AGENTS FOR EXPERIMENT, EXPERIMENTS FOR AGENTS: A TOPOLOGICAL FRAMEWORK FOR AUTOMATED MECHANISM DISCOVERY

Yingjie Zhang
Guanghua School of Management
Peking University
Beijing, China
yingjiezhang@gsm.pku.edu.cn

Weizhang Zhu
Cheung Kong Graduate School of Business
Beijing, China
allen2023cu@link.cuhk.edu.hk

Chun Feng
Cheung Kong Graduate School of Business
Beijing, China
fc13303752056@stu.xjtu.edu.cn

Tianshu Sun
Cheung Kong Graduate School of Business
Beijing, China
tianshusun@ckgsb.edu.cn

ABSTRACT

The transition of AI from static predictors to strategic agents has elevated mechanism design—the engineering of interaction rules, authority flows, and feedback loops—to a central challenge in algorithmic governance. While experimentation serves as the primary instrument for uncovering these dynamics, the traditional manual design process cannot keep pace with the combinatorial complexity of agentic systems. We propose a recursive paradigm of *Agents for Experiment*, employing AI architects to design the protocols used to study AI systems (*Experiments for Agents*). However, without a rigorous ontology, current agentic approaches rely on unstructured text, potentially yielding operationally invalid or scientifically trivial designs. To bridge this gap, we introduce **SEED** (Structural Encoding for Experimental Discovery), a framework that formalizes experimental protocols as computable runtime execution graphs. By decoupling the topological skeleton of an interaction from its semantic context, SEED provides a unified grammar for automated discovery. We demonstrate the framework's utility through three distinct functions: (1) Descriptive Utility, which synthesizes fragmented literature into a standardized topology library; (2) Evaluative Utility, which operationalizes scientific novelty via computable evaluation scores; and (3) Generative Utility, which enables a "Generative Topology Search" algorithm. This allows agentic researchers to systematically identify structural gaps, which are counterintuitive governance architectures adjacent to established science, and propose novel experimental designs. We conclude that SEED transforms mechanism discovery from an artisanal craft into a structured optimization problem, laying the foundation for high-throughput experimental science in the agentic era.

1 INTRODUCTION

The rapid evolution of Large Language Models (LLMs) has transformed AI from static predictive tools into strategic actors. As these agents populate markets, organizations, and decision loops, they introduce a new layer of complexity: they are governed not just by code, but by *mechanisms*—the dynamic architectures of rules, authority, and information flow that determine how human and artificial agents co-adapt. Unlike fixed physical laws, these mechanisms are combinatorial and adaptive; a subtle shift in a feedback loop or a transparency setting can radically alter the emergent outcome of a collaboration (Amershi et al., 2019; Bansal et al., 2019). Consequently, the discovery of robust governance mechanisms has become the central challenge of the agentic era.

Experimentation serves as the "holy grail" for uncovering these mechanisms. Whether evaluating a medical triage assistant or an algorithmic pricing model, the controlled experiment is the most reliable instrument to distinguish causal interaction dynamics from noise (Kohavi et al., 2013). However, a fundamental paradox has emerged: while the systems we study are automated, high-speed, and infinitely complex, the process of experimentation remains manual, labor-intensive, and bounded by human cognition. The design, execution, and iterative improvement of experimental protocols are still performed by human researchers who rely on intuition to navigate a design space that has grown exponentially beyond their reach. We are effectively attempting to map a new continent of interaction dynamics using tools designed for static observations.

To resolve this bottleneck, we propose a recursive paradigm: *Agents for Experiment, Experiments for Agents*. We argue that the scientific process itself must be automated to keep pace with the systems it studies. This goes beyond merely using AI for simulation or data analysis; it requires employing AI agents as experimental architects. Theoretically, autonomous agents possess the computational scale to systematically search the vast landscape of mechanism design, generating hypotheses and optimizing protocols at a speed unattainable by human teams (Yao et al., 2023; Shinn et al., 2023).

Yet, this vision faces a strictly operational barrier. We cannot simply delegate experimental design to unconstrained agents. Without a rigorous structural ontology, it is very likely that an AI asked to "design a mechanism" will hallucinate operationally invalid workflows or fixate on trivial surface-level details, such as prompt phrasing, rather than structural innovation (Cai et al., 2024; Huang et al., 2024). Current experimental designs rely on vague natural language descriptions that machines cannot process rigorously.

To enable true agentic discovery, we must bridge this gap. We introduce **SEED** *(Structural Encoding for Experimental Discovery)*, a framework that formalizes the experiment not as text, but as a computable search space. By providing a standardized grammar for interaction, SEED allows agents to reason about the structure of governance as precisely as they reason about code, transforming mechanism discovery from an artisanal craft into a structured optimization problem.

