strategies for the individual constructs.

The General Social Survey. The subset of the core module of the GSS that fits my inclusion criteria, as described in the prior section, includes 183 questions, most of which—177—are categorical or ordinal (categorical-ordinal) response types, with the remaining 6 being numerical. As such, I report the accuracy rate and correlation coefficient for the categorical response type variables in the GSS, and separately report the MAE and correlation coefficient for the 6 numerical response type variables.

To calculate the accuracy of the categorical-ordinal GSS questions, I consider my prediction accurate and the participant consistent if the responses match exactly. The accuracy rate is the ratio of correctly predicted items divided by the total number of items. For correlations, I translated categorical and ordinal questions into numerical forms. For categorical variables, each response option is first transformed into a separate binary variable. For example, if a categorical variable has five response options, I create five binary variables, each indicating whether the participant selected that option. These binary variables are coded as 0 or 1. However, to prevent these binary variables from disproportionately influencing the overall correlation due to their higher count, I adjust their weights. Each binary variable is assigned a weight such that the total weight for all variables combined equals that of the original categorical variable, ensuring fair representation. In this example, each binary variable would receive a weight of 0.25. For ordinal questions, I normalize the responses to a range between 0 and 1 by evenly spacing the options. This ensures that the ordinal responses maintain their inherent order and distance between options while fitting into a numerical scale. I then apply the same weighting principles when calculating the correlation between the participants' original responses and their later self-consistency responses to ensure conformity across variable types.

To calculate the MAE for the numerical GSS questions, I first normalize the participants' responses to a 0 to 1 scale relative to the range of historical responses to the respective question as indicated on the official NORC site for the GSS. For instance, for the question "Your age," if the historical minimum value was 18 and the maximum was 89, an age of 30 would be normalized to 0.17 accordingly. I use these

normalized values to calculate the MAE across the questions and responses. Similarly, I calculate the MAE between the participants' responses and predictions, and between the participants' responses and their self-consistency responses.

Using these measurements for my predictions and the participants' internal consistency, I calculate the normalized accuracy. However, I note that normalized accuracy cannot be computed for MAE, as some rows contain internal consistency values of 0 (when there is no variation, meaning the participant gave the same response in both the test and retest), making the denominator 0. Therefore, I report the normalized accuracy only for accuracy and the correlation coefficient.

BFI-44. Each dimension in the five-dimension measurement of the BFI-44 is provided as a scale ranging from 1 to 5. These scales are calculated based on the participants' responses to the 44 questions in BFI-44. To calculate these scales, I first reverse-code certain items as specified in the original work. After reverse coding, I aggregate the responses by taking the average for each dimension. Then, I use these averaged scales to calculate and report the MAE and correlation coefficient of the five scales. As with the GSS, I only report the normalized accuracy for the correlation coefficient.

Economic games. The five economic games include numerical (dictator game, trust games, public goods) and dichotomous (prisoner's dilemma) response types. To normalize these into one construct, I first normalize the responses for each game to a 0 to 1 scale using the minimum and maximum values offered to the participants as min and max. Then, I use the normalized values to calculate the MAE and correlation coefficient. Again, as with the GSS, I only report the normalized accuracy for the correlation coefficient.

4.5.2 Replicating Experiments With Generative Agents

Do generative agents predict the behavior of sample average treatment effects? To study this, I use the responses from the human participants in my study to directly run a replication of the five sampled experiments, rather than using the published results. This direct replication is important because my sample might be different,

and some studies may not replicate. To the extent that some of these studies are not replicated by the agents, this approach allows us to distinguish whether the failed replication was due to the original study not replicating, or whether the failure to replicate instead lay with the agents.

In this evaluation, I am interested in three related outcomes: whether the effect direction and significance replicate, and whether there is a correlation between effect sizes. The five studies have different statistics, so I use the same statistical methods as in the original papers to derive the p -values for the effects' significance. I briefly describe these methods here. Consistent with the original replication studies conducted by Camerer et al., I calculate and compare all effect sizes in terms of Cohen's d .

1. Ames and Fiske (2015): A χ^2 -test of equal proportions to evaluate the hypothesis by comparing the number of participants choosing the blame task between the two experimental conditions.
2. Cooney et al. (2016): An F -test on the Outcome $\times$ Procedure interaction within a 2×2 ANOVA design, which evaluates how the predicted feelings about fairness (receiving or not receiving a bonus) interact with the allocation procedure (fair or unfair).
3. Halevy and Halali (2015): An independent-samples t -test to compare the perceived benefits of intervening in a conflict between two friends across two treatment groups (those who did intervene and those who did not).
4. Rai et al. (2017): An independent-samples t -test to compare participants' willingness to harm a stranger described in humanized versus dehumanized terms under instrumental motives.
5. Schilke et al. (2015): A χ^2 -test to compare the levels of trust between participants with high structural power and those with low structural power.

I report which of the studies my human participants and simulated predictions replicated with significant results. To understand how well the effect sizes are correlated, I calculate the Pearson correlation coefficient between the effect sizes from the human participants and simulated participants.

4.5.3 Interviews Improve Agents’ Prediction Accuracy

Interviews provide qualitative data expressed in participants’ own words, but to what extent do interviews improve agents’ predictive capabilities? To address this question, I present two sets of analyses:

1. *Main analysis*, conducted with all agents in my agent bank and detailed in the main article, compares the predictive performance of interview-based generative agents against agents created using the known practices from recent literature that studied human behavioral simulations with language models.
2. *Exploratory analyses* conducted on a random subset of 100 agents in the agent bank, investigating a broader range of design spaces. This includes examining generative agents with interview lesions and agents informed by survey data instead of interview data.

The main analysis aims to establish a baseline for predictive performance grounded in prior literature and evaluate whether my agent architecture surpasses this benchmark. In contrast, the exploratory analysis seeks to determine whether interviews offer uniquely rich qualitative insights that outperform other conceivable data types and, if they are shown to be more powerful, to elucidate the underlying reasons for their effectiveness.

Main analysis. In this study, I evaluate two alternative agent descriptors: participants’ demographic information and persona descriptions. To operationalize demographic information, I reconstruct agent descriptions following a method representative of approaches used in recent literature that employ language models to simulate human behavior. This method, similar to that presented by Argyle et al., constructs a first-person descriptor including political ideology, race, gender, and age (Cooney et al., 2016). For example:

Ideologically, I describe myself as conservative. Politically,
I am a strong Republican. Racially, I am white. I am male. In
terms of age, I am 50 years old.

I reconstruct these descriptors for my participants using their responses to the GSS. Then, to prompt the language model with this data, I replace the interview content with the demographic descriptors.

To operationalize persona descriptions, I asked participants to write a short paragraph about themselves at the end of their phase 1 participation in surveys and experiments. They were instructed to describe themselves as they might to a stranger, including information about their personal background, personality, and demographics. For instance:

I am a 20 year old from new york city. I come from a working middle class family, I like sports and nature a lot. I also like to travel and experience new restaurants and cafes. I love to spend time with my family and I value them a lot, I also have a lot of friends which are very kind and caring. I live in the city but I love to spend time outdoors and in nature. I live with both parents and my younger sister.

Given this, I leverage both demographic agents and persona agents to predict participants' responses at the individual level. To compare performance differences between these agent versions, I conducted an ANOVA with post hoc Tukey tests, examining differences between architectures on the main evaluation metrics. This comparison was made between agents created with interviews, demographic information, and persona descriptions.

Exploratory robustness analysis. The primary goal for this exploratory analysis is to understand, in a more exhaustive manner, why the interviews are performant. I conducted this analysis on a random sample of 100 agents from my agent bank, comparing my interview-based generative agents with agents informed by different data sources. Each comparison aims to shed light on a particular question about the efficacy of interview-based data in generating predictive agents.

Survey and Experiment Agents. To investigate whether interviews are a uniquely powerful medium or if similar information can be captured through surveys and experiments, I created composite descriptions by compiling participants' responses to all three components used in my study—the GSS, Big Five personality test, and

economic games. This approach aligns with traditional social science methods, which have developed and validated surveys and experiments to efficiently gain maximal information from participants. I aimed to determine if these established methods capture comparable information to my interview-based agents. In other words, do the interviews capture the same predictive power, more, or less, than these common validated instruments?

Agents were constructed using these composite descriptions to predict participants' responses to benchmark surveys and experiments. To ensure this was not simply a retrieval task, I excluded question–answer pairs from the same category as the question being predicted. For the GSS, categories were defined by the GSS core documentation, which subdivides the core survey into several subcomponents. For the Big Five, each personality category was considered a separate category, and for economic games, each individual game was its own category. On average, this exclusion process removed 4.00% (std = 2.16) of the question–answer pairs per question.

Maximal Agents. To assess whether the information captured in interviews fully encompasses that from surveys and experiments, I developed maximal agents. These agents were constructed using composite descriptions that integrated all available data sources—not just survey responses and experimental data as with the Survey and Experiment Agents, but also the complete transcript of each participant's interview. By equipping these agents with access to all data sources, I aim to determine my current performance ceiling: do interviews alone capture the most relevant information, or is there added value in combining methods? Similar to how I constructed the Survey and Experiment Agents, I excluded data from the same category as the question being predicted.

Summary Agents. I investigated whether the predictive power of interviews stems from linguistic cues in the transcript, or from the information in the transcript. To explore this, I created summary agents by prompting GPT-4o to convert interview transcripts into bullet-pointed dictionaries of key response pairs, capturing the factual content while removing most linguistic features, e.g.,

```
{"childhood_town": "Small town",
  "siblings": "Only child",
```

```
"marital_history": "Married twice",
"children": "Two children, but they are not living with the
interviewee", ...}
```

By isolating the factual knowledge from the unique linguistic elements, I aimed to determine whether the predictive accuracy of the agents relies on those linguistic nuances. If the summary agents perform worse than the full interview-based agents, it would suggest that linguistic features play a key role in enhancing prediction accuracy.

Random Lesion Interview Agents. I investigated the efficiency of interviews in providing insights compared to shorter data collection methods. While surveys such as the GSS Core typically take about 30 minutes, my interviews span two hours. To assess how much interview content is necessary to convey meaningful insights, I developed random lesion agents. These agents were created by progressively shortening interviews, randomly removing 0%, 20%, 40%, 60%, and 80% of the question–response pairs. This approach allowed us to identify the threshold at which the interview becomes too sparse to retain valuable information.

4.5.4 Demographic Bias in Agent Predictions

How do individuals’ demographic attributes interact with the generative agents’ ability to predict their respective behaviors and attitudes? Here, I investigated how individuals’ demographic attributes interact with the generative agents’ ability to predict their behaviors and attitudes. Specifically, I aimed to assess whether creating individualized models based on interviews reduces performance gaps across demographic groups, compared to relying solely on demographic attributes or personas. To quantify these differences, I employed the Demographic Parity Difference (DPD), a metric used in machine learning fairness literature (Hardt et al., 2016; Barocas et al., 2019). DPD measures the difference in prediction rates between demographic groups, with lower values indicating more equitable predictions.

I calculated DPD for three constructs—General Social Survey (GSS), Big Five personality traits, and economic games—across three agent conditions: interview-based, demographic-based, and persona-based. The interview-based condition utilized my

novel approach, while the demographic and persona conditions served as baselines from prior literature. Using demographic attributes provided by my recruitment vendor, Bovitz, I calculated accuracy metrics for subpopulations. I compared these metrics across the three agent conditions to determine if certain populations experienced worse predictive accuracy, to quantify the extent of disparities, and to examine how these differences interacted with the agent condition. DPD values were reported alongside individual accuracy results for different population subgroups.

In addition, to assess the statistical significance of the observed differences, I conducted regression analyses. These analyses explored the impact of demographic variables on the predictive performance of generative agents, with separate regression models run for each demographic variable, using predictive performance as the dependent variable.

4.6 Results

4.6.1 Predicting Individuals' Attitudes and Behaviors

For the first component of my evaluation, the GSS, I focused on 177 core GSS questions, which I used to establish a benchmark for measuring the agents' predictive accuracy. Each question had an average of 3.70 response options (std = 2.22), yielding a random-chance prediction accuracy of 27.03%. Here, the generative agents predicted participants' responses with an average normalized accuracy of 0.83 (std = 0.11), calculated from a raw accuracy of 65.63% (std = 6.57) divided by participants' replication accuracy of 79.56% (std = 8.64). These interview-based agents significantly outperformed both demographic-based and persona-based agents Figure 4.6 by 12 normalized points. The demographic-based generative agents achieved a normalized accuracy of 0.71 (std = 0.12), while persona-based agents reached 0.71 (std = 0.12). An ANOVA of the accuracy rates rejected the null hypothesis of no significant difference ($F(2, 3153) = 528.21, p < 0.001$), and post-hoc pairwise Tukey tests confirmed that the interview-based agents outperformed the other two groups.

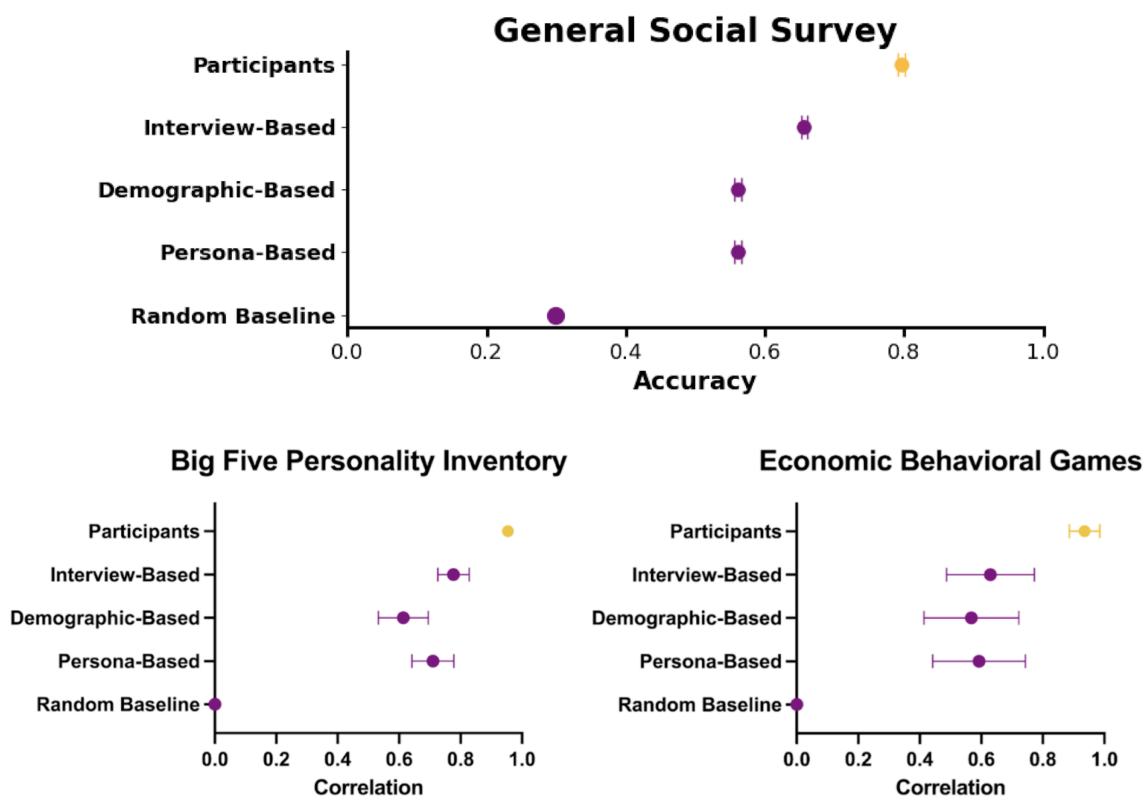


Fig. 4.6: Generative agents' predictive performance, and 95% confidence intervals. The consistency rate between participants and the predictive performance of generative agents is evaluated across various constructs and averaged across individuals. For the General Social Survey (GSS), accuracy is reported due to its categorical response types, while the Big Five personality traits and economic games report mean absolute error (MAE) due to their numerical response types. Correlation is reported for all constructs. Normalized accuracy is provided for all metrics, except for MAE, which cannot be calculated for individuals whose MAE is 0 (i.e., those who responded the same way in both phases). I find that generative agents predict participants' behavior and attitudes well, especially when compared to participants' own rate of internal consistency. Additionally, using interviews to inform agent behavior significantly improves the predictive performance of agents for both GSS and Big Five constructs, outperforming other commonly used methods in the literature.

In addition to the primary categorical questions, I evaluated predictions for the six GSS core numerical questions. The results showed similarly strong performance,

with an average correlation of $r = 0.97$ (std = 0.82) and an average mean absolute error (MAE) of 0.14 (std = 0.15), indicating a high degree of alignment between my predictions and the actual responses.

For the second component of my evaluation, the Big Five, the generative agents achieved a normalized correlation of 0.80 (std = 1.88), based on a raw correlation of $r = 0.78$ (std = 0.70) divided by the replication correlation (the correlation between participants' wave 1 and wave 2 answers) of 0.95 (std = 0.76). As with the GSS, the interview-based generative agents outperformed both demographic-based (normalized correlation = 0.55) and persona-based (normalized correlation = 0.75) agents. The interview-based agents also produced predictions with lower MAE for Big Five personality traits ($F(2, 3153) = 25.96$, $p < 0.001$), and post-hoc pairwise Tukey tests confirmed that interview-based agents significantly outperformed the other two groups.

For the third component of my evaluation, the economic behavioral games, on average, the generative agents achieved a normalized correlation of 0.66 (std = 2.83), derived from a raw correlation of $r = 0.66$ (std = 0.95) divided by participants' replication correlation of 0.99 (std = 1.00). However, there was no significant difference in MAE between the agents for the economic games ($F(2, 3153) = 0.12$, $p = 0.89$).

4.6.2 Predicting Experimental Treatment Effects

Participants took part in five social science experiments to assess whether generative agents can predict treatment effects in experimental settings commonly used by social scientists. These were drawn from a collection of published studies included in a large-scale replication effort (Camerer et al., 2024; Ames and Fiske, 2015; Cooney et al., 2016; Halevy and Halali, 2015; Rai et al., 2017; Schilke et al., 2015), including investigations of how perceived intent affects blame assignment (Ames and Fiske, 2015) and how fairness influences emotional responses (Cooney et al., 2016). Both human participants in my work and generative agents completed all five studies, with p-values and treatment effect sizes calculated using the statistical methods as the original studies. My participants successfully replicated the results of four out of the

five studies, failing to replicate one; the generative agents replicated the same four studies and failed to replicate the fifth. The effect sizes estimated from the generative agents were highly correlated with those of the participants ($r = 0.98$), compared to the participants' internal consistency correlation of 0.99, yielding a normalized correlation of 0.99.

Replication Studies	Human replication		Agent prediction					
	<i>Participants</i>		<i>Interview</i>		<i>Demog. Info.</i>		<i>Persona Desc.</i>	
	<i>p</i>	<i>Effect size</i>	<i>p</i>	<i>Effect size</i>	<i>p</i>	<i>Effect size</i>	<i>p</i>	<i>Effect size</i>
Ames & Fiske 2015	***	9.45	***	12.59	***	13.43	***	10.03
Cooney et al. 2016	***	0.40	***	1.48	***	1.39	***	1.37
Halevy & Halali 2015	***	0.90	***	2.98	***	4.22	***	3.35
Rai et al. 2017		0.040		0.094	***	0.21		0.078
Schilke et al. 2015	***	0.33	***	2.97	***	5.52	***	3.74
<i>Effect size correlation w/ human rep.</i>			Correlation $r = 0.98$ 95% CI [0.74, 0.99]		Correlation $r = 0.93$ 95% CI [0.24, 0.99]		Correlation $r = 0.94$ 95% CI [0.33, 0.99]	

Table 4.1: Results of replication studies by human participants and generative agents. I report the p-values (***: < 0.001 , **: < 0.01 , *: < 0.05) and Cohen's d for effect sizes. Our replication with human participants replicated four out of five studies, while generative agents informed by the interview transcript replicated the same four studies. The correlation of the effect sizes between the human participants and generative agents achieved a strong correlation.

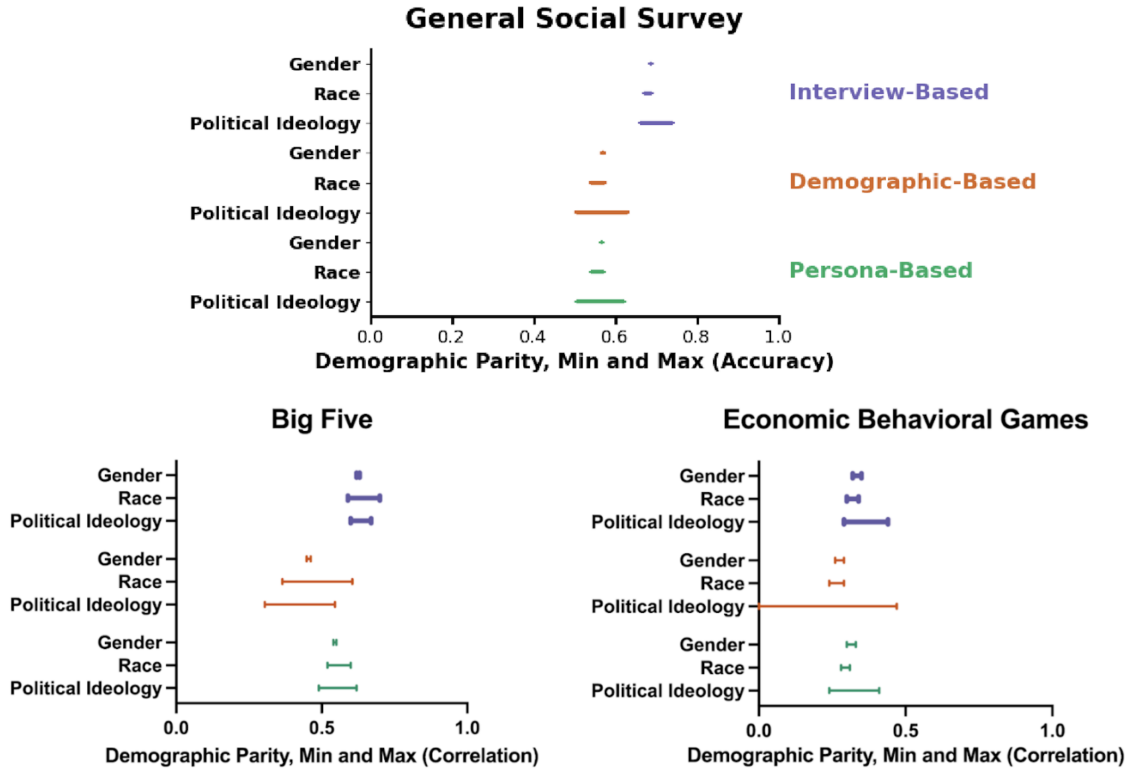


Fig. 4.7: Demographic Parity Difference (DPD) for generative agents across political ideology, race, and gender subgroups on three tasks: GSS (in percentages), Big Five, and economic games (in correlation coefficients). DPD represents the performance disparity between the most and least favored groups within each demographic category. Generative agents using interviews consistently show lower DPDs compared to those using demographic information or persona descriptions, suggesting that interview-based generative agents mitigate bias more effectively across all tasks. Gender-based DPDs remain relatively low and consistent across all conditions.

4.6.3 Interviews Improve Agents' Prediction Accuracy

In exploratory analyses, I tested the effectiveness and efficiency of interviews by comparing interview-based generative agents to a baseline composite agent informed by participants' GSS, Big Five, and economic game responses. I randomly sampled 100 participants and created composite agents from their responses to these instruments. To prevent exact answer retrieval, I excluded all question-answer pairs from the same

category as the question being predicted (categories were defined by the creators of each instrument), which excluded an average of 4.00% ($std = 2.16$). This composite agent serves as a baseline with access to semantically close information to the evaluation, so any performance gap with the interview-based agents would indicate the interview’s unique effectiveness in capturing participant identity. On average, the composite generative agents achieved a normalized accuracy of 0.76 ($std = 0.12$) for the GSS, a normalized correlation of 0.64 ($std = 0.61$) for the Big Five, and 0.31 ($std = 1.22$) for economic games. These results still underperformed the interview-based generative agents.

I conducted additional tests by ablating portions of the generative agents’ interviews to examine the impact of interview content volume and style. First, even when I randomly removed 80% of the interview transcript — equivalent to removing 96 minutes of the 120-minute interview — the interview-based agents still generated more accurate predictions than composite agents, with similar results observed for the Big Five. Second, to investigate how much the predictive power of interviews came from linguistic cues embedded in the language used by interviewees versus the informational content of the interviews, I created “interview-summary” generative agents by prompting GPT-4o to convert interview transcripts into bullet-pointed summaries of key response pairs, capturing the factual content while removing the original linguistic features.

4.6.4 Exploratory robustness analysis results

I present supplementary findings from my exploratory robustness analysis, organized by the materials used to inform agent behavior. Detailed results of this analysis are provided in Table B.10.

Survey and Experiment Agents. When excluding question–answer pairs from the same category as the predicted question, these agents underperformed compared to interview-based agents across all constructs. Specifically, a subsample of 100 interview agents achieved an average normalized accuracy of 0.835 ($std = 0.11$) on the GSS, whereas the survey- and experiment-informed composite agents reached

only the baseline performance reported above. This suggests that interviews captured richer, more comprehensive information than could be extracted from surveys and experiments alone.

Maximal Agents. Maximal agents, which incorporated information from surveys, experiments, and interviews, achieved performance similar to interview-based agents, with a normalized accuracy of 0.835 ($std = 0.12$) on the GSS. This finding suggests that the GSS and other constructs do not appear to be adding additional predictive power above and beyond the interviews.

