Received: 20 November 2024 | Revised: 9 July 2026 | Accepted: 17 July 2026

DOI: 10.1002/smj.70112

RESEARCH ARTICLE

SMS | Strategic Management Journal

WILEY

Simulating strategic interactions with AI agents

Matteo Tranchero[1] | **Cecil-Francis Brenninkmeijer**[2] | **Arul Murugan**[3] | **Abhishek Nagaraj**[4]

[1]The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Fuqua School of Business, Duke University, Durham, North Carolina, USA

[3]School of Information, University of California, Berkeley, Berkeley, California, USA

[4]Haas School of Business, University of California, Berkeley, Berkeley, California, USA

Correspondence
Matteo Tranchero, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA.
Email: mtranc@wharton.upenn.edu

Funding information
Open AI

Abstract

Research Summary: We explore how Large Language Models (LLMs) can serve as synthetic subjects to inform strategy research. We introduce a framework for designing and running simulated experiments with LLM-powered agents. We argue that this approach is useful for rapid, low-cost prototyping of human experiments and for generating novel hypotheses. We apply the framework to the exploration–exploitation dilemma and show that LLM-based experiments reproduce patterns observed among human participants. We then vary parameters and boundary conditions to illustrate how the same setup can support design iteration and surface hypotheses about when and why established results change. In the conclusion, we discuss the promise and limitations of artificial intelligence agents as "model organisms" for strategy.

Managerial Summary: Artificial intelligence (AI) agents are beginning to enter firms as tools that can execute work, from writing code to coordinating complex tasks across systems. This article argues that their value for strategy extends beyond task

automation: AI agents can also be used to simulate strategic interactions and assess how strategies perform under alternative assumptions. In our exploration–exploitation application, these simulations reproduce core patterns from prior human experiments and reveal where those patterns weaken or reverse. Used this way, AI agents can help firms prototype strategic choices, stress-test assumptions, and direct managerial attention toward promising leads before larger commitments of time and effort.

KEYWORDS

AI agents, exploration and exploitation, Large Language Models, strategic interactions, strategy experiments

1 | INTRODUCTION

Whether practiced by managers or studied by scholars, strategy hinges on understanding how actions translate into competitive outcomes in dynamic environments (Leiblein et al., 2018). Just as researchers investigate the returns from research and development (R&D), company executives want to know whether their innovation strategy will boost sales and how competitors are likely to respond. But this is easier said than done. In most business settings, multiple actors shape outcomes, and their decisions are often interdependent. Consider a pharmaceutical firm deciding which drug to develop. Sticking with known molecules may yield predictable but modest returns, while exploring new compounds involves substantial risks. And even if the bet pays off, rivals may quickly follow with cheaper alternatives that draw customers away, ultimately shrinking, not growing, profits.

Strategic interdependencies make it difficult to trace clear causal links between actions and outcomes. The ideal way to isolate mechanisms is through real-world experiments at the company or even industry level, but these are rarely feasible (Adner & Levinthal, 2024). As a result, researchers usually turn to two alternative approaches to run strategy experiments. The first relies on computational models to simulate scenarios that include nonlinear effects and feedback loops, which closed-form models struggle to capture (Davis et al., 2007). These models offer *control*, enabling researchers to tweak key parameters and observe their impact on emergent dynamics (Knudsen et al., 2019). The second involves human-subjects experiments, either in the laboratory or with partner organizations (Di Stefano & Gutierrez, 2019; Levine et al., 2023). This approach emphasizes *realism* by capturing how actors behave under specific conditions, while holding other strategic interactions constant (Chatterji et al., 2016).

Experiments using either agent-based simulations or human subjects offer advantages, but each entails a distinct trade-off between control and realism. On the one hand, the clean abstraction provided by simulated agent-based models (ABMs) increases the distance from real scenarios and complicates translating their insights to managerially relevant situations (Ganco, 2017; Knudsen, 2024). On the other hand, experiments with human participants generate data on real-world behavior but require substantial budgets and long lead times for

implementation, which constrain iteration across alternative designs or choices (Davis et al., 2007). As a result, human-subject experiments are less suited to mapping how outcomes vary across broad parameter spaces.

Recent advances in generative artificial intelligence (AI) herald a new approach that could complement these established methodologies. Large Language Models (LLMs), trained on a vast corpus of human text, can emulate human language with remarkable fluency. This capability also enables them to approximate human behavior to a reasonable degree (Horton, 2023; Tu et al., 2024). As a result, scholars have begun to deploy LLMs as proxies for human behavior, an approach referred to as "synthetic subjects" or "silicon sampling" (Bail, 2024; Dillion et al., 2023). Moreover, LLMs can be instructed to assume a wide range of roles and, as digital artifacts, be embedded in fully manipulable virtual environments. This opens the door to creating ecologies of realistic yet synthetic AI agents that interact within contexts mimicking the strategic interdependencies of real-world settings, offering a novel tool for both research and practice.

In this paper, we propose a general framework for using AI agents as synthetic subjects to study strategic interactions. Our approach is inspired by model organisms in the biomedical sciences: species sharing enough biological similarity to humans that allow researchers to refine initial theories before advancing to human testing (LaFollette & Shanks, 1995). In the same spirit, AI agents can support rapid, low-cost prototyping of experimental designs, while also helping surface unexpected patterns that can be treated as hypotheses for subsequent validation. We then apply our framework to a simple exploration–exploitation problem involving strategic interdependencies. First, we show that our simulated experiments reproduce outcomes from human-subject experiments, suggesting a promising degree of realism for AI agents. Second, we demonstrate how AI agents enable researchers to rapidly and cheaply prototype new experiments. Third, we illustrate how these simulations can surface applicable theoretical constructs that a researcher may not have considered.

In our framework, human strategists define a "starting theory" (Davis et al., 2007) or "thesis" (Ott & Hannah, 2024). Human judgment is required to identify links between actions and outcomes worth exploring, and to specify the relevant payoffs and objectives (Agrawal et al., 2019). Once this theoretical space is articulated, AI agents provide a flexible tool for in silico experimentation. Strategists can instantiate multiple LLM-powered non-deterministic agents and embed them in environments that feature strategic interdependence and feedback loops. They can then manipulate parameters, observe emergent interaction patterns, and generate predictions about the direction and relative strength of causal relationships. Analogous to how the use of model organisms for preclinical screening complements in vitro testing and clinical trials, AI agents can inform the development of ABMs and help prototype human-subject experiments.

Leveraging AI agents as the "laboratory mice" for strategy brings several opportunities. A key strength of LLM-based experiments is their flexibility, as agents can be primed to adopt diverse and nuanced personas. Researchers can easily vary goals and incentives across agents and embed role-specific knowledge or constraints directly in the script. Because the agents can interpret context and generate coherent responses, they can exhibit forms of strategic interaction that are hard to represent in traditional ABMs, where tractability often requires stylized decision-rules. Even when LLM agents merely approximate human responses, the method retains many of the core advantages of computer simulations: it is fast, low-cost, and easily repeatable, which allows researchers to probe a large design space and identify specific hypotheses and boundary conditions that merit subsequent validation.

However, this approach also has important limitations. LLMs can exhibit a rationality bias, deviating from the bounded rationality typical of human decision-making (Hagendorff et al., 2023). They are also prone to hallucination, so their stated motivations may not correspond to stable underlying mechanisms. More fundamentally, observations produced in AI agent simulations are not data from real human subjects. They reflect the behavior of a particular model under a specific prompt and environment design, underscoring the need for validation before drawing substantive managerial or theoretical claims. Even in terms of control, LLMs remain imperfect: they are ultimately black boxes, making it difficult to precisely manipulate their decision-making as traditional ABMs allow. In other words, AI simulations are not a golden mean between computational models and experiments. Instead, we argue that this method provides a novel tool in the researcher's toolkit, with its own advantages and limitations that complement existing approaches.

Our paper makes several contributions. First, we propose a general-purpose framework for simulating strategic interactions with AI agents. Second, we highlight both the potential and the limitations of AI agents for strategy research, building on the rapidly growing literature on AI simulators in computer science and related fields (Aher et al., 2023; Dillion et al., 2023; Horton, 2023; Park et al., 2023). We specify when AI agent simulations are most informative, how their outputs should be interpreted, and what safeguards can improve credibility. As such, we describe best practices for using AI agent simulations as one tool within a broader strategy research toolkit. Finally, our application to the dynamics of exploration and exploitation demonstrates how AI agents can be used in practice, demonstrating the pragmatic utility of our approach and pointing to broader opportunities for strategy research.

2 | AI AGENT SIMULATIONS FOR STRATEGY RESEARCH

2.1 | Does strategy research need a new tool?

Strategic management aims to understand business dynamics and offer normative guidance to improve firm performance (Rumelt et al., 1991). A central challenge, however, is that strategic choices are often deeply interconnected, making it hard to pinpoint which mechanisms drive outcomes. While real-world experimentation remains the gold standard, it is infeasible to do so with many business decisions. As a solution, strategy research is distinctive in its reliance on two complementary methods: in silico experiments with ABMs[1] and randomized controlled trials (RCTs) with human subjects either in the lab or the field (Chatterji et al., 2016; Di Stefano & Gutierrez, 2019; Knudsen et al., 2019). To better understand how these methods are used in research, we manually conducted a systematic review of studies appearing in the *Strategic Management Journal* (SMJ)—the field's longest-running and most representative outlet. Using Scopus, we identified 80 relevant articles published between 1980 and 2025: 49 using ABMs and 31 employing human-subjects experiments (details in Supporting Information S1: Appendix A).

[1]While mathematical models are used in strategy research (e.g., Van den Steen, 2018), they often fall short in capturing the complexity and heterogeneity inherent in strategic interactions (Harrison et al., 2007; Knudsen, 2024). To overcome these limitations, scholars frequently turn to computational simulations when analytical solutions become intractable. However, the distinction between mathematical and computational models is largely one of degree: both rest on the same epistemic logic, but simulations offer greater flexibility to expand state spaces and capture richer patterns of interaction and adaptation (Knudsen et al., 2019).

