NBER WORKING PAPER SERIES

THEORIZING WITH LARGE LANGUAGE MODELS

Matteo Tranchero
Cecil-Francis Brenninkmeijer
Arul Murugan
Abhishek Nagaraj

Working Paper 33033
http://www.nber.org/papers/w33033

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2024

Corresponding author: Abhishek Nagaraj. We thank participants at the Macro Research Lunch at UC Berkeley-Haas and the 2024 Academy of Management for useful feedback. We acknowledge support from OpenAI in the form of computing credits to use their GPT class of models. All errors are our own. Nagaraj received research support in the form of computing credits from OpenAI. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Theorizing with Large Language Models
Matteo Tranchero, Cecil-Francis Brenninkmeijer, Arul Murugan, and Abhishek Nagaraj
NBER Working Paper No. 33033
October 2024
JEL No. M10, O3

ABSTRACT

Large Language Models (LLMs) are proving to be a powerful toolkit for management and organizational research. While early work has largely focused on the value of these tools for data processing and replicating survey-based research, the potential of LLMs for theory building is yet to be recognized. We argue that LLMs can accelerate the pace at which researchers can develop, validate, and extend strategic management theory. We propose a novel framework called Generative AI-Based Experimentation (GABE) that enables researchers to conduct exploratory in silico experiments that can mirror the complexities of real-world organizational settings, featuring multiple agents and strategic interdependencies. This approach is unique because it allows researchers to unpack the mechanisms behind results by directly modifying agents' roles, preferences, and capabilities, and asking them to reveal the explanations behind decisions. We apply this framework to a novel theory studying strategic exploration under uncertainty. We show how our framework can not only replicate the results from experiments with human subjects at a much lower cost, but can also be used to extend theory by clarifying boundary conditions and uncovering mechanisms. We conclude that LLMs possess tremendous potential to complement existing methods for theorizing in strategy and, more broadly, the social sciences.

Matteo Tranchero
The Wharton School
3620 Locust Walk
2204 Sh-Dh
Philadelphia, Penn 19104
mtranc@wharton.upenn.edu

Cecil-Francis Brenninkmeijer
University of California, Berkeley
cecil-francis.b@berkeley.edu

Arul Murugan
University of California, Berkeley
arul@berkeley.edu

Abhishek Nagaraj
Haas School of Business
University of California, Berkeley
2220 Piedmont Ave
Berkeley, CA 94720
and NBER
nagaraj@berkeley.edu

1 Introduction

Strategy research is remarkably diverse, drawing from disciplines as varied as economics, sociology, and psychology. At its core, however, the field is concerned with how agents strategically interact with each other in organizational or market settings. These agents may consist of individuals, such as employees and managers, or organizations, such as firms and startups. The interactions they have are also varied: from more collaborative, like when agents coordinate towards a common outcome, to more rivalrous, when they must compete to discover and exploit new market opportunities. Modeling and analyzing these strategic interactions is at the heart of many research communities in management. The objective is to develop theories that provide convincing explanations for relevant management phenomena and help formulate normative guidance to improve performance (Rumelt et al., 1991).

In this paper, we argue that generative Artificial Intelligence (generative AI) and specifically Large Language Models (LLMs) can be immensely helpful for management researchers focused on theory building. LLMs are already widely diffusing as research tools in the social sciences, primarily to aid with data cleaning, data analysis, and writing (Grimes et al., 2023; Liang et al., 2024; Charness et al., 2023). Researchers have also started to use these models as a sort of "homo silicus" (Horton, 2023), reproducing the results from behavioral research using human subjects (Horton, 2023; Aher et al., 2023; Ashokkumar et al., 2024). Yet, the value of LLMs goes beyond harnessing generative AI to support the research process or replicate known results. Our central contention is that LLMs can become a simulation tool that is particularly well-suited to generate novel theory in management.

We propose that LLMs can be used to create ecologies of AI agents that interact in environments mimicking the strategic interdependencies and complexities of real-world settings. Without the need to mechanistically specify their behavior, researchers can endow these autonomous agents with objectives, preferences, capabilities, and personalities of their choosing. The end result is a fast, robust, and flexible method to generate theoretical predictions under different assumptions. We refer to this framework as Generative AI-Based Experimentation (GABE). When compared with other types of simulations, we argue that GABE has one crucial difference: it relies on AI agents whose behavior is not deterministically pre-specified by the researcher, yet whose reasoning can still be elicited via direct prompting. As we aim to demonstrate, learning the mechanisms behind their actions in this way and tweaking the characteristics of AI agents is a powerful way to build, validate, and extend management theory.

Our framework requires researchers to start with an initial theory about a phenomenon to conceptualize a computational experiment and record the variables of interest (Ludwig and Mullainathan, 2024; Manning et al., 2024; Shrestha et al., 2021; Davis et al., 2007). However, once conceptualized, GABE can augment and strengthen the initial theory by helping test boundary conditions and mechanisms. In so doing, they serve as a useful complement to existing research approaches, like laboratory and field experiments, as well as agent-based models (ABM). On the one hand, experiments that feature multiple human subjects or organizations remain essential for testing theoretical predictions. However, due to the costs and logistical complexities involved, these experiments can rarely be used as an exploratory tool to help researchers theorize. On the other hand, ABMs are a powerful tool for studying the emergent properties of interactive social systems, and as a result, they have enjoyed significant popularity in management. Yet the stylized setups and the large degrees of freedom the researcher enjoys when designing them may compromise realism. GABE can play an intermediate role by simulating cheap and realistic exploratory experiments with agents whose behavior is likely closer to humans than the automatons of traditional simulations.

To demonstrate the potential of this approach, we use GABE to study a phenomenon that is common in management and innovation research: strategic exploration under uncertainty. We focus on a theory that tries to model the "streetlight effect" – namely, the tendency of people to search where there are existing data rather than in the dark where the rewards might lie (Demirdjian et al., 2005). Commentators have noticed this dynamic in high-stakes settings like biomedical research, where firms keep innovating in well-trod areas and neglect potential breakthroughs (Haynes et al., 2018; Bulgheresi, 2016). In a recent paper, Hoelzemann et al. (2024) developed a formal theory to capture this phenomenon and tested its predictions with an online lab experiment where human subjects collectively explored a set of low-, medium- or high-value projects over multiple time periods. The key finding is that participants achieve lower earnings and are less likely to discover the high-value project when they are initially told the location of the medium-value project as compared to when they start exploring without any data. Shedding light on attractive but sub-optimal projects can skew exploration in a sub-optimal direction and deprive agents and society of breakthrough innovations.

We use the GABE framework (powered by OpenAI's GPT-4 model) to extend the Hoelzemann et al. (2024) theory by simulating this same experiment with AI agents. As we will describe in detail, we develop a central experimental engine that feeds multiple AI agents the objectives and rules of the game and coordinates their actions under different experimental conditions. We first use this method to replicate the conditions of the original experiment and reproduce the results derived from human subjects. Our analysis shows that we are able to do so with a remarkable degree of accuracy. Like humans, AI agents also earn less when exploring with the medium-value option revealed as compared to when they explore in the dark. Their qualitative responses are also highly consistent with the behavior of human participants in Hoelzemann et al. (2024), providing face validity for our approach. Reassuringly, these results are not due to the model knowing the results of the paper beforehand, but rather they are endogenously generated by the GABE framework.

Next, we show how AI agents can be used to enrich our theoretical understanding of the streetlight effect beyond the original study. In particular, we run AI-based experiments to introduce several extensions to the original results, including (a) implementing alternate variations in design, (b) relaxing the underlying theoretical assumptions, and (c) incorporating heterogeneous agent preferences, such as risk aversion or pro-sociality. Our objective is to uncover new boundary conditions that delineate when the predictions of the theory are likely to fail, or mechanisms to explain why the streetlight effect might actually prevail. The exploration of mechanisms is greatly aided by the ability to prompt AI agents to justify strategic choices as well as the ability to "fine tune" AI agents to behave in a particular way (for example, to take more risks or care about collective outcomes). These approaches are powerful tools to rationalize the findings and sharpen the mechanisms and boundary conditions of the theory.

Our analysis yields several findings. First, we uncover new potential mechanisms that help to explain the streetlight effect. For instance, when we increase the number of AI agents engaged in search, the LLMs are more likely to gravitate toward the medium-value option. When eliciting agents' motivations, we find that their responses suggest herding behaviors due to social conformity–a mechanism absent from the original theory of Hoelzemann et al. (2024). Second, we uncover additional boundary conditions for the streetlight effect. For instance, when we introduce payoff rivalry, AI agents no longer herd around the safe option for fear of splitting the rewards. We also document the theory's robustness to several assumptions. Adding ambiguity in the distribution of potential payoffs or changing their magnitude has little effect as long as the relative value of options is maintained. Third, we endow a subset of AI agents with heterogeneous preferences and objectives. In so doing, we see that even a small share of agents with alternative preferences can temper the streetlight effect. In a final set of experiments, we simply ask out-of-the-box LLMs to predict the results of the experiment rather than using the GABE approach we propose. We show how this approach has limited power to predict real-world outcomes to the same extent.

Our paper makes several contributions. First, we present GABE as a novel conceptual framework to leverage LLMs as AI agents in management research, positioning them as tools for theory building. We suggest harnessing the ability of LLMs to simulate interactive strategic management settings and generate new theoretical insights by directly engaging with the reasoning processes of AI agents. Second, we show how GABE couples the advantages of traditional agent-based models with the greater realism of LLMs. Researchers can specify the features of the setting and endow agents with preferences or resources, while at the same time relying on the LLM to more instinctively emulate agents' behaviors. This approach enables researchers to run large-scale in silico experiments affordably, trace the logical pathways from observed outcomes back to underlying assumptions, and gain insights into decision-making processes. This level of transparency is particularly valuable because learning the mechanisms behind observed behaviors is fundamental to theory development. Finally, our application to the theory of the streetlight effect exemplifies how GABE can lead to substantial theoretical contributions. The new mechanisms and boundary conditions that we uncovered are intriguing hypotheses for testing in future experimental and empirical work.

We focus our attention on the use of LLMs for theorizing, which plays a central role in strategic management research (Makadok et al., 2018; Shrestha et al., 2021; Sutton and Staw, 1995). Management theorizing often oscillates between induction, which generalizes from observations, and deduction, which develops theories from first principles (Makadok et al., 2018; Choudhury et al., 2021; Davis et al., 2007; Harrison et al., 2007; Shrestha et al., 2021). The observations derived from GABE do not neatly fit into either category. As the product of simulations, they do not constitute "ground truth" evidence from which inductive generalizations can be made. At the same time, they do not provide evidence external to the generating model that can be used to test its predictions. Instead, we suggest these experiments largely facilitate *abductive theorizing*. LLMs can offer verbal explanations for their actions, enabling researchers to reason about the most likely mechanisms for the patterns observed.

One important caveat is in order. It is not our belief that GABE should replace rigorous testing of management theories through careful, experimental study. Despite their general consistency with humans, observations derived using AI agents do not constitute real-world empirical evidence that can be used to falsify theories; rather, they are better thought of as exploratory insights that ultimately need to be validated using traditional experiments with human subjects or observational data. In addition, GABE also suffers from concerns highlighted in other studies that use LLMs as stand-ins for human subjects, such as a potential bias towards rationality and the tendency to hallucinate responses. As the underlying models improve, we expect these issues to improve as well, but the extent to which AI agents can supplant real-world observations is an ongoing topic of research.

The rest of the paper proceeds as follows. Section 2 provides a discussion of how LLMs are currently being used in management research, as well as how they can be used to aid theory building. Section 3 introduces the GABE framework. Section 4 briefly describes the application to the streetlight effect experiment and benchmarks GABE against previous results using human subjects. Section 5 describes how our framework can be deployed to extend the theory of the streetlight effect and develop new mechanisms. Section 6 concludes with a discussion of the potential uses and limitations of our framework in other settings.

2 LLMs as tools for Management Research

2.1 How Do Researchers Use LLMs?

Recent advancements in generative AI are poised to reshape the field of social science. This is especially true with the rise of LLMs, which are deep learning models trained on pre-existing data and capable of producing original content, conditional on a sequence of human prompts. LLMs have captured the attention of the public because of their impressive performance in a variety of domains (Bubeck et al., 2023), such as image generation (Qu et al., 2023) and computer programming (Kazemitabaar et al., 2024).

Increasingly, LLMs are being used as tools for academic research, with their versatility proving to be useful for various tasks (Grimes et al., 2023).[1] For instance, researchers have used LLMs to clean datasets (Chong et al., 2022), perform data analysis (Ma et al., 2023), and generate visualizations (Ye et al., 2024). Beyond data tasks, LLMs have been used extensively in scientific writing, even for generating abstracts and drafts. For instance, it is estimated that up to 20% of the content in computer science conference papers is now substantially AI-modified (Liang et al., 2024). Among other use cases, LLMs have also been leveraged to conduct systematic literature reviews (Agarwal et al., 2024) and refine research instruments such as survey questionnaires (Grimes et al., 2023; Charness et al., 2023). Despite legitimate concerns that LLMs may compromise the quality of research (Lindebaum and Fleming, 2024), it is plausible that automating these tasks will represent a large productivity boon (Korinek, 2024).