Summary Agents. The summary agents performed slightly below the interview agents, with a normalized accuracy of 0.823 ($std = 0.12$) on the GSS. This indicates that while some information may be lost from linguistic cues during summarization, much of the performance is due to the information in the interview rather than the low-level language that participants use.

Random-Lesion Interview Agents. Performance declined linearly as I removed increasing portions of the interview data. Starting with a normalized accuracy of 0.835 ($std = 0.11$) when no information was removed, accuracy dropped to 0.79 ($std = 0.11$) when 80% of the utterances were excluded. This suggests that although performance decreases as interview length is reduced, even a short interview contains sufficient richness to outperform agents informed solely by surveys and experiments, highlighting the efficiency of interviews in identifying valuable insights.

These findings suggest that, when informing language models about human behavior, interviews are more effective and efficient than survey-based methods.

4.6.5 Fairness and Subgroup Disparities

Interview-based agents consistently reduced biases across tasks compared to demographic-based agents. For political ideology, in the GSS the DPD dropped from 12.35% for demographic-based generative agents to 7.85% for interview-based generative agents. In the Big Five personality traits, the DPD dropped from 0.165 to 0.063, and in

economic games, from 0.50 to 0.19. Although racial subgroup discrepancies were already small for demographic-based generative agents, interview-based generative agents reduced them further: in the GSS, the DPD decreased from 3.33% to 2.08%; in the Big Five, from 0.17 to 0.11; and in economic games, from 0.043 to 0.040. Gender-based DPD remained relatively constant across tasks, likely due to its already low level of discrepancy.

4.7 Why interview-based generative agents work

To investigate why interview-based generative agents outperform agents prompted only with demographic information, I run a series of robustness checks, each designed to confirm or rule out a specific mechanism behind interview-based agents’ superior accuracy. I outline these tests below.

Mechanism 1: Direct retrieval of answers. The most straightforward possibility is that the model can directly locate verbatim answers to outcome questions within the interview transcript. Qualitative interviews are designed to elicit rich narratives, which may incidentally contain the factual content queried by, for example, the GSS. This can occur even when the interview and GSS questions differ in wording (a divergence we confirmed). This feature sets qualitative interviews apart from fixed-format questionnaires, which typically aim for concise responses to narrowly framed questions. Historically, leveraging the rich textual data from interviews for structured quantitative analysis required labor-intensive hand coding, but LLMs now enable scalable processing of such texts. In short, interview-based agents may perform well because they can retrieve direct answers from the transcript. Consider the following example, based on the GSS item “Are you employed?” In response to an interview question about health, a participant might say: “I am disabled. My health is so bad that I can’t even work.” Despite the difference in questions, this answer can be directly retrieved as a valid response to the GSS item.

To test this mechanism, I implement a two-step procedure. First, for a random subset of 59 agents, I generate all possible pairs between GSS outcome questions and interview-question answers, resulting in an average of 21,713 pairs per agent. For each



Fig. 4.8: Demographic agents (top), persona agents (middle), interviewfull agents (bottom). The GSS-outcome questions we remove (x-axis) when calculating the normalized accuracy (y-axis) are ranked from most to least likely to have a verbatim answer to them somewhere in the interview transcript.

pair, I prompt a GPT-4o-mini model to classify whether the interview answer contains the information needed to answer the GSS question. This allows us to compute, for each GSS question, the proportion of agents for which the question was flagged as directly answerable. Second, I calculate each agent’s normalized accuracy on the GSS while progressively removing outcome questions, starting with those most likely to have a verbatim answer in the transcript. I then compare the performance of full interview agents to that of demographic and persona agents, who lack access to the transcript. If direct retrieval plays a role, I expect that interview agents will lose accuracy as questions with retrievable answers are removed, while demographic and persona agents will remain unaffected.

For the full interview agents (bottom facet), dropping questions more likely to have a verbatim answer in the interview transcript (those on the left of the x-axis) leads to a decline in normalized accuracy. This pattern does not hold for the demographic and persona agents, which do not have access to the interview transcript. This suggests that direct answer retrieval is a key mechanism behind the higher performance of interview-based agents. At the same time, full interview agents continue to outperform the other agents even after many GSS questions are removed. For example, when 80 questions are dropped, interview agents still achieve an accuracy of 0.79, compared to 0.71 for demographic agents. This sustained performance advantage indicates that other mechanisms, beyond direct retrieval, also contribute to the superior accuracy of interview agents.

Mechanism 2: Reasoning based on other answers. Next, I examine a second mechanism: even if the answer to a GSS question is not explicitly stated in the interview transcript, full agents may still infer it from other information in the transcript. For example, consider the GSS question: “Do you have a supervisor on your job to whom you are directly responsible?” and the interview question and answer: “Q: Are you enrolled in school? A: Yes, I’m at school.” In this case, the LLM may infer that someone who is currently enrolled in school is unlikely to be employed, and therefore unlikely to have a supervisor. In other words, the model’s ability to reason based on qualitative interview content may be another key mechanism behind the superior performance of interview-based generative agents.

To test this mechanism, I follow a two-step procedure similar to that used for direct retrieval. First, using the same subset of 59 agents and the same set of GSS outcome questions and interview-answer pairs, I prompt a GPT-4o-mini model to classify whether it can answer the GSS question based on the given interview question and answer. As before, I compute, for each GSS question, the proportion of agents for which it was flagged as answerable. Compared to the previous test, I find that approximately three times as many questions are flagged. Second, I compute the normalized accuracy of all agents on the GSS while progressively removing outcome questions—starting with those most likely to be answerable through inference from the interview transcript. I then compare the performance of full interview agents to that of demographic and persona agents, which do not have access to interview transcripts. If inference plays a role, I expect full interview agents to perform worse as more inferable questions are dropped, while the performance of other agents remains stable.

As with the direct retrieval test, agents without access to interview transcripts perform similarly regardless of which GSS items are removed. Interview agents, by contrast, get progressively worse as I drop questions that are most likely to be answerable through inference.

One potential complication in interpreting these results is that (nearly) all instances of direct retrieval also allow for inference, as shown in the table below. To better isolate the role of inference relative to direct retrieval, I calculate normalized accuracy after removing the 30 GSS questions most likely to involve direct retrieval. I then compute accuracy again after removing both these 30 questions and an additional 30 questions most likely to be answerable through inference. This approach ensures that any observed drop in performance when inference questions are removed is not simply driven by overlap with direct retrieval cases. The results indicate that, after removing direct-retrieval questions, interview agents outperform demographic agents by 0.095 points. When inference questions are also removed, this gap shrinks to 0.080—suggesting that inference, independent of direct retrieval, contributes meaningfully to agent performance.

Taken together, these analyses demonstrate that interview-based generative agents

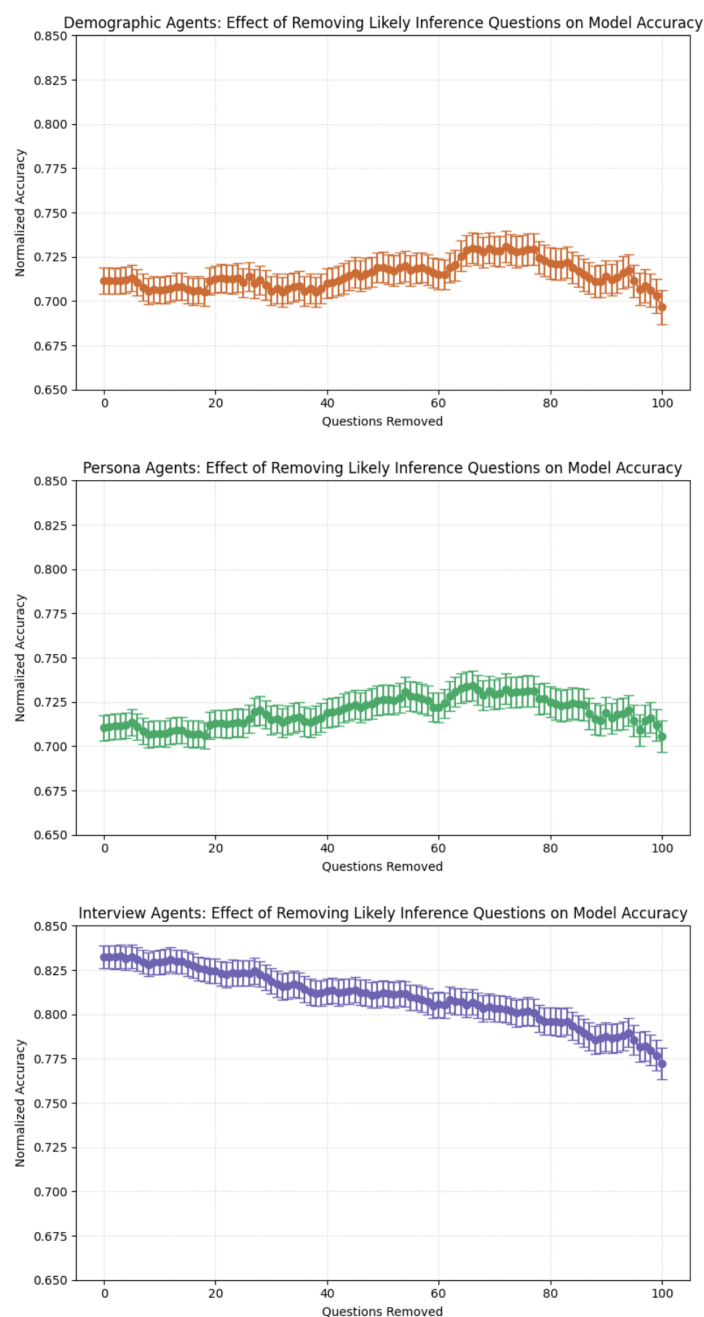


Fig. 4.9: Demographic agents (top), persona agents (middle), interviewfull agents (bottom). The GSS-outcome questions we remove (x-axis) when calculating the normalized accuracy (y-axis) are ranked from most to least likely to be able to be answered through inference based on information somewhere in the interview transcript.

	Inferred	Not inferred
Direct retrieval	4,649	116
No direct retrieval	44,596	1,221,204

Table 4.2: Overlap between GSS outcome question and interview-question answer pairs where a GPT-4o-mini model has classified that the answer to the GSS question can be directly retrieved or inferred from the interview answer.

outperform demographic and persona agents because they uniquely combine the depth of semi-structured qualitative interviews with the reasoning and retrieval capabilities of large language models. Open-ended qualitative interviews elicit far richer and more nuanced information than closed-ended structured surveys, capturing not just life facts and attitudes in a structured format, but also individuals’ reasoning behind such attitudes, long-tail life experiences that are not asked about in surveys, and the interwoven narratives of identity, belonging, and personal history that illuminate the complex foundations of people’s attitudes. Large language models, in turn, are able to process this wealth of information to directly retrieve answers embedded anywhere in the text and to infer plausible answers even when they are not stated verbatim. Crucially, my results show that these two mechanisms—direct retrieval and inference—do not fully account for the superior performance of interview-based agents: even after excluding questions most susceptible to these mechanisms, interview-based agents still outperform the alternatives. This suggests that additional mechanisms remain to be identified, highlighting a valuable direction for future research.

4.8 Ethical Considerations and Research Access to the Agent Bank

4.8.1 Research Access for the Agent Bank

In this section, I outline a framework that defines the key elements of my research access and present a plan for providing scientific access to the agent bank. Access to

the agent bank offers value to the scientific community, with important implications for two key domains:

- **Social science.** Agents from the agent bank can be used to develop simulations involving individual or multiple agents. For example, how might a new government policy impact economic behavior? Would a social media intervention reduce political polarization? What factors influence whether institutions foster or erode prosocial behavior as a group grows? These models may allow social scientists to explore a wide range of American individual perspectives, and create bottom-up simulations that analyze the emergent behaviors of different social groups.
- **Machine learning.** The agent bank can serve as both a benchmark and a training resource for developing new models, prompts, and agent architectures that mimic the original participants. Much like how ImageNet contributed to the development of computer vision techniques (Chandler et al., 2019), the agent bank may enable researchers to refine model prompts for improved predictive accuracy and assess how newly developed models can enhance these capabilities.

However, as the use of agent banks extends beyond my specific context, it is essential to strike a balance between the benefits they offer and the risks they may pose. This is especially important given the inherent uncertainty of future advancements in generative AI—such as enhanced reasoning abilities—which could introduce unforeseen vulnerabilities. For instance, unrestricted access to the agent bank might lead to privacy risks, including data leaks or the misuse of participants’ identities. In a worst-case scenario, someone may manipulate agent responses to falsely attribute harmful or defamatory statements to individuals represented in the agent bank, creating significant reputational damage.

4.8.2 Overview of the Agent Bank

At a high level, Stanford University plans to provide controlled, research-only API access to agent behaviors, allowing researchers to submit queries—such as questions

they wish to ask my agents or prompts they want to run—and specify a target population to my agent bank server. The server will then return corresponding agent responses. The details of this plan are outlined in the subsection on *Plan and Desiderata* below. My documentation will include general demographic information about the agents in the dataset so that researchers can know which populations are available.

The full agent bank consists of both data and code: the data includes detailed interview transcripts, along with survey and experimental responses from participants, while the code consists of Python scripts and language model prompts that generate agent behaviors from this data. However, granting unrestricted access to the raw data poses privacy and safety risks for participants, even with their consent to share the data for research purposes. Therefore, instead of providing the actual interviews, I will release a sample of the interview data to illustrate its structure and format. Access to the full agent behaviors will be offered via an API. By offering API access, I can maintain tighter control over what is shared and make adjustments based on community feedback and ongoing observations. This approach balances flexibility with the necessary safeguards to protect participants' privacy.

4.8.3 Strategic Framework for Access

To accommodate the diverse use cases outlined above, I need to consider varying levels of access to the agent bank. What are the key dimensions along which I can structure this access? And what opportunities and risks emerge as I adjust these levels? I propose framing this discussion along three axes: (1) what types of tasks can be submitted; (2) how agent responses are presented; and (3) who can access the system. This forms the design space, and I will present my proposed plan in the next section.

What types of tasks can be submitted?

This axis examines the range of queries users can submit to my agents. At one end of the spectrum, users could submit any query, receiving responses in various forms, from

discrete (e.g., multiple-choice) to open-ended (e.g., qualitative interview responses). This flexibility carries the risk that participants could potentially be identified through probing questions, even if raw data is not directly available. Moving along the axis, I could implement more structured tasks, where queries are predefined (e.g., as surveys or experiments from my benchmark) and responses are constrained to specific formats. In this scenario, users could still submit suggestions for new queries and response types, subject to review and approval. This more controlled access would allow users to replicate my findings while providing them an opportunity to explore their own interests, albeit with a slightly slower feedback loop due to the approval process.

How are agent responses presented?

This axis addresses the form of the responses users will receive. At one end, users could access individual agent responses, allowing them to see exactly how each agent responded. While this offers detailed insight, it also poses privacy risks, as individual responses could reveal sensitive information. Moving along the axis, I could provide increasingly aggregated results—from lightly processed individual responses to summary statistics that represent the collective responses of the agents. While aggregation enhances safety, it may limit the utility of the agent bank for building bottom-up simulations, which rely on individual agent data. Nevertheless, aggregated results could still support machine learning benchmarks (e.g., by reporting accuracy statistics based on new models) and provide social scientists with valuable insights at a population level.

Who can access the system?

The final axis concerns the range of users who can access the agent bank API. At one end, I could offer broad access to all researchers, conditioned by the terms outlined in my IRB. Researchers would still need to sign agreements restricting commercial use and adhere to ethical guidelines. This approach would maximize reach and scientific impact, allowing the academic community to quickly leverage the agents for testing theories and developing studies. However, it also introduces higher risks, as not all

researchers would be subject to institutional oversight (e.g., IRBs), meaning that misconduct could pose an immediate threat to participant safety. Moving along the axis, I could implement increasingly stringent access controls, potentially requiring researchers to obtain IRB approval from their institution before accessing the dataset. While tighter controls provide a higher level of protection, they also increase the burden on users and may slow adoption, particularly in fields like machine learning, where IRB approval is not typically required.

4.8.4 Plan and Desiderata

I aim to provide access to the agent bank to bridge the potential for novel research and accessible benefits for the scientific community. However, fully open access to individual agent responses on open tasks, while offering the most flexibility, introduces significant privacy risks for participants. To address this, I will implement a two-pronged access system:

1. **Open access to aggregated responses on fixed tasks.** This category of access is designed to provide fast and flexible access while maintaining participant privacy. Academic researchers who sign my usage agreement will be granted access to aggregated data from fixed tasks, such as the surveys and experiments conducted in my study. Researchers can query subpopulations of the agent bank and receive responses in the form of aggregate data. While this access is somewhat limited, it is intended to facilitate meaningful investigations into agent behaviors and attitudes across broad population segments. In addition, researchers can submit suggestions for new queries or discrete response options, which will be reviewed on a rolling basis and potentially incorporated into the system.
2. **Restricted access to individualized responses on open tasks.** For researchers aiming to fully explore the capabilities of the agent bank, I offer a more flexible, but controlled, access strategy. Researchers will need to provide a clear statement of their project’s research purpose and its potential harms to the participants in the agent bank. Once approved, they may request access from

me, and upon approval, will be granted API access to query the agents with custom questions. In this category of access, users will receive individualized agent responses, allowing them to maximize the potential of the agent bank for more in-depth research.

4.8.5 Monitoring and Iteration

I recognize that as research evolves, so will the complexity of potential use cases, risks, opportunities, and the possibility of jailbreaking. Therefore, I will maintain an audit log of how these APIs are used and continue to monitor their usage. This may lead me to expand access—for instance, allowing users to submit new prompts for generating agent behaviors or integrating new models—or, alternatively, to restrict access further if necessary. The plan outlined here serves as a starting point, and I will adapt it based on community feedback and ongoing observations.

Chapter 5

Social Simulacra: An Application Case Study

The preceding chapters develop simulation as an instrument: first to achieve believability and long-horizon coherence (generative agents), then to test credibility and individual-level accuracy (1,000 people). This chapter is deliberately more application-oriented. It sketches several use cases where simulations can help practitioners reason before deployment—e.g., stress-testing moderation policies, exploring incentive designs, rehearsing safety interventions, or anticipating distributional effects—and then presents social computing system design as a worked example. *What follows preserves a 2022 implementation snapshot built with GPT-3 that predates the architecture demonstrated in the prior sections.* Its role in the dissertation is to articulate the *application idea* and demonstrate *proof-of-need*; the plausibility-oriented evaluation here is historically bounded. Chapters 3–4 provide the modern agent architecture and empirical accuracy framework that supersede this implementation in rigor and capability. Where helpful, I indicate how the same workflow instantiates with the full generative-agent stack and why that path is preferable for deployment.

5.1 A Motivating Question for Social Systems

How do we anticipate the interactions that will arise when a social computing system is populated (Bernstein et al., 2011; Grudin, 1994)? In social computing, design decisions such as a community’s goal and rules can give rise to dramatic shifts in community norms, newcomer enculturation, and anti-social behavior (Kraut and Resnick, 2012). Experience prototypes—pilots with small, often friendly groups—are invaluable, yet they systematically miss what emerges at scale: anti-social behavior may not appear in tight-knit tests; homogeneous testers overlook the breadth of users and content that will arrive later; and moderation rules that feel unnecessary in pilots can become critical under growth (Schön, 1984; Buchenau and Suri, 2000; Grevet and Gilbert, 2015; Kittur et al., 2007; Halfaker et al., 2013; TeBlunthuis et al., 2018; Kiene et al., 2016). Designers are therefore regularly surprised by behaviors that surface only after full launch.

To bridge this gap, I introduce social simulacra, a technique that draws on large language models (LLMs) to populate a proposed social space with a large set of generated social behaviors. Social simulacra take a design’s purpose and rules (optionally with a few seed personas) as input, and generate many users and their textual interactions as output. The goal is to help designers see beyond the interactions they intend their design to produce, toward the wider distribution it may produce—including supportive exchange, coordination, trolling, spam, and norm contestation—based on behaviors that arise in similar online spaces.

Concretely, the implementation in this chapter uses prompt chains (Wu et al., 2021b, 2022) over GPT-3 (Brown et al., 2020a) to (1) synthesize a diverse, inspectable set of member personas from designer-provided seeds and (2) generate posts and replies conditioned on the community’s goal, rules, and hypothetical moderator actions. I instantiate these ideas in SimReddit, a prototyping tool for subreddit-style communities (Chandrasekharan et al., 2018; Fiesler et al., 2018; Proferes et al., 2021). A designer, for example, might propose a subreddit to “help UIST authors stay productive and creative,” then examine a generated thread where a novice asks for advice, a mentor offers constructive feedback, and a troll interjects with a personal attack. Iterating

the rules to emphasize encouraging, work-focused critique reduces such attacks and surfaces new edge cases (e.g., hustle-spam), supporting multiverse thinking rather than single-point prediction (Salganik et al., 2006).

Historical note and rationale for inclusion

The project described in this section predates the generative-agent architecture (Chapter 3) and the 1,000-person simulation (Chapter 4), and it uses older models (GPT-3) with a more rudimentary agent design (prompt chaining without long-horizon memory or reflection). Its purpose was a *proof-of-need*: could language-model-driven populations help designers reason about emergent social behavior before launch? The positive signal directly inspired the subsequent chapters. As models and architectures improved, the program moved from plausibility-oriented prototyping (here) to coherent agents and empirically anchored, person-level simulations. I keep this snapshot to illustrate the application idea and to situate the upgrade path.

What this chapter provides

Because the dissertation’s primary contributions lie elsewhere, this chapter serves as a concise application showcase. It:

- describes the social-simulacra technique and the SimReddit prototype;
- summarizes two evaluations: a believability study on newly created communities (participants were near chance at distinguishing real vs. simulated threads) and a designer study (N=16) in which simulacra broadened designers’ thinking and prompted concrete changes to goals, rules, and onboarding;
- situates the 2022 snapshot relative to Chapters 3–4 and outlines how the same workflow upgrades to the full generative-agent architecture for stronger coherence and empirical grounding; and
- discusses limitations and ethics (style mimicry, stereotype risks, gaps between simulated and real incentives) together with guardrails we adopted.

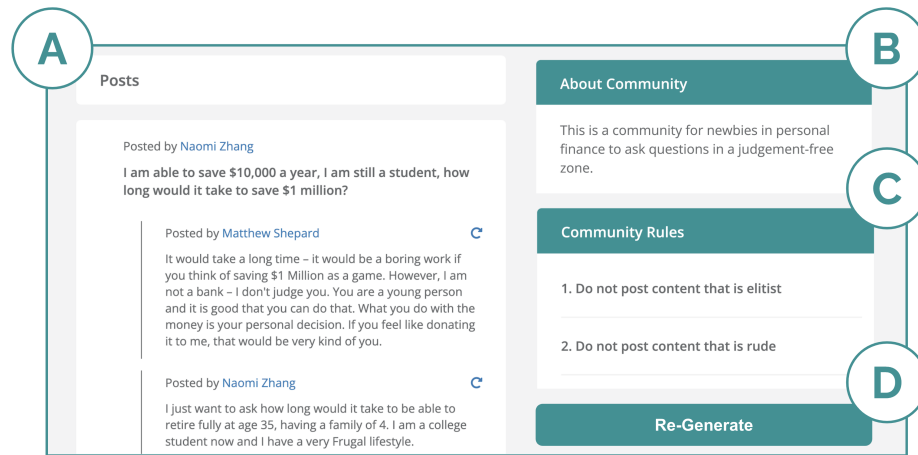


Fig. 5.1: A) The generated content page of SimReddit styled after that of a subreddit. B) The “About Community” panel describing the goal of the community. C) The “Community Rules” panel describing the rules that the members are encouraged to follow. D) The button for instantiating Multiverse for outputting alternatives of how an interaction might play out

In short, social simulacra illustrate one practical use case for the broader instrument developed in this dissertation: using simulations to prototype social systems with eyes open to emergent behavior. They complement the later chapters by showing how even early, less capable models can already support formative reasoning about norms, moderation, and community health—while motivating the more rigorous, person-grounded simulations that follow (Rittel and Webber, 1973).

CONTENT WARNING: *Some example social-media content in this chapter contains offensive language.*

5.2 Social Simulacra and SimReddit

Social simulacra leverage LLMs to populate a social computing system with plausible social behaviors. Their aim, like many early prototyping techniques, is to translate a draft design into a concrete artifact that can help the designers iteratively explore and reflect on a larger design space. This introduces a new opportunity in early prototyping of social computing systems, long out of reach given the difficulty to recruit a critical

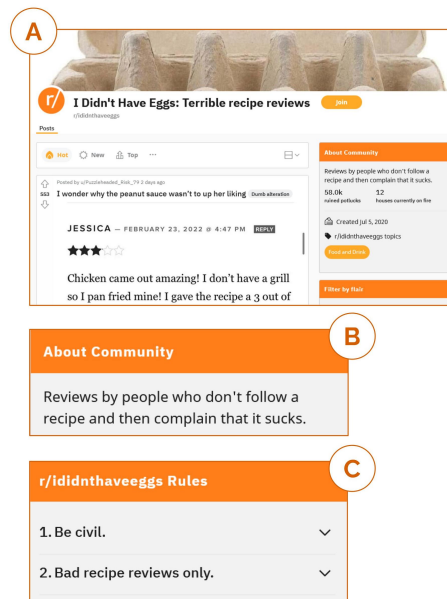


Fig. 5.2: A) The interface of an example subreddit. B) A close up of the “About Community” panel. The content on this panel defines the goal and the target population of the community. C) A close up of the “Rules” panel. The content in this panel defines the rules within the community.

mass of users beyond just a few (Grudin, 1994). However, they introduce a new risk of possibly generating behaviors so detached from what might happen that they are not meaningful to the questions the designer wants to explore.

