Strategy Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Beyond Black Boxes: Designing and Testing Agentic AI Systems for Strategy

Arnaldo Camuffo, Alfonso Gambardella, Saeid Kazemi, Abhinav Pandey

To cite this article:
Arnaldo Camuffo, Alfonso Gambardella, Saeid Kazemi, Abhinav Pandey (2026) Beyond Black Boxes: Designing and Testing Agentic AI Systems for Strategy. Strategy Science 11(1):137-156. https://doi.org/10.1287/stsc.2025.0432

https://pubsonline.informs.org/journal/stsc

STRATEGY SCIENCE
Vol. 11, No. 1, March 2026, pp. 137–156
ISSN 2333-2050 (print), ISSN 2333-2077 (online)

Beyond Black Boxes: Designing and Testing Agentic AI Systems for Strategy

Arnaldo Camuffo,[a,b,*] Alfonso Gambardella,[a,b] Saeid Kazemi,[a,b] Abhinav Pandey[a,b]

[a] Department of Management & Technology, Bocconi University, 20136 Milan, Italy; [b] ION Management Science Lab (IMSL), SDA Bocconi, 20136 Milan, Italy

*Corresponding author

Contact: arnaldo.camuffo@unibocconi.it, https://orcid.org/0000-0001-9039-1057 (AC); alfonso.gambardella@unibocconi.it, https://orcid.org/0000-0002-8714-5813 (AG); saeid.kazemi@unibocconi.it, https://orcid.org/0009-0000-5221-4084 (SK); abhinav.pandey@unibocconi.it, https://orcid.org/0009-0005-4591-5401 (AP)

Received: May 1, 2025
Revised: September 21, 2025; November 30, 2025
Accepted: December 14, 2025
Published Online in Articles in Advance: February 20, 2026

https://doi.org/10.1287/stsc.2025.0432

Abstract. Existing work on AI and strategy examines the effects of general-purpose models, implicitly assuming that strategists will use whatever tools are available and that human-AI interaction is largely a matter of prompting, complacency, or aversion. Inspired by Herbert Simon's work on system architecture, we instead adopt a design view in which purposeful AI system design and use—not mere access to generic models—can itself be a dynamic capability and a source of competitive advantage. Building on this perspective, we develop Aristotle, an agentic multiagent AI system for theory-based strategic decision making, and study how it shapes strategists' reasoning, beliefs, and strategies compared with general AI assistance or no AI assistance. We document the design journey of Aristotle, highlighting design choice trade-offs in strategy framework, human-AI integration, and cost, and then implement a streamlined three-agent version suitable for experimental testing. In a randomized experiment with 976 managers comparing this agentic AI system, general AI (GPT-4o), and a human-only condition, we find that experienced managers achieve quality improvements without confidence inflation, whereas highly educated managers exhibit confidence gains without corresponding quality improvements. We establish user-system-problem fit as a core design dimension requiring alignment between architectural complexity and practitioner expertise. We abductively derive a five-dimensional taxonomy that maps the design space for agentic AI systems and a methodological roadmap that enable researchers and practitioners to experiment with, evaluate, and iteratively improve their AI system design choices.

History: Accepted for the Special Issue: Can AI Do Strategy?

Funding: A. Camuffo and A. Gambardella acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme [Grant 101021061].

Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsc.2025.0432.

Keywords: strategic decision making • artificial intelligence • agentic AI • multiagent AI systems • uncertainty • human-AI interaction

1. Introduction

Organizations are rapidly adopting artificial intelligence (AI) for strategic decision making. We lack systematic guidance—not only on how AI should be used but also on whether it should be purposely designed and customized. Most of the work in strategy and entrepreneurship examines general-purpose models and treats human-AI interaction mainly as a problem of prompting, aversion, or complacency. In this sense, AI is a black-box predictor or a generic conversational assistant. We take a different perspective. We argue that the *purposeful design and use* of AI systems, rather than mere access to generic models, can potentially be a dynamic capability and a source of competitive advantage. In strategic decision making, what matters is not only whether AI is used but which AI system is deployed and how it scaffolds and helps human reasoning in the development of a strategy. Thus, the choice between general-purpose large language models (LLMs) and purposely designed agentic AI systems is itself a strategic decision that shapes how managers decide under uncertainty.

The architecture of complexity (Simon 1962) offers a foundation for designing AI systems (Dehghani and Levin 2024, Miehling et al. 2025). Effective AI systems for strategy should be hierarchical and nearly decomposable, built around stable intermediate artifacts, and constructed so that both internal evolution and

external explanations and representations remain tractable for boundedly rational decision makers (Ethiraj and Levinthal 2004, Csaszar et al. 2024b, Gurzick et al. 2026). Consistent with the emerging conceptualization of organizations as artificial intelligences that encode specific choices about representations, search, and aggregation (Csaszar and Steinberger 2022), we take this design perspective seriously and ask: how should AI systems for strategy be designed, and how do different AI architectures affect strategic theorizing? Keeping general AI—broadly capable LLMs accessed through standard chat interfaces—as the counterfactual, we focus on designing and testing agentic AI—multiagent systems that structure strategic thinking through specialized roles, coordination mechanisms, and explicit cognitive frameworks.

To investigate these questions, we develop *Aristotle*, a multiagent AI system for strategic decision making. We document its design journey around three foundational choices: first, the selection of a strategy framework: Aristotle embodies a theory-based framework (Camuffo et al. 2024), rather than, for example, a resource-based framework, that structures how users formulate, test, and refine causal theories of value; second, the positioning of the system on the automation-augmentation spectrum: Aristotle is designed to augment strategists rather than replace them and therefore operates in "copilot" mode rather than "tool-only" or "autopilot" mode; third, resource constraints: Aristotle takes into account efficiency considerations. Together, these choices lead to a specific agentic AI system for strategy.

To understand whether Aristotle works and how it works relative to general AI, we conducted a randomized controlled trial (RCT) that compares its effects on the beliefs, quality of the theory, and confidence and orientation of decision makers toward AI with those of general AI assistance and no AI assistance. To make the test both rigorous and potentially revelatory, we design a "biased" "high-bar" experiment (Gans 2023) using a "minimum viable" version of Aristotle—a stripped-down system with only three of the original 10 agents. We recruit 976 managers on Prolific to solve a strategic challenge and randomly assign them to one of three conditions: (1) human only (no AI assistance), (2) general AI assistance (GPT-4o accessed via a standard chat interface), or (3) theory-based agentic AI assistance (the three-assistant version of Aristotle). All participants first learn a common theory-based decision framework and then develop and revise a theory of value.

The results reveal *user-system-problem fit* as a central design dimension for multiagent AI systems for strategy. Relative to the human-only condition, both AI conditions strengthen beliefs: participants revise their theories upward in expected success and report higher confidence. However, improvements in theory quality are heterogeneous. General AI produces modest but reliable gains in expert-rated quality and probability of success on average. Agentic AI does not raise expert-rated quality on average but selectively improves quality for experienced managers, who benefit from its theory-based scaffolding without inflation of confidence. Highly educated participants (e.g., PhDs), by contrast, show larger belief updates—especially under agentic AI—without corresponding quality gains, suggesting a potential "expertise trap" in which sophisticated users become more confident without developing better solutions.

Building on this design journey and experimental evidence, we make three contributions. First, we establish user-system-problem fit as a fundamental design dimension: the architectural properties of an AI system for strategy must be matched to the strategic challenges and to the practical domain expertise and behavioral traits of the users to avoid overconfidence, miscalibrated reliance, and potential misuse. Second, we develop a five-dimensional taxonomy—spanning orchestration paradigm, agent architecture, task decomposition, human integration mode, and evaluation mechanisms—that maps the design space for such systems. Third, we provide a methodological roadmap that records design decisions, constraints, and evaluation strategies, allowing researchers and practitioners to move beyond one-shot comparisons with general AI toward systematic experimentation and iterative improvement in AI systems for strategy.

The paper proceeds as follows. Section 2 reviews research on AI and strategy, highlights the predominant focus on general-purpose models, and motivates our design-centric perspective. Section 3 develops our framework for designing AI systems for strategy and clarifies the distinction between general and theory-based agentic AI. Section 4 presents Aristotle and its streamlined experimental implementation. Section 5 describes the randomized experiment and measurement strategy. Section 6 reports the main experimental findings, including heterogeneous effects across user types. Section 7 discusses the findings and their implications, formalizes our design insights into a five-dimensional taxonomy, summarizes contributions and limitations, outlines a future research agenda, and concludes.

2. AI and Strategy: Literature Review

The proliferation of AI is precipitating a paradigm shift in how managers and entrepreneurs formulate and execute strategic decision making and venture creation (Agrawal et al. 2022). Scholars have begun to examine both the promise and the pitfalls of AI in strategic and entrepreneurial contexts (Obschonka and Audretsch 2020). Despite this growing interest, research so far has remained prevalently descriptive and largely speculative. For example, Townsend and

Hunt (2019) contend that AI enhances entrepreneurs' capacity to manage modal uncertainty by improving predictive accuracy and pattern recognition, but they caution that AI should not supplant human judgment and creative insight. Chalmers et al. (2021) similarly offer a descriptive account of the potential of AI to both displace and empower entrepreneurial tasks at all stages of venture development. Shepherd and Majchrzak (2022) extend this view by cataloging specific applications of AI in entrepreneurship while also identifying emergent risks and unintended consequences.

In addition to this predominantly descriptive treatment, strategy scholarship has begun to frame AI as a general-purpose technology (GPT) and, more specifically, an "invention of a method of invention" that can reshape firms' search, recombination, and capability building. This perspective "opens the door" to the possibility that managers can design AI systems and not simply use general AI, thus helping to explain second-order effects of AI that go beyond accuracy gains and highlighting new loci of advantage in how firms design, adapt, and govern AI tools (Cockburn et al. 2018). Relatedly, automation versus augmentation work on AI emphasizes that substitution and complementarity are not mutually exclusive but often coexist, creating a managerial tension that must be actively navigated (Raisch and Krakowski 2021). Recent empirical evidence further suggests that AI adoption can shift sources of competitive advantage toward hybrid human-AI capabilities that are difficult to imitate (Krakowski et al. 2023).

Some insightful empirical studies have contrasted human versus AI performance in generating or evaluating strategic options. Csaszar et al. (2024b) demonstrate that LLMs can match entrepreneurs and investors in assessing new venture concepts. Doshi et al. (2025) reveal that single-shot AI evaluations can be erratic and biased but that multiple AI judgments are reliable when aggregated. Boussioux et al. (2024) offer an intermediate perspective, showing that human-AI collaboration improves the quality—but not the originality—of creative problem solving. Recent empirical developments have also begun to deepen our understanding of the dual promise and danger of AI. Gaessler and Piezunka (2023) find that AI can help decision makers by substituting for scarce human training partners in strategic simulations. However, this substitution may deprive learners of exposure to idiosyncratic errors that foster strategic insight. Krakowski et al. (2023) document concurrent substitution and complementarities: traditional human capabilities wane, even as novel human-AI hybrid competencies emerge as sources of advantage. This pattern resonates with the automation-augmentation paradox in which AI simultaneously substitutes for some human tasks while augmenting higher-order cognition and coordination (Raisch and Krakowski 2021). The interplay between AI and human cognition is further complicated by the unclear ability of AI to mimic causal reasoning. Hünermund et al. (2021) argue that mainstream machine-learning methods, though powerful in pattern recognition, cannot replace theory-driven causal inference, a capacity central to genuine strategic novelty.

A robust behavioral literature on algorithmic aversion shows that users can abandon algorithms after observing even minor errors, preferring inferior human judgment. However, giving users even a slight discretion to adjust or override recommendations substantially mitigates this aversion (Dietvorst et al. 2015, 2018).