Common topics across these papers include cognition and decision-making, competitive and corporate strategy, strategic human capital, adaptation, and knowledge-related themes.[2]

In Panel (a) of Figure 1, we break down the studies by the method used. Although both methods are applied to most questions, notable differences appear across research topics. The most popular areas, such as decision-making and the implementation of strategies in corporate settings, show a relative balance between simulations and human-subject experiments. Other topics, including competition, adaptation, and networks, are more often studied through simulations. We speculate that this pattern reflects structural constraints: it is difficult to experimentally study these dynamics with human subjects, and even more challenging to manipulate competitive environments in ways that align with theoretical models. In contrast, areas like strategic human capital and entrepreneurship are more commonly studied through human-subject experiments—likely because they involve behaviors that are harder to model, and because they aim for empirical, policy-relevant estimates.

Beyond the substantive research topic, we also reviewed each study to identify which dimensions were experimentally manipulated. We inductively developed a taxonomy that distinguishes between manipulations to the agent and changes to the environment, and then applied it to code each study (see Supporting Information S1: Appendix A). We identify three sources of agentic variation (objectives, capabilities, and decision processes) and six sources of environmental variation (population, time horizon, choice landscape, payoff structure, information set, and spatial structure). Overall, we find that computational simulations make use of significantly more variation across both agentic and environmental dimensions. On average, a paper developing an ABM manipulates 3.7 categories, while human experiments vary only 2.1 of them. This pattern supports the view that simulation methods get closer to offering a form of "perfect control" relative to working with human subjects (Harrison et al., 2007; Knudsen, 2024).

Panel (b) of Figure 1 reveals additional heterogeneity across methods. Simulations are more likely to introduce changes to the payoff structure, time horizon, agent population, and agent characteristics. Interestingly, some categories appear less amenable to computational manipulation. This may reflect the mechanistic nature of ABMs, which makes variations in elements such as agent objectives or information sets less theoretically interesting. Human experiments, by contrast, face stricter constraints on what can be manipulated, particularly in the features of the environment or the composition of the actor population. Because causal claims ultimately require evidence from real-world behavior, these constraints limit the ability to validate and calibrate the predictions generated by more flexible ABMs, especially when the underlying design space is high-dimensional. This difficulty helps explain why human-subject experiments are often confined to testing simple relationships, while empirical tests of ABMs remain rare.[3] Taken together, these patterns motivate the development of novel tools that expand the range of feasible manipulations in strategy experiments.

[2]Unlike the field of economics, strategy lacks a standardized classification system such as the *Journal of Economic Literature* (JEL) codes to sort articles by topic. For our purposes, we inductively developed thematic categories to capture each study's primary focus. These topical categories are intended to give a general sense of where ABMs and human experiments tend to be used in strategy research. The full database of the studies we analyzed is available in Supporting Information S1, allowing readers to replicate our analysis or use alternative categorizations.

[3]In the few cases where such experimental translation has been attempted, predictions from ABMs have shown mixed results, especially for widely used frameworks such as NK models (Billinger et al., 2014; Ganco, 2017).

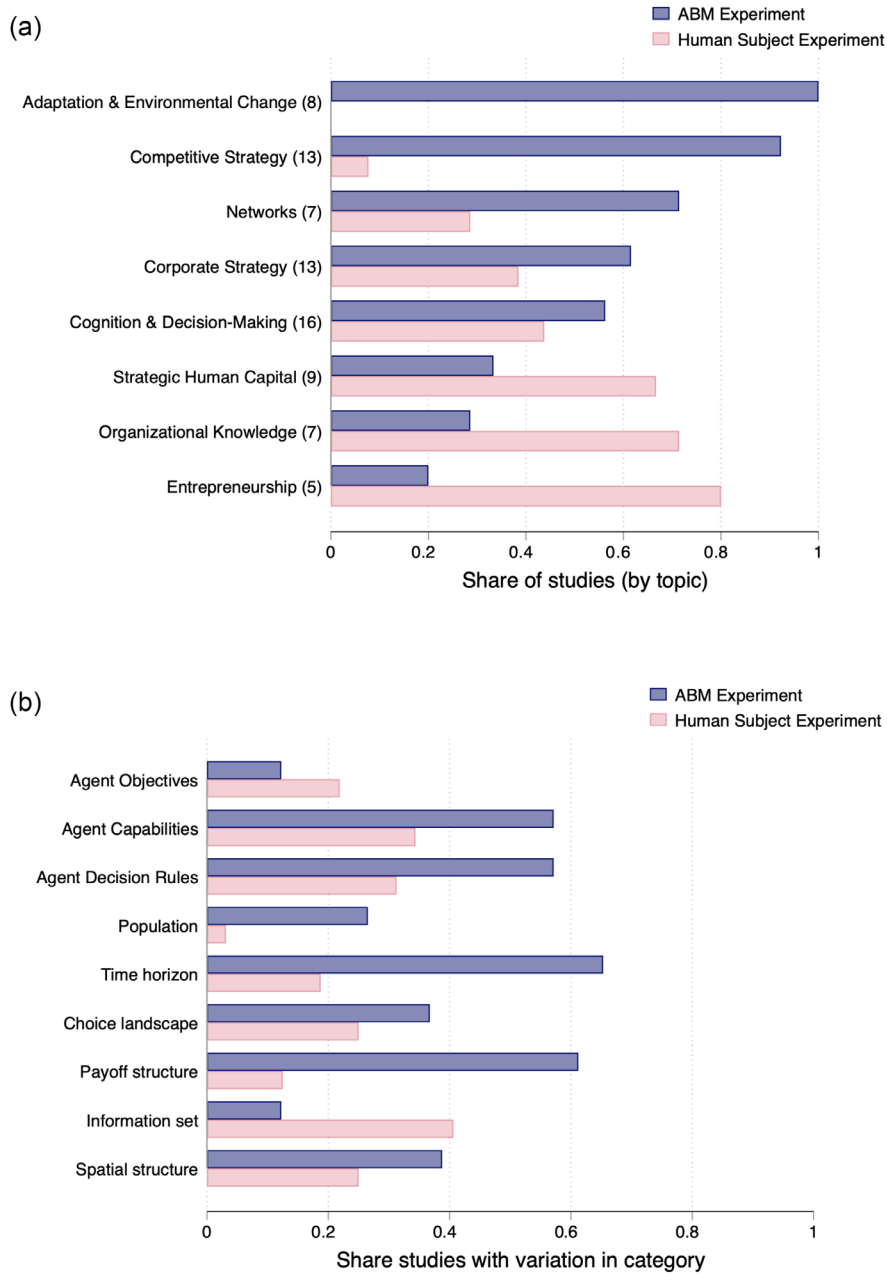

FIGURE 1 Agent-based and human-subject experiments published in the *Strategic Management Journal* (1980–2025). Panel (a): Research topics studied by methodological approach. Panel (b): Types of experimental manipulation by methodological approach. This figure summarizes our review of articles published in the *Strategic Management Journal* that use either simulation-based or human-subject experiments. Panel (a) shows the proportion of papers in each research topic by methodological approach. Panel (b) presents the proportion of studies that experimentally manipulate different categories of variation. For details on our study selection and classification procedures, see Supporting Information S1: Appendix A. ABM, agent-based models.

2.2 | The emergence of AI simulations

Recent advancements in generative AI have opened up many applications across industries. LLMs have gained popularity due to their impressive performance and economic value across domains (Bubeck et al., 2023). Beyond their commercial applications, LLMs are also being adopted as tools in scientific research (e.g., Carlson & Burbano, 2026). As shown in Supporting Information S1: Appendix B, most applications involve supporting research indirectly—through tasks such as dataset cleaning, scientific writing, and refining research instruments. However, LLMs are also beginning to play a more direct role in research, including generating scientific hypotheses (Manning et al., 2024; Si et al., 2024).[4]

We anticipate that a central role of LLMs in strategy research will be as synthetic subjects (Bail, 2024; Grossmann et al., 2023; Park et al., 2023). LLMs are neural networks trained using deep learning on vast amounts of human-generated corpora, enabling them to generate human-like text in response to prompts. This capability follows from how these models are built. LLMs are pre-trained on a vast corpus of human text to predict language. Language encodes much of how people reason, evaluate, and decide. Therefore, learning to model it implicitly leads models to develop internal representations of the social world (Bubeck et al., 2023; Gurnee & Tegmark, 2024). Further, modern models are refined through post-training with reinforcement learning, often with humans in the loop, which rewards outputs that align with human preferences and judgments (Ouyang et al., 2022). The result is a system whose behavior reflects a compressed, recombinable representation of how humans tend to act.

As a result, AI-powered agents exhibit strikingly human-like traits. When evaluated with personality tests and behavioral games, their responses are statistically indistinguishable from those of randomly selected human participants (Mei et al., 2024). Additional studies reveal that LLMs lean on heuristics similar to those used by humans and reproduce their cognitive biases, reasoning errors, and moral intuitions (Hagendorff, 2024; Lampinen et al., 2024). Horton (2023) demonstrates LLMs' strategic acuity by placing them in classic interactions such as the prisoner's dilemma. This methodology is gaining traction in computational social science, where LLMs serve as generative AI simulators that produce "silicon samples," or fully synthetic datasets, to study specific questions (Dillion et al., 2023; Sarstedt et al., 2024).