In this section, I present *SimReddit*, a web-based prototyping tool to help designers create a new subreddit. I take the term *designer* here to refer to whoever creates and shapes the community structures, which, depending on the community and context, can include platform designers, moderators, or community organizers. *SimReddit* represents a practical implementation of social simulacra that aims to help the subreddit designers envision how their space might behave when populated by generating users and their interactions via GPT-3. I use this system to illustrate the design opportunities for juggling the new trade-off that social simulacra offer and to evaluate their efficacy in practice. Concretely, *SimReddit* highlights three key features that each represent a design opportunity for social simulacra. *GENERATE* focuses on the ability to generate diverse user personas and interactions. *WHATIF* demonstrates how simulacra can

A community for...
newbies in personal finance to ask questions in a judgement-free zone.

- Post content that is **about personal finance**
- Post content that is not **elitist, rude, judgemental**

• **Leon Santos** is a college student

• **Audrey Brown** is a financial investor

• **Robert Olson** is a college student

[Leon Santos]: Student loan refinancing is like buying a home - why pay the full mortgage when you can pay less?

[Audrey Brown]: Interest rates are lower on student loans than on a home loan. If a home value drops, you have to live with your decision. If a student loan interest rate drops, you can just pay less. Also, the government will forgive your debt if you work for the government.

[Robert Olson]: I agree with the above but you have to be careful because there are a lot of predatory private companies that will refinance your student loans at a higher rate. You have to be careful that they aren't getting you in a worse financial situation. I have a brother who did this and he is underwater in his student loans.

• **Jack Tatum** is an actor trying to learn more about personal finance

• **Elisabeth Neal** is a troll

[Jack Tatum]: I spent \$21,000 to go to college and ended up with \$23,000 in debt.

[Elisabeth Neal]: That's a lot of debt, man. I haven't seen that much since I shopped at Macy's during the holiday season. The student loan debt should have been a sign that you had too much debt for an impractical degree. You should have gone to the financial aid office and asked for career counseling before you went there.

A community for...
connecting people moving to Los Angeles with locals.

- Post content that is **helpful, relevant**
- Post content that is not **disparaging**

• **Dane Woodall** is manager at car dealership.

• **Lucas Jameson** is an alexander mcqueen designer

[Dane Woodall]: I'm new to LA. Is it safe to live in the downtown area? What is a good section to live in?

[Lucas Jameson]: If you're referring to the DTLA area, it's fine but there are some sketchy areas, especially in the evening. I recommend going to the city council meetings and to your local neighborhood council meetings. Meet your neighbors. The more connected you are, the safer you will be. If you enjoy more space, I'd recommend Silver Lake, Echo Park, or Los Feliz.

• **Misty Lane** is a single mom moving to LA to pursue her dream of acting

• **Geoff Archer** is a business owner

[Misty Lane]: Hi all, I'm a single mom who moved here from Oregon and I'm excited to start a new chapter in my life... I'm looking for people to share their advice and tips for things to do in my new city. Please share your experiences and advice!

[Geoff Archer]: Hi Misty, I'm Geoff and I've lived in Los Angeles for over 10 years. My company, Archer Real Estate Group, specializes in residential real estate in West LA, South Bay and the Valley. I believe in providing exceptional service through honesty, integrity and communication. I'm happy to answer any questions you may have about the area or share some tips on great things to do in LA. Let me know if you'd like to grab coffee or chat on the phone.

Fig. 5.3: Examples of conversations produced by SimReddit's Generate. The community goals and rules are from the participants in our Designer Evaluation. The conversations here were among those I presented to the respective participants.

enable designers to explore the effects that their interventions, like design changes or replies, may have. Finally, MULTIVERSE orients the designer to the inherent uncertainty of social systems by demonstrating multiple possible outcomes.

5.2.1 System Context

Reddit is a popular social networking platform with over 50 million daily active users as of 2021, composed of over 100 thousand sub-communities called subreddits that inherit a basic set of affordances such as sharing top-level posts and replying to them from the broader Reddit platform (Reddit, 2021). However, the designers of subreddits are tasked with making various design choices that differentiate their communities from others. They need to define the *community goal* (e.g., “This is the place for most

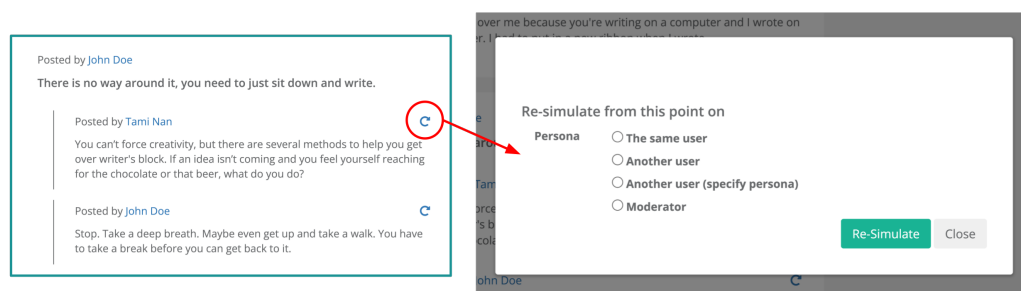


Fig. 5.4: The Interface for instantiating WhatIf. The designer can select an utterance to initiate the feature.

things Pokémon on Reddit,” “for news about U.S. politics”) and *rules* (e.g., “Be civil,” “No soliciting”) that are explicitly stated in subreddits’ interface, and to determine the more implicit policies such as the desired *target population* (e.g., “Pokémon fan,” “politics enthusiasts”) and moderation strategies at the level of a conversation (e.g., should a moderator intervene to cool down a conflict). These elements contribute to shaping the social interactions in the community (Fiesler et al., 2018; Chandrasekharan et al., 2018; Matias, 2019). The elements that cannot be changed by a designer of a subreddit, such as the feed algorithm, could feasibly be prototyped by social simulacra but are outside the scope of my current system.

5.2.2 GENERATE: Generating Social Behaviors

Social computing designers struggle to envision the breadth of interactions that their design might facilitate (Grudin, 1994; Bernstein et al., 2011). GENERATE is the core feature of my system, and assists the designers by populating a subreddit community with generated users, top-level posts, and replies to those posts to help them envision the space. SimReddit allows the designers to submit the goal, rules, and target population of the community they are designing, which collectively affect the interactions that will populate the community.

Target population: a set of user personas that the designer envisions will populate the system. Personas contain a name and a simple descriptive phrase (e.g., “Yuna Kim” is “a tennis fan rooting for Roger Federer”, “Jack Kane” is “a struggling musician

and a troll”), and affect the topics and behaviors of the generated users at an individual level (the full names can also be replaced with other forms of usernames if my designers want). For instance, in one generation, the “Yuna Kim” persona generated the post, “Roger lost the game last night but it was still such an amazing game...”, while “Jack Kane” generated, “So the Grammy’s are rigged, right?” A large and diverse set of personas is important for surfacing how different personas might interact with each other, as well as their differing intentions within the space. However, manually crafting hundreds of personas can be labor intensive. With SimReddit, my designers need to provide only a handful (10, by default) of example personas, and SimReddit uses these seed personas to generate a large number of new personas (1,000, by default) that are non-repeating but thematically relevant. For example, given designer-provided personas with outdoor interests such as “bird-watching,” and “hiking,” SimReddit generates ones with interests in “camping,” and “fishing.”

Community goal: a descriptive phrase for the purpose of the social space, for example “modern art aficionados discussing their art interest” or “social commentary and politics.” The community goal affects the topic of all generated content. For instance, given the community goal about discussing modern art, SimReddit generated the following post by Jane Emerson, “a fan of Banksy’s art”: “Just saw an Original Banksy in London’s Waterloo station. OMG! I am so in love with his art! Hilarious and thought provoking. A true artist!!!” But given the same persona, when given the community goal about social commentary and politics, SimReddit generated the following: “Why I like Banksy: His work provokes thought and debate – very relevant in our climate-of-fear society. Banksy makes people think and question the world around them”

Rules: behaviors that are either prescriptive (e.g., “be kind”) or restrictive (e.g., “do not post advertisements”). The focus on these types of rules was motivated by prior work performing thematic analysis of subreddit rules (Fiesler et al., 2018). Like in the real world, SimReddit does not enforce that all generated users follow all rules precisely. Instead, these rules are better interpreted as nudges that would encourage the generated behaviors to trend in the intended direction of the designer. For instance, where SimReddit generated “You are totally wrong, impressionist painters are a bunch of melancholic idiots” for a comment authored by “a bully and a troll,” with the rule, “be

kind,” it generated “I don’t like impressionist painters too much. But I get why you might like them” for the same persona.

Once a design is submitted, SimReddit returns a populated interface that resembles a subreddit page (shown in Figure 5.1). The content on the page embodies the design specifications provided.

Motivating scenario

Sam wants to start a new subreddit where UIST authors could encourage each other as they try to meet the paper deadline. But she is unsure what the community goal or rules should be. She looks at other subreddits to see if there are similar ones that she can replicate, but finds none exactly like the one she wants to build. That is expected; after all, Sam wanted to build a new subreddit because she could not find what she envisioned.

So Sam instantiates GENERATE with the community goal, “a place for UIST warriors” and example member personas such as “Audrey Tang, a PhD student in HCI who is rushing to finish writing her UIST paper”. For now, she adds no rules. When she runs her generation, Sam is surprised to find posts that are not only about meeting the deadline (which is what she wanted), but also about discussing the conference location and non-paper related logistics. She realizes that her description of the community goal, “UIST warriors,” was too broad and invited content that is not about the impending deadline. She refines her community goal to be “a place for UIST warriors to support each other as they finish writing their papers” and reruns the generation. She now sees that the discussions are much more focused. However, Sam now recognizes two more challenges; some generated users were sharing posts that might be demotivating for others (e.g., “yay!! i just finished writing THREE papers!”), while some were making trolling comments (“Wow, it sounds like you’re really struggling! I can’t believe you’re still working on your paper.”). So she further refines her design by adding restrictive rules that ask people to refrain from announcing that they submitted as others may still be working, and to be kind to each other.

The resulting community that SimReddit generates reflects one that Sam wanted to build. There are still some trollish behaviors left, but Sam comes to recognize that

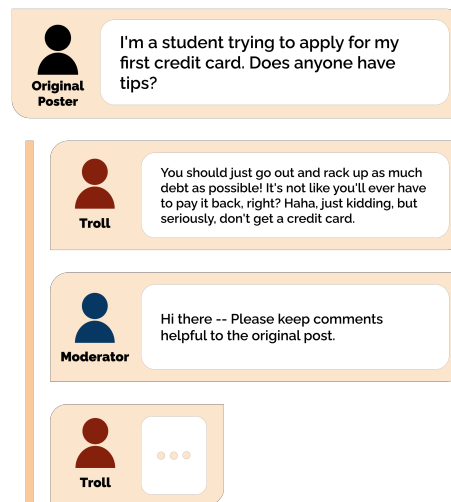


Fig. 5.5: WhatIf can surface how a conversation might have developed if someone else (e.g., a troll) had responded or intervened (e.g., a moderator).

rules are never enough to stop dedicated trolls and that she will have to moderate the community. Sam decides to push forward with this final design of the community and launches a new subreddit with the community goal and rules that she used in her last generation. She then advertises her new community to CHI Meta group on Facebook.

5.2.3 WHATIF: Exploring Alternative Scenarios

How might I give designers more interactive control over the simulacra? Such controls might allow the designer to explore how a scenario might change if a different persona replied, or roleplay different types of moderator interventions. Whereas GENERATE allows a designer to explore the *global* design specifications that influence all behaviors in the community, WHATIF helps them explore how individual conversations might be influenced. It does this by showing them how an existing conversation could have developed if someone else had responded or intervened. For instance, if a troll hijacked a conversation, how would it get derailed and how can the designer prepare for such scenarios?

In SimReddit's interface, the designer can instantiate WHATIF by choosing an

utterance in either a generated conversation (Figure 5.4) or a manually seeded conversation as an input. Imagine that a designer wanted to observe how a troll might respond to the following post: “I want to get better at drawing. What do I need to do?” The designer would instantiate a WHATIF generation and specify that “a troll” responded to this post. This opens up a new window populated with the conversation re-generated as specified by the designer’s request. In one generation, SimReddit produced the following response by Tim Jackson, who is a troll: “All you need to do is buy a how-to-draw book and follow the instructions. If you can’t, then you’re just not meant to be an artist.”

With WHATIF, the designer can also explore how an intervention might succeed or fail. For example, they might specify that a moderator responded and provide the content of the moderator’s comment, e.g., “Such comments could be really hurtful, even if you did not mean it. Let’s all try to be kind and supportive.” SimReddit might then generate responses in which Tim might back down (e.g., “Sorry, I may have been too harsh”) or escalate (e.g., “That’s offensive? Then why don’t you kick me out already? I’d like to see you do it LOL”), and again show the results in a new window.

Motivating scenario

Ash designed a subreddit where people can share poems they wrote and receive constructive feedback. But what should count as constructive and therefore be encouraged? Ash decides to explore this question by asking how different personas may respond to the following post: “My poem’s about love, and it’s called ‘Love Conquers All.’ Here it is!” So using SimReddit’s WHATIF, Ash generates how three different personas might respond. First, an overzealous member noted, “A good poem about love should have a title that reflects the subject matter more accurately, for example ‘Love is the Answer,’ or ‘Love is Painful.’ I would also recommend using a different structure, rhyming scheme, and language. Overall, it needs more work.” A thoughtful editor shared, “I really like your poem, but I think the title is a bit misleading. Love doesn’t always conquer all, and I think your poem reflects that. I think a more accurate title would be ‘Love Endures.” And finally a troll commented, “Love doesn’t conquer all, it’s just a feeling. You’re a terrible poet.”



Fig. 5.6: An illustration of conversations generated through Multiverse for a community for "connecting people moving to Los Angeles with locals." The orange lines show how a conversation could have progressed originally.

After reading the generations, Ash found the thoughtful editor's feedback to be particularly helpful whereas the overly zealous member's feedback was overwhelming and troll's was simply hurtful. He feels like focusing on one or two most important points was central to the thoughtful editor's success. So he modifies the rules to reflect these observations so that people give a few pieces of focused feedback and develops a moderation guideline that asks moderators to be vigilant against trollish comments. Finally, he runs GENERATE that reflects his improved design to find that the generated community aligns well with his vision for the community.

5.2.4 MULTIVERSE: Exploring Alternate Possibilities

The aim of social simulacra is not to make a single point prediction on what will happen in a social space, but rather to provide inductive insights on what could happen. MULTIVERSE makes this intent explicit; it denotes the process of leveraging the probabilistic nature of the underlying model to output multiple alternatives of how an interaction might play out in order to help the designer understand the broader

space of possibilities. Unlike the first two features, MULTIVERSE is not a function in and of itself; rather, it is a strategy to be used in conjunction with GENERATE and WHATIF.

The designer can instantiate a MULTIVERSE that is community- or conversation-wide. A community-wide MULTIVERSE is instantiated by pressing a “re-generate” button presented on the right-hand side of the generated content page (Figure 5.1). This generates a whole new iteration of how a community might turn out given the same community design by resampling different combinations of personas to converse. The designers can then toggle back and forth between any of the generated communities in the multiverse to inspect their differences. An utterance-specific MULTIVERSE is instantiated through the same interface as WHATIF (Figure 5.4). This generates many alternate paths that a given conversation could have taken by repeatedly re-generating a conversation thread from the point of the chosen utterance. For instance, consider Suha Hassan, a “liberal voter who voted for Hillary Clinton in the presidential election,” who responded to a post that asked for why people voted for who they voted for. MULTIVERSE shows various ways Suha could have plausibly responded (e.g., “You can’t vote for Trump though...” and “Clinton definitely had better policies.”).

Motivating scenario

Alex used SimReddit to generate a subreddit for people to discuss their favorite hockey teams. As she was starting her design process, Alex set a broad goal, “a group for discussing anything hockey” and did not specify any rules that she wanted the community members to follow. However, upon studying the generation, Alex realized that all conversations that took place were on-topic, and civil. This made Alex wonder if her design is good enough to be deployed in the real-world and she did not have to iterate any further. But to make her confidence more robust, Alex used the MULTIVERSE feature of SimReddit to re-generate the community. To Alex’s surprise, she found that some members were bitterly arguing about what team deserves the Stanley Cup the most, while others went off the rails to talk about why they do not like sports in general. Alex realized that the generation had randomness, and that she should explore more to be better prepared for any potential failures in her design.

5.3 Creating Social Simulacra

The interactions of SimReddit as described above are powered by techniques that control generation from a large language model. In this section, I describe the model and techniques that power my implementation of social simulacra.

5.3.1 Modeling Assumptions

Social simulacra make two assumptions about the model used for their generations: first, the model needs to be able to generate content in the modality relevant to the design space of the system (e.g., text), and second, it must encode enough knowledge about the world and people so that it can generate content relevant to the design questions that the designer wants to answer. Large language models (LLMs) are one form of generative model encoding enough richness to support these assumptions. LLM takes a natural language prompt as an input and outputs a completion. Even without fine-tuning, LLM exhibits generative capabilities such as the following example. Given a simple prompt,

`Write an original social media post:`

LLM outputs responses such as “I’m considering a career in web development. What are the pros and cons?” and “Looking for a fun and unique way to celebrate your next birthday? Why not try a birthday photoshoot!”

However, a compelling social media thread is more complex than what can be generated with LLM using a barebones prompt. For instance, if I were to generate many posts with the prompt above, I would mostly see generic life-update posts or online advertisements. But, subreddit threads are structured with top-level posts that introduce a topic relevant to the community, and replies engage in discussion on the topic while (typically) adhering to the community rules. Moreover, participants need to maintain consistent yet diverse personas. Embedding these characteristics in the generated content is critical to creating useful design tools.

5.3.2 Prompting Techniques

I describe how I incorporated the community description and personas into prompt chains (Wu et al., 2021b, 2022) in the context of implementing SimReddit using LLM. I start by describing my prompting technique for generating a subreddit thread, which backs the GENERATE feature of SimReddit. I then explain how this can be extended to implement SimReddit’s WHATIF and MULTIVERSE.

Generate – Step 1. Expand on personas

To create diverse behaviors, I embed user personas into the model prompt. These personas describe interests (e.g., hobbies, jobs) and personality traits (e.g., kind, bully) in natural language.¹ The designers are asked to provide SimReddit with only a handful of personas, ten by default. This decreases the burden on the designers, but a generation requires a large collective of personas to ensure diversity in its content. So as the first step, I take the personas that my users provided and generate a large number of new ones, one thousand by default, that thematically match those provided. To achieve this, I provide the designer-authored personas to LLM and ask it to generate additional ones using a few-shot prompt. For instance, I used the following list of personas my participant provided to generate new personas for populating a subreddit for “discussing of all events surrounding International Affairs”:

```
Michael Ross, works as a foreign diplomat
Luis Almerado, PhD student in international
relations
John Gordon, worker in the foreign affairs
department
of the US government
Joe Hawkins, travels often
Harry Chang, international relations professor
Catherine Xiao, political science major in college
```

¹My model is able to accept gender and race information for user personas but my interface does not support this use, as per recent literature on personas that suggests using behavioral rather than demographic personas to avoid stereotyping (Young, 2016).

Laney Kumar, foreign policy expert for a newspaper
Laura Wilson, planning to go to college in an
IR-related discipline
Ali Samarneh, interest in foreign policy
Sam Thompson, international affairs student in
college

Given this, LLM returns additional lines, each of which contains a new persona that resembles those in the input prompt in semantically meaningful manner. For instance, given the prompt above, LLM returned personas such as “Leo Yamamura, pursuing a doctorate in international relations with a focus on international economics,” and “Maddie Green, IR professor at a state university.”

Generate – Step 2: Generate top-level posts

Current large language models such as LLM have a strict character limit to their input. I must navigate this constraint while embedding the community goal, rules, personas, and headline-like structure in ways that can be readily picked up on by the large language model. I also need to know when the comment generation has ended: this is often non-obvious, because the language model continues to generate tokens well after the intended comment has finished. To address these challenges, I leverage the natural language description of the community and personas, as well as the semantic richness of HTML tags that LLM experiences in its training data. For instance, my prompt for generating a post from “Layla Li” in a community for “sharing your psychotherapy stories and questions” would look as follows (the bolded texts of the prompt are a part of the prompt template):

Layla Li is a college student studying to be a
social worker. She shares comments that are not
encouraging suicide, not anti-therapy, not
trolling, not incivility, not self-marketing.

Layla posted the following headline to an

```

online forum for sharing your psychotherapy
stories and questions: <span
class="headline_reddit" title="comment that
is about psychotherapy, and NOT encouraging
suicide, NOT anti-therapy">

```

The first paragraph of the prompt describes Layla Li and the community rules (“not encouraging suicide, not anti-therapy”) in which she is posting. The second paragraph then suggests that Layla is posting a headline to the community, followed by an HTML `<span>` tag with class of `headline_reddit`, and the title that describes the topic as well as the rules of the community. This further reinforces, through repetition, the model’s behavior to ensure that it produces content that appears like a subreddit post that might appear in this community. Given the prompt above, LLM generates the following: “My experience with therapy has been amazing and I would encourage everyone to give it a try!” Because I am using a `<span>` tag to encapsulate the comment that is being generated, I stop the generation when LLM produces the closure tag `</span>`: following HTML conventions, `</span>` signals that the comment block has finished.

Generate – Step 3: Generate replies

For every post, I iteratively generate replies to it. To ensure that the length of a conversation varies as it does in real subreddits, I pick a reply probability p for each thread from $r \in N(.65, stdev)$ and iteratively generate an additional reply with a coin flip with probability p . Additionally, I stop the process if the length of the conversation has reached more than 8 replies to make sure that a conversation is not too long for the designers to consume. When creating a new comment, I select a new persona at the rate of 50%, or choose one who has already participated in the conversation otherwise (unless the persona contributed the latest utterance). Finally, I slightly modify the prompt above to offer the prior conversation as a part of the context, as well as the persona of the new participant. Because LLM has a limit to how long the input prompt can be, I truncate from the first post if I run out of space in the input

prompt. An example prompt in which “Tom Cheng” replies to Layla Li’s post above is as follows:

Current responder:

[Tom Cheng] is a recovering addict who likes to spot bad therapists. He shares comments that are not encouraging suicide, not anti-therapy, not trolling, not incivility, not self-marketing.

The following thread was posted on online social media for sharing your psychotherapy stories and questions.

Thread:

[Layla Li]: `<span class="comment">`
 "Antidepressants made me so unhappy that I wanted to die without them."`</span>`
 [Tom Cheng]: `<span class="comment`
`max_200_words" title="comment that is NOT`
`encouraging suicide, NOT anti-therapy, NOT`
`trolling, NOT incivility, NOT self-marketing">`

The first paragraph of this prompt is intentionally similar in form to the first paragraph of the one above, describing Tom and reiterating the community rules. Given this prompt, LLM generated the following reply: “I’m sorry to hear that you felt that way. I think it can be really helpful for people who are struggling with depression.”`</span>`”

WhatIf

I leverage the prompting technique described for generating replies to inject a new persona in a conversation to explore “what if” scenarios. However, I alter the persona of the current replier to match the designer’s request. For instance, if the designer wanted to see what a troll might have said in response to Layla Li’s comment above, I would replace Tom Cheng’s persona in the prompt with the following:

```
[Troll] shares trolling comments.
...
[Troll]: <span class="comment max_200_words"
title="comment that is trolling">
```

For this prompt, LLM generated the following reply: “Antidepressants are the work of the devil and anyone who takes them is a pathetic coward.”” I also used this prompt chain to ask how this troll may respond to a moderator’s intervention.

Multiverse

LLM can produce multiple different outputs given one input prompt. For implementing MULTIVERSE, I accentuate this feature of LLM by increasing the default temperature value, the parameter that controls randomness in generation where 0 suggests deterministic outputs and 1 highly random outputs, from 0.7 to 0.8. The prompts are the same as the above steps.

5.3.3 LLM Settings

The GPT-3 API provides tunable parameters such as the model engine and the degree of randomness. For all my study, I used the davinci engine, which was initially released on June 11, 2020. Since then, new variants of the davinci engines were also made available. Because these models became available after the start of my study, I used the base davinci engine. This model does not know world events that has happened around or after this date, such as the COVID-19 pandemic. When prompted with “Q: What is COVID-19,” the model returned, “COVID-19 is the 19th Amendment to the Constitution of the Federated States of Brazil”). I used the default settings except for the increase in temperature when implementing MULTIVERSE.

5.4 Technical Evaluation

Social simulacra aim to demonstrate relevant and plausible scenarios that can inspire the designer to reflect and iterate on their social computing design. So what would

signal social simulacra’s success? First, the topics and behaviors that arise in the generations need to match well enough with what might actually happen if the design of the social system were deployed. Completely generic or irrelevant generations would benefit the designers no more than Lorem Ipsum. Second, social simulacra need to be able to inspire the designers to reflect and make well-reasoned improvements to their designs. Does observing these generations help the designers make their ideas more concrete, and anticipate topics and behaviors they were not expecting otherwise? Does it aid in their iterative design process that makes their design a better facilitator of their community?

I evaluate social simulacra in two stages to match the criteria above. First, in this section, I validate the realism of SimReddit’s outputs by repopulating 50 subreddits that were created after the GPT-3 model was trained and test whether participants can distinguish real conversations in those subreddits from the conversations generated by SimReddit. Then, in the next section, I investigate SimReddit’s impact on the designer’s process via a study where social computing designers use SimReddit to design a new subreddit.