2 The SEED Framework: A Grammar for Strategic Experiments

We formalize the proposed framework as Structural Encoding for Experimental Discovery (SEED). This framework represents experimental designs not as procedural narratives, but as graph-structured configurations of actors, interactions, and feedback mechanisms.

2.1 Experimentation as a Graph

Experiments in strategic decision settings are often described in terms of tasks, treatments, and outcomes. While sufficient for execution, such descriptions obscure the relational assumptions that determine how strategic behavior emerges. In experiments involving autonomous agents, outcomes depend not only on *what* decisions are made, but on *who* makes them, *which* information they observe, and *how* feedback shapes subsequent actions. These relationships are typically embedded implicitly in experimental protocols, making them difficult to compare across studies or reason about systematically (Wei et al., 2022; Wang et al., 2023).

A topological abstraction addresses this limitation by shifting attention from *implementations* to *relations*. Rather than encoding experiments as domain-specific procedures, we represent them as configurations of interacting entities connected by information, control, and feedback relations. This perspective allows experimental designs that differ in surface details (e.g., task framing or model implementations) to be recognized as structurally isomorphic when they rely on the same strategic assumptions. Conversely, it highlights when seemingly modest design changes (e.g., changing a "monitor" to a "veto") alter the interaction structure in ways that activate fundamentally different strategic mechanisms.

Drawing on a design-oriented perspective, SEED defines a standardized grammar for representing experimental designs. This abstraction decomposes complex human-agent systems into their constituent atomic structures, allowing researchers to represent the skeleton of an interaction separately from the semantic conditions that govern it.

2.1.1 CORE PRIMITIVES: THE TOPOLOGICAL SKELETON

The foundation of the topology is a directed graph $G = (V, E)$ that represents the static architecture of the decision system. This graph is composed of two primary atomic structures: actors and flows.

Actors (V). The nodes of the graph represent the strategic entities interacting within the system. We distinguish actors by their source of agency and intrinsic constraints:

- *Human Actors (Δ):* Represented by the triangle node, these actors are characterized by subjective preferences and intrinsic motivations. Their behavior is bounded by cognitive fatigue, behavioral biases, and unobservable private values.
- *Agentic Actors ($\bigcirc$):* Represented by the circle node. These are autonomous systems defined by externally specified objectives and computational constraints (e.g., context window limits).

Flows (E). The flows define the connectivity between actors. In SEED, edges do not merely represent connection but define *what* is being exchanged, strictly separating data from governance:

- *Content Flows ($\rightarrow$):* The transmission of high-dimensional information, such as data, predictions, rationale, or generated text. Traversing this flow updates the information state of the receiver but does not force action.
- *Control Flows ($\Rightarrow$):* The transmission of executive authority (e.g., triggers, vetoes, or final decision tokens) that might update the action state (i.e., functional capability Θ_{Func}) of the receiver.
- *Iterative Interactions ($\leftrightarrow^n$):* A bidirectional flow labeled with n. This parameter differentiates a one-shot hand-off ($n = 1$) from a recursive co-evolutionary loop ($n > 1$).

2.1.2 THE SEMANTIC LAYER: ATTRIBUTES & MODERATORS

While the topological skeleton defines the wiring, the Semantic Layer defines the *properties* of the components. We categorize these properties into three classes: intrinsic capabilities, exogenous design parameters, and endogenous interaction states.

Node Attributes. These represent the *intrinsic capacity* of an actor $v \in V$. Each node $v \in V$ is characterized by a multidimensional capability profile Θ and an epistemic state. This allows for the structural modeling of actor heterogeneity across the system:

- *Cognitive Capability (Θ_{Cog}):* Represents reasoning depth. For agents, this maps to foundation model capacity; for humans, it maps to domain expertise.
- *Functional Capability (Θ_{Func}):* Represents the actor's action space, specifically their ability to utilize external tools ($\mathcal{T}$) to intervene in the environment (Schick et al., 2023; Patil et al., 2024; Lewis et al., 2020).
- *Epistemic Attributes (K, C, S):* Semantic labels defining what the actor knows. We distinguish between static *Knowledge* (K), task-specific *Context* (C), and access to real-time *State Information* (S).

Governance Moderators (The Design). These are exogenous parameters set by the architect to condition the rules of the game.