Similarly, recent research documents the potentially negative effects of AI complacency and automation bias—the tendency to overrely on AI recommendations—and how to mitigate it (Harbarth et al. 2025, Romeo and Conti 2025). These findings imply that interface and governance choices (e.g., transparency, auditable steps, bounded user control) are likely to be first-order moderators of general AI's effective use in strategy.

Skeptical perspectives invoke Knightian uncertainty to underscore the predictive limits of AI. Townsend et al. (2025) adapt Knight's framework to show that even advanced AI falters under nonergodic, indeterminate, or radically novel conditions. Sako and Felin (2025) amplify this critique, emphasizing that the reliance of AI on historical data limits its ability to extrapolate to unprecedented futures. An alternative view emerges from Conti and Messinese (2024) and Ramoglou et al. (2025), which demonstrate that AI can expand the epistemic reach of entrepreneurs. By detecting data anomalies and surfacing patterns and mapping them to causal hypotheses, AI tools can selectively enhance performance among highly skilled, frontier-oriented entrepreneurs, acting as a selective complement to human expertise. Moving beyond empirical findings, theoretical contributions are also advancing. Gans (2025) formalizes the notion of AI as a strategic actor, delineating how AI can enhance both decision quality and organizational coordination, provided it establishes credibility and trust. Kemp (2024) develops a situated-AI framework, arguing that firms achieve a sustainable advantage only when AI's agency is embedded in their unique experiential and relational architecture.

In sum, the literature is converging on the view that AI and humans must operate in tandem for strategic decision making. Although human judgment, causal reasoning, and forward-looking thinking remain indispensable for navigating truly uncertain, novel, and strategic environments, it is possible but unclear whether AI can complement and augment strategists in formulating and addressing novel strategic problems. A common feature of the studies we reviewed is that they focus on the effect of general AI, almost implicitly assuming that strategists will use it as it is or become available and that strategist-AI interaction is

driven by complacency or aversion and is a matter of prompting. Complementing this view, we posit that purposeful AI *design* and *use*—not mere access to generic models—can itself represent a dynamic capability and be a source of advantage. This squares with the emerging conceptualization of organizations as artificial intelligence, ensembles of humans and machine agents that encode specific choices about representations, search, and aggregation. Such design choices shape the way human and machine agents jointly search for, represent, and aggregate strategies (Csaszar and Steinberger 2022). In this sense, agentic AI systems for strategy constitute a new form of external representation. Like visual frameworks such as SWOT or Porter's five forces, they serve as cognitive tools that support boundedly rational decision makers—but their effectiveness hinges on matching system properties (usability, malleability) to users' representational capabilities (Csaszar et al. 2024a). This lens underscores that purposeful AI *design* and *use*—not mere access to generic models—can itself be a potential source of advantage, consistent with evidence on new human-AI capabilities and complementarities (Krakowski et al. 2023).

In this study, we are interested in exploring the differential effects of purposely designed AI systems on strategic decision making, compared with general AI assistance or no assistance.

The intuition is that what AI tools strategists will use or how they will purposely design and customize them will make a difference in strategic decision making and potentially represent a source of competitive advantage.

3. Designing AI Systems for Strategy

We distinguish between two types of AI systems relevant to strategic decision making. *General AI* refers to broadly capable language models such as GPT-4o accessed through standard chat interfaces. *Agentic AI* refers to systems purposely designed by humans as autonomous agents capable of perceiving their environment, making decisions, and executing actions to achieve specific goals without direct human intervention (Wooldridge 2009; Russell and Norvig 2016; Shavit et al. 2023; Anthropic 2024b, 2025). Agentic AI systems vary in their degrees of autonomy in decision making, goal-oriented behavior, and adaptability, differing widely in architectural and task complexity. Agent types vary along multiple dimensions—from reactive systems following predefined rules to deliberative agents with rich internal models.

Designing effective agentic AI systems requires making fundamental choices about the purpose of the system, the theoretical framework for strategic decision making to which the system refers, and the interaction between humans and AI. This section describes the design approach underlying Aristotle, an agentic AI system for theory-based strategic decision making. As discussed before, the design of AI systems can fruitfully leverage Simon (1962)'s work on the architecture of complexity. Hierarchy as a desirable architectural property is captured in AI system design through layered systems in which a top-level orchestrator coordinates specialized agents, which, in turn, call tools and data sources. Each level abstracts the complexity below it, as in systems composed of subsystems with their own subsystems (Baldwin et al. 2014). Similarly, near decomposability as a desirable architectural property is captured by agents designed so that most interactions are "inside" a role, whereas communication between roles is through narrow and well-defined interfaces (messages, shared artifacts), that is, strong internal coupling and weaker structured external coupling that makes the overall system more effective, adaptive, and evolvable (Cabigiosu and Camuffo 2016).

In AI system design, workflows and their artifacts are "stable intermediate forms," which let agents work in parallel and humans interrupt, revise, or reuse partial results without collapsing the whole effort. Hierarchical, nearly decomposable AI systems are also characterized by fast local dynamics and slower cross-level dynamics, with complex structures emerging via selective retention of useful partial solutions over time. Finally, hierarchy and near decomposability in AI system design make concise, multilevel output possible, enabling complex internal reasoning to be presented to users in tractable form (Ethiraj and Levinthal 2004).

3.1. Design Approach

Three considerations—all grounded on the above described principles of the architecture of complexity—shaped our approach while designing Aristotle: (1) the underlying framework for strategic decision making (which shapes the decomposition and interdependence structure), (2) the human-AI integration approach (which shapes architectural boundaries and the locus of search), and (3) resource constraints (which determine tractable complexity, internal dynamics, and system evolvability).

3.1.1. Underlying Framework for Strategic Decision Making. LLMs lack specific knowledge of strategy frameworks and the ability to perform genuine causal inference (Hünermund et al. 2021, Felin and Holweg 2024). Agentic AI systems for strategy should therefore embody an explicit framework for strategic decision making. The choice of framework constrains the design choices, determining how the system structures strategic problems and approaches their solution. Alternative strategy frameworks—formulation-implementation, resource-based view, Porter's five forces, etc.—rely on different assumptions and impose different analytical lenses. Aristotle embodies the theory-based approach to

strategic decision making (Felin and Zenger 2009, 2017; Ehrig and Schmidt 2022; Camuffo et al. 2024; Sorenson 2024), which structures how managers formulate, test, select, and revise theories of value through explicit causal reasoning and hypothesis testing.

3.1.2. Human-AI Integration. AI systems span a spectrum: *tool mode* (passive assistance on demand), *copilot mode* (active partnership with shared initiative), and *autopilot mode* (delegated decision authority). The appropriate mode depends on task characteristics—structured tasks suit automation; novel high-stakes decisions require augmentation—and user characteristics such as expertise, cognitive style, and orientation toward AI. For strategy formulation, Aristotle uses the copilot mode, acknowledging the need to preserve managerial judgment while providing structured guidance in theory formulation.

3.1.3. Resource Constraints. Computational cost, response latency, and development complexity constrain viable architectures. AI systems requiring intensive and extensive interaction, extended processing, or substantial resources prove impractical, regardless of the quality of their output. The architecture of complexity (Simon 1962, Ethiraj and Levinthal 2004) offers a guideline: modular agents under hierarchical orchestration. This architecture enables tractability for both system evolution and user comprehension, balancing output quality against operational constraints to achieve the best user-system-problem fit.

3.2. Implementation: Design Choices

We translated these considerations into specific architectural choices. Using the theory-based view as the underlying strategy framework implies a specific type of task decomposition: the "components" of theory-based strategizing are embodied in specialized agents, which produce stable intermediate artifacts (problems, theories, causal structures). Using copilot as orchestration mode preserves user control at the boundary but admits various internal architectures: routing with structured workflow, routing with subagent autonomy, or hybrids thereof. Resource constraints impose parsimony—fewer agents with tractable coordination complexity. This approach aligns with emerging industry standards where orchestrated multiagent workflows are classified as agentic AI systems (Finio and Downie 2025, Microsoft 2025, Sempf and Hooker 2025). Section 4 describes how Aristotle instantiates these choices. Section 7.2 formalizes routing as one of three orchestration paradigms (alongside workflow and autonomous) within a broader architectural taxonomy and develops user-system-problem fit as a guiding principle for matching architectural choices to users and strategic contexts.

4. Aristotle: System Architecture and Implementation

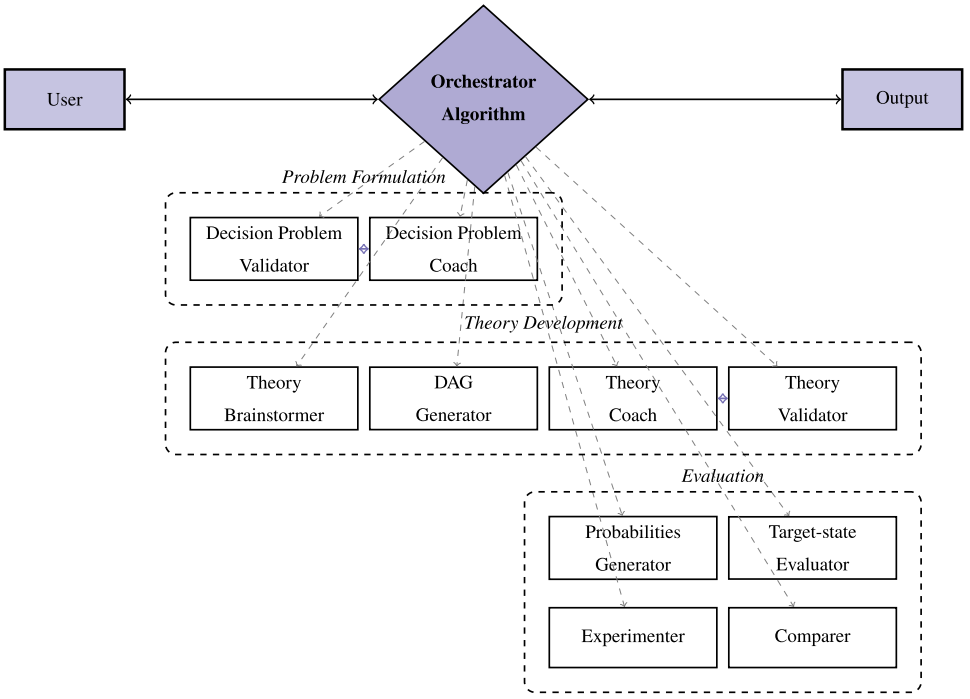
4.1. Aristotle: An Agentic AI System for Strategy

The agentic AI system for theory-based strategic decision making we designed comprises 10 specialized assistants orchestrated by a central coordination algorithm (Figure 1). Each assistant is implemented using OpenAI's Assistants API, providing persistent thread management, context retention, and structured output capabilities. The assistants are organized into three functional modules that, as illustrated above, correspond to the key components of theory-based strategic decision making (Felin and Zenger 2017, Ehrig and Schmidt 2022, Camuffo et al. 2024): (1) strategic problem formulation, (2) theory development, and (3) theory evaluation. The problem formulation module contains two assistants: a decision problem validator ensuring strategic challenges meet the theory-based view criteria and a decision problem coach guiding users in refining problem statements. The theory development module comprises four assistants: the theory brainstormer generates theories, the DAG generator constructs directed acyclic graphs representing conceptual causal structures, the theory coach provides guidance on theory refinement, and the theory validator checks logical consistency and completeness. The evaluation module includes four assistants: the probability generator assists in assigning subjective probabilities to the DAG components, the target-state evaluator assists in prior belief formation assessing the expected value of theories, the experimenter runs tests to validate assumptions, and the comparer facilitates systematic comparison between competing theories. The orchestrator algorithm coordinates these assistants through a decision tree, which routes user queries based on workflow state, guiding users through complete theory-based decision making from initial problem formulation to prior belief formation through theories, experimental testing, belief and theory updating, and theory selection.