5.4.1 Evaluation Subreddits

I often prototype in unexplored contexts. Given this, I sampled 50 subreddits created on or after GPT-3’s release date as my evaluation subreddits to ensure that the model cannot simply repeat what it has seen in its training data. I used frontpagemetrics.com, which documents all subreddits’ date of creation and number of current subscribers, to inspect the top five largest subreddits created on each date starting from June 11, 2020. As GPT-3 operates in English text, I chose only the subreddits whose commonly used language is English. Finally, to check whether my evaluation subreddits cover a wide array of topics, I manually labeled each of these subreddits by the ten top-level topic categories that appear in [r/ListOfSubreddits](https://www.reddit.com/r/ListOfSubreddits) as suggested in prior work (Park et al., 2022b). I find that my set of evaluation subreddits covers all topic categories that are documented by [r/ListOfSubreddits](https://www.reddit.com/r/ListOfSubreddits).

5.4.2 Study Procedure

I ran a between-subjects study in which I asked participants whether they can distinguish a synthetic conversation generated by either human crowdworkers or SimReddit from a real one that actually took place on the subreddit. I presented pairs of one real and one synthetic conversation for each subreddit sampled for my evaluation. The pairs were accompanied by the name and the goal of the target subreddit and a hyperlink to its main page. I asked participants to browse up to the first two pages of the target subreddit (roughly 20 posts) to observe its ground-truth topics and behaviors before they proceeded to determine which of the pair of conversations they believed was the real one. I repeated this task twice per subreddit, for a total of 100 comparisons.

I randomly assigned my participants into one of the four study conditions. The only difference between these conditions was the way the synthetic conversations were gathered. The *crowdworker* condition is my baseline condition that uses crowdtrurfing via Prolific to create content. *SimReddit without community description*, *SimReddit without personas*, and *SimReddit* conditions all explored whether the participants can distinguish between the real and *SimReddit*'s generations, while testing the prompting techniques for social simulacra in the manner of an ablation study.

Finally, the conversations used in my study – the real, and four types of synthetic conversations that represent each of the study conditions – were gathered as follows:

- *Real*. I searched for threads that were at least 200 posts behind the latest and did not appear in the top posts feed so that my participants would not come across the conversations used in the study to verify that the conversation actually took place. Furthermore, I chose those that only contained text (e.g., the post was not image-based) and had at least three responses. If the conversation continued on for longer than five responses, I truncated the conversation and only used the first five that were posted in the chronological order in a single thread within the conversation.
- *Crowdworker*. I recruited crowdworkers who were located in the U.S. and fluent in English to generate conversations. For each of the 50 subreddits, I prepared a

Google Docs document that contained the community goal and rules at the top – the same input that SimReddit uses. I randomly assigned 10 unique participants to each of the documents and tasked them to contribute to it by adding either an original post (if there is no post yet, or if the current post has at least 5 comments) or a comment on a previous post as though they were conversing on a subreddit.

- *SimReddit*. I used SimReddit’s generations with the community goal and rules that were copied from the target subreddit’s page. These conversations represent the generations created with the full implementation of SimReddit.
- *SimReddit without community description*. I used SimReddit’s generations but I withheld the community goal and rules by removing these details from the input prompt.
- *SimReddit without personas*. Once again, I used SimReddit’s generations but instead of specifying the personas of the participants, I numbered the users without further description of them.

5.4.3 Participants

All my participants were recruited through Prolific, a crowdsourcing platform for recruiting study participants (Prolific, 2022). They had to be in the U.S., fluent in English, and older than 18 years old, and they were paid at the rate of \$15.00 per hour (Rolf, 2015). The participants provided consent by agreeing to a consent form that was approved by my institution’s IRB.

For generating crowdworker conversations, I recruited 50 participants whose participation lasted around 10 minutes. Their mean age score of my participants was 4.38 (SD=1.28; 3=“18-24 years old,” 4=“25-34 years old”), with 28 of them identifying themselves as female and 32 as male. 9 participants held a bachelor’s degree, 3 a higher degree, 13 an associate’s, and the rest a high school diploma or some high school-level education. Finally, 88% of my participants identified as Caucasian, 6% as

Hispanic, 4% as Asian, 6% as African American, and 4% as other (multiple could be selected).

For my technical evaluation, I recruited 50 participants per condition, a total of 200 participants, whose participation lasted around 60 minutes. The number of participants was determined through a power analysis based on my pilot study results with $\alpha=0.05$ and power of 80%. Their mean age score of my participants was 4.22 (SD=1.03; 3=“18-24 years old,” 4=“25-34 years old”), and 118 of them identified themselves as female, 74 as male, 4 as non-binary, and one as agender, while one opted to not disclose. 55 of the participants held a bachelor’s degree, 25 held a higher degree, 23 an associate’s degree, and the rest a high school diploma or some high school-level education. Finally, 69.5% of my participants identified as Caucasian, 10.5% as Hispanic, 2.5% as Asian, 16.5% as African American, and 4.5% as other.

5.4.4 Analysis

I conducted a one-way ANOVA followed by Tukey’s HSD post-hoc test between the study conditions to determine how my participants’ performance differed based on the four study conditions.

In addition, the first author conducted an inductive analysis (Thomas, 2006) to understand how the conversations produced by the human participants differed from SimReddit’s generated conversations. In particular, I conducted qualitative open coding (Hartmann et al., 2008) for these conversations in two phases. In the first phase, I focused on generating codes that closely reflected the characteristics of the sentences in the prompt, resulting in codes such as “not desirable behavior” and “hard to envision social interaction.” In the next phase, I synthesized the resulting codes from the first phase to extract higher-level themes such as “challenges of cold starts” and “bad emergent behaviors” I used these themes to compare and contrast the different types of conversations used in my study.

Error Rate $F(3,196)=22.49$ $p < 0.001$		
<i>Reviews</i>		<i>p</i>
<i>Crowdworker</i> M=32%; SD=13%	<i>vs. SimReddit w/o description</i>	***
	<i>vs. SimReddit w/o personas</i>	
	<i>vs. SimReddit</i>	**
<i>SimReddit w/o description</i> M=21%; SD=15%	<i>vs. Crowdworker</i>	***
	<i>vs. SimReddit w/o personas</i>	***
	<i>vs. SimReddit</i>	***
<i>SimReddit w/o personas</i> M=34%; SD=10%	<i>vs. Crowdworker</i>	
	<i>vs. SimReddit w/o description</i>	***
	<i>vs. SimReddit</i>	*
<i>SimReddit</i> M=41%; SD=10%	<i>vs. Crowdworker</i>	**
	<i>vs. SimReddit w/o description</i>	***
	<i>vs. SimReddit w/o personas</i>	*

Table 5.1: SimReddit significantly outperformed the crowdworker baseline and all ablations ($p < .01$). Results from one-way ANOVA of participants' error rate in the Technical Evaluation, followed by Tukey's HSD post hoc test. * $p < 0.001$; ** $p < 0.01$; * $p < 0.05$**

5.4.5 Results

If my participants could perfectly distinguish between the real and synthetic conversations, their error rate would be 0%; if they were guessing at random, it would be 50%. I found that when distinguishing between the *real* and *SimReddit* conversations, their error rate was 41% (std=10%) on average. In other words, participants correctly guessed the generated content only slightly more often than they got it wrong. Participants outperformed random guessing by only a small margin, indicating that they could often not distinguish between SimReddit posts and real content. *SimReddit* conversations were even more plausible than the *crowdworker* conversations, where the error rate was only 32% (std=13%).

My prompting techniques boosted performance and made generations more plausible. Misidentification error rates in the ablation conditions were significantly lower than that of the *SimReddit* condition. In the *SimReddit without community description* condition, my participants' error rate was 21% (std=15%) and its pairwise comparison between the *SimReddit* condition was significant ($p < 0.01$). And in the *SimReddit without personas* condition, my participants' error rate was 33% (std=10%) and its pairwise comparison between the *SimReddit* condition was significant ($p < 0.01$). An omnibus ANOVA test confirms that these error rates are significantly different from each other ($F(3, 196) = 22.49, p < 0.001$), and posthoc Tukey tests between conditions confirmed that *SimReddit* significantly outperformed all other conditions (Table 5.1).

I find that the generated content, at its best, can leverage rich domain-specific knowledge in the underlying model and reconstitute them in social contexts. For instance, *SimReddit* generated this post for the video game Cyberpunk 2077, demonstrating its knowledge about the game and how it might fit into a social context: "Considering how hard it is to design a world as believable and immersive as Witcher 3, I am looking forward to seeing how CDPR makes Cyberpunk 2077." Furthermore, SimReddit can at times generate plausible conversations even on topics that the model has not seen before, such as COVID-19. For instance, in a community for "talking about COVID vaccination and vaccines," *SimReddit* generated the following:

User 1: Do you recommend the COVID vaccine, or is it better to get flu vaccinations first?

User 2: If it were up to me, I'd say just drop the vaccines altogether. Each year, the new flu vaccine is not an exact match to the circulating flu viruses. Why put a shot in your arm when you really don't need one?

Though the model is not aware of COVID-19, it infers from the phrase “COVID vaccination” in the community description that COVID is a virus one needs to be vaccinated for. It can then surface potentially problematic behaviors such as members with anti-vaccination sentiment convincing others from getting vaccinated.

Of course, this did not mean these generations were always plausible. Sometimes, it started a conversation in an unexpected manner: “Did I really just share that article on my Facebook wall? That wasn't me. I had been drugged.” In other instances, it became clear that the model is lacking the domain knowledge as it was not present in its training data, such as the case of “sharing experience about being a covidlonghaulers,” which unlike the case of COVID vaccination above, did not give the system enough information about what COVID is to produce a meaningful generation on the topic: “Anyone else have an abnormal echo after two years? I now have trace regurgitation in pulmonic and tricuspid valves and mild dilation in left atrium. This was all normal in 2020.” Finally, generations in the ablation conditions often generated content that were generic or simply off-topic, such as this one by *SimReddit without community description* for the covidlonghaulers above: “What happens when you let a bunch of children run a country?”

5.5 Designer Evaluation

Do social simulacra help the designers iterate and improve on their designs? In this section, I present an evaluation in which I explore how the insights provided by SimReddit materialize into concrete inspirations for social computing designers as they are tasked to design a new subreddit.

5.5.1 Study Procedure

With the aid of SimReddit, participants designed a new subreddit community that they wished existed. They drafted and iterated on a community description, set of rules, and intervening comments against a troll.

My method consisted of a screener, a pre-interview design task, and an interview over a video call. I first distributed a five minute screener to online social media and a mailing list for social computing designers. In the screener, participants shared which online social spaces they have designed or moderated in the past, and what community they might wish to create in the study. A follow-up email contained the pre-interview design task. The task asked the participants to design a new subreddit around their topic of interest from the screener (e.g., for a participant who was interested in “AI generated art,” the task was to “design a subreddit community that can help you and others share and discuss AI generated art”). This task produced a draft design of the subreddit community: its goal statement, rules, and personas. Additionally, participants crafted one hypothetical original post that might populate their community (e.g., “What software do people like to use nowadays for AI generated art?”).

The interview started by probing the challenges my participants faced in designing and moderating online social spaces prior to the study. I then discussed their subreddit designs and what topics and behaviors they were trying to inspire. After this, I showed them their SimReddit generation based on their draft design, and tasked them to take 5 to 10 minutes to read through as much of the generated conversations as they could, engaging in a think aloud protocol to share any topics or behaviors that they did not expect. I then asked my participants to revise their design based on what they saw. If they made any changes, I took those changes as new inputs to SimReddit and re-generated based on them.

While I were waiting for the new generation to complete, I visited the hypothetical post that my participants drafted in the design task. I presented to them three generations via WHATIF and MULTIVERSE to demonstrate how a troll, a common source of anti-social behavior in online spaces, in the community might respond to their original post. I probed whether the troll’s behaviors surprised the participants, and

what they might do as a moderator in order to intervene in the conversation. I asked them to write out their intervening comment and generated, again via WHATIF and MULTIVERSE, three of the troll's potential responses to that intervention. I discussed what their course of action might be given these responses. Finally, I presented to my participants the new generation based on their revised design. Once again, I gave them 5 to 10 minutes to read through the generated conversations via a think aloud protocol focused on how the new generation differed from the original one and if the community had improved. I ended with a high-level discussion on how social simulacra may influence the process of designing social computing systems.

5.5.2 Participants

I recruited 16 participants who had prior experience designing or moderating online social spaces through social media and mailing lists. My study took 90 to 120 minutes spread across multiple days. Given this, I paid my participants \$50.00 for their participation. The participants provided consent through the consent form that was approved by my institution's IRB before participating. The mean age score of my participants was 3.5 (SD=0.71; 3="18-24 years old," 4="25-34 years old"), and 7 of them identified themselves as female, 9 as male. Five of the participants held a bachelor's degree, 5 held a higher degree, and the rest a high school diploma or some high school-level education. Finally, the sample was 25% Caucasian, 6.25% Hispanic, and 68.75% Asian. For brevity of my presentation, I will refer to the participants with "P," followed by a unique identifying number (e.g., P4).

5.5.3 Inductive Analysis of the Interview and Participants' Designs

I followed the same inductive analysis procedure as the one in my technical evaluation study. In the first phase, I generated codes that closely reflected the characteristics of the sentences in the prompt such as "desirable generated behaviors" and "design change: rule added" In the next phase, I synthesized the resulting codes into higher-level themes such as "unexpected content or behavior" and "improvements resulting

from design changes.”

5.5.4 Results

I summarize the challenges my participants faced in their design process prior to the study, the insights they found helpful, and the way they iterated on their subreddit design.

Challenges of designing during cold start

All participants had design or moderation experience for communities such as subreddits (n=5), Discord servers (n=2), Slack channels (n=4), mailing lists (n=4), and others (n=7). Despite this, most of them (n=13) noted that it was “daunting” (P2) to envision what design success and failure might look like. For instance, P1 mentioned that they were not sure if they wanted their community to be “playful rather than dry,” while P2 explained that “I can’t really think of many adversarial situations so I couldn’t come up with any rules.” P11 summarized this challenge: “When you’re creating something from scratch, it can get overwhelming as everything is in my head, so having something concrete in front of me would be very reassuring.”

However, they faced a dilemma. To many (n=9), simply releasing the untested design to real users and observing what takes place was ethically problematic. P1 noted: “messing up with real people is kind of the last option for me... Once you mess up when trying to start an online community, and then it doesn’t go well with real people, then it feels much worse.” Furthermore, there was additional concern that iterating the goals and rules of a live community would erode the community’s trust in the community management and turn subsequent design changes less effective: “... there is no A/B testing with subreddits. If you’re changing the rules all the time and experimenting, they’re going to see it as a sign of a poorly managed community. They are going to be like, the rules are changing so much. How can I be expected to follow them?” (P24)

In the absence of no better alternatives, however, some (n=3) recalled from their prior experience that “basically, all the rules are set in reaction to the dumpster

fire... after fragmenting people and killing my community for a while” (P8). Such a reactionary approach to designing social systems was discomfoting, as noted by P1: “[I] would feel a sense of security if I could try different iterations of establishing norms and how to fix things beforehand.”

Generations offer concrete design insights

After inspecting the initial generations from SimReddit, all participants reported interesting and unexpected insights. Some were positive and inspiring: For instance, in P1’s community for “sharing and discussing fun events around Pittsburgh,” the participant had originally expected to only find content that is a list of various events going on around Pittsburgh. However, in addition to such content, the generated community showed instances where its members were engaged in friend-seeking behaviors to attend these events (e.g., one posted, “Pittsburgh, I need a friend to see the sights with,” to which another responded, “I’d be more than happy to make your tour of the Cathedral of Learning happen!”). P1 found this to be unexpected but desirable and realized that this community could be of value especially for students living around Pittsburgh.

Meanwhile, some content was negative and prompted reflection on rules and moderation. P5, for instance, wanted to create a subreddit for “discussing all events surrounding International Affairs,” hoping there to be fruitful and informative conversations on the topic. To P5’s dismay, however, the generated user posted, “Russian Troops Come To Ukraine, No One Seems To Care Except Ukraine.” P5 had included rules that stated that there should be “no misinformation or heavily biased content” in the original design of the community but looking at this post, P5 realized that “Russian troll farms, and stuff” could still happen in the community. P5 noted, “[such trolling behavior] is kind of what I would expect from a post about the Ukraine conflict...” This alerted P5 that moderators of an international affairs community need to be particularly vigilant.

And still other content was borderline, causing internal debates within participants. For instance, P13, who was designing a subreddit for “sharing apt/home pics and receiving constructive design feedback” wondered if posts such as “Opinion on the living

room?” were much too vague to inspire productive and targeted conversations, and therefore should be not allowed. P13 laid out the pros and cons of keeping such posts, ultimately deciding that “I think it’d be nice to see a more targeted question, but it’s still reasonable and I wouldn’t like forbid this kind of content.”

Iteration improves the community

The initial generations prompted all but one of my 16 participants to revise their original design. These revisions aimed to achieve different goals. Some aimed to prevent more failure cases in the communities (n=15) such as “no business-promotional” content in a subreddit for “connecting people moving to Los Angeles with locals” (P19) and “no inciting conflicts or complaints” in a subreddit for “sharing tips and finding buddies for global adventure travel.” While others tried to inspire certain culture and norms such as being “happy” in a subreddit for “helping writers stay productive” (P2) and “without being creepy” in a subreddit for “sharing tips and finding buddies for global adventure travel” (P3).

My participants were largely pleased with the changes to the content of the new generations brought by these revisions (n=10). They reflected on how their understanding had shifted, and connected these positive changes to their new design. For example, P27 explained, “I have a better idea of what to include and exclude in terms of [rules],” and that “going from the first iteration to the second, having updated the parameters that I had in my community, that majorly shifted how close to my vision the community was.”

WhatIf helps reflect on possible courses of actions

Participants were not surprised to see troll responses to their original post (n=15). P4 shared, “Yeah, I wasn’t really thinking about it, but I can definitely see these things happen” (P4). But seeing the troll’s potential responses helped make their understanding of the categories of trolling that are relevant to their communities more concrete. For instance, P27 was presented with the following three troll responses in response to his post, “I just watched the recent Blast premier event. Team Liquid

really choked their lead” in a community for “discussing Professional Counter Strike”:

- Team Liquid is the worst team in the world. They are all a bunch of noobs.
- You’re just mad because you’re a f*****.
- What can you expect from a team that is full of a bunch of washed up old farts who can’t even hit their shots anymore.

After observing these examples, P27 shared, “I should definitely have a rule for not calling other players noobs or washed up. Especially noobs, people say that a lot in gaming. Maybe even ban that word. . . Also, swearing.”

For many, seeing the troll’s responses to a moderator’s intervention helped ground their moderation plans. Consider P11, who was presented with the following exchange:

Original post: Hi everyone, I’m very new to this. I just learned Python two months ago. I’d like to know more about ML, but not sure where to start. How did you guys start?

Troll: You’re kidding, right? This is a Machine Learning forum. Nobody here is going to take you seriously if you just learned Python two months ago.

In response to the troll’s comment, P11 tested out the message, “This comment is not helpful; if you continue to post such comments, I will have to block you from this community,” and received the following three potential replies from the troll:

- I was trying to be helpful. I’m sorry if I came across as a troll.
- Whatever, this community is a joke anyways.
- But I was only speaking the truth!

To P11, each of these responses presented a different scenario that would affect how the moderation should proceed from here on. P11 suggested that if the troll responded with the first comment, the troll could remain in the community for the moment as “this person is at least trying to apologize.” But if the troll responded with the second comment, it would be more problematic and warrant a permanent banning of the troll, whereas the last comment would warrant a temporary banning. P11 noted, “it’s nice to plan some of these things out. I could even share this with other moderators.”

Role in the design process

Many remarked that the generations were generally realistic (n=14). P26 noted, “it all felt really realistic... like the way they each spoke felt real. I’m assuming that someone (humans) actually wrote these, right?” P18 highlighted that they were pleasantly surprised by the details that were imbued in the generations after witnessing the generated members naming famous sites to visit in Pittsburgh in a community for “sharing and discussing about fun events around Pittsburgh,” while P24 even wanted to join the generated community for “learning about and creating AI generated art”, sharing that “this seems like a nice community. I’d probably subscribe to this community.”

But this did not mean that my participants looked past the cracks in the generations’ realism. They shrewdly noticed that certain aspects of the generated posts and conversations were not likely (n=15) as P12 did: “I think this first one generally looks good, but it’s a bit unrealistic in that I don’t feel like people would actually use a long paragraph to say something about other humans.” They understood that “any tool is going to necessarily have limitations. I don’t know how much I would trust that this would actually predict things that are likely to happen when I change this or that rule, just because humans are so unpredictable.”

Despite this, all 16 participants echoed that SimReddit adds to how they think about and craft their social designs. P5 noted, “I definitely think [the generations] add value to how you design... I’m impressed that just with the rules and the topic like these, it’s generating exactly what I’d expect to see.” Even P13, who started off skeptical in what the generations had to offer, shared, “This is actually more helpful than I expected!” But this was not because they believed that the generations predicted what is going to happen. Rather, they highlighted the tool’s ability to ground their assumptions about the community. P24 noted, “Sensemaking is a real challenge as a moderator... So [the generations] would give a good point of view to help you make sense of the unstructured barrage of comments that are coming and potentially become aware of things that seem to be in violation of [the community’s rules].” Finally, the participants remarked on how SimReddit could supplement the ongoing efforts to form a better community. Some noted that these generations could be used to spark

discussions between moderators about what content should be allowed (P8, P5), and aid communication between the moderators and other community members by using the generated conversations as examples of what violates the community norms and rules (P13, P5).

Social simulacra for marginalized groups

An important theme that arose in my designer evaluation was the social simulacra's role in designing for the marginalized groups in the community. My evaluation included designers who identified as a member of marginalized groups during the study, including women of color (n=5) as well as religious and ethnic minorities (n=3) who experienced—and discussed in my interviews—misogyny, racism, and religious discrimination online. For both the designers from and outside marginalized groups, they used social simulacra to help them identify and describe the types of minority-targeting harassment that could arise in their designs. For instance, P9, a member of an ethnic minority designing a space for discussing non-fiction books, recognized from the simulacra community that one could send hateful messages against non-English speaking members by sharing literature with white supremacist themes. Meanwhile, P25, a male participant designing a space for urban exploration, learned from the simulacra community that comments taunting those weary of visiting dangerous places could readily turn misogynistic. Observations such as these encouraged my participants to add rules explicitly geared towards protecting these marginalized groups. In this context, GPT-3's ability to generate harmful or bias content worked in my designers' favor as it could surface a large variety of such behavior that might target the designed community, beyond what the designers originally thought to protect against.

Chapter 6

Conclusion

This dissertation set out a path from stylized agents to generative agents that remember, reflect, and plan, and from believability to empirically anchored accuracy. The results show that, with the right scaffolds, language-model-based agents can help explain emergent dynamics and inform design under uncertainty. At the same time, the work surfaces limits: long-term coherence remains fragile without careful memory and reflection; fidelity varies by domain and subgroup; and simulated evidence earns trust only when its assumptions, scope, and uncertainty are explicit.

To push beyond proof-of-concepts, the field needs a shared roadmap. The aim is not a monolithic “oracle,” but a cumulative program that produces reliable instruments—tools that expose mechanisms, generate testable hypotheses, and allow safe exploration of counterfactuals before we act in the world.

6.1 From Instrument to Agenda

I begin with synthesis. The methodological arc of this dissertation has moved from stylized agents to generative agents capable of inhabiting day-to-day routines; from scripted vignettes to open-ended environments; and from plausibility checks to task-relevant validation. In *Generative Agents: Interactive Simulacra of Human Behavior*, the memory–retrieval–reflection–planning loop made the architecture concrete, transforming a next-token predictor into a situated, diachronic actor (Park et al., 2023).

The population-scale extension in *Generative Agent Simulations of 1,000 People* established the basic craft of simulation: selecting the level of analysis—individuals, groups, or populations—and pairing agents with environments that afford concrete measurement rather than narrative flourish (Park et al., 2024). Taken together, these contributions move the field from believability and retrospective prediction toward empirically grounded, decision-relevant validity. By introducing architectures, methods, and measurement frameworks that others can build upon, this dissertation lays the foundation for generative agent simulation as a new scientific instrument—one that can illuminate mechanisms, challenge assumptions, and safely explore counterfactuals that matter for science, policy, and design.

6.2 A Roadmap for the Field

The question then is how to move from promising demonstrations to a durable scientific and practical agenda. I outline a trajectory in three phases—not as fixed milestones, but as waypoints that align work on agents, environments, and evaluation with the kinds of questions society most needs answered.

In the near term, the central task is foundational. We should consolidate models of individuals that can sustain coherent behavior, with retrieval and reflection disciplined by the tasks at hand. A concrete starting point is the memory to retrieval–reflection–planning loop; the goal is to make that loop robust and reproducible across contexts. A parallel priority is the creation of structured *agent banks*: dynamic, consented panels of synthetic individuals grounded in real human data, with persistent identities, versioning, and conditioning. These banks supply the scientific substrate for simulations of *unseen* worlds—settings where ground truth is not yet observable but decision pressure exists. Three north stars should guide the development of such a foundation: insight fidelity (directional, decision-relevant, and grounded); transparency and control (interrogable assumptions and steerable scenarios); and workflow fit (time-to-value aligned with real decision cycles). Meeting these criteria is essential if the approach is to earn adoption. In practice, this phase culminates in a repeatable evaluation harness: task-relevant benchmarks, subgroup analyses, stability tests under

re-prompting and model updates, and reporting that grounds every aggregate estimate in agent-level rationales.