- *Protocols ($\mathcal{P}$):* The logical gates that determine the *admissibility* of a flow. These are hard constraints governing *when* a signal is valid. Common implementations include confidence thresholds (e.g., "Delegate ONLY IF Confidence > 0.8"), time or resource limits (e.g., "Veto allowed within 500ms"), and format constraints (e.g., "Output must be structured JSON") (Geng et al., 2023; Geifman & El-Yaniv, 2019; Mozannar & Sontag, 2020).
- *Incentives (X):* The objective function defining the *payoffs* for each actor or flow (i.e., the execution), used to shape the *preference ordering* over possible outcomes. We distinguish

between outcome-based incentives, such as accuracy bonuses, and process-based incentives, such as risk penalties or speed constraints (Christiano et al., 2017; Achiam et al., 2017).

- *Information Design ($\mathcal{I}$):* The parameters governing the *information set* of the receiver (what is revealed) and the *choice architecture* (how it is presented). This category subsumes transparency, masking, and nudges. Examples include masking (blinding the human to the AI's confidence score to prevent bias) (Zhang et al., 2020; Yin et al., 2019), identity disclosure (explicitly revealing that a partner is an agent) (Shi et al., 2020; Chan et al., 2024), explainability (providing rationale alongside predictions) (Ribeiro et al., 2016; Zhang et al., 2020), and nudges (altering the presentation frame or default options without changing the underlying payoff) (Menon et al., 2020; Karinshak et al., 2023).

Interaction Dynamics (The Relation). These are endogenous variables representing the evolving quality of the connection between actors. These dynamics describe the *alignment* between actors.

- *Psychological Alignment (Ψ):* The *affective bond* between actors. This state vector captures the subjective perception of the relationship, distinct from the information itself. Key variables include trust (the willingness to be vulnerable to the other's actions) and psychological safety (the belief that one can question or reject the partner's signal without negative consequences). Low Ψ degrades the loop by inducing skepticism or defensive compliance (Yin et al., 2019; Bansal et al., 2019).
- *Epistemic Alignment ($\mathcal{E}$):* The *informational consensus*. This measures whether both actors differ in their perception of reality. High alignment implies *confirmation* (shared view), while low alignment implies *conflict* (disagreement), necessitating distinct conflict-resolution mechanisms (Zhang et al., 2020; Bansal et al., 2019).
- *Cognitive Alignment (Ω):* The *process efficiency*. This captures the economic cost of maintaining the connection, represented by the mental workload or computational latency required to parse the signal (Amershi et al., 2019; Kocielnik et al., 2019).

2.1.3 From Atoms to Architectures

While atomic primitives define the vocabulary of interaction, they do not by themselves explain system behavior. To capture the higher-level organization, we need system architecture. In SEED, it is not a static flowchart but a computational object composed of two complementary layers: the execution layer (the mechanism itself) and the wrapper layer (the experimental container).

The Runtime Execution Graph. This layer defines the operational topology of the collaboration. It specifies the precise graph instance instantiated to execute a single task, analogous to the "forward pass" in a neural network. This architectural layer determines the operational sequence, such as a sequence-oriented topology (a static DAG where agents draft and humans review) or an adaptive topology (a dynamic graph where edge weights or routing logic shift based on runtime confidence scores). This layer captures the essential mechanism of the joint system: given an input, how does the signal propagate through the network of actors to produce a decision?

The Meta-Optimization Wrapper. This layer functions as the search space definition. Rather than executing a task, this layer defines the class of possible experimental configurations. It specifies the controllable parameters (defining which topological edges are variable "treatments" versus fixed constraints) and injects observability hooks (logging layers) into the workflow. In algorithmic terms, if the runtime execution graph is a function $y = f(x; G)$, the meta-optimization wrapper is the outer loop that varies the structure of G itself. It allows the agentic researcher to systematically perturb the system, toggling delegation rules, masking information access, or varying interaction depth, to measure sensitivity, causality, and performance boundaries.

2.2 The Topological Design Space

The meta-optimization wrapper navigates a high-dimensional *topological design space*. This space represents the manifold of all valid runtime execution graphs that can be instantiated from our atomic

primitives. To structure the search for novel mechanisms, we categorize this space into two distinct mutation classes:

1. Structural Mutation (Topology Search). This class navigates the design space by perturbing the graph's physical skeleton. Starting from a baseline configuration (typically a dyad), the system applies discrete operations to the two fundamental components of the network:

- *Node Mutation (Composition):* This alters the *scale* and *diversity* of the system. The wrapper can expand the graph by injecting auxiliary nodes (e.g., adding a second "Critic Agent" to create a voting triad) or contract it by pruning nodes (e.g., temporarily removing the Human Δ to test fully autonomous sub-routines).
- *Edge Mutation (Connectivity):* This alters the *wiring* and *hierarchy*. Operations include re-routing information paths or performing authority reversal by flipping the direction of a control flow (changing $\Delta \Rightarrow \bigcirc$ to $\bigcirc \Rightarrow \Delta$). This allows the system to systematically test "Counter-Intuitive Delegation" identifying domains where giving the AI veto power over a human yields superior robustness.