Despite their promise, recent studies also highlight the limitations of AI as a synthetic subject (Aher et al., 2023; Gao et al., 2025). First, LLMs have access to pre-existing research in their training data, raising the possibility that their outputs parrot memorized content. This is problematic when directly prompting an LLM, for instance, to predict the outcome of prospective studies, although recent evidence does seem to downplay this concern (Luo et al., 2025). Second, while LLMs avoid some ethical constraints of human-subjects research, they introduce new concerns, such as the spontaneous emergence of deceptive behavior (Hagendorff, 2024). Finally, AI agents remain black boxes, making it difficult to determine when they provide a credible proxy for human behavior. Their behavior can shift in response to seemingly minor changes in prompts or model versions, and some discrepancies with human behavior appear

[4]Our focus is on how LLMs can be used in research (see Supporting Information S1: Appendix Table B1) and, more generally, strategizing endeavors. However, it must be noted that an emerging literature is exploring the direct impact of AI on firm activities and performance (Csaszar et al., 2024; Dell'Acqua et al., 2023; Doshi et al., 2025). Relatedly, another strand of research studies the extent to which generative AI can replace humans in carrying out a variety of tasks and the related labor-market implications (Eloundou et al., 2024; Felten et al., 2023).

systematic (Jones & Steinhardt, 2022). Determining when an LLM provides a credible proxy for human responses is, therefore, an ongoing research challenge (Broska et al., 2024).

2.3 | AI agents as model organisms for strategy

Given concerns about LLMs' ability to mimic human behavior, it remains an open question whether they are suitable for strategy research. A useful starting point is to draw a parallel with other fields that study highly complex phenomena. One of the clearest examples is biomedicine, which grapples with one of the most intricate and interactive systems we know: the human body.[5] In particular, scientists have long used *model organisms* as tractable experimental systems for examining biological interactions that cannot be easily studied in humans.

Model organisms, such as the lab mouse, are considered sufficiently analogous to humans to generate insights about human biology (LaFollette & Shanks, 1995). Their primary value lies in the lower cost and greater flexibility they offer for experimentation, permitting researchers to screen mechanisms and refine interventions before committing resources to clinical studies. They complement other tools in biomedical research, including in vitro studies focused on isolating mechanisms. At the same time, insights from model organisms are not interpreted as direct causal evidence about humans. Their biology differs from humans in consequential ways, and key mechanisms are often only partially observable, which limits straightforward translation. Even so, model organisms play a useful role by generating testable hypotheses and helping prioritize interventions to evaluate in the clinic.

We argue that LLMs can serve a similar role for strategy as model organisms play in biomedicine. Like laboratory mice, AI agents offer an imperfect but useful approximation for rapidly prototyping human experiments at low cost (Tjuatja et al., 2023). The analogy, however, also highlights a key limitation. A mouse is informative about human biology because it shares an evolutionary history with humans, meaning that several biological mechanisms shaping its behavior and physiology have counterparts in humans. LLMs lack this shared evolutionary basis; instead, their relevance stems from being trained on vast corpora of human-generated text, enabling them to reproduce many regularities in how people reason and respond. Whether those regularities arise from mechanisms comparable to human cognition remains uncertain, limiting the extent to which their behavior can be generalized to humans.

This distinction clarifies the role we envision for AI agent simulations in strategy research. Rather than providing evidence about the mechanisms underlying human behavior, they provide a computational environment for exploring the implications of existing theories. Once a researcher has articulated a starting theoretical argument, AI agents can be used to examine the face validity of the theory, probe alternative assumptions, and experimental designs before committing resources to human subjects or formalizing those extensions in transparent ABMs. In this sense, AI agent simulations complement traditional computational experiments by replacing fully specified behavioral rules with agents that can generate context-sensitive responses, thereby helping researchers identify hypotheses that merit subsequent validation.

[5]Further underscoring the parallel, it is no coincidence that the NK simulation approach (so widely adopted in strategy) was originally developed by Kauffman (1993) in evolutionary biology.

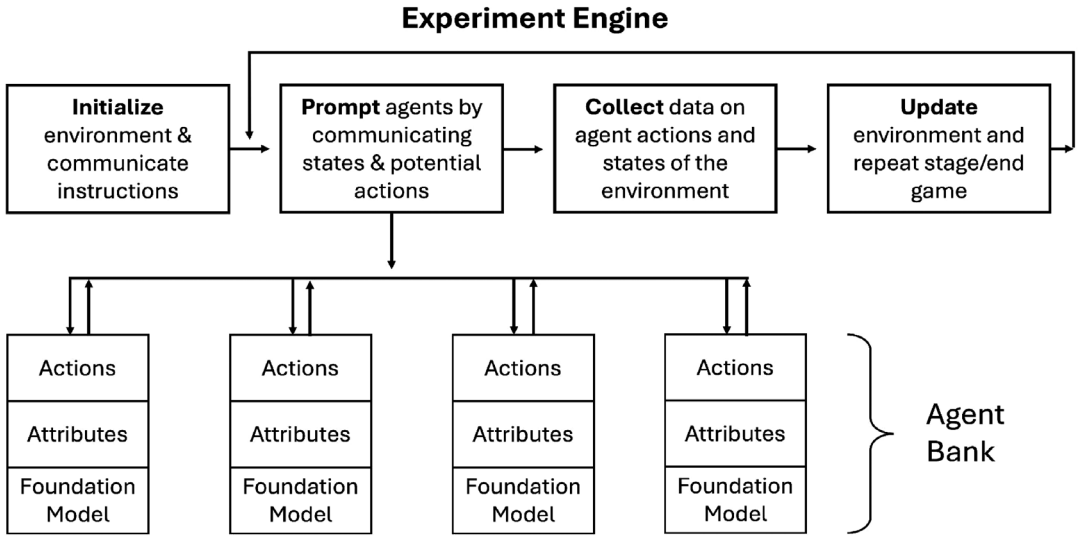

FIGURE 2 Stylized framework to run strategy experiments with artificial intelligence (AI) agents. This figure presents a stylized depiction of the central deterministic engine that powers an interactive simulated experiment. The engine spawns the AI agents (each consisting of an action space, set of attributes, and an underlying foundation model) and then administers the experiment in silico following the four steps above.

3 | A FRAMEWORK FOR SIMULATING STRATEGIC INTERACTIONS WITH AI AGENTS

In this section, our goal is to provide a general framework that enables strategy scholars to simulate experiments where multiple AI agents interact in a rich strategic environment. Our approach is bottom-up, placing micro-level autonomous AI agents in controlled settings where they can interact organically with one another. These interactions can produce emergent behaviors that the researcher did not anticipate from the outset and that are unknown to the AI model itself. Throughout the simulations, each agent's decisions, rationales, and resulting outcomes are carefully recorded, enabling an in-depth study of strategic interactions.

Because these experiments are conducted entirely in silico, they require some basic computational infrastructure (detailed in Supporting Information S1: Appendix C). The researcher must first set up a central (often deterministic) engine that administers the experimental process with AI agents. This engine acts as the conductor of the orchestra, interacting with AI agents, assigning their roles and behaviors, delivering instructions, updating information as the experiment unfolds, and recording outcomes. Figure 2 provides a stylized overview of this process.[6] Because the core functions are agnostic to the specific experiment, the infrastructure is highly adaptable and requires only minor adjustments when applied to new studies. Once in place, the researcher can run a virtually unlimited number of simulations. This engine-based

[6]Researchers have flexibility in how to build the engine and can implement it in nearly any coding environment. One option is to use tools such as Expected Parrot Domain-Specific Language (EDSL), an open-source Python package that connects user prompts to the GPT-4 API (Horton & Horton, 2024). See Supporting Information S1: Appendix C for more details.

approach is well-suited for studying complex, repeated interactions among multiple agents. It can even be adapted to study mixed settings where human and AI agents interact.

Researchers must then specify the key parameters of the experiment. Because the entire setup is a computer simulation, researchers can manipulate all of the degrees of freedom outlined in Panel (b) of Figure 1. This includes variation at the agent level, such as goals, capabilities, and decision-making rules, which can be implemented by prompting the LLMs with additional scripts. It also includes variation in the environment, such as the number of agents, time horizon, choice landscape, payoff structure, information set, and spatial features. In addition, the researcher can determine how the experiment is administered, including whether decisions are made sequentially or simultaneously and whether and how agents are allowed to communicate. The main degrees of freedom available to the researcher are described in Supporting Information S1: Appendix D. As with lab experiments and ABM simulations, these design choices should be grounded in the starting theory that motivates the research question (Davis et al., 2007; Ott & Hannah, 2024).

In terms of data collection, researchers can instrument and record every part of the process, including quantitative data for econometric analysis or qualitative responses. The qualitative data is particularly interesting for hypothesizing what could drive observed results. The simplest approach is to ask the model to explain its decisions through an explicit prompt. These explanations can be enriched through chain-of-thought prompting, which records intermediate reasoning steps and can reveal how decisions unfold (Wei et al., 2022).[7] At the same time, these explanations should not be considered as observational evidence about human cognition. LLMs may not reliably "know" why they produced a given output (not unlike humans). Indeed, recent work shows that chain-of-thought traces need not faithfully represent the computation that actually produced a model's output (Lanham et al., 2023; Turpin et al., 2023). These rationales are therefore best treated as interpretable outputs to be validated, not as a direct window into the model's underlying reasoning. Ultimately, it remains the researcher's responsibility to interpret the results and draw insight from them.

4 | APPLICATION: THE EXPLORATION–EXPLOITATION DILEMMA

4.1 | A theory of exploration with data spillovers

We illustrate the use of AI agents as synthetic subjects through a canonical problem in strategy. In many innovation and entrepreneurship settings, actors must choose from a set of options with uncertain value. A typical example is pharmaceutical firms selecting which genes to target for drug development (Tranchero, 2026). Some options may have known values, in which case the actors must decide whether to exploit them or explore new ones (March, 1991). Although exploring can be an individual decision, it can also depend on the actions of others (Lazer & Friedman, 2007), introducing strategic interdependence. For example, firms frequently learn from their competitors and must decide whether to build on existing innovations or to invest in

[7]Although not yet common, researchers may soon be able to study how outcomes shift in response to direct changes to the foundation model itself. As the field of model interpretability advances, it may become possible to probe specific architectural components or activation patterns and examine how they shape behavior (Lindsey et al., 2025).