4.2. Testing for Implementation

To understand whether Aristotle works and how it works compared with general AI assistance and no AI assistance, we decided to test it. Specifically, we decided to conduct a randomized controlled trial to explore its differential effect on decision makers' beliefs, theory quality, confidence, and orientation toward AI. To ensure rigorous investigation, we designed a biased, high-bar experiment (Gans 2023), slanted toward minimizing false positives. Our idea was that if such an experiment generates a positive signal, then it is worthwhile to further explore the use of agentic AI systems for strategy. Following Agrawal et al. (2025), we used a minimum viable product of Aristotle (a stripped-down

Figure 1. (Color online) Full Aristotle Architecture



version of it) as a high-bar experiment so that the experiment is stacked against success.

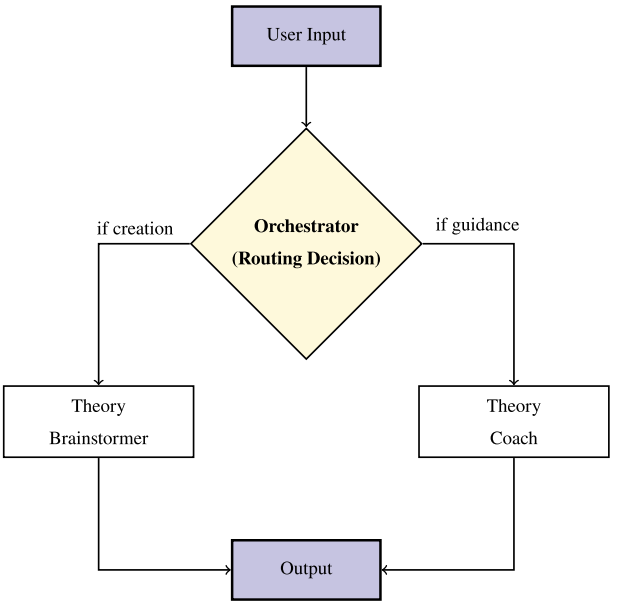
The stripped-down version of Aristotle we tested includes only three of the 10 agents comprised in Aristotle (Figure 2). Coming up with a "minimum viable" version of Aristotle posed an interesting challenge, as it caused us to face the fundamental tensions in the design of agentic AI systems described in the previous section, for example, balancing system architectural complexity to improve system output with system usability and efficiency[1] and, similarly, deciding which functions and features are essential to reflect the key elements of the theory-based view of strategy embodied in Aristotle.

The streamlined, three-agent version of Aristotle we used in the experiment solves these tensions, preserving core agentic features while enabling rigorous and efficient testing. This simplified system comprises (1) an orchestrator routing queries via a classification algorithm, (2) a theory brainstormer generating theories according to a theory-based structure, and (3) a theory coach providing explanations and guidance. Specifically, this implementation focuses exclusively on the generation and refinement of theories—omitting the problem formulation and experimental validation phases.

Whereas the three-agent system does not fully instantiate all the components of the theory-based approach, it allows us to focus on and isolate its key component: theory development. The experiment thus investigates how the use of an agentic AI system focused on theory development affects the formulation of strategies compared with general AI assistance or no assistance.

This process of "design for testing" also made us reflect on our original design choices and provided further insights into the level of architectural complexity that is optimal for different users, strategic problems, and contexts. Testing a simplified version of Aristotle allowed us to explore tractable levels of AI

Figure 2. (Color online) Simplified Experimental Architecture: Three-Assistant System with Autonomous Routing



system complexity. As we will see, the nonsignificant results for some user groups alongside quality improvements for experienced managers suggest that architectural complexity helps selectively—a finding that would remain hidden in tests of the full system. Our approach demonstrates how testability constraints can also generate insights about how to improve AI system design.

The streamlined implementation also revealed insights about agents' coordination. For example, we realized that the type of use drives architecture: comprehensive support for theory-based strategic decision making may require workflow orchestration with complex state management, whereas focused theory generation may succeed with simpler routing-based coordination. We also realized that architecture can be more or less hierarchical and quasidecomposable in different contexts, depending on coordination mechanisms. These insights, which we learned during the experiment, inform the broader taxonomy we develop in Section 7.2.

4.3. Technical Implementation: Models, Infrastructure, and Agentic Behavior

Within the three-agent system, each assistant was implemented using OpenAI's Assistants API with GPT-4o (model version gpt-4o-2024-08-06) as the underlying language model. The Assistants API provides persistent threads that maintain conversation context across interactions, structured outputs that enforce specific response formats such as JSON for routing decisions, and system instructions that enable fine-tuned behavioral specifications. This infrastructure allows each assistant to maintain consistent behavior aligned with its specialized role while preserving context throughout extended participant interactions.

The key agentic feature is the orchestrator's autonomous routing decision. Unlike a standard chatbot that responds uniformly to all inputs, the orchestrator analyzes each user query and decides which specialized assistant should handle the request. This represents a meaningful agency: goal-directed behavior in selecting appropriate response pathways. The orchestrator's decision logic classifies inputs into two categories: requests for theory creation route to the theory brainstormer, whereas requests for explanation or guidance direct to the theory coach. This routing mechanism ensures that participants receive responses tailored to their specific needs at each point in the strategic decision-making process.

This routing mechanism instantiates the distinction between reactive and agentic AI assistance. General-purpose AI responds to each query independently, potentially producing inconsistent guidance throughout the strategic decision process. The agentic system maintains *methodological coherence*: each response is conditioned not just on the immediate query but on the phase of strategic thinking the user engages in. This ensures that theory generation follows consistent causal logic, explanations refer to the same theoretical framework, and revisions build systematically on prior work rather than introducing arbitrary variations.

5. Exploratory RCT

5.1. Rationale and Scope

To explore how AI affects theory-based strategic decision making, our experimental design compares three conditions. In the agentic AI condition, participants used the three-agent system described above. In the general AI condition, participants accessed GPT-4o through a standard chat interface without specialized prompting or routing logic. The fact that both AI conditions used the same underlying language model (GPT-4o) allows us to isolate the effect of multiagent workflow architecture from base model capabilities. In the human-only condition (control), the participants worked independently without AI assistance.

This three-arm experimental design enables the investigation of how general AI and theory-based agentic AI affect strategic decision making differently. However, we need to clarify some important limitations in the scope. The three-agent implementation tests *only one* specific instantiation of theory-based agentic AI for strategy *among many possible designs*.[2]

5.2. Experimental Design

As already mentioned, the experiment was designed as a "biased" or "killer" experiment (Gans 2023), intentionally structured to "make it difficult" to detect treatment effects. This approach aims at ensuring that the insights we get from it are meaningful and do not encourage the pursuit of a nonpromising research avenue. Three design choices make our experimental design slanted toward reducing the likelihood of detecting differential effects of agentic AI for theory-based strategy. First, it involved the use of a simulation platform and a relatively straightforward strategic challenge (reducing restaurant food waste). The low complexity of the strategic problem makes it a high-bar experiment (Agrawal et al. 2025), as the use of general and agentic AI should not make a big difference in this case. Finding any treatment effects under this *extreme* condition makes the potential results of the exploratory experiment more meaningful and insightful. Second, we recruited participants (managers) through Prolific rather than targeting senior executives or managers. It is widely known that this type of participant is less reliable, again decreasing the likelihood of detecting treatment effects. Third, as previously described, we implemented a "minimum viable" version of Aristotle, which also slants the experiment toward difficulty in detecting differential effects of

agentic AI for strategy compared with general AI assistance or no assistance.

The exploratory randomized controlled trial was conducted in December 2024 with 976 participants all exposed to the theory-based approach to strategic decision making. Participants were randomly assigned to three experimental conditions:

- human-only (control condition)
- general AI (standard GPT-4o via Assistants API, labeled as "Aristotle")
- agentic AI (three-assistant system with orchestrator routing, also labeled as "Aristotle")

All participants followed a two-stage experimental protocol implemented on a purpose-designed platform. After completing an online survey asking for baseline measures and randomization, participants viewed a three-minute video that introduced the theory-based approach to strategic decision making. This ensured that all participants, regardless of the condition, had basic familiarity with a common conceptual framework for strategic decision making before beginning the task.

5.2.1. Stage 1: Initial Theory Development. In the first stage, all participants independently worked to develop their initial "theory of value" to address "the food waste challenge." The challenge consists of asking participants to "build a profitable tech startup that reduces restaurant food waste by 50% within 5 years." The participants entered their theory of value[3] (minimum 250 words) in a text field and provided their subjective probability of success (0%–100%) and confidence in their theory (scale of one to seven). These measures correspond to the expected value of the theory and the prior belief in the theory specified in Section 3. This stage was identical across all experimental conditions, establishing baseline measures for each participant's strategic thinking. Figures B1, B2, and B3 in Online Appendix C provide snapshots of the platform interface.

5.2.2. Stage 2: Theory Refinement. The second stage introduced the experimental intervention. Participants were asked to revise their theory of value and update their subjective assessments.[4] In the *human-only* condition, participants independently reviewed their work without any assistance, seeing the same interface as stage 1 and refining their theory of value. In the *general AI* condition, a chat interface appeared on the right side of the screen, providing direct access to GPT-4o, which participants could freely query for assistance. In the *agentic AI* condition, the same chat interface connected the participants to the three-agent AI system (as described in Section 4.2).[5]

5.2.3. Behavioral Distinctions Between Conditions. Although participants in both AI conditions saw an identical chat interface labeled "Aristotle," the underlying systems were different. In the General AI condition, all queries were processed directly by GPT-4o through a standard API call, producing responses typical of a general-purpose language model. These responses may include business advice, strategic suggestions, or analytical frameworks, but without any underlying causal structure or theoretical consistency.

The Agentic AI condition, by contrast, implemented the three-agent version of Aristotle described in Section 4.1.

Importantly, both AI conditions maintained interface behaviors that appeared identical to participants, with the multiagent orchestration happening seamlessly in the background. Both conditions presented AI assistance as "Aristotle" to control for expectation effects, ensuring that any observed differences in outcomes could be attributed to the actual behavioral differences of the systems rather than participant expectations or interface effects.

At the end of the experiment, all participants were briefed, explaining the purpose of the study, and thanked for their participation.

5.2.4. Data Quality and Manipulation Checks. To ensure data quality and participation, we implemented multiple attention checks throughout the experiment. We excluded participants who (1) failed logical consistency checks (e.g., reporting less total work experience than managerial experience; providing spam solutions), (2) provided inconsistent responses on reverse-coded items about AI trust (where the sum of AI aversion and AI preference responses deviated more than two points from the expected value of eight, indicating inattention to reverse coding), or (3) failed comprehension checks about the video content. These exclusion criteria resulted in dropping 373 participants from an initial sample of 1,349, yielding our final analytical sample of 976 participants.[6] To incentivize thoughtful engagement, we informed all participants that the top 33% of performers would receive double compensation.[7]

5.3. Experimental Execution

5.3.1. Recruitment, Randomization, and Sample. We recruited participants on Prolific. The recruitment criteria were (1) located in the United States or United Kingdom (as a proxy for English proficiency), (2) high Prolific approval rate (99%–100%), and (3) education at the level of a college degree or higher. To assign participants to experimental conditions, we used minimized block randomization (batch size = 50) based on three covariates: (1) education level and field, (2) AI aversion, and (3) level of self-confidence. This provided the baseline allocation of participants to the three experimental conditions.