There is one feature of LLMs that may prove to be particularly useful for academic research: they have a remarkable propensity to exhibit human-like traits. For example, a recent study by Mei et al. (2024) subjected LLM chatbots to a personality test and a series of behavioral games, finding that their responses were statistically indistinguishable from randomly picked human subjects. Another body of research finds that LLMs rely on human-like heuristics and are prone to similar cognitive errors, reasoning, and even moral judgment (Dillion et al., 2023; Lampinen et al., 2024; Hagendorff, 2024). Given these similarities, a growing number of researchers are seeking to understand whether there are insights we can learn about

[1] Our focus is on understanding how LLMs can be used in the research process and, more generally, strategizing endeavors. However, it must be noted that a florid literature is exploring the direct impact of AI on firm activities and performance (Berg et al., 2023; Csaszar et al., 2024; Dell'Acqua et al., 2023; Doshi et al., 2024; Jia et al., 2024). 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). Our paper complements these strands of literature by focusing on the use of LLMs to develop strategy theory.

humans just by indirectly studying the behavior of LLMs (Bail, 2024; Grossmann et al., 2023). In this context, LLMs would not be inputs into the research process, but rather the subjects of study, serving as proxies for humans (Manning et al., 2024; Park et al., 2023). While LLMs do not perfectly mirror human decision-making (Tjuatja et al., 2023; Mohammadi, 2024), they might provide a flawed but potentially very useful approximation as "homo silicus" (Horton, 2023).

Indeed, an active body of literature is developing around this promise. Scholars have tried to use LLMs as survey respondents to uncover consumer demand (Brand et al., 2023; Li et al., 2024a) and predict voting behavior (Argyle et al., 2023). However, most research to date has focused on replicating lab-based experiments using LLMs in the place of human subjects. Aher et al. (2023) and Ashokkumar et al. (2024) replicate a suite of behavioral experiments, from the wisdom of crowds to the Milgram shock experiment, finding results consistent with past studies using human subjects. Horton (2023) investigates the strategic capabilities of LLMs by having them participate in games like the prisoner's dilemma or the dictator game. While these games involve two players only, more recent studies have incorporated more players and richer design elements, like information deficiencies and spatial reasoning (Wu et al., 2023; Xu et al., 2023). This bodes well for LLMs' ability to capture the complexities of real-world phenomena studied in strategy.

However, there are important limitations to studies replicating past experiments. First, there is a valid concern that the language models used are simply regurgitating results from famous studies they have been trained on, which presents a risk for generalization. Second, the study authors did not conduct the original experiments, which means their ability to recreate the study conditions is limited by a lack of access to the original protocols. This makes comparing results harder. Finally, and most importantly, these studies are generally not creating new knowledge. Instead of using LLMs to expand or develop theory, these studies have used results validated with human subjects to explore the capabilities of the LLMs themselves. The encouraging results indicate the potential to run new in silico experiments that could help expand our theoretical knowledge, yet with only a couple of recent exceptions (Binz and Schulz, 2023; Li et al., 2024b), these studies have mostly abstained from doing so.

2.2 Management Theorizing With LLMs

While researchers have more recently begun to use LLMs to augment the theorizing process, they have focused on LLM-led hypothesis generation and research ideation. For instance, Manning et al. (2024) study automated hypothesis generation employing LLMs to propose potential causal relationships and design experiments to test those relationships. Doshi and Hauser (2024) find that when fiction writers use generative

AI to obtain ideas for a story, the narratives are evaluated as more novel but tend to be more similar to each other. Similar results are found by Si et al. (2024) when tasking an LLM with generating new research ideas and comparing them with human experts.[2] Researchers have also used LLMs for hypothesis generation in psychology (Tong et al., 2024), materials science (Park et al., 2024), biomedicine (Qi et al., 2024), and mathematical modeling (Shojaee et al., 2024).[3]

While this body of work is intriguing, we believe some level of care is needed to appropriately situate LLMs into the theorizing process. To be sure, there are cases when the theory space is so vast and under-developed that LLMs can generate significant value by suggesting creative directions and novel corners to explore (Tranchero, 2024). However, often researchers will either start from existing theories or develop an initial "simple theory" (Davis et al., 2007). In these cases, the emphasis shifts from novelty to good judgment, and deciding how best to progressively build on established theory. For now, exercising this judgment remains largely the prerogative of humans (Agrawal et al., 2019). We envision an increasingly central role for LLMs in theory development, but one that is steered by the researcher working from a starting theory. This is our primary focus: we are less concerned with the development of de novo theories (Manning et al., 2024) and more with understanding how generative AI and human researchers can work in tandem (Mollick, 2024).

In particular, we believe the researcher can leverage the likeness of AI agents with humans to experiment with extensions to the starting theory. Once the theory is operationalized in an experimental setup, the researcher can simulate it using LLMs as stand-ins for human subjects. Any promising insights gained from this in silico experiment can then later be tested via real-world experiments or observational data. This approach can enhance theory development in two ways. Firstly, it can enable the researcher to theorize about potential underlying causal mechanisms by tweaking micro-level aspects of a complex system (Davis et al., 2007). This is especially true if one takes advantage of LLMs' language generation capabilities and solicits rationales for their decisions, which can help to make sense of the observed behavior.[4] The other benefit of this approach is that it can provide a clearer impression of boundary conditions for models that cannot be solved in a closed form. Articulating all the relevant assumptions and then relaxing them in the simulations can help to both generalize and bound theories (Makadok et al., 2018).

[2] In a related paper, Jia et al. (2024) find that AI can increase human creativity through an alternative channel: automating repetitive tasks while freeing up time to engage with more intellectually challenging tasks, hence stimulating individual creativity. Note that this mechanism does not leverage AI's ability to create per see, but highlights how complementarities between AI and humans can unlock superior innovation.

[3] A related line of work explores the use of LLMs for automated data-driven discovery, including generating novel hypotheses from datasets (Ludwig and Mullainathan, 2024; Majumder et al., 2024; Gu et al., 2024).

[4] While this carries the risk of hallucinations, the researcher can reduce this risk by verifying that the behaviors and rationales are consistent with one another. Our application in the following Sections showcases how this can be done in practice.

How does this approach compare to existing research methods? Laboratory and field experiments remain the gold standard for testing theoretical predictions (Chatterji et al., 2016). These experiments are conducted on the actual subjects of interest (i.e., humans or organizations), so observations represent the "ground truth." This is not true for experiments using LLMs, which can show idiosyncratic behavior (Mohammadi, 2024) or rationality biases (Hagendorff, 2024). However, lab and field experiments are typically conducted when there is already an established theory, which has generated key predictions that are ready to be tested (Card et al., 2011). Real-world experiments are generally far too cumbersome (and expensive) to be used to explore potential extensions in an open-ended manner, which is a crucial part of theory development (Mueller, 2018).[5] As such, LLMs are uniquely suited to the early stages of theory development.

In silico experiments with LLMs also share similarities with traditional simulation methods, such as ABMs. These methods are similarly highly agile, which means they have been used extensively in theory development, not least in strategic management (Ganco and Hoetker, 2009; Harrison et al., 2007). However, they have some drawbacks. It can be challenging to fully capture the complexities of strategic management settings given the constraints imposed by the fixed structures (Arend, 2022), as exemplified by the stylized NK models (Davis et al., 2007; Levinthal, 1997). The naive automaton agents in ABMs need to be exogenously pre-determined by the researcher, which means they do not default to approximating human behavior (Ganco, 2017). Most importantly, these agents cannot verbalize the "thinking" behind their decision-making, which limits the researcher's ability to engage in abductive reasoning (Makadok et al., 2018). None of these challenges apply to in silico experiments using LLMs.

In summary, LLMs have the potential to be used in management research. Going beyond basic applications like data analysis and scientific writing, we believe they can increase the pace of theory development. While researchers have tried placing LLMs in the pilot seat to autonomously generate new research ideas, we argue they are best used for augmenting human-led theorizing. In the following Sections, we develop an application to show how LLMs can help researchers explore extensions to theories with in silico experiments.

3 A Framework for Simulating Experiments

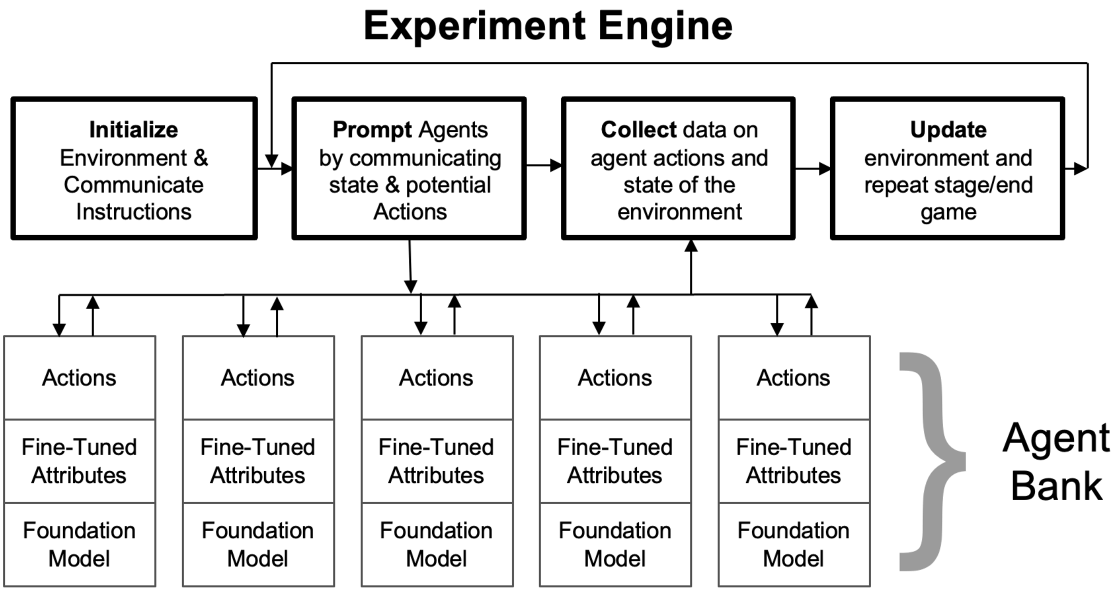
In this section, our goal is to provide a framework that enables management scholars to simulate experiments where multiple LLMs or AI agents interact in a rich strategic environment. We call this framework Generative AI-Based Experimentation, or GABE for short. We envision GABE as a systematic, bottom-up

[5] Moreover, experiments are strictly monitored by university Institutional Review Boards (IRBs) and necessarily constrained from performing a variety of interventions that carry the risk of harming or being unethical.

methodology that deploys micro-level autonomous AI agents in controlled settings, where they are given opportunities to interact organically with each other. These interactions can produce emergent behaviors that are not presumed by the researcher from the outset and are unknown to the individual LLMs. Throughout the simulations, the decisions, rationales, and outcomes of each AI agent are meticulously tracked, enabling an in-depth study of strategic interactions. In this sense, GABE has much in common with more traditional simulation methods (Davis et al., 2007), but it harnesses the power of generative AI to mimic actual human behavior.

Because these experiments are conducted completely in silico, some computational infrastructure is required. In particular, the researcher will need to set up a central deterministic engine that administers the entire experimental process virtually. The engine is tasked with spawning AI agents, defining their roles and behaviors, providing instructions, conveying new information as the experiment proceeds, and recording any outcomes. Figure 1 provides a stylized description of this process. Note that this infrastructure is versatile because the functions are mostly agnostic to the particular experiment, which means only minor tweaking will be needed when tailoring the engine to a new study. Once operational, the researcher can theoretically run an infinite number of simulations. Further, in contrast to studies replicating survey-based experiments (Ashokkumar et al., 2024), this engine-based setup is better suited for studying complex, repeated interactions among multiple agents.

Figure 1: Stylized GABE Engine



Note: This figure presents a stylized depiction of the central deterministic engine that powers an interactive experiment using GABE. The engine spawns the AI agents (each consisting of an action space, fine-tuned personality, and underlying foundation model) and then administers the experiment in silico following the steps above.

The researcher will need to supply the key parameters of each particular experiment. We describe the main degrees of freedom available to the researcher in Appendix A.1. In short, the choices include the features of the strategic environment, such as the choice landscape, the payoff structure, spatial features, and the information set. They also include the properties of the AI agents themselves, such as their goals, capabilities, and action space. Important choices concern the protocols for how decisions are made (e.g. simultaneously or sequentially) and whether to add channels for communication between AI agents. The researcher must also decide which outcomes to collect, such as quantitative data that can be analyzed using traditional econometric techniques, or qualitative responses that can support abductive theorizing. Finally, the researcher must choose which interventions to apply. These could include altering the information structure, agents preferences, payoffs, and communication structures, to name just a few. As with lab experiments involving human subjects and ABM simulations, these parameters should be determined by the starting theory that motivates the research question (Davis et al., 2007).