R&D themselves, knowing that the results may later become public. What strategy should a manager pursue in such settings?

To shed light on this question, a recent paper studied an exploration–exploitation setting with informational spillovers (Hoelzemann et al., 2025). In this paper, the authors present a novel game, where multiple agents explore across projects that could be of either Low (L), Medium (M), or High (H) value. Agents can see the actions of others and learn about project values, but otherwise do not have any information on project quality. In this setting, the authors examine the consequences of shedding light on the value of one project. The counterintuitive result is that public data revelations can make all agents worse off than if no information had been shared. The key mechanism is one where information revelation causes herding and too little exploration compared to settings where agents have to explore in the dark. The authors supported this prediction via an online lab experiment, with groups of undergraduate students who were incentivized to play this game. This work provides a compelling starting theory for studying what is often referred to as the *streetlight effect*, or the tendency to search where data is easiest to access, even when doing so leads to inferior outcomes (Vuculescu, 2017).[8]

We focus on this study for several reasons. First, the online lab experiment features strong strategic interdependencies and uncertainty. These elements are common in real-world settings, making it an ideal test case for evaluating whether LLM-based simulations can contribute meaningfully to strategy research. Second, the experiment relies on a single set of assumptions and parameters, leaving room for further exploration of the parameter space.[9] We take up that task by using our AI agent simulation framework to first replicate and then extend the original findings. Third, the study is grounded in a closed-form mathematical framework. This allows us to benchmark both human and LLM behavior against theoretical optima, providing a precise context for interpretation. Finally, a common concern in studies that replicate human-subject experiments is that LLMs may simply reproduce results seen during training. In our case, we verified that GPT-4 is unfamiliar with the experiment, which increases confidence that our simulations represent out-of-sample predictions.[10]

4.2 | The lab experiment: Searching mountains for hidden gems

The experiment in Hoelzemann et al. (2025) features a group of participants searching across a range of virtual mountains for hidden gems. The setup is shown in Supporting Information S1: Panel (A) of Appendix Figure E1. There are $n = 5$ players and $m = 5$ mountains. Mountains contain topazes with 60% probability and rubies and diamonds with 20% probability each. While the dollar value of each gem varies by round, the diamond is always valued more than the ruby, which in turn is valued more than the topazes. The location of these gems is unknown from the outset, and players only know the probability of finding gems and their values. The experiment consists of two periods: in the first period, each player selects a mountain in sequential order. After everyone

[8]This term comes from the parable of a drunkard who searches for his keys beneath a streetlamp because that is "where the light is," even though he knowingly lost them elsewhere.

[9]An additional advantage is that we have full access to the original materials and results. In most replication studies, researchers must approximate the original design as closely as possible. Here, we can compare results directly and introduce extensions that remain faithful to the underlying model.

[10]We verified this by prompting GPT-4 to describe the experiment in Hoelzemann et al. (2025). The responses were inaccurate and largely hallucinated. This is consistent with the fact that the version of GPT-4 we used was last updated in December 2023, before the working paper was released.

TABLE 1 Framework to run AI agent experiments applied to Hoelzemann et al. (2025).

Parameter	Description
Environment	The engine keeps track of the game state, including which mountains have been explored, which gems lie underneath, and which agents have picked which mountain. *Population*: There are five AI agents. *Time horizon*: There are two periods. *Choice landscape*: There are five choices (i.e., mountains). *Payoff structure*: Each mountain is associated with one of three payoffs, depending on whether it holds a topaz, ruby, or diamond. *Information set*: Agents know the value and distribution of gems, but they do not know the locations. *Spatial features*: There are no spatial features to the environment.
Agent	*Objectives*: Each agent is explicitly instructed to maximize their individual earnings. *Capabilities*: Each agent can pick one mountain at a time. There is no private knowledge. *Decision-rules*: No decision-rules or algorithms for choosing mountains are specified.
Turn-taking procedures	The experiment uses sequential decision-making: each agent is allotted a "turn" to make their decision. This provides an implicit mechanism for agents to coordinate their exploration.
Communication channels	AI agents are not allowed to verbally communicate with each other during the course of the experiment.
Outcomes	We record the numerical earnings of AI agents, whether or not each group made a "breakthrough" and discovered a diamond, and the number of mountains explored. We also solicit AI agents for qualitative explanations behind their decisions.
Interventions	We make isolated changes to the information set. In particular, we have different initial data conditions. In the "no data" condition, the experiment begins with five unknown mountains, while in the other conditions, it begins with one mountain revealed as either high, low, or medium value.

Note: This table illustrates the key parameters of our framework for replicating the online lab experiment by Hoelzemann et al. (2025). See the text and Supporting Information S1: Appendix D for more details.
Abbreviation: AI, artificial intelligence.

has finished selecting, the gems behind the selected mountains will be revealed. In the second period, players choose again, this time with the knowledge of gems revealed in the first period. Each player earns the sum of the payoffs from their choices in both periods. The payoffs are non-rival, which means there is no penalty for choosing the same mountain as other players.

In the baseline condition, participants are not shown any data on the location of the gems. In three other treatment conditions, participants are provided partial initial data. These are depicted in Supporting Information S1: Panel (B) of Appendix Figure E1. In the low-value condition, the location of one topaz is revealed at the beginning of the experiment, while the same happens with the ruby in the medium-value condition and with the diamond in the high-value condition. This design allows researchers to observe how the availability (and nature) of initial payoff-relevant data influences exploration strategies and outcomes. The study involved 180 participants and was conducted over 720 rounds, with each round consisting of five players and two periods.

We reproduce this experiment using our simulation framework as shown in Table 1. Using the setup from Hoelzemann et al. (2025), we have a group of AI agents select which mountains to explore in each round. We build an experimental engine that spawns five AI agents at a time, familiarizes them with the instructions of the experiment, and prompts them sequentially for

their choice of mountains.[11] Once an AI agent chooses a mountain, the code updates the conditions of the environment and informs the other AI agents about this choice. In total, we simulate 500 rounds of the experiment.

4.3 | Comparing human subjects and AI agents

We begin by replicating the original experiment using AI agents. The primary outcome of interest is the group payoff, defined as the total of individual payoffs in a group for a given round. We plot the mean group payoff across experimental conditions, along with 95% confidence intervals (Panel [a] of Figure 3). The original results from the lab experiment appear in the upper left corner (plot i). When participants are shown the location of the medium-valued gem, they tend to earn the least. Further analysis shows that participants avoid the risk of searching for the higher-valued diamond and instead herd on the safer option (Supporting Information S1: Appendix Figure E2). This choice boosts payoffs in the early stages but ultimately leads to lower total earnings by the end of the round (Supporting Information S1:

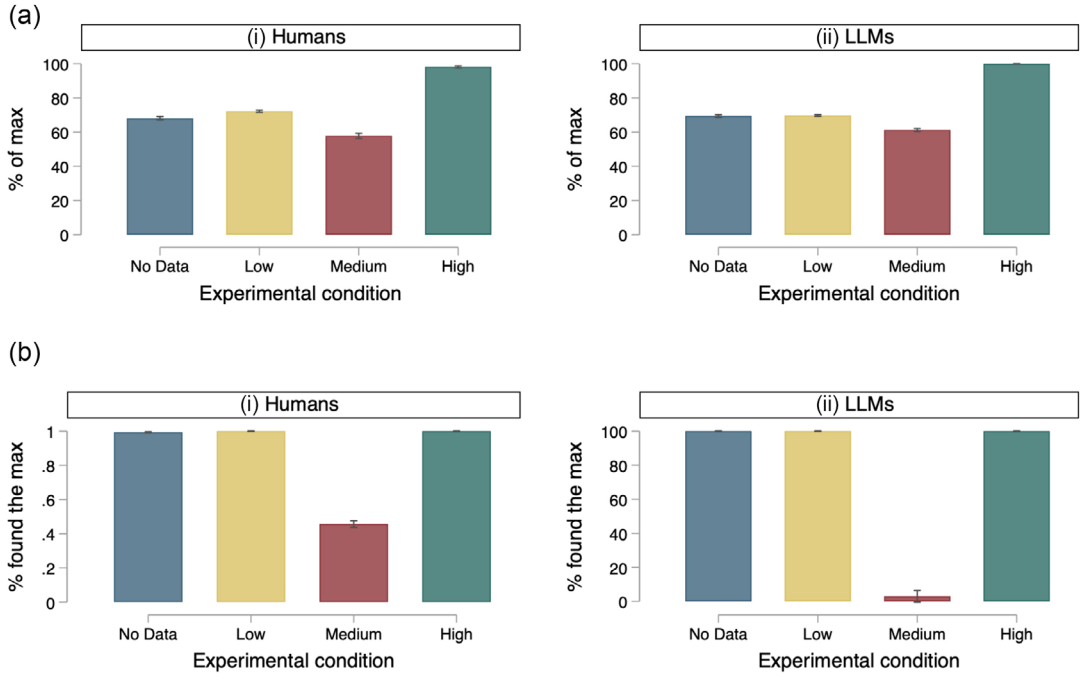

FIGURE 3 Comparing the lab experiments with human subjects and artificial intelligence (AI) agents. This figure compares the results of the lab experiment when using AI agents versus human subjects. In Panel (a), the outcome of interest is the average group earnings (as a percentage of the maximum possible earnings). In Panel (b), the outcome of interest is the likelihood of a breakthrough, which occurs when at least one group member finds the diamond. In each panel, plot (i) shows the results obtained in the original lab experiment using human subjects, while plot (ii) shows the results obtained from the simulated experiments using AI agents. (a) Group payoffs. (b) Group breakthroughs. LLMs, Large Language Models.

[11]See Supporting Information S1: Appendix F for code and other details on implementation.

Appendix Table E1). In contrast, when no information is provided or only the location of a low-value gem is revealed, participants are more likely to continue searching, increasing their chances of uncovering the diamond.