2. Semantic Modulation (Parametric Search). This class alters the *rules* governing that wiring and it applies two types of operations:

- *Parameter Activation (Binary Search):* This determines the *existence* of a constraint. The system toggles specific features on or off (e.g., enabling/disabling masking or activating incentives), to identify the "minimum viable governance" required to stabilize the system.
- *Parameter Calibration (Continuous Search):* This determines the *magnitude* of an active parameter. Once a moderator is active, the system sweeps through continuous value ranges. For example, it may fine-tune the confidence threshold within a Protocol ($\mathcal{P}$) to locate the precise tipping point where automation becomes safer than human control.

3 Operationalizing SEED: From Description to Discovery

Having established SEED as a standardized grammar for human-agent interaction, we now turn to its operational utility. The value of this topological abstraction lies not only in clarifying current research but in enabling a transition from human-led to agent-driven discovery. In this section, we outline how SEED functions as the "operating system" for an automated research engine.

3.1 Descriptive Utility: Mapping the Literature

To demonstrate the operational utility of SEED, we first apply the framework to the existing corpus of human-AI interaction research. We find that the vast majority of empirical designs, despite spanning diverse domains from healthcare to finance, can be precisely described using the runtime execution graphs defined in Section 2.1.3. As summarized in Table 1, contemporary studies generally cluster into two primary topological families: sequence-oriented architectures (static directed graphs, e.g., $\bigcirc \rightarrow \Delta$ for "AI advice") and interactive architectures (dynamic feedback loops, e.g., $\bigcirc \leftrightarrow^{n} \Delta$ for "co-creation"). This verifies that our primitives of content/control flows and feedback loops are sufficient to capture the skeleton of contemporary research.

However, this mapping process reveals a critical insight: while the field has explored both linear and interactive processes, the structural diversity of these topologies remains surprisingly limited. When viewed through the lens of SEED, most studies rely on fixed, dyadic arrangements. The primary scientific contribution in these papers typically lies in manipulating the parameters within these stable structures (e.g., varying model capability (Θ_{Cog}) or toggling governance moderators like incentives (X)) rather than mutating the architecture itself. This suggests that the broader topological design space, which contains complex multi-agent meshes or counter-intuitive authority reversals, remains largely unexplored.

We also acknowledge that real-world workflows can be far more intricate than these clean theoretical models. Yet, SEED handles this complexity through compositionality. Much like a formal grammar, the atomic structures of actors and flows can be recursively assembled to reconstruct even the most

Table 1: Illustrative Examples: Mapping Literature to Basic Topological Graphs

Graph Structure	Workflow Type	Representative Papers
$\bigcirc \rightarrow \Delta$	**AI → Human** (Unidirectional assistance / advice)	*Node addition / AI availability:* (Song et al., 2025; Goh et al., 2025; Argyle et al., 2023; Wan et al., 2024; Freitas et al., 2025; Wang et al., 2025; Meyer et al., 2022; Brynjolfsson et al., 2025) *Information / explanation / uncertainty:* (Schanke et al., 2024; Caplin et al., 2025; Von Zahn et al., 2025; Bayer & Renou, 2024; Gnewuch et al., 2024; You et al., 2022; Reis & et al., 2024; Hou et al., 2021; De Toni & et al., 2024; Alur et al., 2024; Li et al., 2024) *Capability / quality:* (Liel & et al., 2025; Wu et al., 2023; Shin & et al., 2023; Binz et al., 2025) *Context / role / anthropomorphism:* (Siemon et al., 2025; Dennis et al., 2023; Han et al., 2023; Krakowski et al., 2025; Schecter & et al., 2023)
$\Delta \Rightarrow \bigcirc$	**Human → AI** (Delegation / task offloading)	*Delegation and takeover:* (Fügener et al., 2022; Bansak & Paulson, 2024)
$\bigcirc \Rightarrow \Delta$	**AI → Human** (Escalation / return of authority)	*Escalation protocols / delegation dynamics:* (Stelmaszak et al., 2025; Liu et al., 2025)
$\bigcirc \leftrightarrow^{n} \Delta$	**Dynamic interactions** (Iterative collaboration / feedback)	*Multi-round interaction and co-evolution:* (Chen & Chan, 2025; Revilla et al., 2023; Glickman et al., 2025; Niraula et al., 2025; Gonzalez et al., 2025; Treiman & et al., 2024; Lin et al., 2024; Fügener et al., 2021)

labyrinthine multi-stage experiments. Whether a workflow involves a ten-step approval chain or a parallel voting ensemble, it remains a finite assembly of our primitives.