4 Implementing GABE in Management Research: An Application

4.1 The Theory of the Streetlight Effect

We choose to apply GABE in the context of strategic exploration, where agents scour over a fitness landscape in search of valuable discoveries (Levinthal, 1997; Kauffman, 1992). This setting captures many activities in innovation and entrepreneurship, such as venture capital funds deciding which startups to invest in (Lerner and Nanda, 2020), or pharmaceutical companies deciding which genes to target for drug development (Tranchero, 2024). In Hoelzemann et al. (2024), this process is formalized using a strategic multi-armed bandit model: agents explore from a set of risky projects and the value of these projects (low, medium, or high) is only learned upon exploration. The paper theorizes about what happens when external data is introduced on the value of risky projects. In particular, the authors show that when data sheds light on a medium-value opportunity, it can actually reduce welfare compared to when no data is available. This provides a theoretically grounded basis for the "streetlight effect", defined as the tendency for people to search where data is most readily available and convenient, often to their own detriment (Haynes et al., 2018; Bulgheresi, 2016).[6]

We focus on this theory for several reasons. First, we have a closed-form mathematical specification for the underlying theoretical framework. This allows us to precisely contextualize both the human and LLM

[6] This phenomenon takes its name from the parable of a drunkard searching for his keys late at night. He focuses his search underneath a lamp post, since this is "where the light is".

decisions using the calculated theoretical optimums. Second, Hoelzemann et al. (2024) test their theory with human subjects in an online lab experiment with strategic inter-dependencies and uncertainty. These are elements that characterize many real-world settings, so simulating these experiments in silico will be helpful to gauge whether LLMs can be useful for strategic management research.[7] Finally, we can be reasonably confident that our simulations will constitute out-of-sample predictions. There is a legitimate concern in previous studies replicating classic human-subject experiments that the models were simply regurgitating memorized results from the training data. In contrast, GPT-4 is largely unfamiliar with the experiment we chose, helping us bypass this concern.[8]

4.2 The Online Lab Experiment: Searching Mountains for Hidden Gems

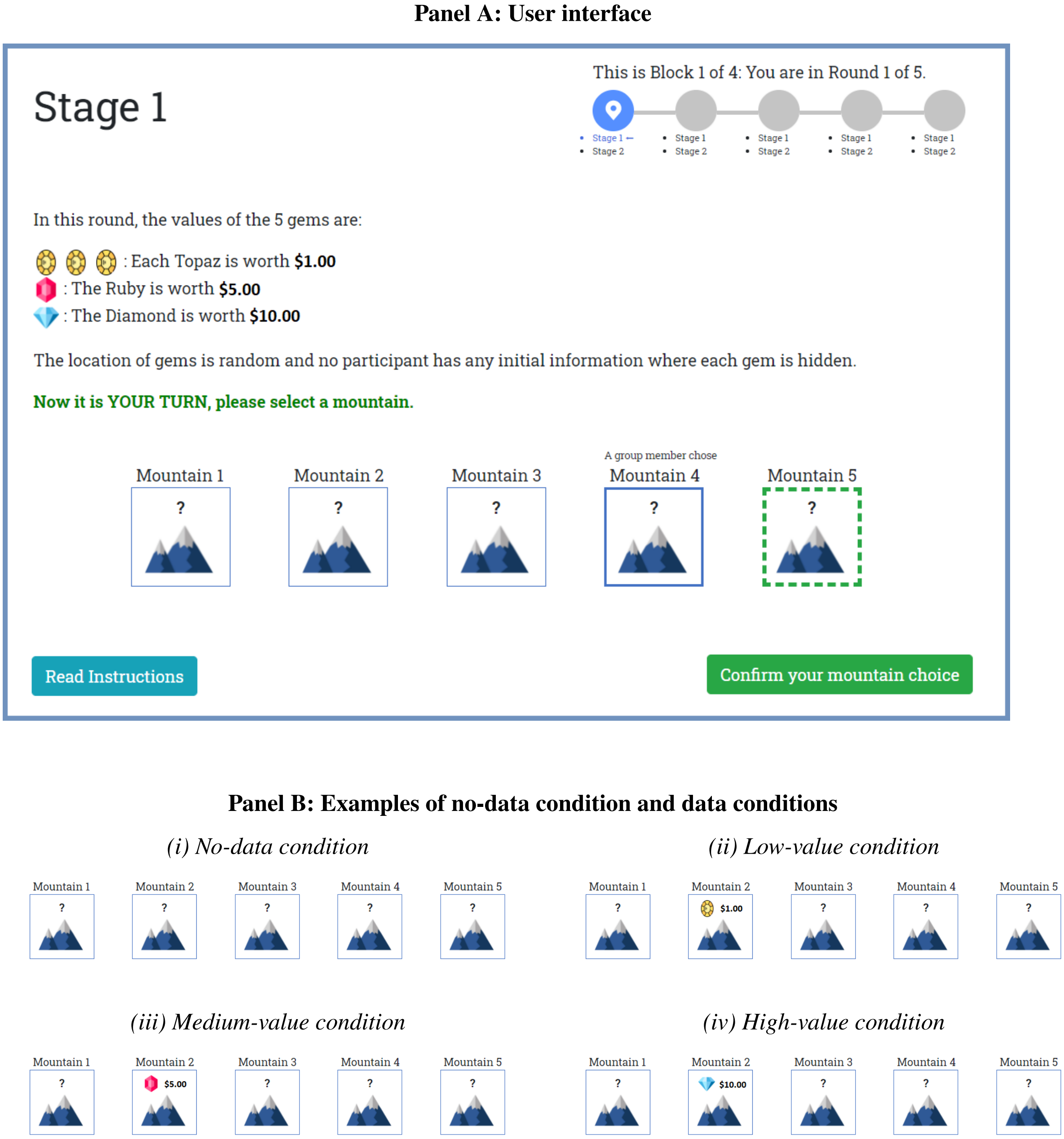
The experiment of Hoelzemann et al. (2024) features a group of participants searching across a virtual range of mountains for hidden gems. The setup is shown in Panel A of Figure 2. There are $n = 5$ players and $m = 5$ mountains. Three of the mountains contain topazes, one mountain contains a ruby, and one mountain contains a diamond. 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. The experiment consists of two periods: in the first period, each player selects a mountain in sequential order. After everyone has finished selecting, the gems behind the selected mountains will be revealed. In the second period, players choose again, this time equipped with the knowledge of gems revealed in the first period. Each player earns the sum total 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. Right from the beginning, players are explicitly told the number of gems and their values.

In the baseline condition, participants are not shown any data on the location of the gems. However, the experiment has several other treatment conditions, each providing partial initial data on payoffs. These are depicted in Panel B of Figure 2. 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 350 participants and was conducted over 1400 rounds, with each round consisting of 5 players and 2 periods.

[7] An additional advantage is that we enjoy unfettered access to the original raw materials and results. In previous studies of this kind, the authors must do their best to faithfully adhere to the original study designs. In our setting, we can be confident to compare results like-for-like, and we can also introduce extensions that are still true to the model at hand.

[8] We verify this by asking GPT-4 to describe the experiment in Hoelzemann et al. (2024). The answers were nonsensical and largely hallucinated. This is not surprising since the version of GPT-4 that we use was last updated in December 2023, while the NBER Working Paper was only published in 2024.

Figure 2: Experimental platform.



Panel A: User interface

Stage 1

This is Block 1 of 4: You are in Round 1 of 5.

In this round, the values of the 5 gems are:

: Each Topaz is worth **$1.00**
: The Ruby is worth **$5.00**
: The Diamond is worth **$10.00**

The location of gems is random and no participant has any initial information where each gem is hidden.

Now it is YOUR TURN, please select a mountain.

Panel B: Examples of no-data condition and data conditions

(i) No-data condition

(ii) Low-value condition

(iii) Medium-value condition

(iv) High-value condition

Note: This figure depicts the software platform used in the online lab experiment of Hoelzemann et al. (2024). Panel A shows how the interface is seen by participants in the no-data condition. The values of the gems for each round are shown in the upper left corner. In this example, the user can see that Mountain 4 has already been picked and decides to select Mountain 5. Panel B depicts the four experimental conditions. For instance, in plot (iii), the mountain uncovering the ruby is revealed at the start of the round. The figure is reproduced with permission from Hoelzemann et al. (2024).

We implement this experiment in silico using GABE. The key parameters that we supply from the GABE framework are shown below in Table 1. We keep the same setup as Hoelzemann et al. (2024). However, this time, it is a group of AI agents selecting which mountains to explore each round. We use Expected Parrot Domain-Specific Language (EDSL), an open-source Python package that provides a wrapper between user prompts and the GPT-4 API (Horton et al., 2024). EDSL allows us to spawn the five AI agents at a time, familiarize them with the instructions of the experiment, and prompt them sequentially for their choice of mountains.[9] Once an AI agent chooses a mountain, the code updates the conditions of the environment and informs the other AI agents about this choice. See Appendix A.2 for an excerpt of the script we use. We simulate 500 rounds of the online lab experiment using AI agents in place of human subjects.

Table 1: Application of GABE to the Streetlight Effect Experiment

Parameter	**Description**
Environment	There are 5 choices (i.e. mountains) available to all AI agents. Each choice is associated with a payoff, depending on whether it holds a topaz, ruby, or diamond. There are no spatial features to the environment. Information is incomplete: agents do not know the locations of gems. 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.
Agent	There are 5 AI agents. Each agent has the same objective: they are explicitly instructed to maximize their individual earnings. There are no differences in capabilities or in action spaces.
Decision-rules	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 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.

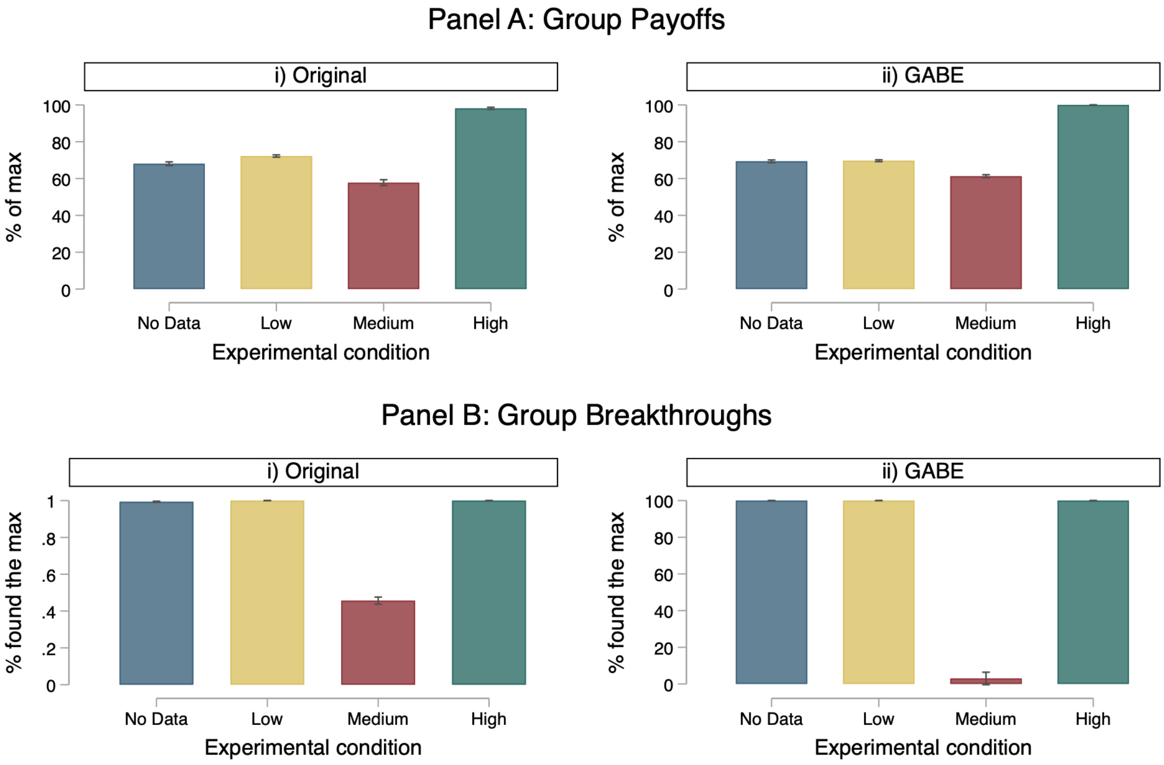
Note: This table illustrates how we specify the key parameters of GABE to replicate the online lab experiment of Hoelzemann et al. (2024). See the text and Appendix A.1 for more details.