We examine how closely AI agents replicate this strategic behavior in plot (ii). The results are strikingly similar, even in this setting with complex interdependence. AI agents earn the least when shown the location of the medium-value gem. Like human participants, they choose not to explore further and instead converge on the safe option (see Supporting Information S1: Appendix Figure E3 and Table E2). When shown only the location of the topaz, or when no information is provided, the agents use the sequential order of choices to explore all options and ultimately achieve higher payoffs. When the diamond's location is revealed, they immediately select it to maximize earnings. In short, LLMs reproduce the same behavioral patterns of human participants and capture the key insight of Hoelzemann et al. (2025): more data is not necessarily beneficial in search.

Our second outcome of interest is whether the group achieves a breakthrough (Panel b), defined as at least one member discovering a diamond. Once again, the patterns are broadly similar. Both AI agents and human participants achieve a breakthrough in nearly every round where the ruby is not revealed, either because they search and find the diamond or because it is revealed to them. However, a key difference emerges when the ruby is shown. In this case, humans achieve a breakthrough in roughly 45% of rounds, while AI agents almost never do. As Hoelzemann et al. (2025) shows, breakthroughs should not occur under rational economic behavior, suggesting that LLMs adhere more closely to theoretical predictions. This finding aligns with other research showing that language models tend to behave more rationally than humans (Hagendorff, 2024), revealing a key limitation. Qualitative interviews from the original experiment indicate that some humans deviated from rational behavior out of boredom, inattention, or randomness, all tendencies that appear absent among AI agents. Even so, the overall patterns across humans and AI agents remain similar.

To further probe AI agents' behavior, we asked each LLM to explain its choices. As one agent put it:

> I choose Mountain 5 because it is revealed to contain a Ruby, which is valued at $6. The certainty of obtaining a Ruby outweighs the risk of potentially finding a Topaz, despite the possibility of finding a Diamond worth $10 in another mountain.

This suggests that LLMs follow the same behavioral logic as human decision-makers. While we cannot rule out the possibility of post hoc rationalization, a plausible explanation is that their choices reflect the stated reasoning—especially given the alignment between their rationales, decisions, and outcomes, and the patterns observed in human participants. Further supporting this interpretation, we reran the experiment using reasoning models that articulate their thought process in real time. These transcripts reveal the same logic unfolding step by step as the model works through the problem.[12] Taken together, this replication serves as a first step

[12]The following is a transcript of one reasoning LLM's internal thought process during the task: ****Evaluating mountain choices****: I need to choose which mountain to pick based on what I've learned. Mountain 1 is revealed to have a Ruby, valued at 8. Since there's only one Ruby, Mountains 2–5 must have the remaining values: one Diamond worth 12 and three Topazes worth 3 each. To maximize the expected value, I should note that picking Mountain 1 guarantees an 8, while the other mountains have a lower expected value of 5.25. Choosing Mountain 1 seems like the best option! ****Deciding on the best choice****: I need to calculate the expected payoff from picking an unknown mountain. By doing the math, I find that 0.25 times 12 equals 3, and 0.75 times 3 equals 2.25, which totals 5.25. This is clearly less than the guaranteed value of 8 from Mountain 1. Therefore, my best rational choice is to select Mountain 1 with the Ruby. I'll respond simply with "Mountain 1" and provide a comment to explain my reasoning.

in assessing how realistically LLMs can simulate human behavior in strategy experiments. By analogy to a machine learning problem, if AI simulators can replicate human responses on held-out data with this level of fidelity, it offers a prima facie case for using them to simulate new, untested experimental designs.

5 | USING AI AGENT SIMULATIONS TO EXPLORE THE PARAMETER SPACE AND FORMULATE NOVEL HYPOTHESES

Our initial data are promising, suggesting that AI agent simulations can match data from human subjects. With that in mind, we now take advantage of the control afforded by the computational setup—both over the agent and the environment characteristics—as well as the speed of running simulations to extend the starting theory by systematically exploring the relevant parameter space. Specifically, we (a) vary the experimental parameters, (b) relax key theoretical assumptions, and (c) introduce heterogeneity in agent preferences. These extensions allow us to probe the boundary conditions of the original experiment and uncover new potential mechanisms. A summary of the full set of extensions and their results is presented in Supporting Information S1: Appendix E.

5.1 | Varying the experimental parameters

The first set of extensions focuses on the experimental setup itself. When researchers test a theory through an experiment, costs usually limit them to trying a fixed number of parameters. Some of these choices are likely innocuous, but others may influence the results in unforeseen ways. To investigate this in the context of Hoelzemann et al. (2025), we conduct 800 additional rounds of AI-based simulations, varying one of three major dimensions at a time.

5.1.1 | Varying population

We begin by varying the number of agents while keeping the number of mountains fixed. To simulate more crowded search environments, we rerun the original experiment with larger groups of AI agents. As shown in plots (i) of Figure 4, increasing the number of participants by increments of 10 does not lead to greater exploration. Even in larger groups, AI agents continue to cluster around the ruby and often fail to make a breakthrough. When asked to explain this behavior, the agents point to a dynamic we had not previously considered, suggesting an intriguing hypothesis: social conformity. For instance, one AI agent explained:

> "Given that 19 agents have already chosen Mountain 1 and each will receive the full value of the gem, it indicates a common strategy to secure a known and relatively high value."

The explanation above mirrors earlier findings that emulation can shape behavior in group settings (Lazer & Friedman, 2007; Puranam & Swamy, 2016). While social conformity is a

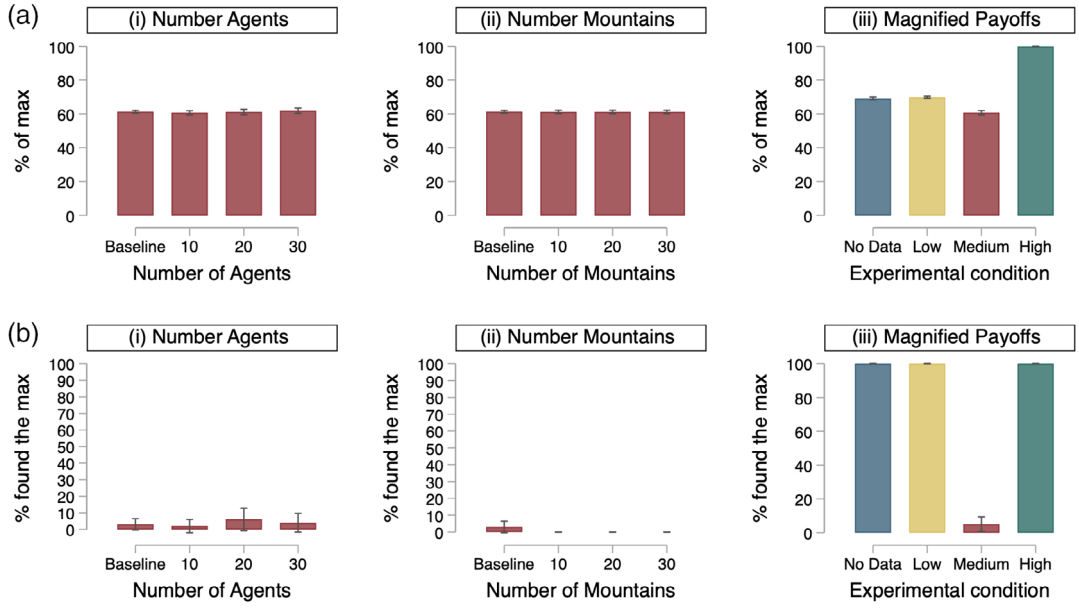

FIGURE 4 Changing the experimental parameters in Large Language Model (LLM)-based experiments. This figure shows how the outcomes of the original experiment change when we adjust the experimental parameters using LLMs. In Panel (a), the outcome of interest is the average group earnings (as a percentage of the maximum possible earnings). In Panel (b), the outcome of interest is the likelihood of a breakthrough, which occurs when at least one group member finds the diamond. In plots (i), we increase the number of agents from the baseline (five agents) by up to 30 agents, holding the number of mountains fixed. We only focus on the medium-value data condition. In plots (ii), we increase the number of mountains from the baseline (five mountains) by up to 30 mountains, holding the number of agents fixed. We once more focus on the medium-value data condition. In plots (iii), we increase the dollar values of gems a millionfold. Here, we look at all four data conditions. (a) Group payoffs. (b) Group breakthroughs.

mechanism well known in the literature, it was absent from the original study of Hoelzemann et al. (2025). This highlights how AI agent simulations can surface meaningful hypotheses worth exploring further with ABMs or human experiments.

5.1.2 | Varying choice landscape

Next, we vary the number of mountains while keeping the number of agents fixed. In this setting, coordinated search no longer guarantees a breakthrough, reflecting larger search spaces where not every option can be explored. We rerun the original experiment with a progressively increasing number of mountains. As shown in plots (ii) of Figure 4, we again find little change in behavior: AI agents continue to cluster around the medium-value option. They are not swayed by the greater absolute number of diamonds or rubies and follow the same decision logic as before. If anything, exploration declines, suggesting that larger search spaces may further reduce experimentation. One potential reason behind this surprising result is that some agents appear to interpret the herd behavior of others as a signal that they have missed valuable information:

> "The fact that many other participants have also chosen Mountain 2 reinforces the idea that it is a preferable choice given the available information."

This interpretation has some precedent in the strategy literature, where firms often struggle to evaluate complex opportunities and rely on heuristics or mimetic isomorphism (Haveman, 1993) to deal with uncertainty. Whether similar dynamics would emerge when human subjects face a larger choice set is a question that could be tested in future laboratory studies.