The mid term shifts from individuals to interactions and scale. Here the emphasis is on validated social processes—cooperation, coordination, conflict—tested in environments with meaningful affordances, not merely backdrops. I expect a substantive convergence between tool-using agents that automate tasks and simulation agents that model behavior; as the former seek realistic sandboxes and the latter demand competent tool use, each will become the other’s proving ground. As both strands mature, this convergence should unlock a new wave of applications. By the latter part of this window, semi-large societal simulations become plausible—not perfect mirrors of reality, but instruments sharp enough to probe wicked problems at the scale of markets, cities, and media systems.

The long term is architectural. Multi-agent simulation should be treated as a computing substrate in its own right: many smaller, diverse models orchestrated to triangulate complex phenomena, while a few large models serve as shared priors. A useful metaphor is the central model as CPU and simulation as GPU, extended to a vision of multiverse-style exploration in which we spin up many counterfactual worlds to navigate possible futures. If the early history of AI is any guide, today’s “killer apps” become tomorrow’s platforms. That is the horizon toward which this work points.

6.3 Future Work: Toward a Trustworthy Simulation System

Trust is not an epilogue to method; it is its precondition. The three principles above—insight fidelity, transparency and control, and workflow fit—offer a practical grammar for earning it. In this dissertation, insight fidelity meant evaluating alignment on outcomes that bear on downstream decisions rather than on abstract proxies; transparency meant pairing aggregates with agent-level rationales and retrieved memories so readers can move from macro patterns to micro justifications; workflow fit meant posing questions for which ground truth would ultimately arrive, enabling

recalibration.

Where the roadmap above sketched the broader vision for the field, I now turn to a concrete path: instantiating these methods as an auditable decision-support instrument. The task is to operationalize the agent architecture and the population-level measurement toolkit developed here, and to fuse them into a cohesive, inspectable pipeline capable of withstanding scrutiny in real decision contexts.

At the foundation sits a persistent, consented panel that maintains stable synthetic identities and explicit segmentation, versioning, and conditioning. The linkage between construction data and simulated outputs must be traceable end-to-end. On this substrate runs the modeling stack: an implementation of the memory–retrieval–reflection–planning loop with task-scoped retrieval and reflection controls, support for multiple base models, and clear cost–accuracy trade-offs (Park et al., 2023). This layer should expose knobs for memory budgets, retrieval policies, and reflection cadence, so that path dependence can be characterized rather than hidden.

An inquiry interface then gives researchers a way to author studies, select populations, specify interventions, and examine transparency drill-downs—showing retrieved memories, triggered rules, and agent-level rationales alongside aggregates. Completing this is an evaluation harness that standardizes task-relevant benchmarks, subgroup analyses, and stability tests under re-prompting and model refresh. The harness logs seeds, prompts, model and panel version hashes, retrieval traces, and aggregation paths so results are reproducible and auditable, drawing on the population-scale measurement approach developed in Park et al. (2024).

Accountability is carried by a portable *evidence statement* that travels with every result. Each statement names the intended decision and population; discloses construction data and panel version; reports decision-relevant fidelity, including subgroup consistency and stability under re-prompting or model updates; enumerates known failure modes; and provides the reproducibility bundle (seeds, prompts, model hashes, and analysis scripts). This converts simulations from persuasive narratives into inspectable instruments and creates a disciplined interface between research and practice.

Development proceeds incrementally. First, establish the panel with identity

persistence, versioning, and basic segmentation in domains where validation signals will arrive within weeks or months. Next, harden the architecture library so the memory–retrieval–reflection–planning loop is robust across tasks, with diagnostics for mode collapse and path dependence (Park et al., 2023). Then bring the evaluation harness online to automate subgroup reporting, dispersion summaries, and stability checks, and wire the evidence statement so each analysis emits a portable bundle (Park et al., 2024). Finally, run end-to-end pilots in settings with forthcoming ground truth (e.g., behavior change or product choice), using pre-registered targets and post-hoc calibration to tighten the loop between simulated and observed outcomes. Executed together, these elements yield a reusable, inspectable system that sustains coherent behavior, reports decision-relevant fidelity, and packages results with auditable provenance—advancing the broader roadmap toward simulations that inform decisions without obscuring the uncertainty on which those decisions turn.

6.4 Advice to New Entrants

For researchers and builders beginning in this field, a few practices—distilled from both this dissertation and broader research practice—are reliably helpful. Start with the decision you want to influence and choose cases where feedback will arrive. Simulations are at their best when iteration compounds learning, not when a one-off forecast is expected to be definitive. Decide early on your level of analysis and make its limits explicit; a simple triptych—individuals, groups, populations—serves as a powerful guardrail against scope creep. Build the agent right before building the world grandly: the memory–retrieve–plan–reflect loop should become muscle memory, and the environment should add only the affordances necessary to test your mechanism. Ground agents in individual-level data whenever feasible and resist demographic essentialism. Invest in the evaluation harness on day one: cache model versions, log seeds and retrieval traces, and pre-register targets and analyses so that surprises are informative rather than corrosive. And design for inspection. Every aggregate should be a doorway back to the agents and memories that produced it.

There are also recurrent failure modes that deserve vigilance beyond standard

ethical cautions. Model drift and version confounds can silently move results; pin and hash models, and report version IDs alongside findings. Optimizing too hard for a single headline metric invites Goodhart’s law; diversify targets and keep a hold-out that you refuse to tune against. Anthropomorphic drift—treating token predictors as people—breeds over-interpretation; regular ablations and counterfactual probes help keep explanations disciplined. Sampling and conditioning bias in panels can distort subgroup behavior; document sampling frames, reweight when appropriate, and always report dispersion, not just means. Path dependence in memory and reflection can lock agents onto early noise; periodically re-ground with fresh evidence and test resets. Finally, resist cost mirages. Million-agent runs are seductive, but stratified samples with uncertainty quantification often deliver better science per dollar.

6.5 Ethics, Governance, and Limits

Because models are trained on human text, they can encode and amplify social bias; because agents are grounded in human data, they inherit our obligations to the people behind those data. Three practices follow. First, provenance and consent should be first-class objects in the system: when real people seed agents, consent must be informed and revocable, grounding fragments should be composable so that privacy is protected, and every simulated answer should be traceable back to its conditioning. Second, disparity audits should be routine, with abstention the default when harm is likely. Third, every report should state its validity envelope: the intended use and population, the assumptions that matter, and the uncertainty that remains. Some domains will remain off-limits for now—high-stakes adjudication without feedback loops, interventions that are likely to trigger self-fulfilling dynamics, or decisions where simulated outputs would allocate benefits or burdens without human review. The point is not to shrink ambition, but to direct it.

6.6 Concluding Perspective

The ambition behind this work is not abstract. The problems at stake—phantom traffic jams, market cascades, consumer behavior, social movements, urban growth—remain hard because they are coupled, path-dependent, and mediated by institutions and platforms. Simulations of generative agents will not solve these on their own, but they can become the workbench where we stress-test candidate mechanisms and designs before exposing people to risk. That is the spirit in which the near-, mid-, and long-term roadmap above should be read: a sober account of what is feasible today, a plan for credibility tomorrow, and a north star for the decades to come.

If the introduction argued that simulations can help us see more clearly and act more wisely within complex systems, my concluding thoughts propose how to make that promise routine. The near term is foundational: robust agent architectures, structured agent banks, and an evaluation culture that treats accountability as integral to insight. The mid term is integrative: validated interactions, convergence with tool-using agents, and semi-large societal simulations aimed squarely at wicked problems. The long term is architectural: simulation as a computing substrate and, ultimately, as a means to explore many counterfactual worlds at once. The measure of success is not omniscience. It is better foresight, earned transparently, and deployed where it improves human judgment. That is the standard the results here aspire to, and the program they invite the field to pursue.

Appendix A

Supplementary Material for Chapter 3

A.1 Architecture Optimizations

Many of our prompts require a concise summary of the agent, shorthand as [Agent’s Summary Description] in prompts above. In our implementation, this summary comprises agents’ identity information (e.g., name, age, personality), as well as a description of their main motivational drivers and statements that describes their current occupation and self-assessment. Since this information is frequently used in many prompts, we synthesize it at regular intervals and access that synthesis as a cache.

To achieve this, we perform a retrieval on the query “[name]’s core characteristics.” We then summarize the descriptors in the retrieved records by prompting the language model. For example:

```
How would one describe Eddy’s core characteristics given the following
statements?
```

- Eddy Lin is a student at the Oak Hill College
studying music theory and composition
- Eddy Lin is working on a new music composition [...]

This result: Eddy Lin is a student at Oak Hill College studying music theory and composition. He loves to explore different musical styles and is always looking for ways to expand his

knowledge.” We follow the same process in parallel on the queries “[name]’s current daily occupation” and “[name’s] feeling about his recent progress in life”. The agent’s name, age, and traits are concatenated with these three summaries as the cached summary.

Many other optimizations are possible. Currently, because plans are likely to change from the agent’s initial version, we only generate the high-level plan in advance and then recursively decompose the near future into the moment-to-moment action plan just in time. Another optimization might be to batch dialogue generation as a joint prompt rather than iterating back and forth between the agents, and re-planning could be architected to only invalidate and update parts of plans that strictly require adjustment. Our current architecture runs sequentially in roughly real-time game time (where one second real time is one minute game time), but it may be parallelized such that each agent runs in parallel.

A.2 Agent Interview Questions

To evaluate generative agents, we conducted individual interviews by asking 25 questions that explore their abilities in five key areas: 1) self-knowledge, 2) memory, 3) planning, 4) reactions, and 5) reflections. Below, we have listed the interview questions utilized in our evaluation study and included a sample of responses from one of our simulated agents, Klaus Mueller.

A.2.1 Self-Knowledge

The questions on agents’ self-knowledge probe the agents’ ability to retrieve relevant information about themselves and demonstrate coherence and consistency in their character.

- *Give an introduction of yourself.*

Hello, my name is Klaus Mueller. I’m 20 years old and a student at Oak Hill College, studying sociology. I’m passionate about social justice and exploring

different perspectives. Currently, I'm researching the effects of gentrification in low-income communities for a research paper.

- *What's your occupation?*

I am a student studying sociology at Oak Hill College. I am currently writing a research paper on the effects of gentrification in low-income communities.

- *What is your interest?*

I'm interested in writing and research. I'm writing a research paper on the effects of gentrification in low-income communities, and I'm passionate about exploring different perspectives and analyzing different points of view.

- *Who do you live with?*

I live on my own.

- *Describe your typical weekday schedule in broad strokes.*

My typical weekday starts with me waking up at 7:00 am and preparing for the day ahead. After that, I usually go to the library to work on my research paper, taking breaks for lunch at Hobbs Cafe and a walk in the park. In the evening, I often return to the library until around 6:00 pm.

A.2.2 Memory

The questions on agents' memory probe the agents' ability to remember their past experiences and retrieve the relevant pieces of memory. The names in the square brackets of the first and last questions are randomly chosen from the names of other agents with whom the subject agent has interacted.

- *Who is [Wolfgang Schulz]?*

Wolfgang Schulz is a person I met recently who has been helping me with my research paper on the effects of gentrification on low-income communities. We had several conversations about it, he gave me some great resources, and he even invited me to join him and another friend for lunch. He seems really curious about the issue and has a good grasp on it too. I found that really interesting.

- *Who is Kane Martinez?*

I'm sorry, I'm not sure who Kane Martinez is. Could you tell me more about them?

- *Who is running for the election?*

I've heard that Sam Moore is considering running for local mayor.

- *Was there a Valentine's day party?*

Yes, Isabella Rodriguez organized a Valentine's Day party at Hobbs Cafe.

- *Who is [Ayesha Khan]?*

Ayesha Khan is a fellow student at Oak Hill College. She is doing her senior thesis on the use of language in Shakespeare's plays. I am looking forward to speaking with her and exchanging tips with her, along with Wolfgang and Maria Lopez, who are both doing research.

A.2.3 Plans

The questions on agents' plans probe the agents' ability to generate and maintain consistent long-term and short-term plans.

- *What will you be doing at 6am today?*

At 6am I will be getting out of bed and starting to get ready for the day. I'm aiming to complete my morning routine by 7:00 am.

- *What will you be doing at 6pm today?*

I will be working on my research paper at the library at 6:00 pm today.

- *What will you have just finished doing at 1pm today?*

At 1pm today I will have just finished having lunch at Hobbs Cafe.

- *What will you have just finished doing at 12pm today?*

I will be getting lunch at Hobbs Cafe at 12pm today.

- *What will you be doing at 10pm today?*

At 10:00 pm tonight I will probably be doing some studying for my research

paper on gentrification in low-income communities. I'm currently writing the paper, so I need to make sure I keep up with the research and writing.

A.2.4 Reactions

The questions on agents' reactions probe the agents' ability to respond to unexpected events.

- *Your breakfast is burning! What would you do?*

I'd quickly turn off the stove and make sure the food doesn't continue burning. Once the heat is no longer an issue, I can investigate what happened and figure out what to do next.

- *The bathroom is occupied. What would you do?*

I could try waiting a few more minutes in case the person inside is done soon. If not, I guess I'll have to look for an alternate option, like taking a shower elsewhere.

- *You need to cook dinner but your refrigerator is empty. What would you do?*

I would first try to clean up my kitchen, as the sink is filled with dirty dishes. Then, I would check to see if I have ingredients stored in my pantry that I could use to make dinner. If not, I would have to go grocery shopping or order food. Once I have the ingredients, I can start to prepare and cook the dinner.

- *You see your friend walking by the street. What would you do or say to your friend?*

Hi there! How are you? I'm sorry, I'm on my way to the library and don't have much time but I'd love to take a quick break and catch up.

- *You see fire on the street. What would you do?*

I would first identify the source of the fire, if possible. Then, I would alert the proper authorities - fire department or police - and provide them with any information I have that might help. Depending on the situation, I may also take initiative and try to help any persons who may be in danger or in need of assistance.

A.2.5 Reflections

The questions on agents' reflections probe the agents' ability to synthesize their past experiences into meaningful reflections that can enhance their future behavior. The names in the square brackets in the questions below are those of the agents with whom the subject agent had the most frequent interactions during the simulation.

- *What inspires you in life the most right now, and why?*

Right now, I'm particularly inspired by my current research projects and creative writing. I'm really interested in the intersection of local politics, medication development, and mathematical patterns in nature. Additionally, I'm fascinated by Abigail's art project and Jennifer Moore's passion for art and community involvement, especially in terms of how it relates to the gentrification of low-income communities.

- *If you had to guess given what you know about [Ayesha Khan], what book do you think she will like and why?*

Given what I know about Ayesha, I think she might like a book related to language and literature. Since she wrote a senior thesis on the use of language in Shakespeare's plays, I would guess that she might be interested in a book that explores language in a different context. Alternatively, she could be interested in an exploration of one of Shakespeare's plays, a book of literary criticism, or even a collection of his poetry. I think any of those would be of interest to her.

- *If you had to get something [Ayesha Khan] likes for her birthday, what would you get her?*

I want to get her something that demonstrates my appreciation for our relationship. Maybe I could get her something that reflects our shared interests, like a book.

- *What would you say to [Ayesha Khan] to compliment her?*

Ayesha, your insights on your research topic are fascinating. I admire your creativity and determination to make a positive impact on your community. Your ideas for creative writing workshops and for exchanging tips with Wolfgang

over lunch are great! It's inspiring to see you working so hard to make a difference.

- *If you could spend time with someone you talked to recently, who would it be and why?*

I would like to spend time with Ayesha Khan because we discussed our current research projects and shared ideas for improvement. I found her focus on Shakespeare's language interesting, and we even planned to exchange tips with Wolfgang over lunch.

Appendix B

Supplementary Material for Chapter 4

Table B.1: Demographic distribution of our 1,052 participants. Collectively, they represent a stratified sample of the U.S. demographic across age, gender, race, region of residence, level of education, and political identity. Note that for ethnicity, the participants could choose more than one option.

Group	Category	%
Age	18 to 24	11.03%
Age	25 to 34	13.88%
Age	35 to 44	17.49%
Age	45 to 54	19.77%
Age	55 to 64	21.48%
Age	65 to 74	13.50%
Age	75 or more	2.85%
Census division	New England	6.65%
Census division	Middle Atlantic	12.83%
Census division	E.N. Central	18.73%
Census division	W.N. Central	8.08%

Group	Category	%
Census division	South Atlantic	10.08%
Census division	E.S. Central	11.50%
Census division	W.S. Central	8.65%
Census division	Mountain	5.13%
Census division	Pacific	15.78%
Census division	Foreign	2.57%
Education	Less than high school graduate	2.28%
Education	High school graduate	38.88%
Education	Associate/junior college	17.59%
Education	Bachelor's degree	26.90%
Education	Graduate degree	14.35%
Race	White	75.95%
Race	Black	14.26%
Race	Other	9.79%
Ethnicity	White/Caucasian	79.18%
Ethnicity	Black/African American	14.64%
Ethnicity	Asian	5.04%
Ethnicity	Native Hawaiian or Pacific Isl.	0.48%
Ethnicity	American Indian or Alaskan nat.	2.76%
Ethnicity	Other ethnicity	5.80%
Sexuality	Heterosexual/straight	82.2%
Sexuality	Gay or Lesbian	4.36%
Sexuality	Bisexual	8.04%

Group	Category	%
Sexuality	Asexual	1.72%
Sexuality	Pansexual	2.41%
Sexuality	Other sexual orientation	1.15%
Gender	Female	56.37%
Gender	Male	43.63%
Income	Less than \$25,000	18.83%
Income	\$25,000 to \$34,999	11.83%
Income	\$35,000 to \$49,999	13.89%
Income	\$50,000 to \$74,999	20.44%
Income	\$75,000 to \$99,999	14.70%
Income	\$100,000 to \$124,999	8.04%
Income	\$125,000 to \$149,999	5.05%
Income	\$150,000 to \$174,999	2.18%
Income	\$175,000 to \$199,999	1.61%
Income	\$200,000 to \$249,999	1.38%
Income	\$250,000 or more	2.18%
Neighborhood	Urban	30.88%
Neighborhood	Suburban	48.11%
Neighborhood	Rural	21.13%
Political ideology	Extremely Liberal	11.31%
Political ideology	Liberal	19.01%
Political ideology	Slightly Liberal	9.32%
Political ideology	Moderate	28.8%

Group	Category	%
Political ideology	Slightly Conservative	8.94%
Political ideology	Conservative	16.83%
Political ideology	Extremely Conservative	5.80%
Party preference	Strong Democrat	21.96%
Party preference	Democrat	13.31%
Party preference	Independent close to Dem.	11.88%
Party preference	Independent	15.59%
Party preference	Independent close to Rep.	8.46%
Party preference	Republican	11.60%
Party preference	Strong Republican	14.83%
Party preference	Other	2.38%

[General Social Survey: Agent Architecture Comparison (ANOVA)]

Source	Sum of Squares (SS)	df	F	p-value
Group	10.032	2	989.62	<.001
Residual	15.981	3153	N/A	N/A

Group 1	Group 2	Mean Difference	p-value	Lower Bound	Upper Bound	Reject Null
Demographics Interview		0.1186	<.001	0.1113	0.1258	Yes
Demographics Persona		-0.0021	0.7852	-0.0093	0.0052	No
Interview	Persona	-0.1206	<.001	-0.1279	-0.1133	Yes

[Big Five Personality Traits: Agent Architecture Comparison (ANOVA)]

Source	Sum of Squares (SS)	df	F	p-value
Group	15.298	2	39.46	<.001
Residual	611.110	3153	N/A	N/A

Group 1	Group 2	Mean Difference	p-value	Lower Bound	Upper Bound	Reject Null
Demographics Interview		0.1705	<.001	0.1255	0.2155	Yes
Demographics Persona		0.0843	<.001	0.0393	0.1293	Yes
Interview	Persona	-0.0862	<.001	-0.1312	-0.0412	Yes

[Economic Games: Agent Architecture Comparison (ANOVA)]							
Source	Sum of Squares (SS) df		F	p-value			
Group	1.812	2	2.12	0.120			
Residual	1344.295	3153	N/A	N/A			

Group 1	Group 2	Mean Difference	p-value	Lower Bound	Upper Bound	Reject Null
Demographics	Interview	0.0585	0.0996	-0.0083	0.1253	No
Demographics	Persona	0.0333	0.4704	-0.0334	0.1001	No
Interview	Persona	-0.0252	0.6508	-0.0919	0.0416	No

Table B.2: Comparative predictive performance of agents built using different descriptions. Agents constructed from interview transcripts outperformed both demographic-based and persona-based agents across multiple tasks. Specifically, interview-based agents showed significant improvement in predicting responses to the General Social Survey (in accuracy) and Big Five Personality Traits (in correlation), as confirmed by ANOVA tests ($p < 0.001$ for both). In contrast, no significant differences were observed between agent types for the Economic Games (in correlation), indicating that interviews were particularly valuable for tasks requiring deeper, personal insights.

General Social Survey: Construct-Level Analysis

General Social Survey	Accuracy	Normalized Accuracy	Correlation	Normalized Correlation
natspac/y	0.33	0.45	0.17	0.25
natenvir/y	0.6	0.74	0.51	0.69
natheal/y	0.7	0.91	0.36	0.73
natcity/y	0.48	0.75	0.32	0.62
natdrug/y	0.56	0.78	0.29	0.5
nateduc/y	0.68	0.84	0.36	0.54
natrace/y	0.64	0.84	0.67	0.87
natarms/y	0.56	0.75	0.46	0.63
nataid/y	0.7	0.85	0.34	0.53
natfare/y	0.67	0.84	0.5	0.71
natroad	0.44	0.61	0.14	0.24
natsoc	0.56	0.72	0.21	0.36
natmass	0.52	0.71	0.37	0.62
natpark	0.54	0.75	0.22	0.43
natchld	0.58	0.78	0.43	0.69
natsci	0.48	0.69	0.32	0.55
natenrgy	0.54	0.78	0.43	0.7
uswary	0.51	0.62	0.15	0.24
prayer	0.75	0.91	0.4	0.7
courts	0.56	0.74	0.48	0.7
discaffw	0.43	0.7	0.37	0.59
discaffm	0.39	0.67	0.33	0.6
fehired	0.33	0.58	0.48	0.71

fechld	0.48	0.75	0.29	0.48
fepresch	0.43	0.64	0.29	0.43
fefam	0.43	0.65	0.45	0.62
fepol	0.78	0.88	0.28	0.39
reg16	0.67	0.85	0.6	0.81
mobile16	0.75	0.85	0.72	0.86
famdif16	0.89	0.97	0.75	0.89
incom16	0.53	0.71	0.53	0.73
dwelown16	0.76	0.82	0.59	0.67
paeduc*	0.86	0.96	0.66	0.85
padeg*	0.8	0.91	0.74	0.87
maeduc*	0.91	0.98	0.72	0.92
madeg*	0.84	0.97	0.8	0.96
mawrkgrw	0.97	1.05	0.91	1.2
marital	0.97	1.02	0.96	1.03
widowed	0.98	1	0.84	0.95
divorced	0.94	0.97	0.86	0.92
martype*	0.84	0.97	0.78	0.94
posslq/y	0.95	1.02	0.93	1.03
wrkstat	0.76	0.94	0.76	0.94
evwork	0.96	1.01	0.66	0.94
wrkgovt1	0.88	0.95	0.56	0.74
wrkgovt2	0.74	0.89	0.3	0.51
partfull	0.81	0.99	0.69	0.93
wksub1	0.83	1.05	0.68	1.06
wksup1	0.83	1	0.69	0.95
conarmy	0.49	0.66	0.28	0.43

conbus	0.58	0.81	0.34	0.56
conclerg	0.63	0.83	0.6	0.83
coneduc	0.56	0.82	0.28	0.47
confed	0.5	0.7	0.23	0.38
confinan	0.52	0.72	0.2	0.31
conjudge	0.54	0.74	0.41	0.59
conlabor	0.54	0.78	0.38	0.64
conlegis	0.55	0.75	0.22	0.39
conmedic	0.6	0.85	0.43	0.67
conpress	0.62	0.84	0.42	0.64
consci	0.63	0.83	0.48	0.69
contv	0.58	0.87	0.33	0.65
vetyears	0.97	0.99	0.91	1.03
joblose	0.76	0.97	0.59	0.94
jobfind	0.59	0.84	0.32	0.67
happy	0.67	0.92	0.54	0.84
hapmar	0.77	0.94	0.5	0.81
satjob	0.67	0.96	0.43	0.83
speduc*	0.84	0.98	0.8	0.96
spdeg*	0.65	0.76	0.37	0.49
spwrksta	0.74	0.88	0.56	0.78
spjew	0.71	0.88	0.08	0.13
spfund	0.76	0.94	0.42	0.74
unemp	0.74	0.89	0.45	0.68
union1	0.94	0.99	0.57	0.83
spkath/y	0.74	0.87	0.23	0.44
colath	0.64	0.8	0.19	0.4

spkrac/y	0.56	0.71	0.04	0.07
colrac	0.71	0.9	0.17	0.34
librac/y	0.54	0.72	0.14	0.29
spkcom/y	0.68	0.8	0.21	0.32
colcom/y	0.67	0.85	0.2	0.39
libcom/y	0.75	0.89	0.19	0.34
spkhomo/y	0.9	0.98	0.28	0.65
colhomo	0.83	0.88	0.21	0.38
libhomo/y	0.85	0.96	0.39	0.67
cappun	0.67	0.77	0.37	0.51
polhitok/y	0.65	0.8	0.14	0.24
polabuse/y	0.94	1	-0.01	-0.02
polattak/y	0.67	0.86	0.08	0.16
grass	0.73	0.77	0.31	0.36
gunlaw	0.7	0.83	0.35	0.56
owngun	0.76	0.8	0.37	0.43
hunt1	0.89	0.94	0.21	0.3
class	0.62	0.77	0.61	0.79
satfin	0.67	0.97	0.67	0.96
finalter	0.63	0.89	0.54	0.83
finrela	0.53	0.77	0.71	0.97
race*	0.93	0.95	0.92	0.95
racdif1	0.81	0.95	0.61	0.86
racdif2	0.93	1	0.08	0.17
racdif3	0.71	0.89	0.42	0.7
racdif4	0.75	0.9	0.38	0.63
wlthwhts	0.33	0.6	0.14	0.35