4.3 Comparing Human Subjects and AI Agents

We begin by comparing the results of using AI agents to conduct the streetlight experiment with those obtained from human participants in online lab experiments. The results of this exercise are presented in Figure 3. The first primary outcome of interest is the group payoff (Panel A), which is the sum total of all individual payoffs in a group for a given round. We plot the mean group payoff across all rounds, as well as the 95% confidence intervals. The original results from the online lab experiment are presented in the upper left corner in plot i), which confirms the key prediction of the streetlight effect. When shown the location of the ruby – the medium-valued gem – human participants tend to earn the least amount of money. Further analysis reveals they avoid risking a search for the (higher-valued) diamond and instead herd around the safer ruby option (Appendix Figure B1). This initially boosts their payoffs in the first period but ultimately leads to lower earnings by the end of the round (Appendix Table B1). In contrast, when the location of the (lowest-valued) topaz is revealed – or when no gem is revealed – humans feel compelled to search for a better option. This leads them to coordinate their search efforts and uncover the diamond.

[9] EDSL also enables us to prime the agents to have different objectives or features. For more details, see Section 5.3.

Figure 3: Comparing the Streetlight Experiment with Human Subjects and AI Agents



Note: This figure compares the results of the streetlight 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 mean likelihood of a breakthrough, which occurs when at least one group member finds the diamond. Plots i) show the results obtained in the original lab experiment using human subjects, while plots ii) show the results obtained from the simulated experiments using AI agents.

How closely do AI agents mimic this strategic behavior? These results are presented in plot ii). We find that the behavior of the AI agents is remarkably similar, even in this socially intricate setting. AI agents tend to earn the least amount of money when shown the location of the ruby. Like humans, they decide not to explore any further, and instead herd around the safer ruby option (see Appendix Figure B2 and Appendix Table B2). When shown only the location of the topaz (or no gems at all), AI agents are also able to effectively coordinate their search efforts, using the sequential order of choice to explore all options and ultimately achieve higher payoffs. When the location of the diamond is revealed, they immediately select it to earn the maximum payoffs. In other words, LLMs reproduce the same patterns and reveal the key, though subtle, insight that more data is not always advantageous in search.

Our other outcome of interest is whether the group achieves a breakthrough (Panel B), which occurs when at least one member discovers a diamond. Once more, the patterns are similar. Both AI agents and humans achieve a breakthrough in nearly every round where a ruby is not revealed (either because they collectively search for the diamond, or it is revealed to them) but are less likely to do so when the ruby is revealed. Yet, one disparity emerges in the latter case: whereas humans achieve a breakthrough at least some of the time (i.e., in roughly 45% of rounds), AI agents almost never do. Indeed, Hoelzemann et al. (2024) proves that we should not expect any breakthroughs to occur from a purely rational perspective, suggesting that LLMs behave more in line with our theoretical expectations. This echoes other research that finds that language models tend to behave more rationally than humans (Hagendorff, 2024). Furthermore, qualitative interviews with humans revealed idiosyncratic reasons for deviating from the ruby: some acted out of boredom, while others chose randomly. These behaviors appear to be largely absent in AI agents.

5 Going Beyond the Original Experiment

Having shown that AI agents can reasonably approximate the behavior of human subjects in our setting, we now introduce a series of exploratory experiments to enrich our understanding of the streetlight effect. In particular, we try extending the original experiment by a) varying the experimental setup, b) relaxing underlying theoretical assumptions, and c) incorporating heterogeneous agent preferences, with a view towards probing the robustness of its boundary conditions and uncovering new potential mechanisms.

5.1 Varying the Experimental Setup

The first set of extensions concerns the setup of the experiment itself. Often, when researchers wish to test a theory in an experiment, they need to first operationalize the theory and specify the key parameters. Accordingly, while the theoretical framework for the streetlight effect in Hoelzemann et al. (2024) is general, the corresponding lab experiment required a series of choices pertaining to the timing of the rounds, the number of agents and mountains, the magnitudes of the payoffs, and even how to best socialize the setting to the participants (e.g., calling risky projects ‘mountains’ and payoffs ‘gems’). While some of these choices were likely inconsequential, others might have influenced the results. How robust were the results to generalizing these decisions? To shed light on these inner workings, we conduct 800 additional rounds of simulated experiments using AI agents, varying one of three major game mechanics at a time.

A. Varying Group Size: We begin by varying the number of agents, keeping the number of mountains fixed.[10] We try running the streetlight experiment with additional AI agents, starting with ten agents and incrementing by 10 until we have thirty AI agents participating in the simulation at a time, to reflect more crowded search settings (Erat and Krishnan, 2012). As plots i) of Figure 4 show, we find that adding more

[10] Having at least as many agents as mountains guarantees that agents would make a breakthrough if they just chose to coordinate. However, enforcing a parity between the number of agents and the number of mountains was a specific design choice of Hoelzemann et al. (2024).

players does not diminish the streetlight effect. Even in large groups, AI agents still tend to cluster around the ruby, failing to make a breakthrough. When we ask them for explanations for this behavior, they allude to a force that we had not previously considered and suggest a new intriguing mechanism for the streetlight effect: social conformity. As one AI agent tells us:

> *"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".*

Another AI agent concurs and expresses the same idea in slightly different terms:

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

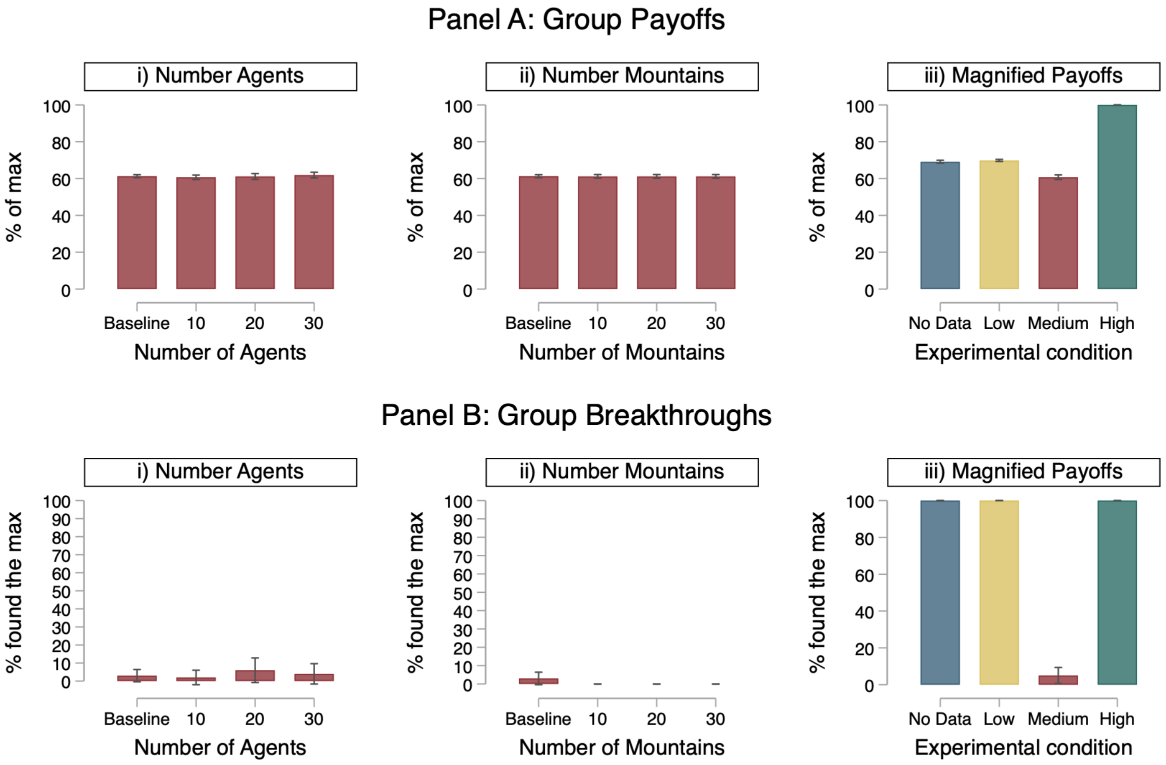
This mirrors other studies where conformity or emulation in group settings drives behavior (Lazer and Friedman, 2007; Puranam and Swamy, 2016). Sometimes, AI agents wrongly infer that the large mass of other agents herding on the same options belies some bit of information that they have missed (Banerjee, 1992). Whether the same dynamics would be at play with human subjects is a theoretically interesting question that can later be tested in a laboratory with humans.[11]

B. Varying Choice Landscape: The next experimental mechanic that we vary is the number of mountains, keeping the number of agents fixed. While the agents are no longer guaranteed to find a breakthrough with coordinated search, this reflects large search spaces where agents cannot search every possibility (LiCalzi and Surucu, 2012; Tranchero, 2024). We try running the streetlight experiment with additional mountains, starting with ten mountains and incrementing by ten until AI agents must choose from thirty mountains at any time. As plots ii) of Figure 4 shows, we once more find no significant change in results: AI agents continue to cluster around the ruby. They remain unswayed by the greater (absolute) number of diamonds (and rubies) and behave according to the original logic. If anything, exploration activity seems to even decrease, suggesting that larger search spaces may pose an additional hindrance in coordinating individual experimentation.

C. Varying Payoff Magnitude: The final experimental mechanic that we vary is the absolute magnitudes of the payoffs. While the relative payoffs of the gems were already pre-determined by the theoretical framework, Hoelzemann et al. (2024) needed to choose the actual dollar amounts. They chose amounts below 15$ per

[11] It could also be that the idiosyncratic behavior of human subjects may be sufficient to unearth a breakthrough in a crowded space, thus counteracting the tendency to herd in settings with strong social influence. This exemplifies well how GABE can lead to uncovering interesting theoretical tensions.

Figure 4: Changing the experimental setup using GABE



Note: This figure shows how the outcomes of the streetlight experiment change when we adjust the experimental setup using GABE. 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 mean 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 (5 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 (5 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.

gem to keep the experiment affordable, but we might wonder how the intensity of the streetlight effect would have changed if players were earning up to several millions in a given round. This would reflect search settings with high stakes in decision-making, such as pharmaceutical innovation and venture capital (Lou and Wu, 2021; Bhatia and Dushnitsky, 2023). Of course, this experiment would be impossible to run with actual human subjects, since they would be entitled to keep the high earnings. But it is one we can simulate with AI agents. We try running the streetlight experiment with the payoffs increased by a millionfold; the results of this exercise are presented in plots iii) of Figure 4. We find similar results to the baseline. This aligns with the theoretical framework of Hoelzemann et al. (2024), which suggests that only the relative values of the gem should matter. To what extent this finding generalizes to humans is open to debate, as the psychological impact of these stakes might differ, but the use of AI agents generates a reasonable prediction that would otherwise be unattainable.

5.2 Relaxing theoretical assumptions

The next set of extensions that we introduce relates to the underlying theoretical framework. There are two key assumptions that are embedded in the framework: that payoffs are non-rivalrous and non-ambiguous. This means that multiple agents who choose the same gem do not have to split the payoff, and the distribution of gems is known from the outset. As a result, the online lab experiments were also designed with these conditions in mind. But we might wonder how strongly the streetlight effect depends on these assumptions, and whether we would still observe any free riding if payoffs were rival and/or ambiguous, as is often the case in many real-world settings (Rahmandad et al., 2021). Knowing this would also give us a more complete understanding of the conditions under which the streetlight effect is likely to emerge. This forms the basis of our next analysis, in which we conduct an additional 1500 rounds of simulated experiments and relax the assumptions one at a time. The baseline case (where payoffs are non-rivalrous and non-ambiguous) is presented in plots i) of Panels A and B in Figure 5.