5.1.3 | Varying payoff structure

Finally, we vary the absolute magnitude of the payoffs. While the theoretical framework in Hoelzemann et al. (2025) relies only on the relative values of the gems, the authors had to select absolute dollar amounts for the lab experiment. They chose amounts below $15 per gem to keep the experiment affordable, but we might wonder how the decisions would change if players were earning up to several million in a given round. Of course, such an experiment would be infeasible with human subjects, but it is easily simulated with AI agents. We rerun the original experiment with payoffs scaled up by a factor of 1 million. The results, shown in plots (iii) of Figure 4, mirror the baseline. This aligns with the theoretical prediction that only the relative value of each gem should influence decisions. Whether this holds for humans is less clear, as the psychological effects of extreme stakes may change reasoning. Still, the ability to simulate such conditions with AI agents offers a valuable base expectation to inform subsequent research.

5.2 | Relaxing theoretical assumptions

The next set of extensions that we explore focuses on the underlying theoretical framework. Two key assumptions are embedded in the original model of Hoelzemann et al. (2025): that payoffs are non-rivalrous and non-ambiguous. Non-rivalry in payoffs implies that multiple agents selecting the same gem do not have to share the reward, while non-ambiguous payoffs assume that the distribution of gem probabilities is known by agents from the outset. These assumptions also guided the design of the original lab experiments. Understanding how sensitive the results are to these assumptions is important for evaluating the robustness of the starting theory and deciding which experiments to run. In particular, it is useful to know whether free riding persists when payoffs are rivalrous or when information is incomplete. To address this, we conduct an additional 1500 rounds of simulated experiments, relaxing each assumption individually as summarized in Supporting Information S1: Appendix Table E3. The baseline case, where payoffs remain non-rivalrous and non-ambiguous, is shown in plots (i) of Panels (A) and (B) in Supporting Information S1: Appendix Figure E4.

We begin by changing the payoff structure to introduce rivalry, meaning that agents who choose the same gem must divide the reward. The results are shown in plots (ii) of Panel (A) and Panel (B). Rivalry reverses the original pattern of findings and reveals a potentially important boundary condition for the theory. Agents no longer earn the least (Panel A), nor do they achieve fewer breakthroughs (Panel B) when the medium-value gem is revealed. Because they can observe others' choices in sequential order, agents often select unchosen mountains to

avoid splitting payoffs. Qualitative responses from AI agents support this behavior and show how competition acts as a strong incentive for exploration. As a result, payoffs begin to converge across the four conditions, though at a significantly lower level. This mirrors the effect of competition and rivalry eroding profits in a market.

Next, we manipulate agents' information set by withholding information about the distribution of gems of each type. We begin by keeping payoffs ambiguous but non-rivalrous. The results of this variation are shown in plots (iii) of Panel (A) and Panel (B). Exploration continues to collapse under these conditions. Group payoffs (Panel A) and breakthroughs (Panel B) closely mirror the baseline: AI agents earn the least and rarely achieve a breakthrough when they know the location of the medium-value outcome. This suggests that the theory may not depend on prior knowledge of the full payoff distribution, but rather on knowledge of the relative value of available options.

Finally, we examine the case where payoffs are both rivalrous and ambiguous. This allows us to assess which assumption has a stronger influence on agent behavior, and whether relaxing both leads to any interaction effects. The results are presented in plots (iv) of Panels (A) and (B). Herding disappears entirely under these conditions, and agents behave almost exactly as they do when payoffs are rival but information is complete. This suggests that rivalry plays a dominant role in shaping behavior and may be the critical condition driving exploratory dynamics. This insight reveals a potential boundary condition that would have been difficult to test without using AI agents as synthetic subjects.

5.3 | Manipulating agent objectives

The final set of extensions focuses on the agents themselves. Hoelzemann et al. (2025) suggest that risk aversion and non-pecuniary objectives may help explain the experimental results and potentially influence the theoretical predictions. Studying such factors in human experiments is difficult, as experimentally manipulating participants' risk tolerance or motivations can be impossible. With AI agents, however, exploration of these aspects becomes feasible. A key feature of LLMs is their capacity to adopt specific preferences, demographics, or personalities through prompt conditioning (Aher et al., 2023; Horton, 2023). This simply involves priming the model with tailored scripts. To test this in practice, we conduct 1000 additional rounds of simulated experiments in which the LLMs are endowed with different preference structures.

We begin by introducing higher risk tolerance. We repeat the original experiment with an increasing share of risk-loving agents, starting with one and adding one at a time until all five agents exhibit risk-seeking behavior.[13] We focus on the condition in which the medium-value gem is revealed. More breakthroughs occur (Figure 5, Panel a) as exploration increases, driven by risk-loving agents (Figure 5, Panel b). To verify that the priming has the intended effect, we review the agents' rationales. As one risk-seeking agent explains:

[13]Before running this extension, we assessed the baseline risk propensity of AI agents using the traditional Holt and Laury task used in behavioral economics. Consistent with previous findings, we confirm that unprimed AI agents tend to be risk-neutral, similar to the average participant in lab studies (Mei et al., 2024). We then repeat the preference elicitation task after priming agents with their respective risk profiles, confirming that the priming significantly alters responses in the expected direction.

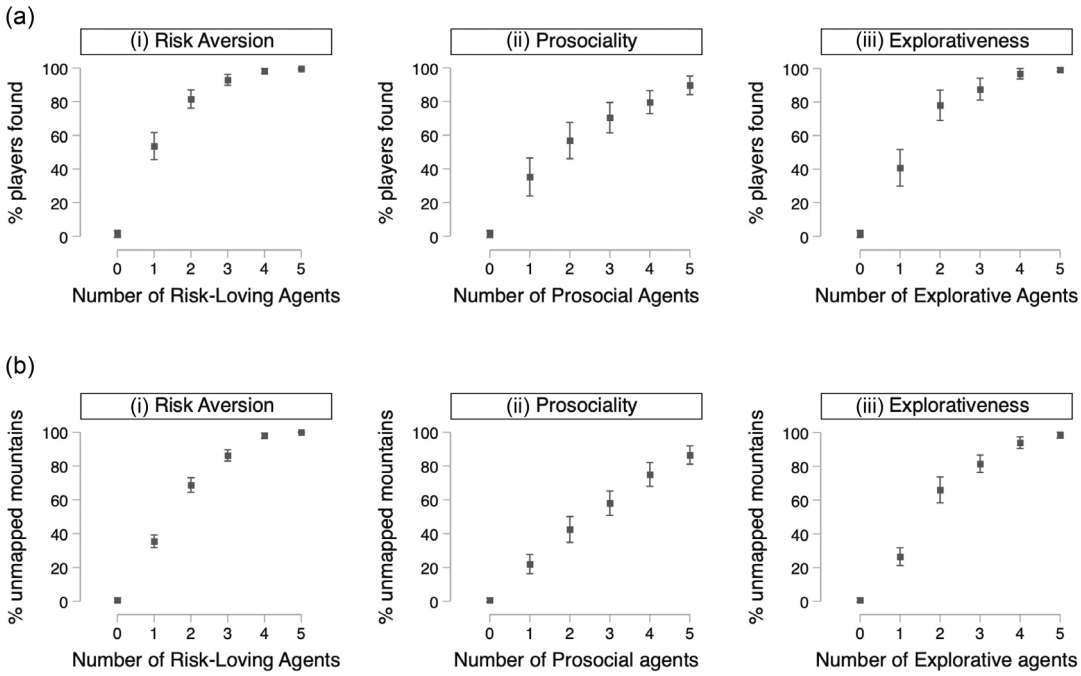

FIGURE 5 Incorporating agent preferences in Large Language Model (LLM)-based simulations. This figure shows how the outcomes of the original experiment change when we incorporate heterogeneous agent preferences using LLMs. In Panel (a), the outcome of interest is the share of agents that achieve a breakthrough in the second period. In Panel (b), the share of agents that chose an unmapped mountain in the first period. In plots (i), we add risk-loving agents, starting with one risk-lover until all five agents are risk-loving. In plots (ii), we add pro-social agents, and in plots (iii), we add agents with a taste for exploration. For all three analyses, we focus on the medium-value data condition. (a) Individual breakthroughs. (b) Individual exploration.

> "Given my risk-loving nature and the information available, I choose Mountain 1. This decision is based on the fact that the Ruby has already been found in Mountain 4, and another agent has chosen Mountain 5. Since the Diamond has not yet been discovered and I prefer taking risks for potentially higher rewards, I opt for Mountain 1, which has not been chosen by any other participant and still holds the possibility of containing the Diamond."

Interestingly, as more agents are primed to take risks, the remaining unprimed agents also begin to explore more. This suggests that changing risk tolerance for a few agents can lead to "coordination spillovers" for the entire group. As one non-risk-loving agent told us, exploring becomes more attractive when the likelihood of collectively uncovering the diamond increases. This implies that only a subset of agents needs to be risk-loving to generate sufficient exploration, a prediction that can be tested in future laboratory experiments.

Next, we introduce motives beyond profit maximization, beginning with pro-social behavior. We rerun the original experiment with varying numbers of pro-social agents, starting with one out of five and incrementing until all are pro-social. These agents are instructed that their success depends on whether others find the diamond, which encourages them to generate

information that maximizes the group's chances of success. We again focus on the condition where the medium-value gem is revealed. The results, shown in plots (ii) of Panel (a) and Panel (b) in Figure 5, show that exploration increases steadily with the number of pro-social agents, as does the rate of breakthrough.

Finally, we introduce an explicit preference for exploration. We rerun the original experiment with varying numbers of explorative agents, starting with one and increasing until all agents are primed to prioritize exploration. These agents are instructed that their sole objective is to find the diamond and that they should continue searching until they succeed. The results, shown in plots (iii) of Panel (a) and Panel (b) in Figure 5, closely resemble those observed with risk-loving agents. As expected, exploration increases steadily with the number of explorative agents.