To illustrate the precision of this mapping, we deconstruct two representative field experiments. These examples demonstrate how SEED differentiates between purely parametric adjustments (i.e., semantic modulation) and fundamental structural changes (i.e., structural mutation).

Demo Case 1: Semantic Modulation (Moderator Intervention). We first reconstruct a standard "voice chatbot" experiment (e.g., similar to Xu et al. (2024)), which employs a 2×2 factorial design manipulating *Identity Disclosure* and *Anthropomorphism*. In SEED, this is **not** a topological change, but a precise modulation of flow-level semantic tags within a fixed sequence-oriented skeleton:

- *Skeleton.* The experiment fundamentally involves a unidirectional transmission from an agent to a human: $\bigcirc \rightarrow \Delta$.
- *Identity Disclosure.* We encode identity disclosure as an information design moderator $\mathcal{I}_{id} \in \{0, 1\}$ attached to the content flow.
- *Nudge.* We encode anthropomorphic framing as a nudge moderator $\mathcal{I}_{nudge} \in \{0, 1\}$ on the same flow.

Result: The experimental design is formalized as $\bigcirc \xrightarrow{\mathcal{I}_{id},\, \mathcal{I}_{nudge}} \Delta$. SEED reveals this as a purely semantic intervention; the underlying topological structure remains a static, isomorphic dyad across all treatments.

Demo Case 2: Topological Mutation (Node Insertion). Next, we reconstruct the financial advisory field experiment by Yang et al. (2026). Unlike the direct dyadic interaction, this study examines a mediated workflow where human bankers curate AI advice before it reaches the customer. SEED captures this as a structural expansion via node insertion:

- *Step 1: Select the Entities.* We distinguish three strategic actors: the AI system ($\bigcirc$), the human banker (Δ_B), and the downstream consumer (Δ_C).
- *Step 2: Modify Connectivity (Node Insertion).* We apply a *Node Insertion* operation. For the treatment group, the direct signal ($\bigcirc \rightarrow \Delta_C$) is re-routed through the banker, establishing a mediated chain ($\bigcirc \rightarrow \Delta_B \rightarrow \Delta_C$).
- *Step 3: Define the Topological Contrast.* The experiment isolates the structural effect by contrasting the *Mediated Graph* (Human-AI Advice) against the *Direct Graph* (Pure AI Advice).

Result: The design is formalized as the mutation from $\bigcirc \rightarrow \Delta_C$ to $\bigcirc \rightarrow \Delta_B \rightarrow \Delta_C$. This represents a *topological mutation*, where the experimental outcome (increased adherence) is attributed specifically to the architectural change of keeping a human in the loop.

3.2 Evaluative Utility: Defining Novelty

To enable autonomous discovery, we must operationalize the concept of "scientific contribution." In SEED, novelty is not a subjective judgment but a computable distance in the topological design space. We define the *Total Distance* between a proposed experiment G_{new} and an existing reference study G_{ref} as a weighted linear combination of their structural and parametric divergence:

$$D(G_{new}, G_{ref}) = w_s \cdot \delta_{struct}(G_{new}, G_{ref}) + w_p \cdot \delta_{param}(G_{new}, G_{ref}). \tag{1}$$

Here, w_s and w_p are hyperparameters for the sensitivity of the research inquiry. Setting $w_s \gg w_p$ configures the system to value mechanism discovery (finding new topologies), whereas setting $w_p > w_s$ configures it for generalizability testing (applying known structures to new contexts).

1. Structural Distance (δ_{struct}): The Weighted Graph Edit Distance. We quantify architectural innovation using a weighted *Graph Edit Distance* (GED), defined as the minimum cost of elementary operations required to transform the topology of G_{new} into G_{ref} (Ranjan et al., 2022; Jain et al., 2024). We impose a potential cost hierarchy to reflect scientific significance, strictly distinguishing between qualitative shifts and quantitative tuning:

- *Topological Operations (High Cost:* Operations that alter the skeleton (V, E), such as node insertion or authority reversal. These are penalized heavily because they fundamentally alter the locus of control, representing the discovery of a distinct mechanism class.
- *Semantic Operations (Low Cost:* Operations that alter the labels or attributes, such as toggling an incentive or enabling identity disclosure. These are assigned lower costs as they represent parametric refinements within the same topological family.

2. Parametric Distance (δ_{param}): The Context Mapping. If the topology is identical, novelty must be sought in the semantic context. A core tenet of SEED is that disparate domains often share identical decision-theoretic properties. To measure this, we map textual domain descriptions into normalized feature vectors $\mathbf{v}$ (e.g., [Stakes, Symmetry, Urgency]).