Our final sample comprises 976 participants. On average, they were 40 years old, and about 47% of them were females. They were highly educated, with

44.34% possessing at least a master's degree. The most common field of higher education was STEM (33% of the sample), followed by business and economics degrees (20%) and humanities (15. 49%). They had, on average, eight years of work experience and four years of managerial experience.

The baseline survey that we administered to all participants before the intervention allowed us to collect preintervention data and check that a wide set of meaningful participant characteristics were balanced across experimental groups. Table 1 reports preintervention balance checks with *t*-tests and *F*-tests of joint orthogonality in the experimental arms, confirming the overall balance between the groups.[8] In Online Appendix A, we include Tables A1 and A2, which present balance checks (across the three experimental conditions and pairwise) for the whole sample. Tables A3 and A4 present the balance checks for the subsample of participants with PhDs, which we will analyze in detail in the next section.

5.3.2. Measures.

5.3.2.1. Dependent Variables. As key dependent variables, we used the prepost intervention changes in the participants' expected probability of success and in their confidence about the theory.

- *Change in expected probability of success*. It is the change before/after the intervention in the "expected value" of entrepreneurial theories as in Camuffo et al. (2024). The prepost difference measures the update of the expected probability of success of participants' theories. The expected probability of success is a discrete variable that can take values between zero and 100.
- *Change in confidence about the theory*. It is the change before/after the intervention in the "confidence about the theory" or the prior belief that participants have about whether their theories are "true," as in Camuffo et al. (2024). The prepost difference measures the update of the participants' "prior-on-prior." Confidence in the theory is a discrete variable measured on a scale of one to seven.

In addition to assessing the effect of exposition to agentic and general AI on subjective expectations, we also assessed their impact on theory *Quality* (Boussioux et al. 2024) using multiple evaluation approaches (Doshi et al. 2025).

We first obtained human-expert assessments from two independent raters with extensive experience in strategic management and entrepreneurship. These experts blindly evaluated all 1,952 theories (976 participants × 2 time periods) on a scale of one to five, applying the rubric developed by Boussioux et al. (2024). The rubric comprises five critical dimensions

Table 1. Average Treatment Effects on Δ*Outcomes*

Dependent variable	Δ*SPS*	Δ*CIT*	Δ*Expert Quality*	Δ*Expert PS*	Δ*Expert CIT*	Δ*LLM Avg Quality*	Δ*LLM Avg PS*	Δ*LLM Avg CIT*
General AI	1.705*	0.205***	0.255***	6.224***	−0.008	0.100***	0.331*	−0.010
	(0.896)	(0.064)	(0.092)	(2.016)	(0.062)	(0.023)	(0.189)	(0.013)
	[0.057]	[0.001]	[0.006]	[0.002]	[0.903]	[0.000]	[0.080]	[0.451]
Agentic AI	3.280***	0.229***	−0.033	−0.851	−0.101*	0.037	0.494***	−0.049***
	(0.922)	(0.063)	(0.085)	(1.799)	(0.061)	(0.023)	(0.191)	(0.014)
	[0.000]	[0.000]	[0.699]	[0.636]	[0.098]	[0.107]	[0.010]	[0.001]
Algo Aversion	−1.176***	−0.074***	0.039	0.839	−0.022	−0.015	−0.084	−0.008
	(0.446)	(0.028)	(0.033)	(0.718)	(0.023)	(0.009)	(0.078)	(0.005)
	[0.009]	[0.008]	[0.246]	[0.243]	[0.344]	[0.101]	[0.279]	[0.132]
Automation Bias	0.904**	0.060**	0.063*	1.524*	0.006	0.017*	0.064	−0.004
	(0.417)	(0.026)	(0.036)	(0.809)	(0.023)	(0.009)	(0.073)	(0.005)
	[0.031]	[0.023]	[0.082]	[0.060]	[0.786]	[0.077]	[0.376]	[0.494]
GenAI Expertise pre	0.174	0.023	−0.019	−0.286	0.024	−0.001	−0.030	0.001
	(0.305)	(0.019)	(0.025)	(0.526)	(0.020)	(0.006)	(0.054)	(0.004)
	[0.568]	[0.231]	[0.436]	[0.586]	[0.224]	[0.899]	[0.574]	[0.799]
edField - Health Sciences	−1.403	0.042	−0.191*	−2.900	−0.023	−0.049*	−0.209	0.015
	(1.134)	(0.074)	(0.113)	(2.495)	(0.079)	(0.029)	(0.212)	(0.017)
	[0.216]	[0.564]	[0.092]	[0.245]	[0.776]	[0.095]	[0.324]	[0.375]
Constant	4.501***	0.074	0.290**	6.470**	−0.163	0.045	0.325	0.005
	(1.729)	(0.108)	(0.131)	(2.724)	(0.107)	(0.031)	(0.278)	(0.022)
	[0.009]	[0.494]	[0.027]	[0.018]	[0.129]	[0.150]	[0.243]	[0.819]
Observations	976	976	975	976	976	976	976	976
R^2	0.039	0.038	0.021	0.024	0.006	0.033	0.012	0.017
Robust SE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Robust standard errors (SEs) are given in parentheses; *p*-values are in brackets. Controls per full-sample balance table.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

that capture different aspects of strategic merit: (1) novelty (how different is it from existing solutions?), (2) feasibility and scalability (how likely is it to succeed, and how scalable is it?), (3) environmental impact (how much does it benefit the planet?), (4) financial impact (what financial value can it create for businesses?), and (5) overall quality (based on the four criteria above, what is the overall quality of the solution?). The experts used the rubric to score each theory's quality on a scale of one to five.

Furthermore, following emerging best practices in LLM-assisted evaluation (Doshi et al. 2025), we used two state-of-the-art reasoning LLMs (GPT-o4-mini and Claude Sonnet 4.5 with extended thinking) to evaluate all theories with the same method defined above (Boussioux et al. 2024). For each theory and dimension, we conducted 10 evaluation runs per model (20 total), setting the output to a precise structured JSON format. We then averaged these scores within and across the models to create aggregated LLM assessments. This multimodel approach helps mitigate individual model biases and provides more robust evaluations.

To assess measurement reliability, we conducted interrater reliability analyses using Cohen's κ coefficients. In Online Appendix E, Table A11 and Table A12, we report reliability between LLM models ($\kappa = 0.070$ using integer matching; $\kappa = 0.008$ using tercile categorization) and between human and LLM assessments ($\kappa = 0.097$ using integer matching; $\kappa = 0.227$ using tercile categorization). In particular, when using tercile categorization, human experts showed fair agreement with the LLM evaluations ($\kappa = 0.227$), with the strongest agreement observed between human experts and Claude ($\kappa = 0.314$). These results suggest that whereas fine-grained rating alignment remains challenging, there is broader categorical agreement on strategic quality between human and AI evaluators. Full details of the evaluation prompts, rubrics, and robustness checks are provided in Online Appendix E.

From these evaluations, we derive two additional dependent variables:

- *Change in overall theory quality: human expert evaluations*. It is the change before/after the intervention in the "quality" of entrepreneurial theories measured as in Boussioux et al. (2024) by human experts (scales of one to five).
- *Change in overall theory quality: LLM evaluations*. It is the change before/after the intervention in the quality of entrepreneurial theories measured as in Boussioux et al. (2024) by LLMs (scales of one to five).

In addition to assessing the quality of the theory, we also asked human experts and LLMs to evaluate the probability of success (scale of 0%–100%) and confidence in the theory (scale of one to seven). We derived two additional dependent variables.

- *Change in subjective probability of success (SPS): human evaluation*. Pre-/postchange in the expected probability of success assessed by the human expert.
- *Change in SPS: LLM evaluation*. Pre-/postchange in the expected probability of success assessed by the LLMs.
- *Change in confidence in theory (CIT): human evaluation*. Pre-/postchange in confidence in the theory assessed by the human expert.
- *Change in CIT: LLM evaluation*. Pre-/postchange in confidence in the theory assessed by the LLMs.

For the above measures, we also examined interrater reliability. Both human experts and LLMs were asked to assess the same metrics: the subjective probability of success (0%–100%) to build a profitable tech startup that reduces restaurant food waste by 50% in five years and their confidence in this assessment (scale of one to seven). We found that human experts and LLMs showed fair agreement when assessing the probability of success ($\kappa = 0.25$), suggesting that they generally align in the strategic viability judgements. However, their confidence scores diverged substantially ($\kappa \approx 0.00$), indicating that humans and LLMs differ fundamentally in the way they express certainty about their evaluations. Interestingly, the two LLMs showed strong consistency with each other in probability assessments ($\kappa = 0.636$), even when they disagreed with human experts.

5.3.2.2. Independent Variables. The assignment to the experimental conditions of the study constitutes the main independent variable. The variable can take three values according to the following three conditions:

- *Human-only*: value = 0 for participants in the control condition.
- *General AI*: value = 1 for participants with access to GPT-4o via Assistants API, presented as "Aristotle" in the interface.
- *Agentic AI*: value = 2 for participants with access to the theory-based multiassistant system detailed in Section 4.2, also presented as "Aristotle" in the interface.

Other variables were used in the analyses as strata/moderators (e.g., level of education) or controls (for unbalanced preintervention covariates). We also used variables related to the interaction between humans and AI based on survey items (scale of one to seven) derived from the empirical literature on AI aversion (Dietvorst et al. 2015, 2018), AI complacency (Harbarth et al. 2025), automation bias (Romeo and Conti 2025), and AI familiarity (Horowitz et al. 2024). Finally, we added some survey questions during the exercises to check attention and effectiveness of the intervention.[9]

6. Exploratory Experimental Evidence

6.1. Descriptive Statistics

Table A1, available in the Online Appendix, reports summary statistics for the pre- and postintervention

samples, overall and by experimental arm. The initial average SPS is 63.3 (standard deviation (SD) = 23.5). CIT averages 5.31 (SD = 1.11). Familiarity with general AI and self-reported prompting skills is moderate (sample means of 4.63 and 4.87, respectively). Algorithmic aversion averages 4.99 (SD = 1.19) and automation bias 4.04 (SD = 1.41). After intervention, SPS increases to 70.5 (22.3), whereas CIT remains essentially unchanged on average (5.26). Familiarity and prompting skills increase slightly. Interestingly, algorithmic aversion decreases (to 4.76), and automation bias increases (to 4.21).

Table A1 in the Online Appendix also reports differences before/after the intervention at the individual level. The average change in SPS is +7.26 percentage points (12.16). In the control group, SPS increases by +5.15 percentage points, providing a baseline for cross-treatment comparisons. The confidence in theory is stable on average (change = −0.05). The correlation matrices (for the pre-, post-, and first-difference values of the variables, given in Online Appendix F) indicate tight associations among evaluations (agreement between external evaluators). At baseline, the quality assessed by experts and the probability of success assessed by experts are correlated at 0.92. The quality assessed by LLMs correlates 0.74 with the quality assessed by experts and 0.75 with the probability assessed by experts. The probability of success assessed with one LLM is correlated 0.89 with the quality assessed with the other LLM. Regarding the first-difference values of the variables, the changes in SPS correlate positively with changes in expert-assessed probability (0.21), expert quality (0.17), LLM quality (0.23), and LLM-assessed probability (0.17), and negatively with changes in algorithmic aversion (−0.13).