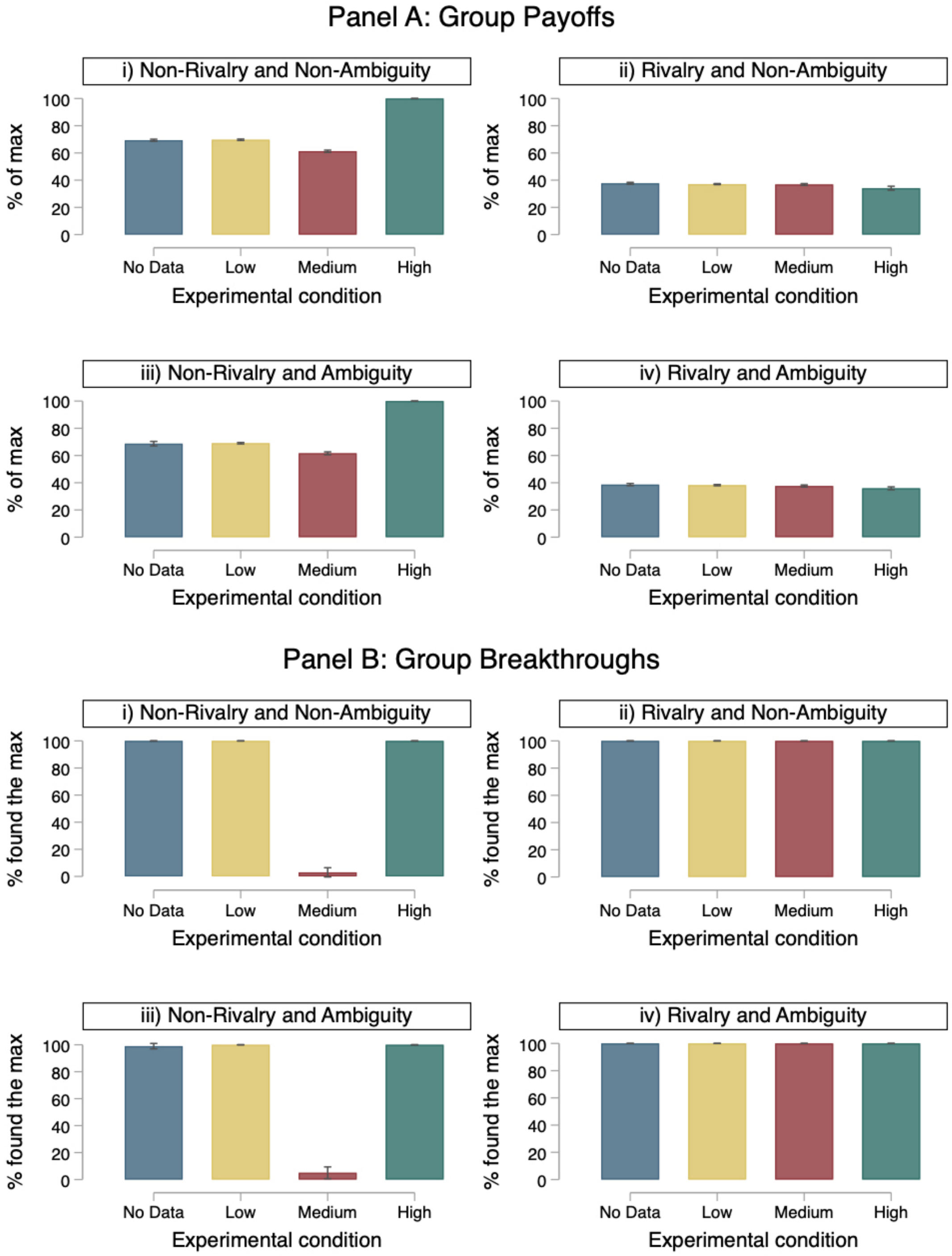
We first try making payoffs rival, which means that any agents who choose the same gem must split the payoffs. The results of this exercise are shown in plots ii) of Panel A and Panel B. Strikingly, we find that introducing rivalry leads the streetlight effect to disappear entirely. This suggests we have uncovered a potentially important new boundary condition for the theory. It is no longer the case that agents make the least money when the ruby is revealed (Panel A), nor do they make fewer breakthroughs (Panel B). In general, because agents can observe the choices of other agents (thanks to the sequential order of choice), they will explore an unchosen mountain simply to avoid splitting rents. The qualitative responses of AI confirm these dynamics and show how competition acts as a powerful incentive for risky exploration. As one AI agent tells us:

> *I chose Mountain 5 to avoid competition and potentially discover the Diamond or another Topaz, maximizing my potential earnings by not sharing the already known Ruby value in Mountain 3 with another agent.*

As a result, we also find that payoffs begin to equalize across the four conditions, although they are significantly diminished, which is analogous to competition eliminating profits in a market.

Next, we try making payoffs ambiguous, which means we no longer inform the agents how many gems of each kind there are from the outset. We begin by keeping payoffs non-rival. The results of this exercise are shown in plots iii) of Panel A and Panel B. Interestingly, we continue to find strong evidence of the streetlight effect even when we introduce ambiguity. The group payoffs (Panel A) and group breakthroughs (Panel B)

Figure 5: Relaxing Theoretical Assumptions using GABE



Note: This figure shows how the outcomes of the streetlight experiment change when we relax key theoretical assumptions. 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 mean likelihood of a breakthrough, which occurs when at least one group member finds the diamond. Plots i) show the baseline results, where there is no rivalry or ambiguity in payoffs. In plots ii), we introduce rivalry in payoffs, which means agents who choose the same gem do not have to split the pay. In plots iii), we introduce ambiguity in payoffs, which means the distribution of gems is known from the outset. In plots iv), we introduce both rivalry and ambiguity in payoffs.

are nearly identical to the baseline: AI agents earn the least money when shown the location of the ruby, and they almost never achieve a breakthrough. This indicates that the streetlight effect may not, in fact, depend on prior information about the distribution of payoffs, but only about the relative value of payoffs.

Finally, we try making payoffs both rivalrous and ambiguous. This allows us to see which assumption exerts a greater force on the behavior of AI agents, or if there is any interaction effect that results from relaxing both assumptions. The results of this exercise are shown in plots iv) of Panels A and B. Interestingly, the streetlight effect once more disappears entirely, and the agents behave almost identically to the case where payoffs are rival but non-ambiguous. This suggests that rivalry exerts a much stronger force when acting upon the participants and that, of the two, it might be the more relevant precondition. This provides an important boundary condition that was hard to foresee before experimenting with GABE.

5.3 Manipulating agent preferences and objectives

The final set of extensions concerns the goals and preferences of agents. Hoelzemann et al. (2024) speculate that risk-aversion and other alternative decision rules could have some role to play in explaining the experimental results, and possibly even the strength of the streetlight effect. However, studying this in an experimental context is challenging, as varying the risk tolerance of human subjects can be impractical. In contrast, this process becomes feasible with AI agents: one notable feature of LLMs is their ability to be endowed with specific preferences, views, demographics, or personalities from the outset (Horton, 2023; Aher et al., 2023), which would simply involve priming them with additional scripts. To explore whether this is true in practice, we run an additional 1000 rounds of simulated experiments where we endow the LLMs with different preferences.

To begin with, we try introducing higher risk tolerance. We run the baseline streetlight experiment with risk-loving agents, starting with one risk-loving agent and incrementing by one until all AI agents are risk-lovers.[12] Since we are interested in the intensity of the streetlight effect, we once more focus only on the data condition where the ruby is revealed. We find that exploration monotonically increases with the number of risk-loving agents (Figure 6 Panel B), and consequently, more players are able to locate a breakthrough (Figure 6 Panel A). To verify that the risk endowment is working as intended, we review the agents' rationales, which are quite instructive. As one risk-lover tells us:

[12] Before running this extension, we have also assessed the risk propensity of AI agents at baseline. Using the traditional Holt & Laury task (Hoelzemann et al., 2024), we confirm research showing how AI agents based on LLM tend to be risk-neutral as the average participant in academic lab studies (Mei et al., 2024). We also repeat the same preference elicitation task after priming AI agents with their respective risk profiles, finding that our priming significantly changes how agents respond to the Holt & Laury task in the expected direction.

> *"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, we find that when more agents are primed to take risks, the remaining agents (who were not primed) follow suit and increase their exploration. In other words, risk tolerance can trigger an implicit coordination process. As one non-risk-lover tells us, it makes more sense to explore when they are now highly likely to collectively uncover the diamond:

> *"I choose Mountain 1 because Mountains 3, 4, and 5 have already been chosen by other participants. Given that Mountain 2 is the only other remaining option aside from Mountain 1, I randomly select Mountain 1.*

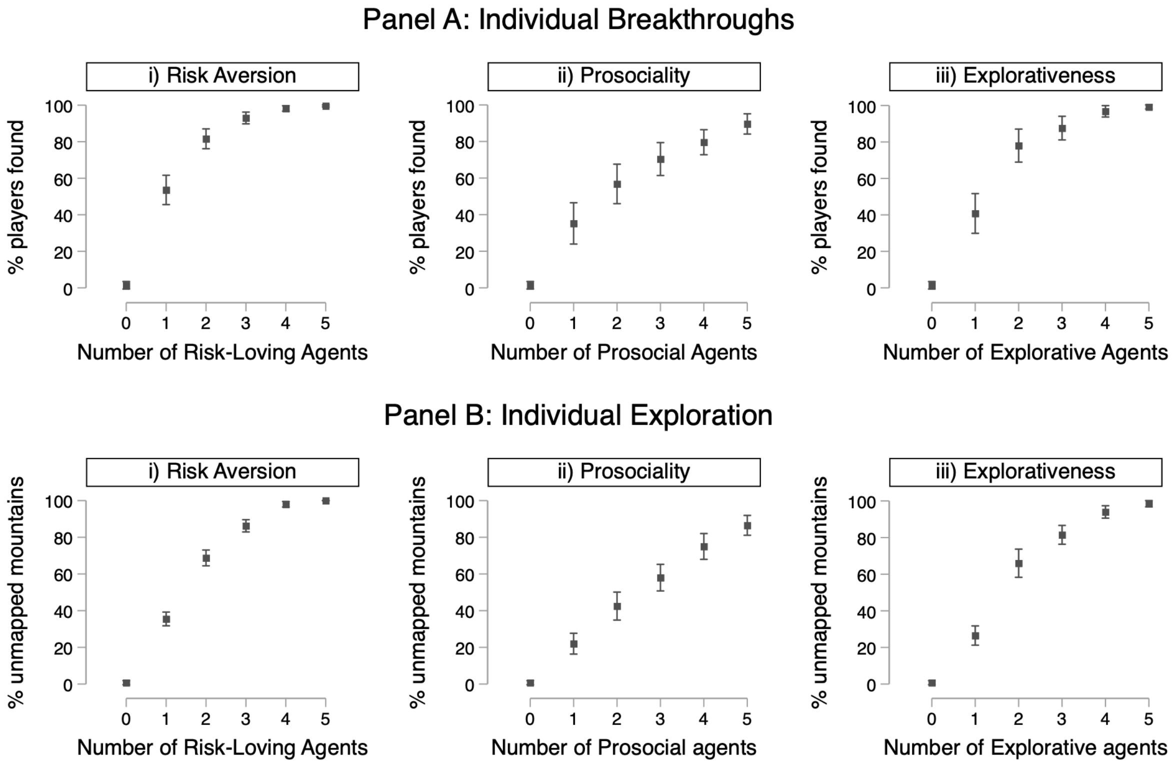
This means the number of risk-lovers would need to grow less than proportionally to produce enough exploration, a prediction we can later test in the laboratory.

Next, we try introducing pro-sociality. This time, we run the baseline streetlight experiment with pro-social agents, starting with one pro-social agent (out of the five players) and incrementing until all five AI agents are pro-social. We tell the primed agents they are only successful if the others choose the diamond, which means they should try to generate information that gives the group the best chance of discovering a diamond. We once more focus only on the data condition where the ruby is revealed. The results of this exercise are presented in plots ii) of Panel A and Panel B in Figure 6. Like before, we see that exploration monotonically increases with the number of pro-social agents, as does the share of agents that find the diamond. Once again, the rationales are instructive. As one pro-social agent tells us,

> *"Since I am pro-social and aim to give other participants the chance to discover a diamond, I will choose a mountain that has not been selected by others yet. All other agents have chosen Mountain 4, which contains a ruby. To maximize the group's potential earnings and to explore the possibility of finding a diamond, I will choose Mountain 1.*

Finally, we try introducing a taste for exploration. We run the baseline streetlight experiment with explorative agents, starting with one explorative agent and incrementing until all AI agents are explorative. We explicitly tell primed agents that their sole objective is to find the diamond and that they should continue exploring

Figure 6: Incorporating Agent Preferences into GABE



Note: This figure shows how the outcomes of the streetlight experiment change when we incorporate heterogeneous agent preferences using GABE. 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.

until they do so. The results of this exercise are presented in plots iii) of Panel A and Panel B in Figure 6. The results are nearly identical to the case with risk-loving agents, and we see that exploration (unsurprisingly) increases monotonically with the number of explorative agents.

This completes our extensions to the original streetlight experiment using GABE. A summary of the full set of results is presented in Table 2. Before we provide further discussion, we briefly explore whether these results were also achievable through direct elicitation using GPT-4.