For example, consider two studies: *Bank Loan Approval* and *University Admission*. While semantically distinct, SEED maps both to the vector `[High Stakes, Asymmetric Info, Binary Choice]`. Consequently, $\delta_{param} \approx 0$. This metric prevents the agent from "discovering" the same mechanism twice by simply changing the cover story; true parametric novelty requires traversing the vector space to a functionally distinct region.

The Novelty Score ($\mathcal{N}$). Finally, we define the scalar *Novelty Score* of a proposed design relative to the entire body of prior knowledge. Let $\mathcal{L} = \{G_1, \ldots, G_n\}$ represent the set of all known experimental designs in the literature (the Topology Library). The novelty of a new design is its distance to its *nearest neighbor* in $\mathcal{L}$:

$$\mathcal{N}(G_{new}|\mathcal{L}) = \min_{G_\ell \in \mathcal{L}} D(G_{new}, G_\ell). \tag{2}$$

This formulation provides the explicit objective function for the Meta-Optimization Wrapper. Rather than relying on random generation, the agentic researcher solves a constrained maximization problem over the topological design space $\mathbb{G}$ with certain complexity penalty $\mathcal{C}(G)$:

$$G^* = \operatorname*{argmax}_{G \in \mathbb{G}} \left[\mathcal{N}(G|\mathcal{L}) - \lambda \cdot \mathcal{C}(G)\right]. \tag{3}$$

3.3 Generative Utility: The Agentic Discovery Loop

We now operationalize the SEED framework by defining the *Generative Topology Search* algorithm. This enables the transition from the static evaluation of novelty to the active discovery of experimental designs. The process functions as a "human-in-the-loop" search engine, where the researcher defines the starting region, and the agent navigates the topological boundaries. As illustrated in Figure 1, this navigation is governed by a structured prompting protocol that forces the agent to interact with the external Topology Library and Solver tools rather than relying on unconstrained generation.

Phase 1: Knowledge Initialization (The Pre-Computed Prior). To ground the agent, we first construct the *Topology Library* ($\mathcal{L}$). This is a pre-computed vector database derived from the existing literature. An ingestion agent parses prior studies, discarding surface-level narratives to extract the underlying runtime execution graph G_{paper}. This transforms the messy history of the field into a structured map, $\mathcal{M}$, where every node is a known experimental design. Note that this step is performed once; the resulting library $\mathcal{L}$ serves as the immutable "ground truth" against which all new ideas are measured (Lewis et al., 2020; Ma et al., 2023).

Phase 2: Constrained Neighbor Generation (The Explorer). The search for new designs is not an unconstrained random walk, which would yield valid but irrelevant structures. Instead, it is a directional search initiated by researcher constraints $\mathcal{C}$ (e.g., Target Domain = "Medical Triage", Focus = "Delegation Mechanisms").

- **Parent Selection:** Agent retrieves the set of "Parent Graphs" $\{G_{parent}\} \subset \mathcal{L}$ that satisfy constraints $\mathcal{C}$.
- **Mutation:** Agent applies the editable operators (defined in Section 2.2) to these parents to generate a candidate set $\mathcal{S}_{cand}$. For example, if the parent is a standard "AI Advice" graph ($\bigcirc \rightarrow \Delta$), the agent explores immediate neighbors: adding a loop ($\bigcirc \leftrightarrow \Delta$), adding a monitor ($\bigcirc \xrightarrow{M} \Delta$), or parameter scaling ($n = 1 \rightarrow n = 10$).
- **Result:** This produces a set of plausible candidates, which are experimental designs that are structurally rooted in proven science but contain a specific mutation.

Phase 3: Candidate Ranking & Human Selection (The Solver). The agent must now filter the "trivial" mutations (e.g., $n = 1 \rightarrow n = 2$) from the "scientific" ones. We formulate this as a Pareto-Optimization problem. The agent first scores each candidate $G' \in \mathcal{S}_{cand}$ using a composite utility function:

$$U(G') = \alpha \cdot \text{Novelty}(G'|\mathcal{L}) + \beta \cdot \text{Coherence}(G'), \tag{4}$$

where novelty is derived from Section 3.2. High novelty rewards structural mutations (e.g., authority reversal). Coherence is an LLM-based heuristic score that evaluates if the mutation makes sense within the researcher's context (e.g., "Is 10 rounds of feedback realistic for a surgeon?").

Finally, the system outputs the top-k candidates that maximize this utility. Crucially, the human researcher serves as the final gatekeeper. Rather than automatically deploying these designs, the system presents the ranked proposals to the expert, who judges their practical feasibility and ethical implications before instantiation.