5.4 | Comparing agentic simulations to direct elicitation of results

So far, we have used LLMs to run experiments using AI agents as synthetic subjects, applying a bottom-up approach that parallels ABMs in spirit. However, another possibility is to use LLMs to predict the outcomes of experiments directly in "one shot." LLMs have become increasingly capable of handling complex tasks, and their performance continues to improve. In this setting, an LLM could generate outcome predictions along with explanations for its reasoning. This approach would require even less effort and fewer resources than running simulations with synthetic subjects. Of course, the feasibility of this simpler idea depends entirely on the model's ability to accurately predict the results of complex social interactions from the outset.

We asked GPT-4 to predict the outcomes of the original baseline experiment from Hoelzemann et al. (2025). We provided the model with the description of the experiment and asked it to predict the results. The results of this exercise are presented in Supporting Information S1: Appendix Table E4. While direct elicitation occasionally yields reasonable predictions, overall accuracy remains modest. More critically, GPT-4 does not capture the core mechanisms underlying the theory and overlooks the key insight that more data can lead to worse outcomes. By contrast, using AI agents within our experimental framework more effectively replicates the original findings, providing further support for the bottom-up simulation approach. We also find that GPT-4's direct predictions fail to align with the results of several simulation-based extensions, reinforcing the limitations of relying solely on outcome prediction instead of a computational experiment.

6 | DISCUSSION

In this paper, we argue that using AI agents as synthetic subjects in simulated experiments offers a powerful new tool for strategy research. We introduce a framework for experimenting with AI agents and apply it to a theory of exploration in the presence of data spillovers. Replicating the online lab experiment from Hoelzemann et al. (2025), we find that LLMs can approximate human behavior in group settings with strategic interdependencies. We then extend the original theory through a series of exploratory simulations. These identify plausible boundary conditions and connect the resulting patterns to known theoretical mechanisms. In either case, the method's value lies less in discovering entirely new mechanisms than in helping to generate hypotheses that merit empirical validation.

Several takeaways are worth noting. In our setting, LLMs exhibited a meaningful degree of behavioral realism as compared to human subjects. This is promising, since the key question is whether synthetic subjects' behavior offers a reasonable approximation of how humans would act in the same strategic context.[14] However, the degree of alignment between LLMs and humans is likely context-dependent, as documented by previous studies (Tjuatja et al., 2023). Even in our application, we noted key points of divergence, with LLMs displaying a bias toward rationality. As such, human-subject experiments remain the gold standard. Furthermore, as others have emphasized in the context of traditional models (Knudsen et al., 2019), LLM-based simulations do not constitute empirical data for hypothesis testing. Rather, their value lies in leveraging their quasi-realism to prototype or to generate extensions and refinements to the starting setup.

Although not evident in our application, additional limitations of LLMs warrant closer scrutiny in the context of strategy research. Because LLMs are trained on text drawn from the internet, they disproportionately reflect English-speaking, urban populations (Dillion et al., 2023). Unlike the "WEIRD" bias in human-subject research that can be addressed through a diverse subject pool (Henrich et al., 2010), this bias is embedded in the foundational model itself, potentially limiting its ability to simulate environments or represent actors with limited digital data (Tao et al., 2024). This is especially relevant for subfields such as international business in non-Western settings, or research on privately held companies, including startups and family enterprises. It might be possible to generate diversity in model responses by prompting for specific populations, but this technique is still untested. A second concern involves cognitive bias. LLMs exhibit framing and anchoring effects that, while similar to human tendencies, may distort outcomes in strategic experiments (Jones & Steinhardt, 2022).

How these known biases affect the hypotheses generated by simulations remains unclear. Recent evidence suggests that LLM-generated research hypotheses can be as novel and useful as those proposed by expert researchers (Si et al., 2024). At the same time, a model skewed in known ways may generate hypotheses that inherit that skew (Messeri & Crockett, 2024). LLMs can also fail in unexpected ways, in part because the world models implicit in their outputs are less coherent than they appear (Vafa et al., 2024). For these reasons, we view the hypotheses surfaced by our simulations as leads whose value rests not on the model being unbiased, but on their subsequent validation. As with any exploratory tool, researcher discretion, domain knowledge, and prior evidence should guide which surprises are worth pursuing.

Despite these challenges, a key lesson from our study is that the validity of LLM-based simulations can be tested. First, borrowing a concept from machine learning, access to a prior "hold-out" dataset was crucial. It provided a benchmark for evaluating outcomes and a reference point for simulating new experimental conditions. Without this anchor, it would have been difficult to justify extending the analysis to untested scenarios. This approach parallels best practices in agent-based modeling, where researchers typically begin by replicating known results before introducing additional complexity (Davis et al., 2007). Second, directly eliciting qualitative explanations from the models proved valuable. These rationales often revealed reasoning patterns similar to those of human subjects, suggesting that the alignment was not coincidental. Still, such responses may reflect hallucination or post hoc justification. For this reason, the researcher's judgment remains critical.

[14]We are indebted to an anonymous reviewer for raising this point and encouraging us to frame AI agents as synthetic subjects.

We also found that LLM-based simulations offer a moderate degree of control. We were able to introduce meaningful variation to systematically link actions and outcomes, much like in ABMs. Changes to the strategic environment—such as the choice landscape or payoff structure—were easy to implement by directly modifying our script. We also suspect that some categories of agentic variation difficult to represent in ABMs may be more readily accessible through LLMs. These include variation in reasoning styles, shifts in moral framing, or sensitivity to social norms. Indeed, most ABMs in our literature review emphasized technical, easily parameterized constructs, such as performance aspirations or risk preferences. At the same time, the use of a foundation model (in our case, OpenAI's GPT-4) introduces limits to this control. Unlike ABMs, where agents are fully specified by the researcher, LLM agents are opaque and not entirely programmable. This complexity may contribute to their behavioral realism, but it also constrains our ability to isolate mechanisms. For this reason, ABMs remain key when the goal is to transparently identify the essential mechanisms that can explain observed patterns.

Table 2 translates this discussion into a set of concrete tasks for which AI agent simulations may complement human-subject experiments and ABMs. The table is intended to offer a starting point for thinking about where AI simulations may be most useful. Before a human-subject experiment is run, synthetic subjects can be used to pilot a protocol, assess whether instructions and incentives capture the intended constructs, and narrow the set of moderators and boundary conditions that warrant experimental attention. After a human-subject experiment, the same setup can help probe why a pattern emerged and generate alternative hypotheses consistent with the observed responses. A similar logic applies to ABMs. Because AI agents can be assigned rich profiles and behavioral dispositions through relatively simple prompting, they make it easier to vary cognitive assumptions and heterogeneity without having to hard-code them. Once an ABM has been built, AI agents can be used to stress-test whether core results persist under richer behavioral assumptions and to translate model predictions into candidate experimental manipulations.

Where might the use of AI agents add the most value in strategy research? While the answer is necessarily speculative, our review points to several promising directions. Some domains, such as competitive dynamics and adaptation to environmental change, have long relied on simulation methods. Revisiting these findings with AI-powered agents that do not rely on prespecified behavioral rules, but instead display elements of human-like adaptability, could yield meaningful insights. Other areas, including entrepreneurship and strategic human capital, have yet to fully exploit the variation that simulation enables. Here, AI agent experiments create new possibilities. For instance, researchers could create AI-managed organizations to evaluate candidate résumés, replicating the design of prior audit studies. They could then manipulate applicant attributes to detect subtle forms of bias that would be costly and time-consuming to identify using human subjects. We also see promise in domains where outcomes hinge on relational governance, such as joint ventures or supplier relationships. AI agents could be used to simulate negotiation dynamics, to then examine changes in trust and opportunism before moving to empirical validation. The potential set of applications is substantial.

As LLMs continue to improve, AI-powered agents in strategy may come to play a role analogous to model organisms in biomedical science. Just as animal models became a central research tool, LLM-based simulations could complement both ABMs and human-subject experiments by enabling rapid, low-cost screening of mechanisms, design choices, and boundary conditions. Realizing this vision, however, will require sustained engagement from the strategy research community. In biology, the use of lab mice became possible through their standardization as a research tool (Hedrich, 2004). Institutions such as the Jackson Laboratory played a

TABLE 2 How artificial intelligence (AI) agent simulations can complement human-subject experiments and traditional agent-based models (ABMs).

A. Human-subject experiments			
Typical task	**Typical research challenges**	**Utility of AI agent simulations**	**Limitations of AI agent simulations**
Prototyping the experimental setup	Small experimental choices can meaningfully affect results (e.g., small wording changes); pilots are costly, and flaws may only become apparent late in the experimental process	Before running the experiment, iterate over experimental choices through exploring variations on stimuli and procedures (e.g., generate alternative wordings, controls, and attention checks)	Prototyping results may shift with prompts, sampling settings, or model choices and updates; promising design choices might not replicate with human subjects
Exploration of parameter space	Many plausible moderators and boundary conditions; factorial designs are often infeasible given limited sample sizes	Before running the experiment, low-cost iteration across parameterizations to prioritize which moderators and conditions merit testing with human subjects	Patterns may be unstable across seeds, prompts, or model versions; any inference about human behavior requires subsequent testing with human participants
Replication of experimental results to check robustness or surface "surprises"	Replications with human subjects are expensive and time-consuming; procedural details are often underspecified in the original study; discrepancies with published results are hard to diagnose	After running the experiment, AI simulations can identify fragile design elements and help design follow-up tests and robustness checks	Apparent replications may reflect prior exposure to the study in training data; unexpected patterns require researcher judgment and, if pursued, validation with human evidence
Explore risky or infeasible experimental manipulations	Some agentic or environmental characteristics cannot be randomized; some experimental conditions may be unethical, risky, or too costly to implement	Safely explore designs that include ethically or practically infeasible conditions, and use these patterns to motivate feasible empirical tests using field and archival data or simply generate hypotheses	Outputs can be sensitive to prompts and foundational model; behavioral opacity complicates interpretation; credible conclusions require triangulation with real-world evidence
B. Agent-based models (ABMs)			
Typical task	**Typical research challenges**	**Utility of AI agent simulations**	**Limitations of AI agent simulations**
Easy manipulation of agents or environments	Implementing many variants of a simulation (e.g., different agent behavioral rules) is time-consuming because each	After developing the ABM, easy priming of agent types or environments with short scripts (e.g., "be pro-social") to explore candidate	Construct validity is fragile because prompts may bundle multiple traits, vary with wording, and interact with model version; the