5.4 Comparing GABE to Direct Elicitation

So far, we have used LLMs to simulate individual human subjects, which is analogous to the bottom-up approach of agent-based modeling. But we might wonder if there was an easier way to achieve the same insights and extend management theory. In particular, if LLMs are already quite advanced at performing complex tasks, and they are only getting better over time, perhaps we can just ask an LLM to predict the

Table 2: Summary of results using GABE

Intervention	Description	Rounds (#)	Cost ($)	Hypothetical Payment ($)	Summary of findings
Baseline	The original experiment with 5 players and 5 mountains	500	250	50,000	Players herd around the suboptimal option, lowering payoffs
Extension 1: Varying the experimental setup					
Varying group size	We vary the size of groups from 5 agents to 30 agents	150	300	60,000	No change in exploration: social conformity reinforces the decision to herd around suboptimal option
Varying choice landscape	We vary the number of mountains from 5 mountains to 30 mountains	150	75	15,000	No change in exploration: agents unswayed by greater number of absolute options
Varying payoff magnitude	We amplify gem values one millionfold	500	250	-	No change in exploration: Streetlight Effect depends on relative magnitudes of payoffs
Extension 2: Relaxing theoretical assumptions					
Relaxing payoff rivalry	Agents who choose the same gem must split the payoffs	500	250	50,000	Rivalry breaks the Streetlight Effect: agents no longer exhibit herding and want to avoid splitting payoffs
Relaxing payoff ambiguity	We no longer inform the agents how many gems of each kind there are.	500	250	50,000	No change in exploration: Streetlight Effect does not depend on prior information about distribution
Relaxing payoff rivalry and ambiguity	Agents must split payoffs, and we no longer inform agents of the distribution	500	250	50,000	Rivalry exhibits a greater force than ambiguity, continuing to break the Streetlight Effect
Extension 3: Manipulating agent preferences					
Incorporating risk-aversion	Some portion of the group is now risk-loving, varying from 1 agent to all 5 agents.	500	250	50,000	Risk aversion breaks the Streetlight Effect - exploration increases with the number of risk-lovers
Incorporating pro-sociality	Some portion of the group is now pro-social, varying from 1 agent to all 5 agents.	250	125	25,000	Pro-sociality breaks the Streetlight Effect - exploration increases with the number of pro-social agents
Incorporating explorativeness	Some portion of the group is now explorative, varying from 1 agent to all 5 agents.	250	125	25,000	Explorativeness breaks the Streetlight Effect - exploration increases with the number of explorative agents
Total		**3800**	**2125**	**$375,000**	*(Excludes payoffs from amplified case)*

Note: This table provides an overview of all the results from this study. Column 1 lists the extensions we introduce. Column 2 provides a description of each extension. Column 3 indicates the number of rounds simulated. Column 4 shows the approximate total cost in GPT-4 credits as of July 2024. Column 5 highlights the amount we would have needed to pay to human subjects. Column 6 summarizes the key result. Row 2 covers the baseline replication of the original lab experiment. Rows 4-6 cover changes in the experimental setup. Rows 8-10 cover key theoretical assumptions being relaxed. Rows 12-14 cover heterogeneous agent preferences being introduced.

outcomes of the experiments outright (Horton, 2023). Not only could it provide point estimates, but the LLM could also explain its reasoning, which could then be used to verify the results and potentially even explore underlying mechanisms. If this were true, the researchers would save themselves even more time, effort, and resources than if they were to simulate experiments using GABE. Of course, this depends entirely on the ability of language models to accurately predict the outcomes of complex social interactions.

To explore the extent to which this is currently feasible, we tasked GPT-4 with predicting the outcomes of the baseline streetlight experiment. The results of this exercise are presented in Appendix Table B3. We find that while direct elicitation can sometimes produce reasonable predictions, the overall accuracy of these predictions remains modest at best. More importantly, GPT-4 fails to grasp the core mechanisms underlying the theory and consequently misses the subtle (but core) insight that more data is not always better. It is also clear that using AI agents per the GABE framework outperforms direct elicitation in replicating the original experiment, which lends further support to our bottom-up methodology. Finally, we confirm that the predictions of GPT-4 also do not align with the results of GABE for several of the extensions that we ran.

6 Discussion

In this study, we have shown how Generative AI-Based Experimentation (GABE) can be a powerful tool for extending management theory. Firstly, we validate that LLMs can closely approximate human behavior in strategic group settings. In our replication of the lab experiment in Hoelzemann et al. (2024), AI agents showed a similar tendency to cluster around safer but sub-optimal choices. This demonstrates GABE's potential to extend the theory of the streetlight effect by simulating exploratory experiments. More importantly, we actively implement several of these extensions. We first demonstrate the ease with which GABE enables modifications to the experimental setup, such as group sizes, the choice landscape, and payoff magnitudes. Then, we show how GABE can be used to relax underlying theoretical assumptions, such as payoff rivalry and ambiguity. Finally, we demonstrate how to incorporate heterogeneous preferences using GABE, priming individual AI agents to be risk-averse, pro-social, or have a taste for adventure. By leveraging LLMs' language generation capabilities to interpret these simulations in the appropriate manner, we uncover intriguing new mechanisms, like social conformity and emulation, as well as plausible boundary conditions, like rivalry and widespread risk aversion.

These findings from the streetlight experiment provide compelling proof of concept for using GABE to extend management research. Nevertheless, we expect GABE to have broader applicability, and expanding the list of use cases remains a promising avenue for further research. While our focus has been on organizational search and strategic exploration, other social phenomena are equally ripe for investigation. For instance, researchers could use GABE to deepen our understanding of large-scale discrimination. This might entail creating AI-managed organizations and having them evaluate candidate resumes, mimicking previous audit studies. As in this study, researchers could then implement various extensions, like manipulating applicant attributes to identify subtle biases that would be costly and time-consuming to detect with human participants. There may also be significant potential to use GABE to simulate field experiments. For instance, researchers could use GABE to model interactions between authentic economic agents (e.g., farmers, traders, and consumers), allowing researchers to explore the impact of field interventions (e.g., subsidies, pricing strategies, or market conditions) prior to real-world implementation. The more studies of this kind are conducted, the better we will understand both the unique strengths of the GABE framework, as well as its inherent limitations.

At the same time, we recognize this vision will require sufficient future engagement with GABE from the management research community. As of now, the lack of familiarity with transformer-based language models and the significant technical barriers to entry may hinder such engagement. Therefore, in future work, we hope to directly address this gap. While the engine that we programmed to implement the GABE framework is specifically designed for the experiments presented in this paper, it will become necessary to develop more standardized tools and protocols that can be easily adapted by other researchers to study a wider range of settings. With this in mind, we plan to release software that automates the process of simulating multi-agent strategy experiments for researchers who are mostly unfamiliar with LLMs. This should help to significantly increase broader engagement with GABE from the management community.

Our hope is that this framework marks a meaningful step forward in integrating generative AI into management research, moving beyond conventional uses of LLMs such as data processing and scientific writing. By enabling researchers to simulate complex social interactions at a fraction of the time and cost of traditional experiments and using AI agents that are more verbally responsive than traditional simulation methods, GABE has the potential to improve the way management scholars approach the stage of theory development. While much work remains to be done, the findings from our study suggest that AI agents can serve as a powerful catalyst for advancing our understanding of strategic interactions in large-scale social systems.

References

Agarwal, S., I. H. Laradji, L. Charlin, and C. Pal (2024): "LitLLM: A Toolkit for Scientific Literature Review," *arXiv preprint arXiv:2402.01788.*

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

Aher, G., R. I. Arriaga, and A. T. Kalai (2023): "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," *arXiv preprint.*

Arend, R. J. (2022): "Balancing the perceptions of NK modelling with critical insights," *Journal of Innovation and Entrepreneurship*, 11, 23.

Argyle, L. P., E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate (2023): "Out of one, many: Using language models to simulate human samples," *Political Analysis*, 31, 337–351.

Ashokkumar, A., L. Hewitt, I. Ghezae, and R. Willer (2024): "Predicting Results of Social Science Experiments Using Large Language Models," *Working Paper.*

Bail, C. A. (2024): "Can Generative AI improve social science?" *Proceedings of the National Academy of Sciences*, 121.

Banerjee, A. V. (1992): "A simple model of herd behavior," *The Quarterly Journal of Economics*, 107, 797–817.

Berg, J. M., M. Raj, and R. Seamans (2023): "Capturing value from artificial intelligence," *Academy of Management Discoveries*, 9, 424–428.

Bhatia, A. and G. Dushnitsky (2023): "The Future of Venture Capital? Insights Into Data-Driven VCs," *California Management Review*.

Binz, M. and E. Schulz (2023): "Turning large language models into cognitive models," *arXiv preprint arXiv:2306.03917*.

Brand, J., A. Israeli, and D. Ngwe (2023): "Using GPT for market research," *Harvard Business School Marketing Unit Working Paper*.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, et al. (2023): "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *arXiv preprint*.

Bulgheresi, S. (2016): "Bacterial cell biology outside the streetlight," *Environmental Microbiology*, 18, 2305–2318.

Card, D., S. DellaVigna, and U. Malmendier (2011): "The role of theory in field experiments," *Journal of Economic Perspectives*, 25, 39–62.

Charness, G., B. Jabarian, and J. A. List (2023): "Generation next: Experimentation with AI," *NBER Working Paper w31679*.

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

Chong, D., J. Hong, and C. D. Manning (2022): "Detecting label errors by using pre-trained language models," *arXiv preprint arXiv:2205.12702*.

Choudhury, P., R. T. Allen, and M. G. Endres (2021): "Machine learning for pattern discovery in management research," *Strategic Management Journal*, 42, 30–57.

Csaszar, F. A., H. Ketkar, and H. Kim (2024): "Artificial Intelligence and Strategic Decision-Making: Evidence from Entrepreneurs and Investors," *Available at SSRN 4913363*.

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

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

Demirdjian, D., L. Taycher, G. Shakhnarovich, K. Grauman, and T. Darrell (2005): "Avoiding the "streetlight effect": tracking by exploring likelihood modes," *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, 1, 357–364.

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

Doshi, A. R., J. J. Bell, E. Mirzayev, and B. Vanneste (2024): "Generative Artificial Intelligence and Evaluating Strategic Decisions," *Available at SSRN 4714776*.

Doshi, A. R. and O. P. Hauser (2024): "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Science Advances*, 10, eadn5290.

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

Erat, S. and V. Krishnan (2012): “Managing delegated search over design spaces,” *Management Science*, 58, 606–623.

Felten, E. W., M. Raj, and R. Seamans (2023): “Occupational heterogeneity in exposure to generative ai,” *Available at SSRN 4414065.*

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

Ganco, M. and G. Hoetker (2009): “NK modeling methodology in the strategy literature: Bounded search on a rugged landscape,” in *Research Methodology in Strategy and Management*, Emerald Group Publishing Limited, 237–268.

Grimes, M., G. Von Krogh, S. Feuerriegel, F. Rink, and M. Gruber (2023): “From scarcity to abundance: Scholars and scholarship in an age of generative artificial intelligence,” *Academy of Management Journal*, 66, 1617–1624.

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

Gu, K., R. Shang, R. Jiang, K. Kuang, R.-J. Lin, D. Lyu, Y. Mao, Y. Pan, T. Wu, J. Yu, et al. (2024): “BLADE: Benchmarking Language Model Agents for Data-Driven Science,” *arXiv preprint.*

Hagendorff, T. (2024): “Deception abilities emerged in large language models,” *Proceedings of the National Academy of Sciences*, 121.

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

Haynes, W. A., A. Tomczak, and P. Khatri (2018): “Gene Annotation Bias Impedes Biomedical Research.” *Scientific Reports*, 8.

Hoelzemann, J., G. Manso, A. Nagaraj, and M. Tranchero (2024): “The streetlight effect in data-driven exploration,” *NBER Working Paper w32401.*

Horton, J., A. Filippas, and R. Horton (2024): “EDSL: Expected Parrot Domain Specific Language for AI Powered Social Science,” Whitepaper, Expected Parrot.

Horton, J. J. (2023): “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” *NBER Working Paper w31122.*

Huang, J.-t., E. J. Li, M. H. Lam, T. Liang, W. Wang, et al. (2024): “How Far Are We on the Decision-Making of LLMs? Evaluating LLMs’ Gaming Ability in Multi-Agent Environments,” *arXiv preprint.*

Jia, N., X. Luo, Z. Fang, and C. Liao (2024): “When and how artificial intelligence augments employee creativity,” *Academy of Management Journal*, 67, 5–32.

Kauffman, S. A. (1992): “Origins of order in evolution: self-organization and selection,” in *Understanding origins: Contemporary views on the origin of life, mind and society*, Springer, 153–181.

Kazemitabaar, M., R. Ye, X. Wang, A. Z. Henley, P. Denny, M. Craig, and T. Grossman (2024): “CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs,” *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Korinek, A. (2024): “LLMs Level Up—Better, Faster, Cheaper: June 2024 Update to “Generative AI for Economic Research: Use Cases and Implications for Economists”,” *Journal of Economic Literature*, 61, 4.

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

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

Lerner, J. and R. Nanda (2020): "Venture capital's role in financing innovation: What we know and how much we still need to learn," *Journal of Economic Perspectives*, 34, 237–261.

Levinthal, D. A. (1997): "Adaptation on rugged landscapes," *Management Science*, 43, 934–950.

Li, P., N. Castelo, Z. Katona, and M. Sarvary (2024a): "Frontiers: Determining the validity of large language models for automated perceptual analysis," *Marketing Science*, 43, 254–266.

——— (2024b): "Frontiers: Determining the validity of large language models for automated perceptual analysis," *Marketing Science*, 43, 254–266.

Liang, W., Y. Zhang, Z. Wu, H. Lepp, W. Ji, X. Zhao, H. Cao, S. Liu, S. He, Z. Huang, et al. (2024): "Mapping the increasing use of llms in scientific papers," *arXiv preprint arXiv:2404.01268*.

LiCalzi, M. and O. Surucu (2012): "The power of diversity over large solution spaces," *Management Science*, 58, 1408–1421.