6.2. Average Treatment Effects

We use a standard regression approach to experimental analysis (Gerber and Green 2017). Given our focus on before/after changes at the individual level, we use cross-sectional ordinary least squares (OLS) regressions on the first differences to estimate and compare the effects of the three experimental conditions (*Human-only*,[10] *General AI*, and *Agentic AI*) on the dependent variables. The model specification includes as control variables the unbalanced covariates of interest before the intervention (as in Online Appendix A), and we report robust standard errors. Throughout the analyses presented below, we analyze the treatment effects on the dependent variables described in Section 5 labeled as follows: (1) Δ SPS (change in subjective probability of success), (2) Δ CIT (change in confidence in theory), (3) Δ expert quality (change in theory quality evaluated by expert), (4) Δ expert SPS (change in theory probability of success evaluated by expert), (5) Δ expert confidence (change in confidence in theory evaluated by expert), and (6) Δ LLM quality (change in theory quality evaluated by LLMs), Δ LLM SPS (change in theory probability of success evaluated by LLMs), and Δ LLM confidence (change in confidence in theory evaluated by LLMs).

Table 1 reports full-sample estimates ($N = 976$). Relative to the *Human-only* condition, *General AI* increases Δ SPS by +1.7 percentage points ($p = 0.057$) and Δ CIT by +0.21 ($p < 0.01$). *Agentic AI* increases Δ SPS by +3.28 percentage points ($p < 0.01$) and Δ CIT by +0.23 ($p < 0.01$). Thus, both treatments raise belief-based outcomes, with a larger average effect of Agentic AI on SPS.

Moving to human expert evaluations, *General AI* slightly increases quality and probability of success (Δ expert quality +0.26; $p < 0.01$) and Δ expert SPS +6.2 percentage points ($p < 0.01$). *Agentic AI* instead neither improves expert-rated quality nor expert-rated SPS (coefficients not significant) and shows a small negative effect on Δ expert confidence (−0.10, $p < 0.10$). This lack of gains occurs despite participants in both treatment groups significantly increasing automation bias (*General AI*: +0.371, $p < 0.01$; *Agentic AI*: +0.183, $p < 0.05$) and *General AI* reducing algorithmic aversion (−0.344, $p < 0.01$). The weak but positive association between self-reported AI automation bias and performance improvements ($\beta = 0.063$, $p < 0.10$) contradicts theoretical predictions (Parasuraman and Manzey 2010).

Regarding LLM evaluations, *General AI* modestly increases quality (Δ LLM quality (+0.10, $p < 0.01$), but not probability of success. *Agentic AI* increases the quality assessed with LLM (Δ LLM SPS by +0.494 percentage points; $p < 0.010$) and slightly reduces LLM-assessed confidence in theory (Δ LLM confidence (−0.049, $p < 0.01$) without increasing LLM-assessed quality.

Overall, relative to the human-only condition, AI assistance sizably and reliably amplifies belief-based outcomes (subjective probability of success and confidence in theory). The effect of agentic AI is larger than that of general AI. The average treatment effects of AI assistance on theory quality and probability of success are instead heterogeneous across evaluators: the human expert evaluations show some quality and success probability gains for General AI, whereas Agentic AI yields larger success probability from LLM evaluations but no significant expert-quality increase in this setting.

These results align with those of Csaszar et al. (2024b) for General AI, which also improved confidence and performance in our study. However, theory-based Agentic AI produced a different pattern—primarily behavioral changes without quality improvements—revealing that the type of AI system used critically shapes whether the increase delivers substantive versus merely behavioral or psychological benefits.

A potential mechanism underlying this puzzling result is "cognitive miserliness" (Fiske and Taylor 2020)—the tendency to offload cognitive effort when AI is available (Deng and Deng 2025). In addition, our results show that AI systematically amplifies rather than mitigates overconfidence bias. To shed light on this phenomenon, we examined the before/after intervention changes in AI automation bias and aversion of the participants (Table 2) under different experimental conditions. Participants exposed to general AI consistently increase automation bias (+0.371, $p < 0.01$) and reduce algorithmic aversion (−0.344, $p < 0.01$).

Table 2. Average Treatment Effects on Δ*Automation Bias* and Δ*AI Aversion*

	All observations		PhD subsample		Non-PhD subsample		*High Mng Exp* (≥10 years)		*Low Mng Exp* (<10 years)	
Variable	Δ*Automation Bias*	Δ*Algo Aversion*	Δ*Automation Bias*	Δ*Algo Aversion*	Δ*Automation Bias*	Δ*Algo Aversion*	Δ*Automation Bias*	Δ*Algo Aversion*	Δ*Automation Bias*	Δ*Algo Aversion*
General AI	0.371***	−0.344***	0.587**	−0.054	0.330***	−0.401***	0.319*	−0.180	0.416***	−0.395***
	(0.080)	(0.084)	(0.280)	(0.249)	(0.082)	(0.089)	(0.184)	(0.199)	(0.090)	(0.094)
	[0.000]	[0.000]	[0.038]	[0.830]	[0.000]	[0.000]	[0.085]	[0.368]	[0.000]	[0.000]
Agentic AI	0.183**	−0.151	0.105	0.199	0.163*	−0.214**	0.116	−0.059	0.222**	−0.187*
	(0.085)	(0.093)	(0.264)	(0.273)	(0.090)	(0.100)	(0.204)	(0.213)	(0.094)	(0.100)
	[0.031]	[0.103]	[0.691]	[0.467]	[0.071]	[0.033]	[0.573]	[0.782]	[0.018]	[0.062]
Δ*SPS*	0.006	−0.008**	0.019	−0.025*	0.006	−0.008**	0.013	−0.023***	0.005	−0.007
	(0.004)	(0.004)	(0.011)	(0.013)	(0.004)	(0.004)	(0.010)	(0.009)	(0.004)	(0.004)
	[0.115]	[0.027]	[0.101]	[0.060]	[0.118]	[0.049]	[0.201]	[0.009]	[0.228]	[0.124]
Δ*CIT*	0.071	−0.085*	−0.071	0.016	0.075	−0.076	0.047	−0.135	0.067	−0.050
	(0.049)	(0.048)	(0.148)	(0.083)	(0.050)	(0.056)	(0.129)	(0.148)	(0.053)	(0.052)
	[0.144]	[0.079]	[0.636]	[0.849]	[0.137]	[0.180]	[0.717]	[0.362]	[0.208]	[0.333]
AI Expertise (Pre)	−0.070***	0.147***	—	—	—	—	—	—	—	—
	(0.023)	(0.024)								
	[0.002]	[0.000]								
Δ*AI Expertise*	—	—	0.768***	−0.605***	—	—	—	—	—	—
			(0.232)	(0.161)						
			[0.001]	[0.000]						
edField: Health Sciences	−0.143	0.158	—	—	—	—	—	—	−0.082	−0.024
	(0.128)	(0.126)							(0.144)	(0.136)
	[0.263]	[0.210]							[0.569]	[0.859]
edField: Other	—	—	—	—	−0.008	−0.188	—	—	—	—
					(0.123)	(0.138)				
					[0.948]	[0.176]				
edLevel: Bachelors	—	—	—	—	—	—	0.205	−0.602***	—	—
							(0.155)	(0.162)		
							[0.188]	[0.000]		
Job: Accounting & Fin	—	—	0.156	−0.125	—	—	—	—	—	—
			(0.303)	(0.301)						
			[0.607]	[0.679]						
Job: IT	—	—	0.230	−0.053	—	—	—	—	—	—
			(0.416)	(0.296)						
			[0.582]	[0.858]						
Job: Strategy & Planning	—	—	—	—	—	—	0.135	0.662***	—	—
							(0.212)	(0.214)		
							[0.525]	[0.002]		
Industry: Manufacturing	—	—	—	—	—	—	−0.157	0.147	—	—
							(0.216)	(0.197)		
							[0.468]	[0.457]		
Constant	0.261**	−0.697***	−0.245	0.109	−0.076	0.052	−0.198	0.094	−0.094	0.060
	(0.127)	(0.139)	(0.166)	(0.145)	(0.060)	(0.065)	(0.145)	(0.167)	(0.065)	(0.065)
	[0.040]	[0.000]	[0.142]	[0.455]	[0.207]	[0.426]	[0.173]	[0.574]	[0.149]	[0.359]
Observations	976	976	113	113	863	863	184	184	792	792
R^2	0.048	0.076	0.249	0.145	0.032	0.042	0.057	0.208	0.038	0.032
Robust S.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Robust standard errors (SEs) are given in parentheses; *p*-values are in brackets. All: Full-sample balance controls. *PhD/No-PhD/High Mng Exp/Low Mng Exp*: Respective balance table controls.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

Participants exposed to agentic AI instead show smaller increases in automation bias (+0.183, $p < 0.05$) without significantly changing algorithmic aversion, suggesting that its theory-based approach may partially mitigate blind reliance.

6.3. Heterogeneous Treatment Effects

To discern meaningful differences between participants, we also examined the heterogeneity of treatment effects by education (PhDs) and managerial experience (high ≥10 years), using the same estimation approach (OLS cross-sectional regression on first differences) specified above.

- *High-education (PhD) subsample* ($N = 113$; Table 3): *General AI* increases Δ SPS by +4.71 percentage points ($p < 0.05$). In addition, its effect on quality and probability of success (externally assessed) improves significantly compared with full-sample estimates (Δ expert quality +0.55, $p < 0.10$; Δ expert SPS +15.9 percentage points, $p < 0.05$; and a marginal increase in Δ expert confidence, +0.32, $p < 0.10$). *Agentic AI* produces a large behavioral effect (Δ SPS +8.42 percentage points, $p < 0.01$) without the corresponding quality improvements. Surprisingly, PhDs show no significant change in automation bias if assigned to the agentic AI condition, whereas those in the general AI group increase their automation bias (+0.587, $p < 0.05$).
- *High managerial experience subsample* (≥10 years; $N = 184$; Table 4): *Agentic AI* shows an association with quality improvements for experienced managers (Δ expert quality +0.421, $p < 0.05$) without increased self-reported automation bias (+0.116, $p = 0.573$). Experienced managers also show positive changes in subjective assessments (Δ SPS +3.982 percentage points, $p < 0.10$; Δ CIT +0.385, $p < 0.05$) and LLM-assessed quality (Δ LLM quality +0.112, $p < 0.05$). *General AI* produces smaller effects for this group (Δ CIT +0.230, $p < 0.10$) without significant quality gains assessed by experts.