System Prompt: Topological Discovery Agent

Role: You are an Expert Mechanism Designer utilizing the SEED framework.
External Tools Available:

- `Library.retrieve(`$\mathcal{L}$`)`: Access to the vector database of prior studies.
- `Solver.evaluate(`G'`)`: Utility function calculating Novelty and Coherence.

Context: You are analyzing the parent graph "Standard AI Advice" ($\bigcirc \rightarrow \Delta$) in the domain of [Medical Triage].
Task: Query the Library for neighbors, then generate 3 adjacent experimental designs by applying Atomic Mutations.

Constraints: 1. *Structural:* Apply exactly ONE topological edit (e.g., Reverse Edge, Add Loop, Mask Info). 2. *Parametric:* The setting must remain [High Stakes, Time Pressure]. 3. *Objective:* Maximize the Novelty Score returned by the Solver tool.

Output Format: 1. `Mutation_Type`: [e.g., Authority Reversal] 2. `New_Topology`: [SEED Notation, e.g., $\Delta \xrightarrow{D} \bigcirc$] 3. `Rationale`: Why does the Solver rate this as a high-value structural gap?

Figure 1: An Illustrative Generative Prompt Template (for Phase 2).

4 Discussions and Conclusion

SEED establishes a fundamental bridge between the scientific necessity of understanding AI systems and the methodological innovation required to study them. While *Experiments for Agents* serve as the critical engine for mechanism discovery, revealing the invisible rules of governance in complex systems, it is the paradigm of *Agents for Experiment* that transforms this process from an artisanal craft into a scalable science. By formalizing the topological design space, SEED provides the computational scaffolding for this shift, ensuring that our methods of inquiry evolve as rapidly as the agentic systems we seek to understand.

From Strategies to Game Forms. Current research is often trapped in a local minimum of parameter tuning, optimizing prompts or model weights within fixed interaction loops. SEED elevates the abstraction level from *parameters* to *architectures*. In the language of mechanism design, this enables agents to search the space of game forms, altering the information sets, authority flows, and sequence of moves, rather than merely optimizing strategies within a fixed game. This capability is critical for uncovering structural gaps: counter-intuitive governance structures (e.g., reverse-veto or asymmetric oversight) that human designers, biased by convention, frequently overlook.

Toward a Computable Science of Interaction. A persistent barrier to cumulative knowledge is the lack of a unified grammar: identical mechanisms are often obscured by disparate terminology across disciplines, from economics to computer science. SEED resolves this "Tower of Babel" by standardizing experiments into a graph-theoretic format. This transforms the literature from a collection of fragmented prose into a structured database, enabling automated meta-analysis. Future agentic systems could query the global topology library not just for text matches, but for structural invariants, answering high-level questions such as: *"In which topological configurations does system robustness consistently decouple from individual agent accuracy?"*

The Shift from Architect to Governor. The power to automatically generate mechanisms introduces new responsibilities. An agent maximizing novelty might propose operationally efficient but ethically hazardous topologies (e.g., removing human oversight in high-stakes loops). Consequently, the role of the human researcher must shift from *Architect* (generating ideas) to *Governor* (defining constraints) (Sun et al., 2023). Future work must focus on integrating formal verification into the SEED solver, ensuring that all agent-generated topologies satisfy mathematically defined safety properties before instantiation.

Ultimately, this framework lays the groundwork for a future where human experts and agentic architects collaborate to explore the vast, unseen topology of strategic mechanisms. By treating experiments not as loose prose but as computable graphs, we accelerate the pace of discovery, moving closer to a science that can keep pace with the rapid evolution of AI itself.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. ICML 2017.

Rohan Alur, Manish Raghavan, and Devavrat Shah. Human expertise in algorithmic prediction. In *NeurIPS*, 2024. doi: 10.52202/079017-4384.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, 2019. CHI 2019.

LP Argyle, CA Bail, EC Busby, JR Gubler, T Howe, C Rytting, T Sorensen, and D Wingate. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. In *Proceedings of the National Academy of Sciences of the United States of America*, 2023.

Kirk Bansak and Elisabeth Paulson. Public attitudes to artificial intelligence taking over high-impact decisions. *Proceedings of the National Academy of Sciences of the United States of America*, 2024. doi: 10.1073/pnas.2404201121.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2019. HCOMP 2019.

RC Bayer and L Renou. Interacting with man or machine: When do humans reason better? *Management Science*, 2024.

M Binz, X Qian, W Zhao, Z Zhang, J Chen, J Song, Z Du, and M Bansal. Controllable and interpretable retrieval augmentation for large language models. *Nature*, 2025. doi: 10.1038/s41586-025-08681-0.