Lindebaum, D. and P. Fleming (2024): "ChatGPT undermines human reflexivity, scientific responsibility and responsible management research," *British Journal of Management*, 35, 566–575.

Lou, B. and L. Wu (2021): "AI on Drugs: Can Artificial Intelligence Accelerate Drug Development? Evidence from a Large-Scale Examination of Bio-Pharma Firms," *Management Information Systems Quarterly*, 45, 1451–1482.

Ludwig, J. and S. Mullainathan (2024): "Machine learning as a tool for hypothesis generation," *The Quarterly Journal of Economics*, 139, 751–827.

Ma, P., R. Ding, S. Wang, S. Han, and D. Zhang (2023): "InsightPilot: An LLM-empowered automated data exploration system," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 346–352.

Majumder, B. P., H. Surana, D. Agarwal, B. D. Mishra, A. Meena, et al. (2024): "DiscoveryBench: Towards Data-Driven Discovery with Large Language Models," *arXiv preprint*.

Makadok, R., R. Burton, and J. Barney (2018): "A practical guide for making theory contributions in strategic management," *Strategic Management Journal*, 39, 1530–1545.

Manning, B. S., K. Zhu, and J. J. Horton (2024): "Automated social science: Language models as scientist and subjects," *NBER Working Paper w32381*.

Mei, Q., Y. Xie, W. Yuan, and M. O. Jackson (2024): "A Turing test of whether AI chatbots are behaviorally similar to humans," *Proceedings of the National Academy of Sciences*, 121.

Mohammadi, B. (2024): "Wait, It's All Token Noise? Always Has Been: Interpreting LLM Behavior Using Shapley Value," *arXiv preprint*.

Mollick, E. (2024): *Co-Intelligence*, Random House UK.

Mueller, J. (2018): "Finding new kinds of needles in haystacks: Experimentation in the course of abduction," *Academy of Management Discoveries*, 4, 103–108.

Park, J. S., J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein (2023): "Generative Agents: Interactive Simulacra of Human Behavior," *arXiv preprint arXiv:2304.03442*.

Park, Y. J., D. Kaplan, Z. Ren, C.-W. Hsu, C. Li, H. Xu, S. Li, and J. Li (2024): "Can ChatGPT be used to generate scientific hypotheses?" *Journal of Materiomics*, 10, 578–584.

Puranam, P. and M. Swamy (2016): "How initial representations shape coupled learning processes," *Organization Science*, 27, 323–335.

Qi, B., K. Zhang, K. Tian, H. Li, Z.-R. Chen, S. Zeng, E. Hua, H. Jinfang, and B. Zhou (2024): "Large Language Models as Biomedical Hypothesis Generators: A Comprehensive Evaluation," *arXiv preprint*.

Qu, L., S. Wu, H. Fei, L. Nie, and T.-S. Chua (2023): "LayoutLLM-T2I: Eliciting Layout Guidance from LLM for Text-to-Image Generation," *Proceedings of the 31st ACM International Conference on Multimedia*, 643–654.

Rahmandad, H., J. Denrell, and D. Prelec (2021): "What makes dynamic strategic problems difficult? Evidence from an experimental study," *Strategic Management Journal*, 42, 865–897.

Rumelt, R. P., D. Schendel, and D. J. Teece (1991): "Strategic management and economics," *Strategic Management Journal*, 12, 5–29.

Shojaee, P., K. Meidani, S. Gupta, A. B. Farimani, and C. K. Reddy (2024): "Llm-sr: Scientific equation discovery via programming with large language models," *arXiv preprint*.

Shrestha, Y. R., V. F. He, P. Puranam, and G. von Krogh (2021): "Algorithm supported induction for building theory: How can we use prediction models to theorize?" *Organization Science*, 32, 856–880.

Si, C., D. Yang, and T. Hashimoto (2024): "Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers," *arXiv preprint*.

Sutton, R. I. and B. M. Staw (1995): "What theory is not," *Administrative Science Quarterly*, 371–384.

Tjuatja, L., V. Chen, S. T. Wu, A. Talwalkar, and G. Neubig (2023): "Do LLMs exhibit human-like response biases? A case study in survey design," *arXiv preprint*.

Tong, S., K. Mao, Z. Huang, Y. Zhao, and K. Peng (2024): "Automating Psychological Hypothesis Generation with AI: Large Language Models Meet Causal Graph," *arXiv preprint*.

Tranchero, M. (2024): "Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation," *University of Pennsylvania*.

Wu, Y., X. Tang, T. M. Mitchell, and Y. Li (2023): "Smartplay: A benchmark for LLMs as intelligent agents," *arXiv preprint*.

Xu, Y., S. Wang, P. Li, F. Luo, X. Wang, W. Liu, and Y. Liu (2023): "Exploring large language models for communication games: An empirical study on werewolf," *arXiv preprint*.

Ye, Y., J. Hao, Y. Hou, Z. Wang, S. Xiao, Y. Luo, and W. Zeng (2024): "Generative ai for visualization: State of the art and future directions," *Visual Informatics*.

A Deeper Dive Into Generative-AI Based Experimentation (GABE)

A.1 Key Parameters for Experiments

Environment: The researcher must specify the common set of conditions that the agent encounter during an in silico experiment. The following conditions are especially relevant to strategic management:

1 **Choice landscape:** The array of options available to all agents. With GABE, the computational environment could theoretically support an unlimited number of choices.

2 **Payoff structure:** The magnitude of rewards associated with each choice in a given choice landscape. With GABE, the computational environment could theoretically support unlimited combinations of earning schedules. Researchers can also easily add other complex dynamics, like variable rewards or payoff rivalry.

3 **Spatial features:** The physical layout of the experiment, including the positions of agents. With GABE, the computational environment supports highly tunable search landscapes that might capture the spatial features usually represented in NK models (Kauffman, 1992; Levinthal, 1997).

4 **Information set:** The degree to which the elements above are common knowledge. For instance, agents may know the options available, but not the payoffs associated with each option.

Agents: The researcher must also specify the number of agents and the characteristics of each agent. Agents may differ in the following ways:

1 **Goals:** Each agent may have a different objective or set of preferences. For example, agents can be explicitly instructed to profit-maximize or to be pro-social and value collective welfare in their decisions. Alternatively, the agents can be assigned specific roles, which might contain implicit objectives that the language model needs to infer. For example, a "student agent" might focus on maximizing learning, while a "firm agent" might choose to maximize profits.

2 **Capabilities:** Each agent may have a different set of capabilities or resources to achieve their objective. For instance, an agent might incur lower costs when carrying out certain actions as a result of their skills, or a firm might have slack resources to invest in certain opportunities.

3 **Action space:** Each agent may be able to take a different course of action at any given point in time. This may dynamically change over the experiment, perhaps as a result of learning, or it could vary by agent as a function of their capabilities.

Decision rules: The researcher must specify how decisions are made by agents. There are two primary methodologies they can choose from:

1 **Sequential decision-making:** Agents are allotted a "turn" to make their decision. After each turn, the state space is updated and the experiment proceeds until a predetermined endpoint has been reached. The researcher would need to decide how turns are allotted, e.g., using a random order or perhaps having an independent LLM interact with each agent, and then decide (Manning et al., 2024).

2 **Simultaneous decision-making:** Agents take actions at the same time within a round. After each round, the state space is updated and the experiment proceeds until a predetermined endpoint has been reached. The researcher would simply need to decide the number of rounds (Huang et al., 2024).

Communication channels: The researcher must decide whether the AI agents can speak with each other during the course of the experiment (or engage in other forms of interaction, such as trading). This functionality must be programmed into the engine but remains feasible with LLMs (Xu et al., 2023).

Outcomes: The researcher must specify which outcomes to record as a function of the research question and the theoretical constructs explored.

1 **Quantitative data:** This includes participant choices, their earnings as a function of collective decisions, the time or effort taken to make a decision, or any other measurable outcome. The engine is also able to calculate more complex corollary measures. This data could then be analyzed using traditional econometric techniques for rigorous inference.

2 **Qualitative responses:** This primarily includes the explanations that AI agents provide for their decisions, but could include anything in light of LLMs' language comprehension and generation capabilities (e.g. their subjective experiences navigating a certain environment). Collecting these responses is essential for eliciting mechanisms (see section 2.2), as well as for verifying the experiments are working as intended (e.g. seeing whether an agent is embodying an assigned role).

Interventions: These are isolated changes to any element of the game environment or the agents themselves. With GABE, it becomes possible to pull levers that are not normally available in real-world lab experiments, such as manipulating an agent's preferences or tolerance for risk. Once all the building blocks have been specified, the researcher can begin simulating the experiment, adjusting one element at a time and tracing how the outcomes of interest evolve.

A.2 Excerpt of Script Given to AI Agents

Welcome. This is an experiment in the economics of decision-making. If you pay close attention to these instructions, you can earn a significant amount of money paid to you at the end of the experiment. Following these instructions, you will be asked to make some choices. There are no correct choices. Your choices depend on your preferences and beliefs, so different participants will usually make different choices. You will be paid according to your choices, so read these instructions carefully and think before you decide.\n",

"The Basic Idea:\n",

"There are {len(MOUNTAINS_DISTRIBUTION)} mountains and each of them hides one type of gem, which can only be found by exploring the mountain. There are 3 types of gems hidden in the {len(MOUNTAINS_DISTRIBUTION)} mountains: Diamonds, Rubies, and Topazes. The exact values of the topazes, rubies, and diamonds vary across rounds but the diamonds are always worth more than the rubies and the rubies are always worth more than the topazes. You choose which mountains to explore and the value of the gems you find are your earnings in dollars. Your objective is to maximize your own earnings.\n",

"The Mapped Mountain:\n",

"At the beginning of each round, one mountain will be randomly selected to be mapped and its gem value will be revealed to all participants. Each participant will be able to see the same gem contained by the mountain. The mountain chosen for mapping is random and changes in each round. Besides the value of the mapped mountain, no participant has any other initial information in Stage 1 on the location of gems.\n",

"How Participants Choose Mountains:\n",

"There are {len(agent_ids)} participants including you in total. In each round, participants choose which mountain to explore. The choice does not happen simultaneously, but participants choose sequentially, one after the other, according to a random order. You can choose to explore any mountain you wish or select the mapped mountain. If you choose the same mountain chosen by other participants, each of you will receive the full value uncovered. Similarly, if someone else chooses the same mountain that you previously chose, you will still receive the full gem's value (and so will the other participant(s) who chose it). This means that payoffs are non-rival and there is no penalty in choosing the same mountain as other players. To repeat, no participant has any private information in Stage 1 on the location of the gems, besides the common knowledge about the mapped mountain.\n",

"Each Round Has 2 Stages:\n",

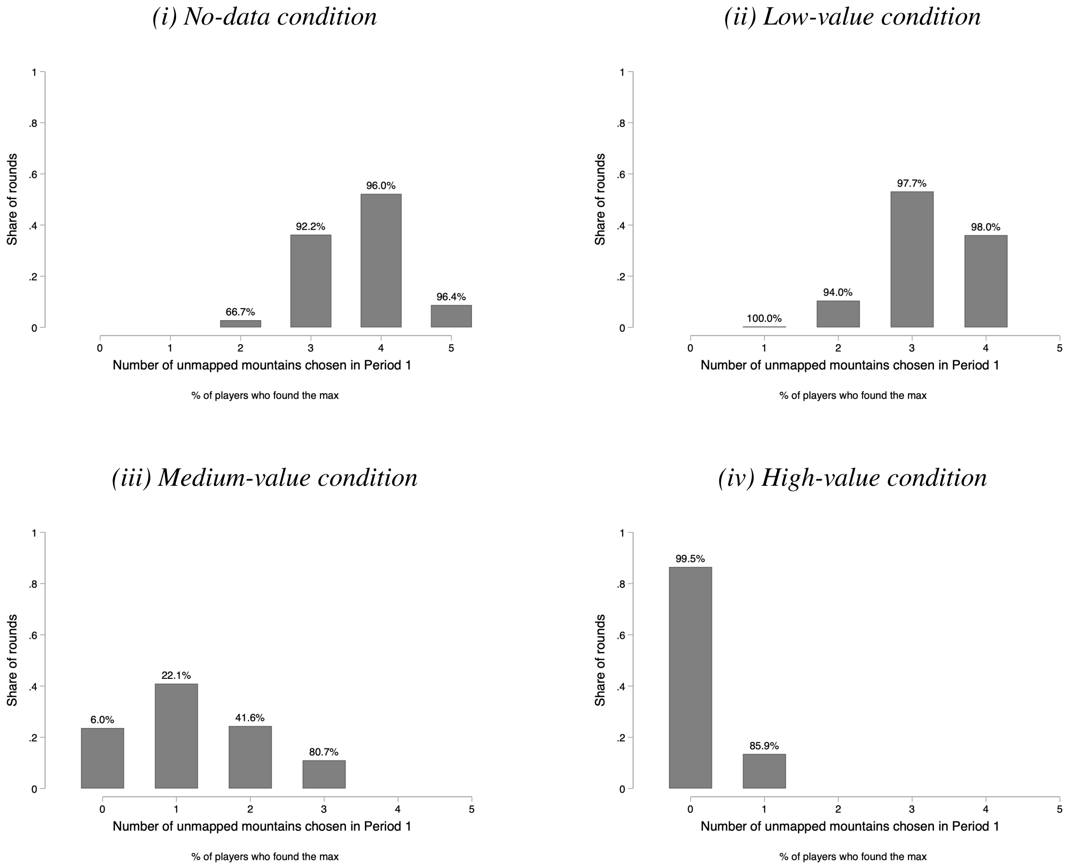
"A round consists of 2 stages. At the beginning of a new round, gems are randomly allocated to the {len(MOUNTAINS_DISTRIBUTION)} different mountains. The position of gems will not be reset between the two stages in a round. Then, before Stage 1 begins, one mountain will be mapped and its value revealed to everyone. In Stage 1, all participants sequentially choose one mountain to explore. Before choosing a mountain, you will see which mountains have been selected by the other participants in your group who chose before you. You can choose the same mountain chosen by other participants or a different mountain. At the end of Stage 1, the gems hidden in each mountain selected by all participants in Stage 1 are revealed, and you earn the value of the gem hidden in the mountain you chose. In Stage 2, you can again choose any of the same {len(MOUNTAINS_DISTRIBUTION)} mountains; that is, you can either choose the same mountain of Stage 1 or switch to another one. The position of gems remains the same as in Stage 1, but this time you will also see the gems located in the mountains revealed in Stage 1 in addition to the mapped mountain. At the end of Stage 2, the gems hidden in each mountain selected by all participants in Stage 2 are revealed, and you earn the value of the gem hidden in the mountain you chose in Stage 2. Your total earnings for the round equal the sum of the value of the gem you found in Stage 1 and the value of the gem you found in Stage 2. Again, if multiple players choose the same mountain, they all receive its full value.\n",