6.4. Human-AI Interaction Patterns

To better understand how the participants used AI and what they find most helpful, we studied their queries. Specifically, we analyzed 1,273 messages from 603 participants (609 threads), classifying user intent into seven categories to understand the typical questions that users were asking. Classification was LLM-based (GPT-5.1 and Claude Sonnet 4.5; 10 repeats per model).[11] Task delegation captures autopilot behavior (offloading the entire task), evaluation seeking and refinement request capture copilot behavior (iterative

Table 3. Heterogeneous Treatment Effects on ΔOutcomes: PhD Subsample

Dependent variable	Δ*SPS*	Δ*CIT*	Δ*Expert Quality*	Δ*Expert PS*	Δ*Expert CIT*	Δ*LLM Avg Quality*	Δ*LLM Avg PS*	Δ*LLM Avg CIT*
General AI	4.713**	0.115	0.547*	15.882**	0.318*	0.186**	0.139	0.025
	(2.218)	(0.173)	(0.329)	(7.368)	(0.189)	(0.082)	(0.605)	(0.039)
	[0.036]	[0.508]	[0.099]	[0.033]	[0.095]	[0.024]	[0.819]	[0.527]
Agentic AI	8.422***	−0.064	0.164	4.983	−0.003	0.078	0.314	−0.019
	(2.326)	(0.203)	(0.241)	(5.663)	(0.224)	(0.067)	(0.548)	(0.045)
	[0.000]	[0.754]	[0.498]	[0.381]	[0.988]	[0.243]	[0.568]	[0.671]
Automation Bias	1.467*	0.019	0.059	1.134	−0.055	0.014	0.052	0.020*
	(0.761)	(0.076)	(0.093)	(2.087)	(0.072)	(0.030)	(0.242)	(0.012)
	[0.057]	[0.808]	[0.526]	[0.588]	[0.449]	[0.635]	[0.830]	[0.090]
Job - Accounting & Fin	1.375	−0.025	−0.307	−6.774	−0.157	−0.032	0.780*	−0.005
	(2.170)	(0.180)	(0.341)	(6.706)	(0.195)	(0.064)	(0.470)	(0.041)
	[0.528]	[0.891]	[0.370]	[0.315]	[0.421]	[0.620]	[0.100]	[0.904]
Δ*GenAI Expertise*	−4.313**	0.116	−0.027	1.866	0.073	0.012	0.612	−0.082***
	(2.161)	(0.145)	(0.179)	(4.155)	(0.140)	(0.059)	(0.423)	(0.027)
	[0.049]	[0.425]	[0.880]	[0.654]	[0.605]	[0.834]	[0.151]	[0.003]
Job - IT	−1.729	0.134	−0.244	2.780	0.474	0.262	1.866	−0.087*
	(7.621)	(0.435)	(0.421)	(9.275)	(0.294)	(0.207)	(1.347)	(0.050)
	[0.821]	[0.758]	[0.563]	[0.765]	[0.110]	[0.209]	[0.169]	[0.086]
Constant	2.475**	0.236**	0.228	2.380	−0.131	0.005	−0.260	0.000
	(0.974)	(0.103)	(0.142)	(2.958)	(0.121)	(0.026)	(0.271)	(0.026)
	[0.013]	[0.025]	[0.112]	[0.423]	[0.278]	[0.833]	[0.339]	[0.990]
Observations	113.000	113.000	113.000	113.000	113.000	113.000	113.000	113.000
R^2	0.151	0.021	0.044	0.076	0.050	0.111	0.098	0.111
Robust SE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Robust standard errors (SEs) are given in parentheses; *p*-values are in brackets. Controls per PhD balance table.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

Table 4. Heterogeneous Treatment Effects on ΔOutcomes: High Managerial Experience ≥ 10 Years

Dependent variable	Δ*SPS*	Δ*CIT*	Δ*Expert Quality*	Δ*Expert PS*	Δ*Expert CIT*	Δ*LLM Avg Quality*	Δ*LLM Avg PS*	Δ*LLM Avg CIT*
General AI	0.774	0.229*	0.203	6.247	0.020	0.039	−0.230	−0.006
	(1.773)	(0.137)	(0.218)	(4.806)	(0.147)	(0.059)	(0.452)	(0.034)
	[0.663]	[0.097]	[0.353]	[0.195]	[0.891]	[0.506]	[0.611]	[0.852]
Agentic AI	3.982*	0.385**	0.421**	5.174	−0.184	0.112**	0.634	−0.054
	(2.158)	(0.152)	(0.210)	(4.053)	(0.150)	(0.051)	(0.480)	(0.038)
	[0.067]	[0.012]	[0.046]	[0.203]	[0.223]	[0.029]	[0.188]	[0.157]
edField - Health Sciences	−1.797	−0.019	0.134	0.099	0.098	0.036	0.420	0.017
	(1.750)	(0.123)	(0.168)	(3.595)	(0.123)	(0.047)	(0.376)	(0.029)
	[0.306]	[0.874]	[0.424]	[0.978]	[0.425]	[0.453]	[0.265]	[0.566]
Job - Strategy & Planning	−1.733	−0.240**	0.198	1.400	−0.145	−0.120	−0.980	0.098**
	(2.313)	(0.119)	(0.328)	(5.489)	(0.159)	(0.077)	(0.656)	(0.040)
	[0.455]	[0.045]	[0.547]	[0.799]	[0.364]	[0.123]	[0.137]	[0.015]
Industry - Manufacturing	−1.632	−0.126	−0.337	−10.090	−0.188	0.011	0.899	0.031
	(2.215)	(0.154)	(0.294)	(6.384)	(0.142)	(0.096)	(0.684)	(0.038)
	[0.462]	[0.416]	[0.254]	[0.116]	[0.190]	[0.911]	[0.190]	[0.411]
Constant	5.641***	0.131	−0.061	2.815	−0.069	0.050	0.462	−0.032
	(1.265)	(0.101)	(0.162)	(3.296)	(0.119)	(0.035)	(0.329)	(0.028)
	[0.000]	[0.196]	[0.706]	[0.394]	[0.563]	[0.154]	[0.162]	[0.243]
Observations	184.000	184.000	184.000	184.000	184.000	184.000	184.000	184.000
R^2	0.030	0.052	0.033	0.027	0.022	0.047	0.054	0.047
Robust SE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Robust standard errors (SEs) are given in parentheses; *p*-values are in brackets. Controls per *High-MngExp* balance table.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

validation and improvement), and the remaining categories include "information seeking," "clarification," "acknowledgment," and "other." The participants averaged 2.1 queries; 39% asked only one question. Single-query users favored task delegation (39%) over copilot engagement (22%); multiquery users showed the reverse (27% versus 30%), suggesting distinct orientations from the outset. Task delegation decreased from 32% (first query) to 19% (third query), whereas copilot behavior increased from 27% to 37%—a shift from autopilot to iterative engagement. Participation diverged according to the characteristics of the users, aligned with the heterogeneous treatment effects. Highly educated participants (PhDs) under the Agentic AI condition showed elevated autopilot behavior (34% versus 22%) and reduced copilot behavior (36% versus 41%)—delegating the strategy task to AI without iteration. Experienced managers showed the opposite: elevated copilot behavior in Agentic AI (36% versus 26%).

7. Discussion

7.1. Interpreting the Experimental Evidence

Our experiment reveals that AI effectiveness depends on matching system architecture to user expertise (and other characteristics), given the strategic problem. The human-only condition showed learning effects—increased confidence and performance—aligned with Heshmati and Csaszar (2024), who show that taking a strategy course improves both. General AI produced a similar pattern, increasing both confidence (+1.71 percentage points, $p = 0.057$) and quality (+0.255, $p < 0.01$). Agentic AI, instead, increased confidence, but performance varied by user type. Both AI treatments increased automation bias (General AI: +0.371, $p < 0.01$; Agentic AI: +0.183, $p < 0.05$), with General AI also reducing algorithmic aversion (−0.344, $p < 0.01$). Participants seem to develop "functional dependence"—they rely on AI while maintaining sufficient critical distance to extract value.

Agentic AI overall exhibited selective effectiveness. Experienced managers (≥10 years) achieved quality improvements (+0.421, $p < 0.05$) without confidence inflation, demonstrating adaptive expertise (Chaturvedi et al. 2025). Their interaction patterns reflected this: elevated copilot behavior (+9.4 percentage points in evaluation and refinement), suggesting they used the system to validate rather than replace their judgment. Less experienced managers showed the strongest increases in automation bias (General AI: +0.416, $p < 0.001$; Agentic AI: +0.222, $p < 0.05$) with mixed performance, suggesting overreliance without critical evaluation. Managers with PhDs presented a puzzling pattern: maximum confidence inflation (+8.42 percentage points, $p < 0.01$) without quality improvements. Analysis of user interactions (Section 6.4) provides indicative behavioral evidence: PhDs in Agentic AI showed elevated task delegation (34% versus 22%), with 64% asking only a single query (versus 34% in General AI), consistent with autopilot-style engagement.

Several mechanisms may explain these effects, such as *mirroring* (surface-level validation when AI uses familiar causal language), *cognitive miserliness* (Deng and Deng 2025) (offloading effort to plausible-sounding outputs), or *false familiarity* with theoretical frameworks. The negative correlation between AI expertise[12] and confidence inflation (−4.31, $p < 0.05$) suggests these effects diminish with AI exposure (Horowitz et al. 2024).

Our results suggest that user characteristics, AI system design choices, and the nature of strategic problems idiosyncratically interact to shape outcomes. Given the strategic problem at hand, our results identify an effective design configuration for experienced managers: an agentic AI system characterized by a modular architecture with routing-based orchestration in copilot mode. But other sets of design choices are likely to best fit users with different characteristics and different strategic problems.

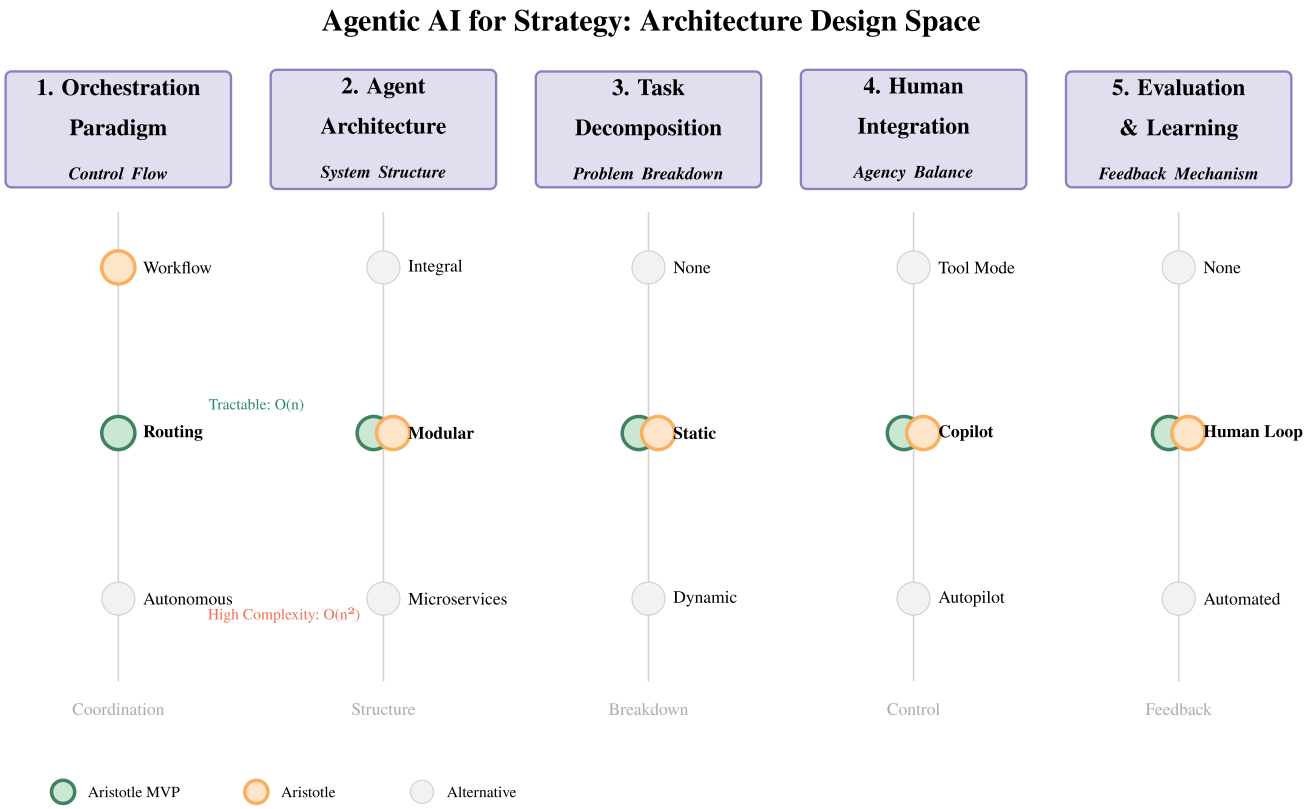
7.2. A Taxonomy of Agentic AI System Architectures for Strategy

When designing Aristotle, as illustrated in Section 3, we started with three design considerations—underlying framework, human-AI integration, and resource constraints—each grounded in the principles of the architecture of complexity. After designing the complete 10-agent system and testing a stripped-down version of it, we learned how user characteristics, system architecture, and problem complexity interact—the "user-system-problem" fit dimension developed in Section 7.