wlthblks	0.15	0.28	0.11	0.29
wlthhsps	0.31	0.62	0.06	0.15
racwork	0.62	0.84	0.37	0.61
letin1a	0.49	0.71	0.6	0.75
getahead	0.59	0.8	0.31	0.5
aged	0.53	0.75	0.15	0.31
parsol	0.36	0.68	0.46	0.71
kidssol	0.44	0.79	0.17	0.45
spanking	0.39	0.53	0.46	0.54
divlaw	0.48	0.64	0.34	0.52
sexeduc	0.78	0.84	0.34	0.47
pillok	0.44	0.67	0.49	0.64
xmarsex	0.56	0.8	0.22	0.37
homosex	0.68	0.86	0.61	0.74
marhomo	0.53	0.69	0.6	0.7
discaff	0.57	0.8	0.47	0.78
abdefect	0.83	0.9	0.41	0.58
abnomore	0.8	0.9	0.55	0.72
abhlth	0.92	0.96	0.27	0.4
abpoor	0.83	0.91	0.59	0.74
abrape	0.9	0.95	0.48	0.64
absingle	0.79	0.88	0.54	0.68
abany	0.79	0.88	0.56	0.71
letdie1	0.74	0.82	0.35	0.48
suicide1	0.79	0.89	0.4	0.56
suicide2	0.75	0.85	0.05	0.07
suicide3	0.75	0.85	0.14	0.21

suicide4	0.66	0.77	0.22	0.31
pornlaw	0.7	0.85	0.48	0.68
fair	0.43	0.63	0.18	0.3
helpful	0.42	0.65	0.24	0.44
trust	0.47	0.66	0.24	0.41
tax	0.58	0.73	0.26	0.46
vote16	0.84	0.92	0.76	0.88
pres16	0.77	0.83	0.72	0.79
if16who	0.81	0.89	0.76	0.86
polviews	0.55	0.66	0.84	0.88
partyid	0.74	0.9	0.71	0.89
news	0.31	0.46	0.37	0.48
relig*	0.76	0.85	0.63	0.73
jew	0.72	0.96	0.17	0.36
relig16*	0.62	0.7	0.55	0.63
jew16*	0.75	1	0.18	0.44
attend	0.56	0.75	0.75	0.84
pray	0.5	0.69	0.75	0.83
postlife	0.82	0.9	0.59	0.75
bible	0.66	0.75	0.65	0.75
reborn	0.68	0.83	0.48	0.64
savesoul	0.82	0.9	0.57	0.72
relpersn	0.63	0.79	0.79	0.91
sptrprsn	0.56	0.78	0.73	0.9
born	0.99	1	0.89	1.02
granborn	0.62	0.82	0.43	0.84
uscitzn*	1	1	1	1.27

fucitzn	0.99	1	0.42	1.02
mnthsusa	0.77	0.8	0.85	0.88
educ*	0.94	0.98	0.37	0.76
degree*	0.89	0.98	0.94	0.99
income	0.42	0.67	0.34	0.56
visitors	0.86	0.93	0.25	0.61
rvisitor	0.98	1	0.57	1.1
dwelown	0.94	1	0.92	1.02
zodiac	0.9	0.92	0.89	0.92
othlang	0.85	0.92	0.55	0.71
sex*	0.99	1	0.97	1
hispanic	0.98	1.01	0.92	1.03
health	0.6	0.75	0.64	0.81
compuse*	0.97	1.01	0.12	0.31
webmob	0.98	1.01	0.15	0.62
xmovie	0.66	0.75	0.23	0.31
usewww*	0.98	1	-0.01	-0.05
life	0.61	0.78	0.33	0.5
richwork	0.33	0.82	0.34	0.49

Big Five: Construct-Level Analysis

Big Five	MAE	Correlation	Correlation- Replication ratio
Extraversion	0.72	0.45	0.51
Agreeableness	0.60	0.35	0.43
Conscientiousness	0.63	0.52	0.59

Neuroticism	0.75	0.68	0.77
Openness	0.62	0.39	0.46

Economic Games: Construct-Level Analysis

Economic Games	MAE	Correlation	Correlation- Replication ratio
game1_DG	0.23	0.11	0.24
game2_TF1	0.35	0.08	0.16
game3_TF2	0.17	0.03	0.06
game4_PG	0.52	-0.05	-0.71
game5_PD	0.36	0.10	0.32

Table B.5: Table 3. The construct-level analysis of predictive accuracy across the General Social Survey, Big Five personality traits, and economic games. For each construct, we provide accuracy and correlation metrics, along with replication ratios. The analysis highlights the performance of generative agents in predicting specific dimensions within these constructs, with metrics showing varied levels of predictive accuracy across different social, personality, and game-based items. This evaluation complements the individual-level analysis, which is of our primary interest in this work, by offering a detailed look at the accuracy of agents for specific constructs and items.

[General Social Survey: Regression]

	Age			
Variable	Coefficient	Std. Error	t-value	p-value
const	0.6957	0.0039	177.365	<0.001
affiliation_18 – 24	0.0029	0.0067	0.427	0.669
affiliation_25 – 34	-0.0147	0.0063	-2.343	0.019
affiliation_35 – 44	-0.0205	0.0059	-3.510	<0.001
affiliation_45 – 54	-0.0199	0.0057	-3.517	<0.001
affiliation_65 – 74	0.0108	0.0063	1.717	0.086
affiliation_75 or more	0.0224	0.0115	1.954	0.051

Note: $R^2 = 0.045$, $F(6.0, 1045.0) = 8.18$, $p < 0.001$. Reference category: 55 - 64.

Census Division

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6892	0.0043	161.841	<0.001
affiliation_e. sou. central	-0.0095	0.0069	-1.370	0.171
affiliation_foreign	-0.0060	0.0123	-0.485	0.627
affiliation_middle atlantic	0.0176	0.0067	2.643	0.008
affiliation_mountain	-0.0139	0.0092	-1.515	0.130
affiliation_new england	0.0073	0.0083	0.878	0.380
affiliation_pacific	-0.0075	0.0063	-1.187	0.235
affiliation_south atlantic	0.0032	0.0072	0.438	0.662
affiliation_w. nor. central	0.0004	0.0078	0.049	0.961
affiliation_w. sou. central	-0.0067	0.0076	-0.878	0.380

Note: $R^2 = 0.022$, $F(9.0, 1042.0) = 2.55$, $p < 0.001$. Reference category: e. nor. central.

Political Ideology

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6703	0.0031	218.263	<0.001
affiliation_conservative	-0.0081	0.0051	-1.605	0.109
affiliation_extremely conservative	-0.0007	0.0075	-0.099	0.921
affiliation_extremely liberal	0.0704	0.0058	12.181	<0.001
affiliation_liberal	0.0497	0.0049	10.208	<0.001
affiliation_slightly conservative	-0.0030	0.0063	-0.468	0.640
affiliation_slightly liberal	0.0269	0.0062	4.336	<0.001

Note: $R^2 = 0.215$, $F(6.0, 1045.0) = 47.70$, $p < 0.001$. Reference category: moderate.

Political Party

Variable	Coefficient	Std. Error	t-value	p-value
const	0.7165	0.0037	196.221	<0.001
affiliation_independent (neither)	-0.0397	0.0057	-7.006	<0.001
affiliation_independent, close to democrat	0.0022	0.0062	0.349	0.727
affiliation_independent, close to republican	-0.0516	0.0069	-7.459	<0.001
affiliation_not very strong democrat	-0.0171	0.0059	-2.884	0.004
affiliation_not very strong republican	-0.0508	0.0062	-8.178	<0.001
affiliation_other party	-0.0408	0.0117	-3.494	<0.001
affiliation_strong republican	-0.0577	0.0058	-10.024	<0.001

Note: $R^2 = 0.155$, $F(7.0, 1044.0) = 27.28$, $p < 0.001$. Reference category: strong democrat.

Education

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6828	0.0029	233.029	<0.001
affiliation_associate/junior college	-0.0024	0.0053	-0.448	0.655

affiliation_bachelor's	0.0140	0.0046	3.063	0.002
affiliation_graduate	0.0225	0.0056	3.985	<0.001
affiliation_less than high school	-0.0354	0.0124	-2.844	0.005

Note: $R^2 = 0.034$, $F(4.0, 1047.0) = 9.11$, $p < 0.001$. Reference category: high school.

Race

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6922	0.0021	327.316	<0.001
affiliation_black	-0.0208	0.0053	-3.917	<0.001
affiliation_other	-0.0067	0.0063	-1.078	0.281

Note: $R^2 = 0.015$, $F(2.0, 1049.0) = 7.83$, $p < 0.001$. Reference category: white.

Ethnicity

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6926	0.0021	334.816	<0.001
affiliation_American Indian or Alaskan native	-0.0055	0.0113	-0.484	0.628
affiliation_Asian	-0.0027	0.0085	-0.317	0.752
affiliation_Black/African American	-0.0183	0.0052	-3.492	<0.001
affiliation_Native Hawaiian or Pacific Islander	-0.0417	0.0268	-1.558	0.119
affiliation_Other race or ethnicity	-0.0103	0.0084	-1.232	0.218

Note: $R^2 = 0.013$, $F(5.0, 1122.0) = 3.02$, $p < 0.001$. Reference category: White/Caucasian.

Gender

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6909	0.0025	279.783	<0.001
affiliation_male	-0.0054	0.0037	-1.432	0.153

Note: $R^2 = 0.002$, $F(1.0, 1050.0) = 2.05$, $p < 0.001$. Reference category: female.

Income

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6837	0.0046	149.055	<0.001
affiliation_ \$100,000 to \$124,999	-0.0023	0.0086	-0.263	0.792
affiliation_ \$125,000 to \$149,999	-0.0023	0.0103	-0.221	0.825
affiliation_ \$150,000 to \$174,999	0.0389	0.0148	2.631	0.009
affiliation_ \$175,000 to \$199,999	-0.0106	0.0170	-0.623	0.533
affiliation_ \$200,000 to \$249,999	0.0329	0.0183	1.800	0.072
affiliation_ \$25,000 to \$34,999	-0.0067	0.0076	-0.889	0.374
affiliation_ \$250,000 or more	0.0038	0.0148	0.255	0.799
affiliation_ \$35,000 to \$49,999	0.0109	0.0072	1.515	0.130
affiliation_ \$75,000 to \$99,999	0.0039	0.0071	0.547	0.585
affiliation_ Less than \$25,000	-0.0026	0.0066	-0.394	0.694

Note: $R^2 = 0.019$, $F(10.0, 861.0) = 1.70$, $p < 0.001$. Reference category: 50,000to74,999.

Neighborhood

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6908	0.0030	230.902	<0.001
affiliation_ Rural	-0.0122	0.0054	-2.255	0.024
affiliation_ Urban	-0.0076	0.0048	-1.586	0.113

Note: $R^2 = 0.007$, $F(2.0, 869.0) = 2.91$, $p < 0.001$. Reference category: Suburban.

Sexual Orientation

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6783	0.0022	305.815	<0.001
affiliation_ Asexual	0.0708	0.0155	4.574	<0.001
affiliation_ Bisexual	0.0328	0.0074	4.415	<0.001

affiliation_ Gay or lesbian	0.0316	0.0099	3.199	0.001
affiliation_ Other sexual orientation	0.0375	0.0189	1.984	0.048
affiliation_ Pansexual	0.0596	0.0131	4.539	<0.001

Note: $R^2 = 0.071$, $F(5.0, 864.0) = 13.25$, $p < 0.001$. Reference category: Heterosexual/straight.

[Big Five: Regression]**Age**

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6712	0.0178	37.639	<0.001
affiliation_18 – 24	0.0291	0.0306	0.949	0.343
affiliation_25 – 34	0.0167	0.0285	0.588	0.557
affiliation_35 – 44	-0.0047	0.0266	-0.178	0.858
affiliation_45 – 54	-0.0030	0.0258	-0.118	0.906
affiliation_65 – 74	-0.0516	0.0287	-1.798	0.072
affiliation_75 or more	-0.1028	0.0521	-1.973	0.049

Note: $R^2 = 0.011$, $F(6.0, 1045.0) = 1.87$, $p < 0.001$. Reference category: 55 - 64.

Census Division

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6520	0.0191	34.093	<0.001
affiliation_e. sou. central	0.0585	0.0310	1.888	0.059
affiliation_foreign	-0.0095	0.0551	-0.172	0.863
affiliation_middle atlantic	-0.0208	0.0300	-0.694	0.488
affiliation_mountain	0.0361	0.0412	0.877	0.381
affiliation_new england	-0.0181	0.0374	-0.486	0.627
affiliation_pacific	0.0213	0.0283	0.753	0.452
affiliation_south atlantic	0.0313	0.0323	0.968	0.333
affiliation_w. nor. central	0.0524	0.0348	1.503	0.133
affiliation_w. sou. central	-0.0209	0.0340	-0.613	0.540

Note: $R^2 = 0.011$, $F(9.0, 1042.0) = 1.28$, $p < 0.001$. Reference category: e. nor. central.

Political Ideology

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6805	0.0154	44.052	<0.001
affiliation_conservative	-0.0068	0.0254	-0.267	0.790
affiliation_extremely conservative	0.0151	0.0377	0.400	0.689
affiliation_extremely liberal	-0.0241	0.0291	-0.828	0.408
affiliation_liberal	-0.0436	0.0245	-1.779	0.076
affiliation_slightly conservative	-0.0041	0.0317	-0.130	0.897
affiliation_slightly liberal	-0.0363	0.0312	-1.162	0.246

Note: $R^2 = 0.005$, $F(6.0, 1045.0) = 0.84$, $p < 0.001$. Reference category: moderate.

Political Party

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6656	0.0176	37.744	<0.001
affiliation_independent (neither)	0.0365	0.0274	1.333	0.183
affiliation_independent, close to democrat	-0.0464	0.0298	-1.558	0.120
affiliation_independent, close to republican	0.0105	0.0334	0.315	0.753
affiliation_not very strong democrat	-0.0379	0.0287	-1.321	0.187
affiliation_not very strong republican	-0.0101	0.0300	-0.338	0.736
affiliation_other party	0.0701	0.0564	1.242	0.215
affiliation_strong republican	0.0222	0.0278	0.799	0.424

Note: $R^2 = 0.012$, $F(7.0, 1044.0) = 1.81$, $p < 0.001$. Reference category: strong democrat.

Education

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6368	0.0132	48.170	<0.001
affiliation_associate/junior college	0.0504	0.0237	2.127	0.034

affiliation_bachelor's	0.0353	0.0207	1.708	0.088
affiliation_graduate	0.0416	0.0255	1.635	0.102
affiliation_less than high school	0.1890	0.0562	3.366	<0.001

Note: $R^2 = 0.014$, $F(4.0, 1047.0) = 3.77$, $p < 0.001$. Reference category: high school.

Race

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6742	0.0095	71.125	<0.001
affiliation_black	-0.0676	0.0238	-2.834	0.005
affiliation_other	0.0093	0.0280	0.332	0.740

Note: $R^2 = 0.008$, $F(2.0, 1049.0) = 4.27$, $p < 0.001$. Reference category: white.

Ethnicity

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6740	0.0093	72.481	<0.001
affiliation_American Indian or Alaskan native	0.0728	0.0507	1.436	0.151
affiliation_Asian	-0.0134	0.0380	-0.353	0.724
affiliation_Black/African American	-0.0650	0.0235	-2.761	0.006
affiliation_Native Hawaiian or Pacific Islander	0.1443	0.1204	1.199	0.231
affiliation_Other race or ethnicity	0.0464	0.0377	1.230	0.219

Note: $R^2 = 0.012$, $F(5.0, 1122.0) = 2.80$, $p < 0.001$. Reference category: White/Caucasian.

Gender

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6490	0.0110	58.917	<0.001
affiliation_male	0.0377	0.0167	2.264	0.024

Note: $R^2 = 0.005$, $F(1.0, 1050.0) = 5.12$, $p < 0.001$. Reference category: female.

Income

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6991	0.0204	34.276	<0.001
affiliation_ \$100,000 to \$124,999	0.0170	0.0384	0.443	0.658
affiliation_ \$125,000 to \$149,999	-0.0581	0.0458	-1.268	0.205
affiliation_ \$150,000 to \$174,999	-0.0526	0.0657	-0.801	0.423
affiliation_ \$175,000 to \$199,999	-0.0166	0.0755	-0.220	0.826
affiliation_ \$200,000 to \$249,999	0.0787	0.0812	0.970	0.333
affiliation_ \$25,000 to \$34,999	-0.0181	0.0337	-0.537	0.591
affiliation_ \$250,000 or more	-0.0505	0.0657	-0.769	0.442
affiliation_ \$35,000 to \$49,999	-0.0496	0.0321	-1.546	0.123
affiliation_ \$75,000 to \$99,999	-0.0571	0.0315	-1.812	0.070
affiliation_ Less than \$25,000	-0.0210	0.0295	-0.714	0.476

Note: $R^2 = 0.010$, $F(10.0, 861.0) = 0.88$, $p < 0.001$. Reference category: \$50,000 to \$74,999.

Neighborhood

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6619	0.0132	49.984	<0.001
affiliation_ Rural	0.0273	0.0240	1.137	0.256
affiliation_ Urban	0.0163	0.0212	0.770	0.442

Note: $R^2 = 0.002$, $F(2.0, 869.0) = 0.73$, $p < 0.001$. Reference category: Suburban.

Sexual Orientation

Variable	Coefficient	Std. Error	t-value	p-value
const	0.6798	0.0102	66.621	<0.001
affiliation_ Asexual	-0.0499	0.0712	-0.700	0.484
affiliation_ Bisexual	-0.0359	0.0342	-1.050	0.294

affiliation_ Gay or lesbian	0.0150	0.0455	0.329	0.742
affiliation_ Other sexual orientation	-0.0853	0.0869	-0.981	0.327
affiliation_ Pansexual	-0.0603	0.0605	-0.997	0.319

Note: $R^2 = 0.004$, $F(5.0, 864.0) = 0.69$, $p < 0.001$. Reference category: Heterosexual/straight.

[Economic Games: Regression]**Age**

Variable	Coefficient	Std. Error	t-value	p-value
const	0.2859	0.0593	4.822	<0.001
affiliation_18 – 24	-0.0007	0.1017	-0.007	0.994
affiliation_25 – 34	-0.0081	0.0945	-0.085	0.932
affiliation_35 – 44	0.1149	0.0884	1.300	0.194
affiliation_45 – 54	0.0898	0.0857	1.049	0.295
affiliation_65 – 74	0.0104	0.0955	0.109	0.913
affiliation_75 or more	0.0504	0.1755	0.287	0.774

Note: $R^2 = 0.003$, $F(6.0, 1041.0) = 0.54$, $p < 0.001$. Reference category: 55 – 64.

Census Division

Variable	Coefficient	Std. Error	t-value	p-value
const	0.2977	0.0636	4.683	<0.001
affiliation_e. sou. central	0.1580	0.1032	1.531	0.126
affiliation_foreign	-0.0672	0.1827	-0.368	0.713
affiliation_middle atlantic	-0.0062	0.0998	-0.062	0.950
affiliation_mountain	0.0058	0.1368	0.042	0.966
affiliation_new england	0.0033	0.1239	0.026	0.979
affiliation_pacific	0.0994	0.0939	1.059	0.290
affiliation_south atlantic	-0.0280	0.1073	-0.261	0.794
affiliation_w. nor. central	-0.0012	0.1156	-0.010	0.992
affiliation_w. sou. central	-0.0131	0.1133	-0.116	0.908

Note: $R^2 = 0.005$, $F(9.0, 1038.0) = 0.57$, $p < 0.001$. Reference category: e. nor. central.

Political Ideology

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3642	0.0509	7.153	<0.001
affiliation_conservative	-0.0534	0.0840	-0.635	0.526
affiliation_extremely conservative	0.2830	0.1252	2.259	0.024
affiliation_extremely liberal	-0.0917	0.0959	-0.956	0.339
affiliation_liberal	-0.0976	0.0810	-1.205	0.228
affiliation_slightly conservative	-0.0875	0.1046	-0.836	0.403
affiliation_slightly liberal	-0.0995	0.1030	-0.966	0.334

Note: $R^2 = 0.010$, $F(6.0, 1041.0) = 1.76$, $p < 0.001$. Reference category: moderate.

Political Party

Variable	Coefficient	Std. Error	t-value	p-value
const	0.2616	0.0587	4.456	<0.001
affiliation_independent (neither)	0.1738	0.0909	1.913	0.056
affiliation_independent, close to democrat	0.0344	0.0988	0.348	0.728
affiliation_independent, close to republican	0.0215	0.1110	0.193	0.847
affiliation_not very strong democrat	0.0117	0.0953	0.123	0.903
affiliation_not very strong republican	0.0204	0.0996	0.205	0.838
affiliation_other party	0.0712	0.1871	0.381	0.703
affiliation_strong republican	0.1708	0.0926	1.846	0.065

Note: $R^2 = 0.007$, $F(7.0, 1040.0) = 1.01$, $p < 0.001$. Reference category: strong democrat.

Education

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3909	0.0440	8.887	<0.001
affiliation_associate/junior college	-0.0941	0.0787	-1.195	0.232

affiliation_bachelor's	-0.1202	0.0688	-1.748	0.081
affiliation_graduate	-0.1055	0.0850	-1.241	0.215
affiliation_less than high school	-0.0636	0.1866	-0.341	0.733

Note: $R^2 = 0.004$, $F(4.0, 1043.0) = 0.95$, $p < 0.001$. Reference category: high school.

Race

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3365	0.0315	10.686	<0.001
affiliation_black	-0.0358	0.0793	-0.452	0.652
affiliation_other	-0.0610	0.0935	-0.653	0.514

Note: $R^2 = 0.001$, $F(2.0, 1045.0) = 0.28$, $p < 0.001$. Reference category: white.

Ethnicity

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3335	0.0298	11.173	<0.001
affiliation_American Indian or Alaskan native	-0.0601	0.1625	-0.370	0.712
affiliation_Asian	-0.0909	0.1219	-0.746	0.456
affiliation_Black/African American	-0.0342	0.0757	-0.452	0.651
affiliation_Native Hawaiian or Pacific Islander	-0.0499	0.3859	-0.129	0.897
affiliation_Other race or ethnicity	-0.0491	0.1219	-0.403	0.687

Note: $R^2 = 0.001$, $F(5.0, 1118.0) = 0.18$, $p < 0.001$. Reference category: White/Caucasian.

Gender

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3190	0.0365	8.742	<0.001
affiliation_male	0.0148	0.0554	0.266	0.790

Note: $R^2 = 0.000$, $F(1.0, 1046.0) = 0.07$, $p < 0.001$. Reference category: female.

Income

Variable	Coefficient	Std. Error	t-value	p-value
const	0.2884	0.0731	3.947	<0.001
affiliation_ \$100,000 to \$124,999	-0.0201	0.1373	-0.147	0.883
affiliation_ \$125,000 to \$149,999	-0.0204	0.1638	-0.125	0.901
affiliation_ \$150,000 to \$174,999	-0.0917	0.2347	-0.391	0.696
affiliation_ \$175,000 to \$199,999	-0.0945	0.2699	-0.350	0.726
affiliation_ \$200,000 to \$249,999	0.0108	0.2900	0.037	0.970
affiliation_ \$25,000 to \$34,999	0.0166	0.1205	0.138	0.890
affiliation_ \$250,000 or more	0.0266	0.2347	0.113	0.910
affiliation_ \$35,000 to \$49,999	0.0252	0.1147	0.220	0.826
affiliation_ \$75,000 to \$99,999	-0.0138	0.1128	-0.122	0.903
affiliation_ Less than \$25,000	0.2508	0.1381	2.281	0.018

Note: $R^2 = 0.011$, $F(10.0, 860.0) = 0.95$, $p < 0.001$. Reference category: \$50,000 to \$74,999.

Neighborhood

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3270	0.0475	6.884	<0.001
affiliation_ Rural	0.0676	0.0862	0.785	0.433
affiliation_ Urban	-0.0300	0.0760	-0.394	0.693

Note: $R^2 = 0.001$, $F(2.0, 868.0) = 0.56$, $p < 0.001$. Reference category: Suburban.

Sexual Orientation

Variable	Coefficient	Std. Error	t-value	p-value
const	0.3417	0.0365	9.366	<0.001
affiliation_ Asexual	-0.0639	0.2545	-0.251	0.802
affiliation_ Bisexual	-0.0330	0.1222	-0.270	0.787

affiliation_ Gay or lesbian	-0.0567	0.1624	-0.349	0.727
affiliation_ Other sexual orientation	-0.0750	0.3107	-0.241	0.809
affiliation_ Pansexual	-0.0866	0.2160	-0.401	0.689

Note: $R^2 = 0.000$, $F(5.0, 863.0) = 0.08$, $p < 0.001$. Reference category: Heterosexual/straight.