TABLE 2 (Continued)

B. Agent-based models (ABMs)			
Typical task	**Typical research challenges**	**Utility of AI agent simulations**	**Limitations of AI agent simulations**
	variant's logic must typically be specified in detail	behavioral regimes or extensions worth modeling	same label (e.g., "pro-social") may not map to a stable or interpretable decision rule
Introducing more behaviorally complex agents	ABMs often rely on simple, mechanistic heuristics; modeling cognition, norms, and role-based reasoning may miss important behavioral complexity	After developing the ABM, insert richer microfoundations (e.g., moral framing or deliberative styles) and translate them into explicit candidate rules for incorporation and comparison within ABMs	These behaviors may reflect conventions in the corpus of training data rather than the target population; additional richness can reduce interpretability; stability across model versions is uncertain
Connect ABM predictions to real-world scenarios	Model validation is difficult and often omitted; different mechanisms can fit the same empirical phenomenon; links from ABM constructs to observable measures are often ad hoc	After developing the ABM, use to translate ABM constructs and predicted outcomes into realistic contexts and concrete manipulations, helping identify feasible proxies and validation designs with data	Suggested proxies may be infeasible, noisy, or conceptually misaligned; testing ABM predictions still requires actual real-world data from the lab or the field

Note: This table clarifies where Large Language Model-powered simulations using AI agents as synthetic subjects are most useful as complements to human-subject experiments and ABMs. To enable replicability of AI agent simulations, we recommend reporting details on the model and version, prompt templates, sampling settings, and number of runs, as we did in Supporting Information S1: Appendix D.

critical role by providing widely available mouse strains and codifying best practices for experimental use. In the same spirit, progress in strategy will depend on shared standards for reporting and reproducibility. We hope the framework offered here and the code reported in Supporting Information S1: Appendix F serve as a foundation for applying AI agent simulations to deepen the theoretical richness and practical relevance of strategy research.

ACKNOWLEDGMENTS

We thank participants at the Macro Research Lunch (UC Berkeley-Haas), 2024 and 2025 Academy of Management, Faculty Research Lunch (UC Berkeley-Haas), FOOLs (Wharton), HIVA seminar (KU Leuven), 2024 Conference on AI, Machine Learning and Business Analytics (Yale SOM), AI & Strategy Seminar, GenAI Lab seminar (TUM School of Management), AI and Entrepreneurship Workshop (HEC Paris), 2025 Utah Winter Strategy Conference, and INSEAD for useful feedback. All errors are our own.

FUNDING INFORMATION

We acknowledge support from OpenAI in the form of computing credits to use their GPT class of models.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Matteo Tranchero https://orcid.org/0009-0002-8477-817X
Abhishek Nagaraj https://orcid.org/0000-0001-9049-0522

REFERENCES

Adner, R., & Levinthal, D. A. (2024). Strategy experiments in nonexperimental settings: Challenges of theory, inference, and persuasion in business strategy. *Strategy Science*, *9*, 311–321.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, *47*, 1–6.

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *arXiv Preprint*.

Bail, C. A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences of the United States of America*, *121*, e2314021121.

Billinger, S., Stieglitz, N., & Schumacher, T. R. (2014). Search on rugged landscapes: An experimental study. *Organization Science*, *25*, 93–108.

Broska, D., Howes, M., & Van Loon, A. (2024). The mixed subjects design: Treating large language models as potentially informative observations. *Sociological Methods & Research*, *54*, 00491241251326865.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint*.

Carlson, N., & Burbano, V. (2026). The use of LLMs to annotate data in management research: Guidelines and warnings. *Strategic Management Journal*, *47*, 699–725.

Chatterji, A. K., Findley, M., Jensen, N. M., Meier, S., & Nielson, D. (2016). Field experiments in strategy research. *Strategic Management Journal*, *37*, 116–132.

Csaszar, F. A., Ketkar, H., & Kim, H. (2024). Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. *SSRN*. https://ssrn.com/abstract=4913363

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing theory through simulation methods. *Academy of Management Review*, *32*, 480–499.

Dell'Acqua, F., Mcfowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. In *Harvard Business School TOM Working Paper*. Harvard Business School.

Di Stefano, G., & Gutierrez, C. (2019). Under a magnifying glass: On the use of experiments in strategy research. *Strategic Organization*, *17*, 497–507.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*, 597–600.

Doshi, A. R., Bell, J. J., Mirzayev, E., & Vanneste, B. S. (2025). Generative artificial intelligence and evaluating strategic decisions. *Strategic Management Journal*, *46*(3), 583–610. https://ssrn.com/abstract=4714776

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, *384*, 1306–1308.

Felten, E. W., Raj, M., & Seamans, R. (2023). Occupational heterogeneity in exposure to generative AI. *SSRN*. https://ssrn.com/abstract=4414065

Ganco, M. (2017). NK model as a representation of innovative search. *Research Policy*, *46*, 1783–1800.

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences of the United States of America*, *122*, e2501660122.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*, 1108–1109.

Gurnee, W., & Tegmark, M. (2024, May). Language models represent space and time. In *International Conference on Learning Representations* (Vol. 2024, pp. 2483–2503).

Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *121*, e2317967121.

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, *3*, 833–838.

Harrison, J. R., Lin, Z., Carroll, G. R., & Carley, K. M. (2007). Simulation modeling in organizational and management research. *Academy of Management Review*, *32*, 1229–1245.

Haveman, H. A. (1993). Follow the leader: Mimetic isomorphism and entry into new markets. *Administrative Science Quarterly*, *38*, 593–627.

Hedrich, H. (2004). *The laboratory mouse*. Academic Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83.

Hoelzemann, J., Manso, G., Nagaraj, A., & Tranchero, M. (2025). The streetlight effect in data-driven exploration. In *NBER Working Paper w32401*. National Bureau of Economic Research.

Horton, J., & Horton, R. (2024). EDSL: Expected parrot domain specific language for AI powered social science. White Paper, *Expected Parrot*. https://docs.expectedparrot.com/en/latest/papers.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from Homo silicus? In *NBER Working Paper w31122*. National Bureau of Economic Research.

Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems*, *35*, 11,785–11,799.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford University Press.

Knudsen, T. (2024). Agent-based modelling for strategy. In *The Oxford handbook of agent-based computational management science*. Oxford University Press.

Knudsen, T., Levinthal, D. A., & Puranam, P. (2019). A model is a model. *Strategy Science*, *4*, 1–3.

LaFollette, H., & Shanks, N. (1995). Two models of models in biomedical research. *The Philosophical Quarterly*, *45*, 141–160.

Lampinen, A. K., Dasgupta, I., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., Mcclelland, J. L., & Hill, F. (2024). Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, *3*, pgae233.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., … Perez, E. (2023). Measuring faithfulness in chain-of-thought reasoning. *arXiv*. arXiv:2307.13702.

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly*, *52*, 667–694.

Leiblein, M. J., Reuer, J. J., & Zenger, T. (2018). What makes a decision strategic? *Strategy Science*, *3*, 558–573.

Levine, S. S., Schilke, O., Kacperczyk, O., & Zucker, L. G. (2023). Primer for experimental methods in organization theory. *Organization Science*, *34*, 1997–2025.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … Batson, J. (2025). On the biology of a large language model. *Transformer Circuits Thread*.

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., Marinazzo, D., Nicholas, J., Salatiello, A., Sucholutsky, I., Minervini, P., Razavi, S., Rocca, R., Yusifov, E., Okalova, T., … Love, B. C. (2025). Large language models surpass human experts in predicting neuroscience results. *Nature Human Behaviour*, *9*, 305–315.

Manning, B. S., Zhu, K., & Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. In *NBER Working Paper w32381*. National Bureau of Economic Research.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2*, 71–87.
Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences of the United States of America*, *121*, e2313925121.
Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58.
Ott, T. E., & Hannah, D. P. (2024). On the origin of entrepreneurial theories: How entrepreneurs craft complex causal models with theorizing and data. *Strategy Science*, *9*, 461–482.
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.
Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv Preprint*.
Puranam, P., & Swamy, M. (2016). How initial representations shape coupled learning processes. *Organization Science*, *27*, 323–335.
Rumelt, R. P., Schendel, D., & Teece, D. J. (1991). Strategic management and economics. *Strategic Management Journal*, *12*, 5–29.
Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, *41*, 1254–1270.
Si, C., Yang, D., & Hashimoto, T. (2024). Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv Preprint*.
Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*, pgae346.
Tjuatja, L., Chen, V., Wu, S. T., Talwalkar, A., & Neubig, G. (2023). Do LLMs exhibit human-like response biases? A case study in survey design. *arXiv Preprint*.
Tranchero, M. (2026). *Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation*. University of Pennsylvania.
Tu, X., Zou, J., Su, W., & Zhang, L. (2024). What should data science education do with large language models? *Harvard Data Science Review*, *6*.
Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952–74965.
Vafa, K., Chen, J. Y., Rambachan, A., Kleinberg, J., & Mullainathan, S. (2024). Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, *37*, 26941–26975.
Van den Steen, E. (2018). The strategy in competitive interactions. *Strategy Science*, *3*, 574–591.
Vuculescu, O. (2017). Searching far away from the lamppost: An agent-based model. *Strategic Organization*, *15*, 242–263.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tranchero, M., Brenninkmeijer, C.-F., Murugan, A., & Nagaraj, A. (2026). Simulating strategic interactions with AI agents. *Strategic Management Journal*, 1–27. https://doi.org/10.1002/smj.70112