Based on this, the AI design principles and dimensions highlighted in Sections 3 and 4, and the recent acquisitions of the agentic AI literature (Anthropic 2024a, b, 2025; Abou Ali et al. 2025; Sapkota et al. 2025), we abductively developed a taxonomy mapping the design space for agentic AI systems (Sætre and Van de Ven 2021, Miller et al. 2025). This taxonomy defines the possible architectures for AI strategy assistants and might help strategists situate their own designs within this broader space. The taxonomy defines a design space that logically extends the design approach described in Section 3. The taxonomy is characterized by five dimensions: (1) *task decomposition* (how strategic workflow segments into agent responsibilities), (2) *orchestration paradigm* (how agents coordinate), (3) *human-AI integration* (what mode is chosen along the spectrum tool mode: passive assistance on demand, copilot mode (active partnership with shared initiative), and autopilot mode (delegated decision authority)), (4) *agent architecture* (the complexity-tractability trade-off), and (5) *evaluation and learning* (feedback mechanisms critically shape system effectiveness and user-system-problem fit) (Chakraborty et al. 2025).[13]

7.2.1. The Architecture Design Space. Figure 3 presents the five-dimensional taxonomy characterizing the design space for agentic AI systems for strategy.

Figure 3. (Color online) Taxonomy of Agentic AI Systems for Strategy



The figure visually represents the possible architectures for agentic AI systems for strategy and helps strategists visualize their own designs within this broader space. The figure also allows us to trace our design journey. Aristotle (in orange) is characterized by workflow orchestration and higher architectural complexity (10 agents, etc.). When we started to think about its implementation, we had to deal with resource constraints—workflows proved too rigid; microservice architectures exceeded computational budgets. Testing required (and was purposely designed as) an MVP of Aristotle (in green)—three agents with routing orchestration. One dimension emerged through building: static task decomposition (after dynamic routing created delays).

7.2.2. Dimensional Analysis. Each dimension of the taxonomy represents a potential set of architectural design choices with implications for the AI system's behavior.

The orchestration paradigm captures how control flows through the system (Tran et al. 2025). The mechanism for coordinating agent activities ranges from rigid workflow systems that enforce methodological discipline, through flexible routing approaches that preserve user agency, to fully autonomous systems that adapt dynamically. Our experiment demonstrates that routing provides an effective middle ground—maintaining sufficient structure to guide users while avoiding the brittleness of predetermined workflows.

Near decomposability captures the coupling structure between the components of the system (Ethiraj and Levinthal 2004, Derouiche et al. 2025). System components' degree of coupling varies from low to high, moving from *integral* designs through *modular* architectures with specialized agents (the minimal version of Aristotle we used in the experiment) to *microservice* approaches where agents spawn subagents dynamically. Coordination complexity scales dramatically across this spectrum, from O(n) for modular systems to $O(n^2)$ for microservices—a critical consideration for practical deployment.

Task decomposition captures how strategic problems become stable intermediate forms: how AI systems break down strategic problems ranges from *none* (treating each query holistically) through *static* decomposition (predetermined task allocation) to *dynamic* adaptation based on problem characteristics. Our static approach in Aristotle proved adequate for relatively simple problems, but may limit applicability to more exploratory strategic challenges.

Human-AI integration captures the architectural boundaries, that is, where human agency meets system agency. The relative distribution of agency between the human and AI systems varies from *tool mode* (AI as passive assistant) through *copilot* collaboration (our approach) to *autopilot* delegation (He et al. 2025). For strategic work—inherently creative and contextual—the copilot paradigm seems especially desirable, as it preserves human judgment while providing structured support.

The evaluation and learning dimension captures adaptability and evolvability, that is, how change options are built into the system and how the system improves through feedback (Ethiraj and Levinthal 2004, Chakraborty et al. 2025). Systems range from having *no* evaluation mechanisms through *human-in-the-loop* feedback (our approach) to fully *automated* assessment. In AI systems for strategy, given their exploratory purposes and the importance of contextual factors, human evaluation remains essential for meaningful evaluation and learning.

7.2.3. Empirical Grounding. Our experimental findings (Section 5) provide an instantiation of the five dimensions of the taxonomy. Experienced managers achieved higher-quality improvements with agentic AI without confidence inflation, suggesting that the "copilot plus routing configuration" supports calibrated use among domain experts." Highly educated participants (PhD) using agentic AI showed confidence gains without quality improvements, revealing a misfit between AI system design and users lacking practical grounding. These patterns of user-system-problem fit help to make sense of the design options available and of the corresponding trade-offs.

7.3. Contributions and Limitations

Building on our design journey and experimental evidence, this study makes three contributions to the literature on AI systems for strategy. First, we establish user-system-problem fit as a fundamental design dimension: the architectural properties of an AI system for strategy must be matched to the strategic challenges and to the practical domain expertise and behavioral traits of the users to avoid overconfidence and miscalibrated reliance. Second, we offer a five-dimensional taxonomy of AI systems for strategy—spanning orchestration paradigm, agent architecture, task decomposition, human integration mode, and evaluation mechanisms—that maps the design space and allows strategists to identify possible design configurations. Third, we provide a methodological roadmap for systematic AI agent design. By documenting our design journey—from conceptual architecture through foundational choices to experimental testing—we illustrate how to navigate trade-offs in AI system design choices. Our case exemplifies principled simplification that trades theoretical completeness for empirical clarity, enabling researchers to move beyond prompt optimization toward designing systems where architectural elements produce measurable effects.

From a practical point of view, our study highlights that implementing AI systems for strategic decisions must ensure user-system-problem fit. Deployment decisions require full awareness of three interacting elements: (1) who receives the tool, (2) what problems it addresses, and (3) how the system is designed. Our taxonomy (Section 7.2) reveals the architectural choices available across these dimensions; the task of practitioners is to ensure that their choices are internally consistent. When mismatches occur—for example, users becoming disengaged or overconfident without performance gains, or sophisticated tools underperforming simpler alternatives—the appropriate response is to revisit design choices rather than assume universal AI benefits. Standard self-report metrics prove insufficient for detecting such mismatches; organizations should instead rely on behavioral indicators—decision revision rates, time allocation patterns, information search behaviors—to assess AI's actual influence on decision quality.

Our study suffers from several limitations. In addition to its limited external validity, it does not investigate the effects of variation on task complexity (the nature of the strategic problems to solve). Furthermore, it also captures only short-term effects, potentially missing longer-term adaptation, learning curves, or calibration processes that would emerge through repeated AI interaction over a longer period of time. Another limitation regards model specificity. Our findings are specific to GPT-4o; newer models with enhanced reasoning capabilities or different training approaches might produce different interaction patterns. Other limitations pertain to the experimental study.

For example, whereas our experiment employed two human expert evaluators whose assessments were corroborated by LLM evaluations, a larger number of human evaluators would strengthen confidence in quality assessments. Similarly, we cannot determine whether observed confidence increases ultimately prove adaptive (encouraging beneficial experimentation) or harmful (leading to inadequate decision scrutiny) in extended use. The nature of the strategic problem and the characteristics of the sample in the experiment also limit the external validity and reliability. Our self-selected Prolific sample may differ from "real-world" managers in terms of technological comfort and decision-making approaches, and our strategy challenge (technology startup for food waste reduction) may not fully capture real-world strategic problems and stakes. With regard to potential measurement error, self-reported automation bias and algorithmic aversion may inadequately capture AI's actual influence and might represent an alternative explanation of the disconnect between different outcomes.

7.4. Future Research Agenda

The proposed taxonomy, combined with our empirical findings, enables a systematic research agenda on how to purposely design and use AI for strategy.

7.4.1. User Fit. An important set of questions concerns user characteristics and cognitive mechanisms: how managers' representational capability—their ability to select and deploy appropriate analytical frameworks—interacts with AI assistance and whether AI compensates for skill deficits or amplifies existing differences among decision makers (Csaszar et al. 2024a, b). Research should investigate how individual differences (expertise, cognitive style, domain knowledge) interact with architectural choices to predict performance outcomes. Equally important are questions about how managers optimally allocate cognitive resources between independent reasoning and AI interrogation, what factors determine belief updating versus skepticism, and how mental models of AI capabilities evolve with experience.

7.4.2. System Design. Researchers should investigate the optimal design of AI-augmented strategy tools, particularly regarding malleability: highly flexible tools may overwhelm less capable users, whereas overly constrained ones may limit expert strategists (Gurzick et al. 2025). Experimental studies could empirically test different AI system design configurations—for instance, comparing routing versus workflow orchestration for entrepreneurial strategy formulation or strategy visioning for established firms. Research on dynamic adaptation could help to elicit the mechanisms that enable AI systems to adjust their architecture in response to task characteristics or user needs. Hybrid approaches also merit attention: could systems combine multiple positions on single dimensions—for example, offering both workflow and routing modes depending on user preference or task requirements? Investigation should address how different implementations of agentic AI—beyond our minimal, theory-based three-agent architecture—affect learning and performance outcomes, exploring different underlying strategy frameworks, agent typologies, orchestration strategies, and integration of capabilities into general AI systems. The development of measurement frameworks would enable standardized evaluation across dimensions, facilitating cumulative knowledge building. As Sapkota et al. (2025) emphasize, predicting the complexity implications of design choices before implementation remains a critical methodological challenge.

7.4.3. Problem Fit. A critical question is when managers, strategy frameworks, and agentic AI truly complement each other rather than create false confidence. Our study underlines the importance of exploring the design of purposely designed, firm-specific agentic AI

systems as dynamic capabilities and sources of competitive advantage. AI integration also carries risks, including overreliance that erodes critical thinking, rejection of assistance when outputs challenge priors, and biases in training data that skew decisions.

7.5. Conclusion

We asked how AI systems for strategy should be designed and how architecture affects strategic decision making. The answer is contingent: design choices must fit the users and problems involved. We call this user-system-problem fit. If purposeful AI design can be a source of competitive advantage, then strategists must learn to be designers of AI systems, not merely users of them. As agentic AI systems become more capable and widespread, we expect this design capability to become increasingly consequential.

Endnotes

[1] Experimentally testing the full-fledged version of Aristotle would be challenging and hardly feasible on a large scale because of extended interaction times, computational requirements, and other constraints.

[2] We do not claim that other agentic AI systems for strategy, even theory based, would produce similar effects. Furthermore, we do not argue that routing-based agency represents the only or optimal form of AI support to strategic decision making.

[3] They were asked to articulate a theory of value—a causal narrative explaining how the venture would create value by addressing the food waste problem. This differs from a detailed business model in that it focuses on the underlying causal logic (e.g., "predictive analytics reduces overordering, which cuts waste and improves margins") rather than operational specifics (pricing, channels, organizational structure). This distinction aligns with the theory-based approach's emphasis on causal reasoning over procedural details (Felin and Zenger 2017).

[4] To ensure meaningful engagement, participants had to spend a minimum of two minutes on the task (they could not proceed on the platform before then).

[5] We explicitly instructed participants in the control group (human-only condition) not to use AI. Whereas we do not know whether they complied or not, potential noncompliance makes our estimates of the treatment effects conservative.

[6] Attrition did not differ by treatment. Of 373 exclusions, two participants maintained 0% SPS throughout both stages (indicating nonengagement).

[7] The top 33% of performers received double compensation (incentive identical across conditions). Both AI interfaces were labeled "Aristotle" to control for expectation effects; participants engaged meaningfully with the systems.

[8] In Section 6, we present analyses on subsamples, including strata used to stratify randomization. Balance checks were also conducted in this case. All regressions include unbalanced covariates as controls.