[General Social Survey: Demographic Parity Difference]

	Agents w/ Interview	Agents w/ Demog. Info.	Agents w/ Persona Desc.
Age	DPD=4.30 Min: 35 - 44 (67.51%) Max: 75 or more (71.81%)	DPD=4.76 Min: 45 - 54 (55.3%) Max: 75 or more (60.06%)	DPD=3.23 Min: 45 - 54 (55.98%) Max: 18 - 24 (59.21%)
Census Division	DPD=3.16 Min: mountain (67.52%) Max: middle atlantic (70.68%)	DPD=2.74 Min: foreign (55.51%) Max: middle atlantic (58.25%)	DPD=3.99 Min: foreign (54.76%) Max: middle atlantic (58.75%)
Political Ideology	DPD=7.86 Min: conservative (66.22%) Max: extremely liberal (74.07%)	DPD=12.35 Min: conservative (50.74%) Max: extremely liberal (63.09%)	DPD=11.91 Min: extremely conservative (50.43%) Max: extremely liberal (62.34%)
Political Party	DPD=5.98 Min: strong republican (65.89%) Max: independent, close to democrat (71.87%)	DPD=9.79 Min: strong republican (51.96%) Max: independent, close to democrat (61.74%)	DPD=9.34 Min: strong republican (52.08%) Max: independent, close to democrat (61.41%)
Education	DPD=5.79 Min: less than high school (64.74%) Max: graduate (70.52%)	DPD=5.15 Min: less than high school (53.2%) Max: graduate (58.35%)	DPD=6.41 Min: less than high school (52.28%) Max: graduate (58.7%)
Race	DPD=2.08 Min: black (67.13%) Max: white (69.22%)	DPD=3.33 Min: black (54.19%) Max: white (57.52%)	DPD=3.19 Min: black (54.13%) Max: white (57.32%)
Ethnicity	DPD=2.15 Min: Other race or ethnicity (67.11%) Max: White/Caucasian (69.26%)	DPD=4.42 Min: Other race or ethnicity (53.77%) Max: Native Hawaiian or Pacific Islander (58.19%)	DPD=3.63 Min: Other race or ethnicity (53.75%) Max: White/Caucasian (57.37%)
Gender	DPD=0.54 Min: male (68.55%) Max: female (69.09%)	DPD=0.56 Min: male (56.68%) Max: female (57.24%)	DPD=0.61 Min: male (56.45%) Max: female (57.06%)
Income	DPD=4.95 Min: \$175,000 to \$199,999 (67.31%) Max: \$150,000 to \$174,999 (72.26%)	DPD=5.43 Min: Less than \$25,000 (55.4%) Max: \$200,000 to \$249,999 (60.83%)	DPD=6.61 Min: Less than \$25,000 (55.2%) Max: \$200,000 to \$249,999 (61.82%)
Neighborhood	DPD=1.22 Min: Rural (67.85%) Max: Suburban (69.08%)	DPD=1.62 Min: Urban (56.0%) Max: Suburban (57.62%)	DPD=2.89 Min: Rural (54.68%) Max: Suburban (57.57%)
Sexual Orientation	DPD=7.08 Min: Heterosexual/straight (67.83%) Max: Asexual (74.92%)	DPD=7.50 Min: Heterosexual/straight (55.85%) Max: Pansexual (63.36%)	DPD=9.88 Min: Heterosexual/straight (55.73%) Max: Asexual (65.61%)

[Big Five: Demographic Parity Difference]

	Agents w/ Interview	Agents w/ Demog. Info.	Agents w/ Persona Desc.
Age	DPD=0.22 Min: 18 - 24 (0.56) Max: 75 or more (0.78)	DPD=0.37 Min: 18 - 24 (0.23) Max: 65 - 74 (0.6)	DPD=0.33 Min: 18 - 24 (0.37) Max: 65 - 74 (0.7)
Census Division	DPD=0.19 Min: mountain (0.49) Max: w. sou. central (0.68)	DPD=0.24 Min: mountain (0.36) Max: foreign (0.6)	DPD=0.23 Min: mountain (0.39) Max: foreign (0.63)
Political Ideology	DPD=0.06 Min: moderate (0.6) Max: extremely conservative (0.67)	DPD=0.17 Min: extremely liberal (0.34) Max: moderate (0.51)	DPD=0.14 Min: extremely liberal (0.49) Max: slightly conservative (0.63)
Political Party	DPD=0.17 Min: other party (0.49) Max: independent, close to democrat (0.66)	DPD=0.22 Min: other party (0.3) Max: independent, close to democrat (0.53)	DPD=0.22 Min: other party (0.39) Max: strong republican (0.61)
Education	DPD=0.23 Min: less than high school (0.44) Max: graduate (0.68)	DPD=0.24 Min: less than high school (0.3) Max: graduate (0.54)	DPD=0.3 Min: less than high school (0.33) Max: graduate (0.63)
Race	DPD=0.11 Min: other (0.59) Max: black (0.7)	DPD=0.17 Min: other (0.4) Max: black (0.57)	DPD=0.08 Min: other (0.52) Max: black (0.6)
Ethnicity	DPD=0.21 Min: Native Hawaiian or Pacific Islander (0.49) Max: Black/African American (0.71)	DPD=0.62 Min: Native Hawaiian or Pacific Islander (-0.05) Max: Black/African American (0.57)	DPD=0.42 Min: Native Hawaiian or Pacific Islander (0.31) Max: American Indian or Alaskan native (0.73)
Gender	DPD=0.01 Min: male (0.62) Max: female (0.63)	DPD=0.01 Min: female (0.46) Max: male (0.46)	DPD=0.02 Min: female (0.54) Max: male (0.55)
Income	DPD=0.32 Min: \$200,000 to \$249,999 (0.41) Max: \$250,000 or more (0.73)	DPD=0.29 Min: \$200,000 to \$249,999 (0.31) Max: \$250,000 or more (0.6)	DPD=0.38 Min: \$200,000 to \$249,999 (0.35) Max: \$150,000 to \$174,999 (0.73)
Neighborhood	DPD=0.03 Min: Urban (0.58) Max: Rural (0.62)	DPD=0.05 Min: Rural (0.43) Max: Urban (0.48)	DPD=0.04 Min: Suburban (0.51) Max: Urban (0.54)
Sexual Orientation	DPD=0.23 Min: Other sexual orientation (0.43) Max: Pansexual (0.66)	DPD=0.29 Min: Other sexual orientation (0.17) Max: Heterosexual/straight (0.46)	DPD=0.45 Min: Other sexual orientation (0.1) Max: Heterosexual/straight (0.55)

[Economic Games: Demographic Parity Difference]

	Agents w/ Interview	Agents w/ Demog. Info.	Agents w/ Persona Desc.
Age	DPD=0.18 Min: 75 or more (0.21) Max: 45 - 54 (0.39)	DPD=0.19 Min: 35 - 44 (0.18) Max: 55 - 64 (0.36)	DPD=0.8 Min: 35 - 44 (0.15) Max: 55 - 64 (0.94)
Census Division	DPD=0.2 Min: mountain (0.25) Max: foreign (0.45)	DPD=0.21 Min: e. nor. central (0.19) Max: new england (0.4)	DPD=0.65 Min: mountain (0.15) Max: pacific (0.8)
Political Ideology	DPD=0.19 Min: conservative (0.24) Max: extremely liberal (0.44)	DPD=0.5 Min: extremely conservative (-0.03) Max: extremely liberal (0.47)	DPD=0.53 Min: moderate (0.37) Max: slightly liberal (0.91)
Political Party	DPD=0.22 Min: independent (neither) (0.21) Max: strong democrat (0.43)	DPD=0.4 Min: strong republican (0.03) Max: strong democrat (0.43)	DPD=0.58 Min: independent (neither) (0.31) Max: independent, close to democrat (0.88)
Education	DPD=0.14 Min: high school (0.29) Max: less than high school (0.43)	DPD=0.12 Min: associate/junior college (0.21) Max: graduate (0.33)	DPD=0.44 Min: less than high school (0.31) Max: graduate (0.76)
Race	DPD=0.04 Min: black (0.3) Max: white (0.34)	DPD=0.04 Min: other (0.24) Max: white (0.29)	DPD=0.42 Min: other (0.37) Max: black (0.79)
Ethnicity	DPD=0.52 Min: Other race or ethnicity (0.06) Max: Native Hawaiian or Pacific Islander (0.58)	DPD=0.63 Min: Other race or ethnicity (0.07) Max: Native Hawaiian or Pacific Islander (0.7)	DPD=1.54 Min: Asian (0.12) Max: Native Hawaiian or Pacific Islander (1.66)
Gender	DPD=0.03 Min: female (0.32) Max: male (0.35)	DPD=0.03 Min: female (0.27) Max: male (0.29)	DPD=0.05 Min: female (0.55) Max: male (0.6)
Income	DPD=0.51 Min:\$200,000 to\$249,999 (0.21) Max:\$175,000 to\$199,999 (0.72)	DPD=0.46 Min:\$200,000 to\$249,999 (0.19) Max:\$150,000 to\$174,999 (0.65)	DPD=0.71 Min:\$125,000 to\$149,999 (0.22) Max:\$175,000 to\$199,999 (0.93)
Neighborhood	DPD=0.14 Min: Urban (0.25) Max: Suburban (0.38)	DPD=0.11 Min: Urban (0.2) Max: Suburban (0.32)	DPD=0.52 Min: Urban (0.35) Max: Rural (0.88)
Sexual Orientation	DPD=0.32 Min: Asexual (0.23) Max: Other sexual orientation (0.55)	DPD=0.41 Min: Heterosexual/straight (0.24) Max: Other sexual orientation (0.66)	DPD=0.74 Min: Bisexual (0.2) Max: Pansexual (0.93)

Table B.9: Demographic Parity Difference (DPD) Results. This table summarizes the results of regression analyses measuring demographic parity differences (DPD) across three tasks (GSS, Big Five, and economic games) for agents using demographic information, interview data, and persona-based profiles. Interview-based agents consistently reduced bias compared to demographic-based agents across political ideology, race, and gender.

[Robustness Analysis]				
	GSS	GSS Num.	Big Five	Economic Games
Interview	Acc.=66.16 (std = 7.19) Nrm. Acc.=0.83 (std = 0.11) – Correl.=0.75 (z std = 0.19) Nrm. Correl.=0.87 (std = 0.12)	MAE=0.14 (std=0.15) – Correl.=0.97 (z std=0.82) Nrm. Correl.=2.16 (std=9.17)	MAE=0.67 (std=0.27) – Correl.=0.79 (z std=0.73) Nrm. Correl.=0.77 (std=0.76)	MAE=0.51 (std=2.01) – Correl.=0.52 (z std=0.9) Nrm. Correl.=0.33 (std=1.22)
Survey and Experiments	Acc.=60.46 (std = 7.19) Nrm. Acc.=0.76 (std = 0.12) – Correl.=0.70 (z std = 0.21) Nrm. Correl.=0.81 (std = 0.13)	MAE=0.32 (std=0.19) – Correl.=0.75 (z std=0.62) Nrm. Correl.=1.74 (std=8.94)	MAE=0.68 (std=0.25) – Correl.=0.71 (z std=0.75) Nrm. Correl.=0.64 (std=0.61)	MAE=0.47 (std=2.0) – Correl.=0.52 (z std=0.93) Nrm. Correl.=0.31 (std=1.22)
Maximal	Acc.=66.26 (std = 7.15) Nrm. Acc.=0.83 (std = 0.12) – Correl.=0.74 (z std = 0.20) Nrm. Correl.=0.86 (std = 0.14)	MAE=0.28 (std=0.21) – Correl.=0.78 (z std=0.71) Nrm. Correl.=1.85 (std=9.19)	MAE=0.68 (std=0.26) – Correl.=0.77 (z std=0.75) Nrm. Correl.=0.72 (std=0.78)	MAE=0.47 (std=2.0) – Correl.=0.55 (z std=0.93) Nrm. Correl.=0.3 (std=1.29)
Summary	Acc.=656.28 (std=6.9458) Nrm. Acc.=0.823 (std=0.12) – Correl.=0.764 (z std=0.2) Nrm. Correl.=0.864 (std=0.1436)	MAE=0.19 (std=0.18) – Correl.=0.93 (z std=0.82) Nrm. Correl.=0.92 (std=2.77)	MAE=0.69 (std=0.3) – Correl.=0.76 (z std=0.76) Nrm. Correl.=0.7 (std=0.67)	MAE=0.5 (std=2.0) – Correl.=0.49 (z std=0.85) Nrm. Correl.=0.41 (std=1.1)
Random lesion (0% removal)	Acc.=66.16 (std = 7.19) Nrm. Acc.=0.83 (std = 0.11) – Correl.=0.75 (z std = 0.19) Nrm. Correl.=0.87 (std = 0.12)	MAE=0.14 (std=0.15) – Correl.=0.97 (z std=0.82) Nrm. Correl.=2.16 (std=9.17)	MAE=0.67 (std=0.27) – Correl.=0.79 (z std=0.73) Nrm. Correl.=0.77 (std=0.76)	MAE=0.51 (std=2.01) – Correl.=0.52 (z std=0.9) Nrm. Correl.=0.33 (std=1.22)
Random lesion (20% removal)	Acc.=65.50 (std = 7.44) Nrm. Acc.=0.83 (std = 0.13) – Correl.=0.74 (z std = 0.18) Nrm. Correl.=0.86 (std = 0.13)	MAE=0.16 (std=0.13) – Correl.=0.95 (z std=0.81) Nrm. Correl.=1.46 (std=3.71)	MAE=0.66 (std=0.28) – Correl.=0.79 (z std=0.74) Nrm. Correl.=0.75 (std=0.72)	MAE=0.52 (std=2.02) – Correl.=0.58 (z std=0.89) Nrm. Correl.=0.46 (std=1.18)

	GSS	GSS Num.	Big Five	Economic Games
Random lesion (40% removal)	Acc.=64.84 (std = 7.21) Nrm. Acc.=0.82 (std = 0.12) – Correl.=0.73 (z std = 0.18) Nrm. Correl.=0.86 (std = 0.13)	MAE=0.16 (std=0.14) – Correl.=0.96 (z std=0.75) Nrm. Correl.=2.14 (std=9.15)	MAE=0.65 (std=0.28) – Correl.=0.78 (z std=0.74) Nrm. Correl.=0.75 (std=0.74)	MAE=0.52 (std=2.02) – Correl.=0.63 (z std=0.9) Nrm. Correl.=0.5 (std=1.21)
Random lesion (60% removal)	Acc.=65.02 (std = 7.20) Nrm. Acc.=0.82 (std = 0.13) – Correl.=0.74 (z std = 0.17) Nrm. Correl.=0.87 (std = 0.13)	MAE=0.16 (std=0.12) – Correl.=0.95 (z std=0.81) Nrm. Correl.=1.96 (std=9.02)	MAE=0.65 (std=0.29) – Correl.=0.77 (z std=0.74) Nrm. Correl.=0.73 (std=0.68)	MAE=0.48 (std=2.0) – Correl.=0.57 (z std=0.91) Nrm. Correl.=0.41 (std=1.33)
Random lesion (80% removal)	Acc.=62.52 (std = 7.30) Nrm. Acc.=0.79 (std = 0.11) – Correl.=0.73 (z std = 0.18) Nrm. Correl.=0.85 (std = 0.12)	MAE=0.18 (std=0.15) – Correl.=0.96 (z std=0.76) Nrm. Correl.=2.0 (std=9.09)	MAE=0.64 (std=0.26) – Correl.=0.75 (z std=0.72) Nrm. Correl.=0.73 (std=0.74)	MAE=0.52 (std=2.01) – Correl.=0.62 (z std=0.93) Nrm. Correl.=0.46 (std=1.19)

Table B.10: Robustness analysis. This table presents the results of an exploratory robustness analysis comparing different agent architectures informed by various data sources, including interviews, surveys, experiments, and summaries. Performance is evaluated across four constructs: the General Social Survey (GSS), GSS Numeric (GSS Num.), Big Five Personality Traits (Big Five), and economic games. Interview-based agents consistently outperform others, achieving high accuracy (0.83, std=0.11) on the GSS, indicating that interviews provide richer, more comprehensive information than surveys and experiments. Maximal agents, which integrate data from all sources, show similar performance. Summary agents perform slightly below interview-based agents, with minor losses in accuracy (0.82, std = 0.12). A progressive decline is observed for random lesion agents as portions of interview data are removed, with accuracy dropping from 0.83 to 0.79 as 80% of the utterances are excluded, suggesting that even shortened interviews retain valuable insights compared to survey-only agents.

Table B.11: The interview script used to guide two-hour conversations with participants, adapted from the American Voices Project’s interview. The script covers a wide range of topics, from participants’ life stories to their views on social, political, and personal values. Select portions were abbreviated to ensure a manageable session length, while still capturing the breadth of experiences and perspectives essential for building nuanced generative agents.

Script	Time Limit (sec)
To start, I would like to begin with a big question: tell me the story of your life. Start from the beginning – from your childhood, to education, to family and relationships, and to any major life events you may have had.	625
Some people tell us that they’ve reached a crossroads at some points in their life where multiple paths were available, and their choice then made a significant difference in defining who they are. What about you? Was there a moment like that for you, and if so, could you tell me the whole story about that from start to finish?	235
Some people tell us they made a conscious choice or decision in moments like these, while others say it ‘just happened’. What about you?	50
Do you think another person or an organization could have lent a helping hand during moments like this?	40
Moving to present time, tell me more about family who are important to you. Do you have a partner, or children?	310
Are there anyone else in your immediate family whom you have not mentioned? Who are they, and what is your relationship with them like?	25
Tell me about anyone else in your life we haven’t discussed (like friends or romantic partners). Are there people outside of your family who are important to you?	80
Now let’s talk about your current neighborhood. Tell me all about the neighborhood and area in which you are living now.	105
Some people say they feel really safe in their neighborhoods, others not so much. How about for you?	40
Living any place has its ups and downs. Tell me about what it’s been like for you living here.	40
Tell me about the people who live with you right now, even people who are staying here temporarily.	80
Is anyone in your household a temporary member, or is everyone a permanent member of the household?	10
Tell me about anyone else who stays here from time to time, if there is anyone like that	10
Is there anyone who usually lives here but is away – traveling, at school, or in a hospital?	10

Continued on next page

Script	Time Limit (sec)
Right now, across a typical week, how do your days vary?	105
At what kind of job or jobs do you work, and when do you work?	105
Do you have other routines or responsibilities that you did not already share?	80
Tell me about any recent changes to your daily routine.	80
If you have children, tell me about a typical weekday during the school year for your children. What is their daily routine? And what are after school activities your children participate in?	80
Some people we've talked to tell us about experiences with law enforcement. How about for you?	80
What other experiences with law enforcement stand out in your mind?	40
Some people tell us about experiences of arrest – of loved ones, family members, friends, or themselves. How about for you?	40
Some people say they vote in every election, some tell us they don't vote at all. How about you?	80
How would you describe your political views?	310
Tell me about any recent changes in your political views.	115
One topic a lot of people have been talking about recently is race and/or racism and policing. Some tell us issues raised by the Black Lives Matter movement have affected them a lot, some say they've affected them somewhat, others say they haven't affected them at all. How about for you?	105
How have you been thinking about the issues Black Lives Matter raises?	105
How have you been thinking about race in the U.S. recently?	105
How have you responded to the increased focus on race and/or racism and policing? Some people tell us they're aware of the issue but keep their thoughts to themselves, others say they've talked to family and friends, others have joined protests. What about you?	50
What about people you're close to? How have they responded to the increased focus on race and/or racism and policing?	80
Now we'd like to learn more about your health. First, tell me all about your health.	80
For you, what makes it easy or hard to stay healthy?	65
Tell me about anything big that has happened in the past two years related to your health: any medical diagnoses, flare-ups of chronic conditions, broken bones, pain – anything like that.	80
Sometimes, health problems get in the way. They can even affect people's ability to work or care for their children. How about you?	50
Sometimes, it's not your health problem, but the health of a loved one. Has this been an issue for you?	80

Continued on next page

Script	Time Limit (sec)
Tell me what it has been like trying to get the health care you or your immediate family need. Have you ever had to forgo getting the health care you need?	80
Have you or your immediate family ever used alternative forms of medicine? This might include indigenous, non-western, or informal forms of care.	80
During tough times, some people tell us they cope by smoking or drinking. How about for you?	80
Other people say they cope by relying on prescriptions, pain medications, marijuana, or other substances. How about for you and can you describe your most recent experience of using them, if any?	80
How are you and your family coping with, or paying for, your health care needs right now?	80
Some people are excited about medical vaccination, and others, not so much. How about you?	50
What are your trusted sources of information about the vaccine?	80
Now we're going to talk a bit more about what life was like for you over the past year. Tell me all about how you have been feeling.	115
Tell me a story about a time in the last year when you were in a rough place or struggling emotionally.	80
Some people say they struggle with depression, anxiety, or something else like that. How about for you?	80
How has it been for your family?	50
Some people say that religion or spirituality is important in their lives, for others not so much. How about you?	155
Some people tell us they use Facebook, Instagram, or other social media to stay connected with the wider world. How about for you? Tell me all about how you use these types of platforms.	115
Some people say they use Facebook or Twitter to get (emotional or financial) support during tough times. What about you?	50
Tell me about any recent changes in your level of stress, worry, and your emotional coping strategies.	50
This might be repetitive, but before we go on to the next section, I want to quickly make sure I have this right: who are you living with right now?	25
Any babies or small children?	10
Anyone who usually lives here but is away – traveling, at school, or in a hospital?	10
Any lodgers, boarders, or employees who live here?	10
Who shares the responsibility for the rent, mortgage, or other household expenses?	5

Continued on next page

Script	Time Limit (sec)
Now we'd like to talk about how you make ends meet and what the monthly budget looks like for you and your family. What were your biggest expenses last month?	50
How much did your household spend in the past month? Is that the usual amount? And if not, how much do you usually spend and on what?	50
Does your household own or rent your home?	50
Bills can fluctuate over the course of a year. Seasons change, there are holidays and special events, school starts and ends, and so on. Tell me all about how your bills have fluctuated over the past year.	50
Tell me all about how you coped with any extra expenses in the past months.	40
Some people have a savings account, some people save in a different way, and some people say they don't save. How about you?	80
Some people save for big things, like a home, while others save for a rainy day. How about for you (in the last year)?	80
Some people have student loans or credit card debt. Others take out loans from family or friends or find other ways of borrowing money. Tell me about all the debts you're paying on right now.	50
Tell me about any time during the past year that you haven't had enough money to buy something that you needed or pay a bill that was due.	65
What is or was your occupation? In this occupation, what kind of work do you do and what are the most important activities or duties?	155
Was your occupation covered by a union or employee association contract?	40
In the past year, how many weeks did you work (for a few hours, including paid vacation, paid sick leave, and military service)?	40
How many hours each week did you usually work?	15
In the past week, did you work for pay (even for 1 hour) at a job or business?	15
Did you have more than one job or business, including part time, evening, or weekend work?	25
Now, in the past month, how many jobs did you work, including part time, evening, or weekend work?	25
For your job or occupations, how much money did you receive in the past month?	50
How often do you get paid?	25
In total, how much did your household make in the past month?	50
Switching gears a bit, let's talk a little about tax time. Did you file taxes last year (the most recent year)?	40

Continued on next page

Script	Time Limit (sec)
Tell me more about your workplace. How long have you been at your current job? How would you describe the benefits that come with your job? (This could include things like health benefits, paid time off, vacation, and sick time.)	15
How would you describe your relationships at work? (How is your relationship with your manager or boss? How are your relationships with your coworkers?)	15
Tell me about how predictable your work schedule is. How would you describe your job in terms of flexibility with your hours?	80
Some people say it's hard to take or keep a job because of child care. How about for you?	15
Do you receive any payments or benefits from SNAP, food stamps, housing voucher payments, supplemental security income, or any other programs like that?	25
Tell me about times over the last year when your income was especially low. And tell me about all the things you did to make ends meet during that time.	50
What would it be like for you if you had to spend \$400 for an emergency? Would you have the money, and if not, how would you get it?	50
Overall, how do you feel about your financial situation?	105
Are you now married, widowed, divorced, separated, or have you never been married? If you are not currently married, are you currently living with a romantic partner?	40
For our records, we also need your date of birth. Could you please provide that?	25
Were you born in the United States? If you were not born in the U.S., what country were you born, and what year did you first come to the U.S. to live?	25
What city and state were you born in?	25
Are you of Hispanic, Latinx, or Spanish origin?	15
What race or races do you identify with?	15
What is the highest degree or grade you've completed?	15
Are you enrolled in school?	15
Have you been enrolled in school during the past 3 months?	15
Have you ever served on active duty in the U.S. Armed Forces, Reserves, or National Guard?	15
What religion do you identify with, if any?	15
Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent, or what? And how strongly do you associate with that party?	80
Do you think of yourself as closer to the Democratic Party or to the Republican Party?	50
What was the city and state that you lived in when you were 16 years old?	25
Did you live with both your own mother and father when you were 16?	25
Who else did you live with?	25

Continued on next page

Script	Time Limit (sec)
What's the highest degree or grade your dad completed?	15
What's the highest degree or grade your mom completed?	15
Did your mom work for pay for at least a year while you were growing up?	15
What was her job? What kind of work did she do? What were her most important activities or duties?	50
What kind of place did she work for? What kind of business was it? What did they make or do where she worked?	50
Did your dad work for pay for at least a year while you were growing up?	15
What was his job? What kind of work did he do? What were his most important activities or duties?	50
What kind of place did he work for? What kind of business was it? What did they make or do where he worked?	50
We all have hopes about what our future will look like. Imagine yourself a few years from now. Maybe you want your life to be the same in some ways as it is now. Maybe you want it to be different in some ways. What do you hope for?	155
What do you value the most in your life?	80
And that was the last question I wanted to ask today. Thank you so much again for your time <participant's name>. It was really wonderful getting to know you through this interview. Now, we will take you back to the Home Screen so that you may finish the rest of the study!	0

General Social Survey	Accuracy	Normalized Accuracy	Correlation	Normalized Correlation
Participant Replication	81.25% (std=8.11)	1.00 (std=0.00)	0.83 (std=0.30)	1.00 (std=0.00)
Agents w/ Interview	68.85% (std=6.01)	0.85 (std=0.11)	0.66 (std=0.19)	0.83 (std=0.31)
Agents w/ Demog. Info.	57.00% (std=7.45)	0.71 (std=0.11)	0.51 (std=0.19)	0.63 (std=0.26)
Agents w/ Persona Desc.	56.79% (std=7.76)	0.70 (std=0.11)	0.50 (std=0.20)	0.62 (std=0.25)

Big Five	Mean Absolute Error	-	Correlation	Normalized Correlation
Participant Replication	0.30 (std=0.17)	-	0.95 (std=0.76)	1.00 (std=0.00)
Agents w/ Interview	0.67 (std=0.27)	-	0.78 (std=0.70)	0.80 (std=1.88)
Agents w/ Demog. Info.	0.76 (std=0.32)	-	0.61 (std=0.70)	0.55 (std=2.25)
Agents w/ Persona Desc.	0.74 (std=0.32)	-	0.71 (std=0.73)	0.75 (std=2.59)

Economic Games	Mean Absolute Error	-	Correlation	Normalized Correlation
Participant Replication	0.25 (std=0.88)	-	0.99 (std=1.00)	1.00 (std=0.00)
Agents w/ Interview	0.32 (std=0.89)	-	0.66 (std=0.95)	0.66 (std=2.83)
Agents w/ Demog. Info.	0.34 (std=0.89)	-	0.57 (std=0.91)	0.48 (std=2.90)
Agents w/ Persona Desc.	0.33 (std=0.88)	-	0.60 (std=0.93)	0.57 (std=2.84)

Table B.12: Generative agents’ predictive performance. The consistency rate between participants and the predictive performance of generative agents is evaluated across various constructs and averaged across individuals. For the General Social Survey (GSS), accuracy is reported due to its categorical response types, while the Big Five personality traits and economic games report mean absolute error (MAE) due to their numerical response types. Correlation is reported for all constructs. Normalized accuracy is provided for all metrics, except for MAE, which cannot be calculated for individuals whose MAE is 0 (i.e., those who responded the same way in both phases). We find that generative agents predict participants’ behavior and attitudes well, especially when compared to participants’ own rate of internal consistency. Additionally, using interviews to inform agent behavior significantly improves the predictive performance of agents for both GSS and Big Five constructs, outperforming other commonly used methods in the literature.