"Payment:\n",

"At the end of the round, you will be paid an amount equivalent to the sum of payoffs you earned in Stage 1 and Stage 2. This protocol of determining payments suggests that you should choose in each Stage knowing that your choice directly determines your payment because the dollar value of the gems you select will directly translate into your earnings. \n"

B Additional Figures and Tables

Figure B1: Number of unknown mountains chosen in period 1 by human subjects



(i) No-data condition

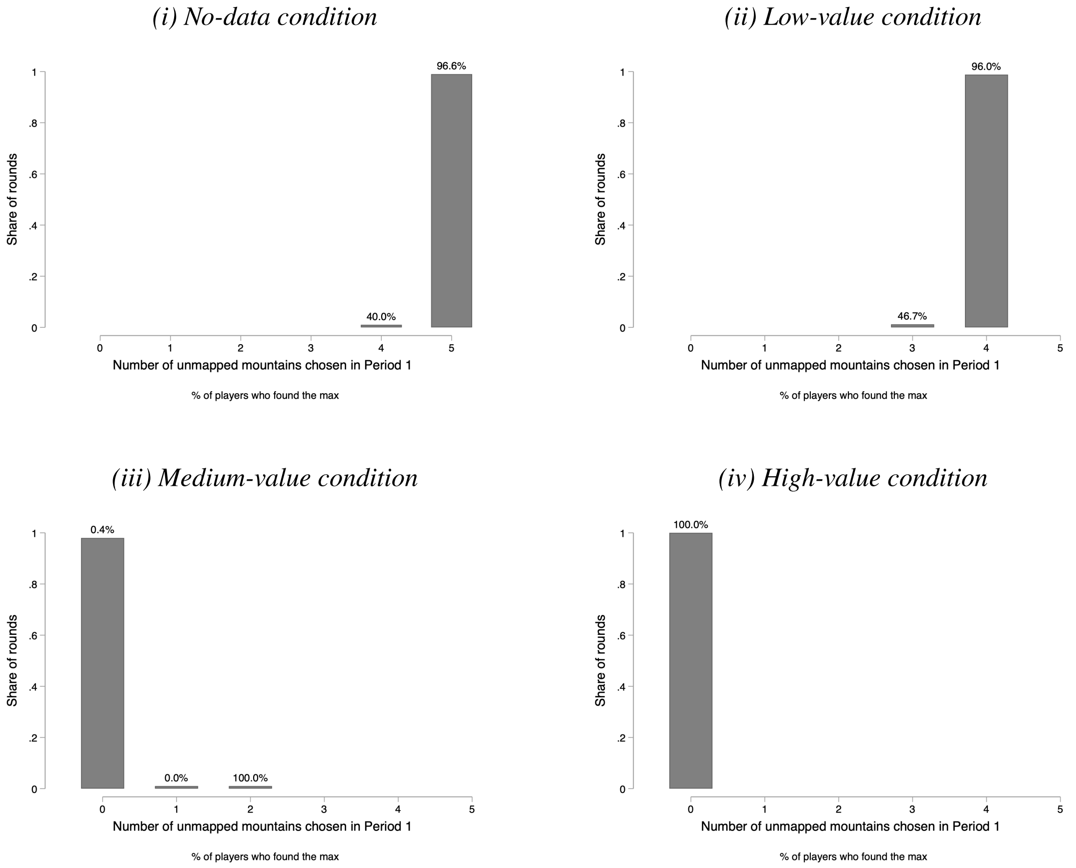
(ii) Low-value condition

(iii) Medium-value condition

(iv) High-value condition

Note: Each plot represents the empirical frequencies of rounds for each possible number of unknown options chosen in period 1, shown separately by experimental condition. The text written on top of each bar shows the share of participants who found the maximum payoff in period 2. The figure is reproduced with permission from Hoelzemann et al. (2024).

Figure B2: Number of unknown mountains chosen in period 1 by AI agents (baseline)


(i) No-data condition

(ii) Low-value condition

(iii) Medium-value condition

(iv) High-value condition

Note: Each plot represents the empirical frequencies of rounds for each possible number of unknown options chosen in period 1, shown separately by experimental condition. The text written on top of each bar shows the share of participants who found the maximum payoff in period 2.

Table B1: Quantitative results of experiment with human subjects

Panel A: Round-level Outcomes

	Individual payoff	I(Individual found max)	I(Group found max)
	(1)	(2)	(3)
High	6.682***	0.039**	0.003
	(0.143)	(0.011)	(0.006)
Low	0.889***	0.037**	0.006
	(0.096)	(0.011)	(0.007)
Medium	-2.261***	-0.645***	-0.539***
	(0.214)	(0.028)	(0.039)
Constant	13.349***	0.968***	1.012***
	(0.163)	(0.018)	(0.020)
Round order FE	Yes	Yes	No
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	7000	7000	1400

Panel B: Analysis of Mechanisms

	Exploration	Individual payoff		I(Individual found max)		I(Group found max)	
	Round	Period 1	Period 2	Period 1	Period 2	Period 1	Period 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
High	-75.059***	6.499***	0.191*	0.784***	0.037**	0.310***	0.003
	(1.770)	(0.097)	(0.073)	(0.008)	(0.011)	(0.015)	(0.006)
Low	5.744***	0.664***	0.225**	0.055***	0.036**	0.122***	0.006
	(1.300)	(0.074)	(0.062)	(0.007)	(0.011)	(0.020)	(0.007)
Medium	-34.130***	1.251***	-3.511***	-0.134***	-0.644***	-0.444***	-0.539***
	(2.450)	(0.122)	(0.142)	(0.007)	(0.027)	(0.029)	(0.039)
Constant	83.977***	3.610***	9.717***	0.187***	0.963***	0.752***	1.012***
	(2.045)	(0.104)	(0.117)	(0.008)	(0.018)	(0.018)	(0.020)
Round Order FE	No	Yes	Yes	Yes	Yes	No	No
Block order FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1400	7000	7000	7000	7000	1400	1400

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. Estimates from OLS models. In Panel A, the sample in Columns 1 and 2 is at the participant-round level (5 participants × 1400 rounds). The sample in Column 3 is at the group-round level (1400 rounds). *Individual payoff*= participant-level round payoffs in Canadian dollars; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the round. The excluded category captured by the constant is the condition without data.

In Panel B, the sample in Column 1 is at the group-round level (1400 rounds). The sample in Columns 2, 3, 4, 5 is at the participant-period level (5 participants × 1400 periods of each type). The sample in Columns 6 and 7 is at the group-period level (1400 periods of each type). *Exploration*= share of unknown mountains explored in the round; *Individual payoff*= participant-level period payoffs in Canadian dollars; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant in the period; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the period. The table is reproduced with permission from Hoelzemann et al. (2024).

Table B2: Quantitative results of experiment with AI agents (baseline)

Panel A: Round-level Outcomes

	Individual payoff	I(Individual found max)	I(Group found max)
	(1)	(2)	(3)
High	6.777***	0.022***	-0.001
	(0.183)	(0.007)	(0.002)
Low	0.067	-0.011	0.000
	(0.207)	(0.008)	(0.002)
Medium	-1.805***	-0.964***	-0.969***
	(0.182)	(0.008)	(0.017)
Constant	15.423***	0.978***	1.000***
	(0.176)	(0.007)	(0.001)
Round order FE	Yes	Yes	No
Payoff structure FE	Yes	Yes	Yes
Observations	2500	2500	500

Panel B: Analysis of Mechanisms

	Exploration	Individual payoff		I(Individual found max)		I(Group found max)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Round	Period 1	Period 2	Period 1	Period 2	Period 1	Period 2
High	-100.000***	6.569***	0.208***	0.802***	0.040***	0.010	0.000
	(0.085)	(0.183)	(0.049)	(0.018)	(0.009)	(0.010)	(0.002)
Low	-0.100	0.077	-0.009	0.007	-0.006	0.002	0.000
	(0.115)	(0.205)	(0.057)	(0.021)	(0.011)	(0.012)	(0.001)
Medium	-98.250***	2.258***	-4.063***	-0.196***	-0.946***	-0.980***	-0.970***
	(0.806)	(0.179)	(0.060)	(0.018)	(0.010)	(0.014)	(0.017)
Constant	100.000***	4.531***	10.892***	0.198***	0.960***	0.990***	1.000***
	(0.049)	(0.176)	(0.049)	(0.018)	(0.009)	(0.010)	(0.001)
Round order FE	No	Yes	Yes	Yes	Yes	No	No
Payoff structure FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	500	2500	2500	2500	2500	500	500

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease class level in parentheses. Estimates from OLS models. In Panel A, the sample in Columns 1 and 2 is at the participant-round level (5 participants × 500 rounds). The sample in Column 3 is at the group-round level (500 rounds). *Individual payoff*= participant-level round payoffs in Canadian dollars; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the round. The excluded category captured by the constant is the condition without data. See text for more details.

In Panel B, the sample in Column 1 is at the group-round level (500 rounds). The sample in Columns 2, 3, 4, 5 is at the participant-period level (5 participants × 500 periods of each type). The sample in Columns 6 and 7 is at the group-period level (500 periods of each type). *Exploration*= share of unknown mountains explored in the round; *Individual payoff*= participant-level period payoffs in Canadian dollars; *I(Individual found max)*:0/1=1 if the location of the maximum was found by the participant in the period; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the period.

Table B3: Comparing GABE to direct elicitation

				Mean group payoff (%)			Mean group breakthrough (%)		
Configuration	**Data**	**Rivalry**	**Ambiguity**	**Humans**	**AI Agents**	**Prediction**	**Humans**	**AI Agents**	**Prediction**
1	None	FALSE	FALSE	68	69	75	99	100	80
2	Low			72	70	62	100	100	80
3	Medium			58	61	75	46	3	80
4	High			98	100	95	100	100	100
5	None	TRUE	FALSE	-	38	45	-	100	40
6	Low			-	37	62	-	100	80
7	Medium			-	37	75	-	100	40
8	High			-	34	75	-	100	100
9	None	FALSE	TRUE	-	69	50	-	99	20
10	Low			-	69	62	-	100	40
11	Medium			-	62	62	-	5	40
12	High			-	100	75	-	100	100

Note: This table shows the results when we try asking GPT-4 to predict the outcomes of the Streetlight experiment. Column 1 lists the specific prediction task. Column 2 lists the data condition. Columns 3 and 4 list which theoretical assumptions apply. Column 7 shows GPT-4's direct prediction for the average group earnings (as a percentage of the maximum possible earnings). This is compared with the outcomes derived using human subjects (Column 5) and using AI agents (Column 6). Similarly, Column 10 shows GPT-4's direct prediction for the mean likelihood of a breakthrough, which occurs when at least one group member finds the diamond. This is compared with the outcomes derived using human subjects (Column 8) and using AI agents (Column 9). In Rows 1-4, we show the predictions for the baseline replication. In Rows 5-8, we show the predictions for the extension where we introduce payoff rivalry. In Rows 5-8, we show the predictions for the extension where we introduce payoff ambiguity.