[9] We excluded participants who failed attention and manipulation checks from the analyses.

[10] Working as the baseline, the omitted category.

[11] Classification details, including all the user queries, are publicly available as codebase at https://github.com/abhinavbocconi/AI-TheoryBasedDecisions_LLMEvaluation.

[12] Self-assessment of prompting skills and GenAI familiarity.

[13] The design approach maps onto the five-dimensional extended design space as follows. The underlying strategy framework shapes both task decomposition and the orchestration paradigm. Human-AI integration maps to the human integration dimension. Resource constraints inform agent architecture. Evaluation and learning emerged abductively from our experimental findings.

References

Abou Ali M, Dornaika F, Charafeddine J (2025) Agentic AI: A comprehensive survey of architectures, applications, and future directions. *Artificial Intelligence Rev.* 59(1):11.

Agrawal A, Gans J, Goldfarb A (2022) *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence* (Harvard Business Review Press, Boston).

Agrawal A, Camuffo A, Gans JS, Scott E, Stern S (2025) *The Foundations of Bayesian Entrepreneurship* (MIT Press, Cambridge, MA).

Anthropic (2024a) Building effective agents. Accessed November 29, 2025, https://www.anthropic.com/research/building-effective-agents.

Anthropic (2024b) Claude's extended thinking. Accessed November 29, 2025, https://www.anthropic.com/news/visible-extended-thinking.

Anthropic (2025) Introducing Claude 4. Accessed November 29, 2025, https://www.anthropic.com/news/claude-4.

Baldwin C, MacCormack A, Rusnak J (2014) Hidden structure: Using network methods to map system architecture. *Res. Policy* 43(8):1381–1397.

Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR (2024) The crowdless future? Generative AI and creative problem-solving. *Organ. Sci.* 35(5):1589–1607.

Cabigiosu A, Camuffo A (2016) Measuring modularity: Engineering and management effects of different approaches. *IEEE Trans. Engrg. Management* 64(1):103–114.

Camuffo A, Gambardella A, Pignataro A (2024) Theory-driven strategic management decisions. *Strategy Sci.* 9(4):382–396.

Chakraborty S, Pourreza M, Sun R, Song Y, Scherrer N, Huang F, Bedi AS, et al. (2025) On the role of feedback in test-time scaling of agentic AI workflows. Preprint, submitted April 2, https://arxiv.org/abs/2504.01931.

Chalmers D, MacKenzie NG, Carter S (2021) Artificial intelligence and entrepreneurship: Implications for venture creation in the fourth industrial revolution. *Entrepreneurship Theory Practice* 45(5):1028–1053.

Chaturvedi A, Dasgupta M, Yadav N (2025) Tech-driven transformation: Unravelling the role of artificial intelligence in shaping strategic decision-making. *Internat. J. Human-Comput. Interaction* 41(19):12305–12324.

Cockburn IM, Henderson R, Stern S (2018) The impact of artificial intelligence on innovation: An exploratory analysis. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 115–146.

Conti A, Messinese D (2024) The selective tailwind effect of A.I. on startups: Predictions and anomalies. Preprint, submitted October 24, https://doi.org/10.2139/ssrn.4958898.

Csaszar FA, Steinberger T (2022) Organizations as artificial intelligences: The use of artificial intelligence analogies in organization theory. *Acad. Management Ann.* 16(1):1–37.

Csaszar FA, Hinrichs N, Heshmati M (2024a) External representations in strategic decision-making: Understanding strategy's reliance on visuals. *Strategic Management J.* 45(11):2191–2226.

Csaszar FA, Ketkar H, Kim H (2024b) Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. *Strategy Sci.* 9(4):322–345.

Dehghani N, Levin M (2024) Bio-inspired AI: Integrating biological complexity into artificial intelligence. Preprint, submitted November 22, https://arxiv.org/abs/2411.15243.

Deng Z, Deng Z (2025) Becoming a cognitive miser? Antecedents and consequences of addictive ChatGPT use. *Soc. Sci. Medicine* 383:118467.

Derouiche H, Brahmi Z, Mazeni H (2025) Agentic AI frameworks: Architectures, protocols, and design challenges. Preprint, submitted August 13, https://arxiv.org/abs/2508.10146.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych.: General* 144(1):114–126.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Sci.* 64(3): 1155–1170.

Doshi AR, Bell JJ, Mirzayev E, Vanneste BS (2025) Generative artificial intelligence and evaluating strategic decisions. *Strategic Management J.* 46(3):583–610.

Ehrig T, Schmidt J (2022) Theory-based learning and experimentation: How strategists can systematically generate knowledge at the edge between the known and the unknown. *Strategic Management J.* 43(7):1287–1318.

Ethiraj SK, Levinthal D (2004) Bounded rationality and the search for organizational architecture: An evolutionary perspective on the design of organizations and their evolvability. *Admin. Sci. Quart.* 49(3):404–437.

Felin T, Holweg M (2024) Theory is all you need: AI, human cognition, and causal reasoning. *Strategy Sci.* 9(4):346–371.

Felin T, Zenger TR (2009) Entrepreneurs as theorists: On the origins of collective beliefs and novel strategies. *Strategic Entrepreneurship J.* 3(2):127–146.

Felin T, Zenger TR (2017) The theory-based view: Economic actors as theorists. *Strategy Sci.* 2(4):258–271.

Finio M, Downie A (2025) What is AI agent orchestration? Accessed November 29, 2025, https://www.ibm.com/think/topics/ai-agent-orchestration.

Fiske STT, Taylor SE (2020) *Social Cognition: From Brains to Culture* (SAGE Publications Ltd., London).

Gaessler F, Piezunka H (2023) Training with AI: Evidence from chess computers. *Strategic Management J.* 44(11):2724–2750.

Gans JS (2023) Experimental choice and disruptive technologies. *Management Sci.* 69(11):7044–7058.

Gans JS (2025) AI as strategist. NBER Working Paper No. 33650, National Bureau of Economic Research, Cambridge, MA.

Gerber AS, Green DP (2017) Field experiments on voter mobilization: An overview of a burgeoning literature. Banerjee AV, Duflo E, eds. *Handbook of Economic Field Experiments*, vol. 1 (North Holland, Amsterdam), 395–438.

Gurzick D, Joshi MP, Gurzick M, Csaszar F, Jia N (2026) AI-augmented strategic tools for strategy formulation and implementation: Revisiting traditional strategy tools & frameworks. Csaszar FA, Nick AM, Jia N, eds. *Handbook of Artificial Intelligence and Strategy* (Edward Elgar, Cheltenham, UK), 47–81.

Harbarth L, Göβwein E, Bodemer D, Schnaubert L (2025) (Over)trusting AI recommendations: How system and person variables affect dimensions of complacency. *Internat. J. Human-Comput. Interaction* 41(1):391–410.

He G, Hemmer P, Vössing M, Schemmer M, Gadiraju U (2025) Fine-grained appropriate reliance: Human-AI collaboration with a multi-step transparent decision workflow for complex task decomposition. Preprint, submitted January 19, https://arxiv.org/abs/2501.10909.

Heshmati M, Csaszar FA (2024) Learning strategic representations: Exploring the effects of taking a strategy course. *Organ. Sci.* 35(2):453–473.

Horowitz M, Kahn L, Macdonald J, Schneider J (2024) Adopting AI: How familiarity breeds both trust and contempt. *Internat. J. Human-Comput. Interaction* 41(1):1721–1735.

Hünermund P, Kaminski J, Schmitt C (2021) Causal machine learning and business decision making. Preprint, submitted June 24, http://dx.doi.org/10.2139/ssrn.3867326.

Kemp A (2024) Competitive advantage through artificial intelligence: Toward a theory of situated AI. *Acad. Management Rev.* 49(3):618–635.

Krakowski S, Luger J, Raisch S (2023) Artificial intelligence and the changing sources of competitive advantage. *Strategic Management J.* 44(6):1425–1452.

Microsoft (2025) AI agent orchestration patterns. Accessed November 29, 2025, https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/ai-agent-design-patterns.

Miehling E, Ramamurthy KN, Varshney KR, Riemer M, Bouneffouf D, Richards JT, Dhurandhar A, et al. (2025) Agentic AI needs a systems theory. Preprint, submitted February 28, https://arxiv.org/abs/2503.00237.

Miller CC, Chattopadhyay P, Bamberger P, Rockmann K (2025) Leveraging empirical abduction to bridge the rigor–Relevance divide: Celebrating 10 years of academy of management discoveries. *Acad. Management Discoveries* 11(3):325–331.

Obschonka M, Audretsch DB (2020) Artificial intelligence and big data in entrepreneurship: A new era has begun. *Small Bus. Econom.* 55:529–539.

Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: An attentional integration. *Human Factors* 52(3):381–410.

Raisch S, Krakowski S (2021) Artificial intelligence and management: The automation–augmentation paradox. *Acad. Management Rev.* 46(1):192–210.

Ramoglou S, Chandra Y, Jin Q (2025) Opportunity search in the era of GenAI: Navigating uncertainty in an expanding universe of imaginable but unknowable futures. *J. Management Stud.*, ePub ahead of print November 6, https://doi.org/10.1111/joms.70011.

Romeo G, Conti D (2025) Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI Soc.* 41:259–278.

Russell SJ, Norvig P (2016) *Artificial Intelligence: A Modern Approach* (Pearson Education, Hoboken, NJ).

Sætre AS, Van de Ven A (2021) Generating theory by abduction. *Acad. Management Rev.* 46(4):684–701.

Sako M, Felin T (2025) Does AI prediction scale to decision making? *Comm. ACM* 68(4):18–21.

Sapkota R, Roumeliotis KI, Karkee M (2025) AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *Inform. Fusion* 126:103599.

Sempf A, Hooker A (2025) Agentic AI patterns and workflows on AWS. Accessed November 29, 2025, https://docs.aws.amazon.com/prescriptive-guidance/latest/agentic-ai-patterns/introduction.html.

Shavit Y, Agarwal S, Brundage M, Adler S, O'Keefe C, Campbell R, Lee T, et al. (2023) Practices for governing agentic AI systems. Research paper, OpenAI, San Francisco.

Shepherd DA, Majchrzak A (2022) Machines augmenting entrepreneurs: Opportunities (and threats) at the nexus of artificial intelligence and entrepreneurship. *J. Bus. Venturing* 37(4): 106227.

Simon HA (1962) The architecture of complexity. *Proc. Amer. Philos. Soc.* 106(6):467–482.

Sorenson O (2024) Theory, search, and learning. *Strategy Sci.* 9(4): 372–381.

Townsend DM, Hunt RA (2019) Entrepreneurial action, creativity, & judgment in the age of artificial intelligence. *J. Bus. Venturing Insights* 11:e00126.

Townsend DM, Hunt RA, Rady J, Manocha P, Hyeong JJ (2025) Are the futures computable? Knightian uncertainty and artificial intelligence. *Acad. Management Rev.* 50(2):415–440.

Tran KT, Dao D, Nguyen MD, Pham QV, O'Sullivan B, Nguyen HD (2025) Multi-agent collaboration mechanisms: A survey of LLMs. Preprint, submitted January 10, https://arxiv.org/abs/2501.06322.

Wooldridge M (2009) *An Introduction to Multiagent Systems* (John Wiley & Sons, Hoboken, NJ).

Arnaldo Camuffo is a professor of management at Bocconi University and codirector of ION Management Science Lab.

Alfonso Gambardella is a professor of management at Bocconi University and codirector of ION Management Science Lab.

Saeid Kazemi holds a PhD from Bocconi University and is a research fellow at the ION Management Science Lab.

Abhinav Pandey is a PhD candidate at Bocconi University and research fellow at the ION Management Science Lab.