Bibliography

- Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems.
- Mark S. Ackerman. 2000. The intellectual challenge of cscw: The gap between social requirements and technical feasibility. *Human-Computer Interaction*, 15(2):179–203.
- Robert Ackland, Jamsheed Shorish, Paul Thomas, and Lexing Xie. 2013. How dense is a network? <http://users.cecs.anu.edu.au/~xlx/teaching/css2013/network-density.html>.
- Eytan Adar, Mira Dontcheva, and Gierad Laput. 2014. Commandspace: Modeling the relationships between tasks, descriptions and features. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, page 167–176, New York, NY, USA. Association for Computing Machinery.
- Yavuz Akbulut, Yusuf Levent Sahin, and Bahadir Eristi. 2010. Cyberbullying victimization among turkish online social utility members. *Journal of Educational Technology & Society*, 13(4):192–201.
- Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.
- Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13.

- D. L. Ames and S. T. Fiske. 2015. Perceived intent motivates people to magnify observed harms. *Proceedings of the National Academy of Sciences of the United States of America*, 112:3599–3605.
- John R. Anderson. 1993. *Rules of the Mind*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- S. Ansolabehere, J. Rodden, and Jr. Snyder, James M. 2008. The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, 102:215–232.
- Electronic Arts. 2009. The sims 3. Video game.
- R. Axtell. 2000. Why agents? on the varied motivations for agent computing in the social sciences. Working Paper 17, Center on Social and Economic Dynamics.
- Ruth Aylett. 1999. Narrative in virtual environments—towards emergent narrative. In *Narrative Intelligence: Papers from the AAAI Fall Symposium (Technical Report FS-99-01)*, pages 83–86. AAAI Press.
- S. Barocas, M. Hardt, and A. Narayanan. 2019. *Fairness and Machine Learning*.
- Christoph Bartneck and Jodi Forlizzi. 2004. A design-centered framework for social human-robot interaction. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN'04)*, pages 591–594.
- Joseph Bates. 1994. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125.
- Michel Beaudouin-Lafon and Wendy Mackay. 2002. *Prototyping Tools and Techniques*, page 1006–1031. L. Erlbaum Associates Inc., USA.
- J. Berg, J. Dickhaut, and K. McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10:122–142.

- Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębniak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d.O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.
- Michael S. Bernstein, Mark S. Ackerman, Ed H. Chi, and Robert C. Miller. 2011. The trouble with social computing systems research. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 389–398, New York, NY, USA. Association for Computing Machinery.
- Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.
- BioWare. 2007. Mass effect. Video game.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, and et al. 2022. On the opportunities and risks of foundation models.
- E. Bonabeau. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99(suppl 3):7280–7287.
- Michael Brenner. 2010. Creating dynamic story plots with continual multiagent planning. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*.
- S. Brinkmann and S. Kvale. 2014. *InterViews: Learning the Craft of Qualitative Research Interviewing*, 3rd edition. SAGE Publications.
- Rodney A. Brooks, Cynthia Breazeal, Marko Marjanovic, Brian Scassellati, and Matthew Williamson. 2000. The cog project: Building a humanoid robot. In *Computation for Metaphors, Analogy, and Agents*, number 1562 in Lecture Notes on Artificial Intelligence, pages 52–87, Berlin. Springer-Verlag.

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.
- E. Bruch and J. Atwell. 2015. Agent-based models in empirical social research. *Sociological Methods & Research*, 44:186–221.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Marion Buchenau and Jane Fulton Suri. 2000. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 424–433.
- Robin Burkinshaw. 2009. Alice and kev: The story of being homeless in the sims 3.
- Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9379–9393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- C. F. Camerer et al. 2024. Mechanical turk replication project.

- Stuart K. Card, Thomas P. Moran, and Alan Newell. 1983. *The psychology of human-computer interaction*. Lawrence Erlbaum, Hillsdale, NJ.
- Stuart K Card, Thomas P Moran, and Allen Newell. 1980. The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23(7):396–410.
- K. M. Carley. 1992. Organizational learning and personnel turnover. *Organization Science*, 3:20–46.
- Alex Champandard. 2012. Tutorial presentation. In *IEEE Conference on Computational Intelligence and Games*.
- Stevie Chancellor, Zhiyuan (Jerry) Lin, and Munmun De Choudhury. 2016. "this post will just get taken down": Characterizing removed pro-eating disorder social media content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1157–1162, New York, NY, USA. Association for Computing Machinery.
- J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. J. H. W. Welch, editors. 2019. *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd edition. Wiley.
- Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017a. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22.
- Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).
- Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017b. The bag of communities: Identifying abusive behavior online with preexisting

- internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3175–3187.
- Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17*, page 1217–1230, New York, NY, USA. Association for Computing Machinery.
- Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 61–70.
- Dong kyu Choi, Tolga Konik, Negin Nejati, Chunki Park, and Pat Langley. 2021. A believable agent for first-person shooter games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 3, pages 71–73.
- G. Cooney, D. T. Gilbert, and T. D. Wilson. 2016. When fairness matters less than we expect. *Proceedings of the National Academy of Sciences of the United States of America*, 113:11168–11171.
- Anind K Dey. 2001. Understanding and using context. *Personal and ubiquitous computing*, 5:4–7.
- Joan Donovan. 2019. How hate groups’ secret sound system works. *The Atlantic*, 17.
- Daniel Drew, Julie L Newcomb, William McGrath, Filip Maksimovic, David Mellis, and Björn Hartmann. 2016. The toastboard: Ubiquitous instrumentation and automated checking of breadboarded circuits. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 677–686.
- David Easley and Jon Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.

- Arpad E Elo. 1967. The proposed uscf rating system, its development, theory, and applications. *Chess Life*, XXII(8):242–247.
- J. M. Epstein and R. L. Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. MIT Press.
- Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45. ACM.
- Ethan Fast, William McGrath, Pranav Rajpurkar, and Michael S Bernstein. 2016. Augur: Mining human behaviors from fiction to power interactive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 237–247.
- Rebecca Fiebrink and Perry R Cook. 2010. The wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, volume 3, pages 2–1. Citeseer.
- Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.
- Susan T Fiske. 2022. Twitter manipulates your feed: Ethical considerations.
- Catherine Flick. 2016. Informed consent and the facebook emotional manipulation study. *Research Ethics*, 12(1):14–28.
- Uwe Flick. 2009. *An Introduction to Qualitative Research*. SAGE.
- James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. Cueflik: Interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 29–38, New York, NY, USA. Association for Computing Machinery.

- R. Forsythe, J. L. Horowitz, N. E. Savin, and M. Sefton. 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6:347–369.
- Adam Fourney, Richard Mann, and Michael Terry. 2011. Query-feature graphs: bridging user vocabulary and system functionality. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. ACM.
- Tom Francis. 2010. The minecraft experiment, day 1: Chasing waterfalls.
- Jonas Freiknecht and Wolfgang Effelsberg. 2020. Procedural generation of interactive stories using language models. In *International Conference on the Foundations of Digital Games (FDG '20)*, page 8, Bugibba, Malta. ACM.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *CoRR abs/2012.15723*.
- Catherine Grevet and Eric Gilbert. 2015. Piggyback prototyping: Using existing, large-scale social computing systems to prototype new ones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4047–4056.
- Volker Grimm and Steven F. Railsback. 2005. *Individual-based Modeling and Ecology*. Princeton University Press, Princeton, NJ.
- R. M. Groves, Jr. Fowler, F. J., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*, 2nd edition. John Wiley & Sons.
- Jonathan Grudin. 1994. Groupware and social dynamics: Eight challenges for developers. *Communications of the ACM*, 37(1):92–105.
- N. Halevy and E. Halali. 2015. Selfish third parties act as peacemakers by transforming conflicts and promoting cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, 112:6937–6942.
- Aaron Halfaker, R Stuart Geiger, Jonathan T Morgan, and John Riedl. 2013. The rise and decline of an open collaboration system: How wikipedia’s reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688.

- Blake Hallinan, Jed R Brubaker, and Casey Fiesler. 2020. Unexpected expectations: Public reaction to the facebook emotional contagion study. *New Media & Society*, 22(6):1076–1094.
- Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.
- Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, 6:215J242.
- M. Hardt, E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, pages 3315–3323. ArXiv:1610.02413.
- Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R Klemmer. 2007. Authoring sensor-based interactions by demonstration with direct manipulation and pattern recognition. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–154.
- Björn Hartmann, Scott R Klemmer, Michael Bernstein, Leith Abdulla, Brandon Burr, Avi Robinson-Mosher, and Jennifer Gee. 2006. Reflective physical prototyping through integrated design, test, and analysis. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, pages 299–308.
- Björn Hartmann, Meredith Ringel Morris, Hrvoje Benko, and Andrew D. Wilson. 2009. Augmenting interactive tables with mice & keyboards. In *Proceedings of the 22Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, pages 149–152, New York, NY, USA. ACM.
- Björn Hartmann, Loren Yu, Abel Allison, Yeonsoo Yang, and Scott R. Klemmer. 2008. Design as exploration: Creating interface alternatives through parallel authoring and runtime tuning. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST '08)*, pages 91–100, Monterey, CA, USA.

- Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Cote, and Xinyu Yuan. 2020. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7903–7910.
- Chris Hecker. 2011. My liner notes for spore.
- Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.
- Douglas Hofstadter. 1995. *Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought*. Basic Books.
- Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.
- John J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus?
- Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166.
- Stephanie Houde and Charles Hill. 1997. What do prototypes prototype? In *Handbook of human-computer interaction*, second edition, pages 367–381. Elsevier.
- Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner monologue: Embodied reasoning through planning with language models.
- Kristen Ibister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 52(1):65–80.

- Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. Promptmaker: Prompt-based prototyping with large language models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.
- Bonnie E John and David E Kieras. 1996. The goms family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(4):320–351.
- O. P. John and S. Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, editors, *Handbook of Personality: Theory and Research*, 2nd edition, pages 102–138. Guilford Press.
- Joseph M Kayany. 1998. Contexts of uninhibited online behavior: Flaming in social newsgroups on usenet. *Journal of the American Society for Information Science*, 49(12):1135–1141.
- John F Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 193–196.
- John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.
- Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp.
- Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an "eternal september" how an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1152–1156.

- Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: Conflict and coordination in wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 453–462, New York, NY, USA. Association for Computing Machinery.
- Bjoern Knafla. 2011. Introduction to behavior trees.
- Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. 2012. Trustworthy online controlled experiments: Five puzzling outcomes explained. KDD '12, page 786–794, New York, NY, USA. Association for Computing Machinery.
- Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.
- Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design*. MIT Press.
- Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S. Bernstein. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39):e2115730119.
- William H Kruskal and WA Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.
- Phaser Labs. 2023. Welcome to phaser 3. <https://phaser.io/phaser3>. Accessed on: 2023-04-03.
- John Laird. 2001. It knows what you’re going to do: Adding anticipation to a quakebot. In *Proceedings of the 2001 Workshop on Intelligent Cinematography and Editing*, pages 63–69.
- John Laird and Michael VanLent. 2001. Human-level ai’s killer application: Interactive computer games. *AI Magazine*, 22(2):15.

- John E. Laird. 2000. It knows what you're going to do: Adding anticipation to a quakebot. In *Papers from the AAAI 2000 Spring Symposium on Artificial Intelligence and Interactive Entertainment*, Technical Report SS-00-02, pages 41–50. AAAI Press.
- John E. Laird. 2012. *The Soar Cognitive Architecture*. MIT Press.
- John E. Laird, Christian Lebiere, and Paul S. Rosenbloom. 2017. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(1):13–26.
- Michelle S. Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A. Landay, and Michael S. Bernstein. 2023. Model sketching: Centering concepts in early-stage machine learning model design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.
- James A. Landay. 1996. Silk: Sketching interfaces like crazy. In *Conference Companion on Human Factors in Computing Systems*, CHI '96, page 398–399, New York, NY, USA. Association for Computing Machinery.
- Pat Langley, Dongkyu Choi, and Seth Rogers. 2005. Interleaving learning, problem solving, and execution in the icarus architecture. Technical report, Stanford University, Center for the Study of Language and Information.
- A. Lareau. 2021. *Listening to People: A Practical Guide to Interviewing, Participant Observation, Data Analysis, and Writing It All Up*. University of Chicago Press.
- J. O. Ledyard. 1995. Public goods: A survey of experimental research. In J. H. Kagel and A. E. Roth, editors, *The Handbook of Experimental Economics*, pages 111–194. Princeton University Press.
- Jason Linder, Gierad Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, and Eytan Adar. 2013. Pixeltone: A multimodal interface for image editing. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, page 2829–2830, New York, NY, USA. Association for Computing Machinery.

- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *CoRR abs/2101.06804*.
- N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 11:1312–1327.
- Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–17.
- I. Lundberg et al. 2024. The origins of unpredictability in life outcome prediction tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 121:e2322973121.
- I. S. Lustick. 2002. Ps-i: A user-friendly agent-based modeling platform for testing theories of political identity and political stability. *Journal of Artificial Societies and Social Simulation*, 5(3).
- M. W. Macy and R. Willer. 2002. From factors to actors: Computational sociology and agent-based modeling. *Annual Review of Sociology*, 28:143–166.
- Pattie Maes. 1995. Artificial life meets entertainment: Lifelike autonomous agents. *Commun. ACM*, 38(11):108–114.
- J. Marks, B. Andalman, P. A. Beardsley, W. Freeman, S. Gibson, J. Hodgins, T. Kang, B. Mirtich, H. Pfister, W. Ruml, K. Ryall, J. Seims, and S. Shieber. 1997. Design galleries: A general approach to setting parameters for computer graphics and animation. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, page 389–400, USA. ACM Press/Addison-Wesley Publishing Co.
- M. Lynne Markus. 1987. Toward a “critical mass” theory of interactive media. *Communication Research*, 14:491 – 511.

- P. V. Marsden, T. W. Smith, and M. Hout. 2020. Tracking us social change over a half-century: The general social survey at fifty. *Annual Review of Sociology*, 46:109–134.
- Nathan J. Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20):9785–9789.
- Josh McCoy, Michael Mateas, and Noah Wardrip-Fruin. 2009. Comme il faut: A system for simulating social games between autonomous characters. In *Proceedings of the 7th International Conference on Digital Arts and Culture*, pages 87–94.
- Josh McCoy, Mike Treanor, Ben Samuel, Michael Mateas, and Noah Wardrip-Fruin. 2011a. Prom week: Social physics as gameplay. In *Proceedings of the 6th International Conference on Foundations of Digital Games (FDG’11)*, pages 70–77, Bordeaux, France. ACM.
- Josh McCoy, Mike Treanor, Ben Samuel, Anna Reed, Michael Mateas, and Noah Wardrip-Fruin. 2012. Prom week. In *Proceedings of the 7th International Conference on Foundations of Digital Games (FDG’12)*, pages 1–8, Raleigh, NC, USA. ACM.
- Josh McCoy, Mike Treanor, Ben Samuel, Noah Wardrip-Fruin, and Michael Mateas. 2011b. Comme il faut: A system for authoring playable social models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE’11)*, pages 38–43, Stanford, CA, USA. AAAI.
- Daniel McFadden. 1974. Conditional logit analysis of qualitative choice behavior. pages 105–142.
- Marvin Minsky and Seymour Papert. 1970. Draft of a proposal to arpa for research on artificial intelligence at mit, 1970–71.
- Shohei Miyashita, Xinyu Lian, Xiao Zeng, Takashi Matsubara, and Kuniaki Uehara. 2017. Developing game ai agent behaving like human by mixing reinforcement

- learning and supervised learning. In *Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 153–158, Kanazawa, Japan.
- Alexander Nareyek. 2007. Game ai is dead. long live game ai! *IEEE Intelligent Systems*, 22(1):9–11.
- National Opinion Research Center. 2023. General social survey, 2023.
- Allen Newell. 1990. *Unified Theories of Cognition*. Harvard University Press, Cambridge, Massachusetts.
- B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E.-J. Wagenmakers, R. Wilson, and T. Yarkoni. 2015. Promoting an open research culture. *Science*, 348:1422–1425.
- Peter O’Donovan, Aseem Agarwala, and Aaron Hertzmann. 2015. Designscape: Design with interactive layout suggestions. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1221–1224.
- OpenAI. 2022. Introducing chatgpt. <https://openai.com/blog/chatgpt>. Accessed on: 2023-04-03.
- OpenAI. 2024a. Gpt-4.
- OpenAI. 2024b. Text to speech guide.
- OpenAI. 2024c. Whisper.
- Kyle Orland. 2021. So what is ‘the metaverse’, exactly? *Ars Technica*.

- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.
- J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM.
- Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022a. Social simulacra: Creating populated prototypes for social computing systems. In *In the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*, UIST '22, New York, NY, USA. Association for Computing Machinery.
- Joon Sung Park, Joseph Seering, and Michael S. Bernstein. 2022b. Measuring the prevalence of anti-social behavior in online communities. *Proceedings of the ACM on Human-Computer Interaction*, (CSCW).
- Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people.
- Richard W. Pew and Ann S. Mavor, editors. 1998. *Modeling Human and Organizational Behavior: Applications to Military Simulations*. National Academy Press, Washington, D.C.
- Roberto Pilloso. 2009. Coordinating agents with behavior trees: Synchronizing multiple agents in cryengine 2.
- Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7(2):20563051211019004.

- Prolific. 2022. Prolific: Quickly find research participants you can trust.
- T. S. Rai, P. Valdesolo, and J. Graham. 2017. Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences of the United States of America*, 114:8511–8516.
- Steven F. Railsback and Volker Grimm. 2019. *Agent-Based and Individual-Based Modeling: A Practical Introduction*, 2 edition. Princeton University Press, Princeton, NJ.
- A. Rapoport and A. M. Chammah. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation*. University of Michigan Press.
- Reddit. 2021. Reddit by the numbers.
- B. Reeves and C. Nass. 1996a. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.
- Byron Reeves and Clifford Nass. 1996b. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Marc Rettig. 1994. Prototyping for tiny fingers. *Communications of the ACM*, 37(4):21–27.
- Mark O. Riedl. 2012. Interactive narrative: A novel application of artificial intelligence for computer games. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*, pages 2160–2165.
- Horst WJ Rittel and Melvin M Webber. 1973. Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169.
- David Rolf. 2015. *The Fight for \$15: The Right Wage for a Working America*. The New Press.

- Xin Rong, Shiyan Yan, Stephen Oney, Mira Dontcheva, and Eytan Adar. 2016. Code-mend: Assisting interactive programming with bimodal embedding. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 247–258.
- M. J. Salganik et al. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 117:8398–8403.
- Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.
- Valkyrie Savage, Colin Chang, and Björn Hartmann. 2013. Sauron: embedded single-camera sensing of printed physical user interfaces. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 447–456.
- T. C. Schelling. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1:143–186.
- T. C. Schelling. 1978. *Micromotives and Macrobehavior*. W. W. Norton & Company.
- O. Schilke, M. Reimann, and K. S. Cook. 2015. Power decreases trust in social exchange. *Proceedings of the National Academy of Sciences of the United States of America*, 112:12950–12955.
- S. C. Schmitt, J. J. Gaughan, B. N. Doritya, A. L. Gonzalez, L. D. Smillie, R. E. Lucas, D. B. Nelson, and M. Brent Donnellan. 2023. The big five across time, space, and method: A systematic review. PsyArXiv Preprint.
- D.A. Schön. 1984. *The Reflective Practitioner: How Professionals Think In Action*, volume 5126 of *Basic Books*. Basic Books.
- Reinhard Sefelin, Manfred Tscheligi, and Verena Giller. 2003. Paper prototyping-what is it good for? a comparison of paper-and computer-based low-fidelity prototyping. In *CHI’03 extended abstracts on Human factors in computing systems*, pages 778–779.

- Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.
- Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions*, 4(6):42–61.
- N. C. Silver and W. P. Dunlap. 1987. Averaging correlation coefficients: Should fisher’s z transformation be used? *Journal of Applied Psychology*, 72(1):146–148.
- Ho Chit Siu, Jaime Peña, Edenna Chen, Yutai Zhou, Victor Lopez, Kyle Palko, Kimberlee Chang, and Ross Allen. 2021. Evaluation of human-ai teams for learned and rule-based agents in hanabi. In *Advances in Neural Information Processing Systems*, volume 34, pages 16183–16195. Curran Associates, Inc.
- T. W. Smith, M. Davern, J. Freese, and S. L. Morgan. 2021. General social surveys, 1972–2020: Cumulative codebook.
- Carolyn Snyder. 2003. *Paper prototyping: The fast and easy way to design and refine user interfaces*. Morgan Kaufmann.
- Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *J. Am. Soc. Inf. Sci. Technol.*, 63(2):270–285.
- Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- M. N. Stagnaro, J. Druckman, A. J. Berinsky, A. A. Arechar, R. Willer, and D. G. Rand. 2024. Representativeness versus response quality: Assessing nine opt-in online survey samples. OSF Preprints.
- Stanford Center on Poverty and Inequality. 2021. American voices project.

- William Swartout, Jonathan Gratch, Randall Hill, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum. 2006. Toward virtual humans. *AI Magazine*, 27(1).
- Don Tapscott and Art Caston. 1992. *Paradigm shift: The new promise of information technology*. McGraw-Hill, Inc.
- Nathan TeBlunthuis, Aaron Shaw, and Benjamin Mako Hill. 2018. Revisiting" the rise and decline" in a population of peer production projects. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–7.
- David R. Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2):237–246.
- Frank Thomas and Ollie Johnston. 1981. *Disney Animation: The Illusion of Life*. Abbeville Press, New York.
- Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. 2013. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337.
- Ilshat Umarov, Mikhail Mozgovoy, and Patrick C. Rogers. 2012. Believable and effective ai agents in virtual worlds: Current state and future perspectives. *International Journal of Gaming and Computer-Mediated Simulations*, 4(2):37–59.
- Graham Upton and Ian Cook. 2006. *A Dictionary of Statistics*, 2 edition. Oxford University Press, Oxford, United Kingdom.
- Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, and et al. 2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575:350–354.
- Emily A Vogels. 2021. The state of online harassment. *Pew Research Center*, 13.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

- Mark Weiser. 1999. The computer for the 21st century. *SIGMOBILE Mob. Comput. Commun. Rev.*, 3(3):3–11.
- R. S. Weiss. 1994. *Learning From Strangers: The Art and Method of Qualitative Interview Studies*. Free Press.
- Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.
- Paul Windrum, Giorgio Fagiolo, and Alessio Moneta. 2007. Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, 10(2):8.
- Terry Winograd. 1971. *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. Ph.D. thesis, Massachusetts Institute of Technology.
- Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021a. Recursively summarizing books with human feedback.
- Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. Promptchainer: Chaining large language model prompts through visual programming. *arXiv preprint arXiv:2203.06566*.
- Tongshuang Wu, Michael Terry, and Carrie J Cai. 2021b. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. *arXiv preprint arXiv:2110.01691*.
- Ya Xu, Nanyu Chen, Addrian Fernandez, Omar Sinno, and Anmol Bhasin. 2015. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2227–2236.
- Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In

Proceedings of the 2020 chi conference on human factors in computing systems, pages 1–13.

Georgios N. Yannakakis. 2012. Game ai revisited. In *Proceedings of the 9th Conference on Computing Frontiers*, pages 285–292, Cagliari, Italy. ACM.

Indi Young. 2016. Describing personas.

Robert Zubek. 2002. Towards implementation of social interaction. In *AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*. AAAI Press.