E Brynjolfsson, D Li, and LR Raymond. Generative ai at work. *National Bureau of Economic Research*, 2025.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. Large language models as tool makers. In *International Conference on Learning Representations (ICLR)*, 2024.

A Caplin, D Deming, and A Kapor. The abcs of who benefits from working with ai: Evidence from a controlled experiment. *Management Science*, 2025.

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024. FAccT 2024.

Z Chen and J Chan. Large language model in creative work: The role of collaboration modality and user expertise. *Management Science*, 2025.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. NeurIPS 2017.

A De Toni and et al. The double-edged roles of generative ai in the creative process. *Information Systems Research*, 2024.

AR Dennis, A Lakhiwal, and A Sachdeva. Ai agents as team members: Effects on satisfaction, conflict, trustworthiness, and willingness to work with. *Journal of Management Information Systems*, 2023.

Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni. Ai companions reduce loneliness. *Journal of Consumer Research*, 2025. doi: 10.1093/jcr/ucaf040. Advance article.

A Fügener, J Grahl, A Gupta, and W Ketter. Will humans in the loop become more cooperative? *Journal of Management Information Systems*, 2021.

A Fügener, J Grahl, A Gupta, and W Ketter. Cognitive challenges in human-artificial intelligence delegation. *Information Systems Research*, 2022.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. ICML 2019.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured nlp tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. EMNLP 2023.

D Glickman, S Goel, K Heller, and J Kleinberg. How human–ai collaboration changes decision making over time. *Nature*, 2025.

U Gnewuch, S Morana, and M Adam. More than a bot? the impact of disclosing human involvement in hybrid service agents. *Information Systems Research*, 2024.

E Goh, RJ Gallo, K Brizzi, E Ferrara, and et al. Gpt-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nature Medicine*, 2025.

C Gonzalez, B Mellers, and P Tetlock. A cognitive approach to human–ai interactive forecasting. *Nature*, 2025.

J Han, Y Tan, and Y Zhang. The effect of ai disclosure on user trust and reliance. *Management Science*, 2023.

Y Hou, Z Wang, and Y Zhang. Overreliance on algorithmic advice: Evidence and interventions. *Management Science*, 2021.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Eeshaan Jain, Indradyumna Roy, Saswat Meher, Soumen Chakrabarti, and Abir De. Graph edit distance with general costs using neural set divergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. NeurIPS 2024.

Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2023. PACM HCI (CSCW) 2023.

Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, 2019. CHI 2019.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013. KDD 2013.

S Krakowski, D Haftor, J Luger, and et al. Human-centered artificial intelligence: A field experiment. *Management Science*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. NeurIPS 2020.

X Li, M Yang, and T Sun. The impact of ai advice and framing on human decisions. *Information Systems Research*, 2024.

B Liel and et al. Gpt-4 assistance for clinical reasoning: Evidence from controlled experiments. *Management Science*, 2025.

H Lin, AJ Berinsky, DG Rand, and et al. Decision-oriented dialogue with human–ai systems. *Nature Human Behaviour*, 2024.

Junming Liu, Wei Thoo Yue, Alvin Chung Man Leung, and et al. Find the good. seek the unity: A hidden markov model of human–ai delegation dynamics. *MIS Quarterly*, 2025.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5303–5315, 2023. doi: 10.18653/v1/2023.emnlp-main.322.

Sanju Menon, Weiyu Zhang, and Simon T. Perrault. Nudge for deliberativeness: How interface features influence online discourse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)*, 2020. CHI 2020.

Julien Meyer, April Khademi, Bernard Têtu, and et al. Impact of artificial intelligence on pathologists' decisions: an experiment. *Journal of the American Medical Informatics Association*, 2022. doi: 10.1093/jamia/ocac103.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. ICML 2020.

Dipesh Niraula, Shreyas Kannan, Andrew Qian, and et al. Intricacies of human–ai interaction in dynamic settings. *Nature Communications*, 2025.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. NeurIPS 2024.

Rishabh Ranjan, Siddharth Grover, Sourav Medya, Venkat Chakravarthy, Yogish Sabharwal, and Sayan Ranu. Greed: A neural framework for learning graph distance functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. NeurIPS 2022.

A Reis and et al. Who is the advisor? identity and performance disclosure in human–ai advice taking. *Management Science*, 2024.

Elena Revilla, María Jesús Saenz, Matthias Seifert, and Ye Ma. Human–artificial intelligence collaboration in prediction: A field experiment in the retail industry. *Journal of Management Information Systems*, 2023. doi: 10.1080/07421222.2023.2267317.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016. KDD 2016.

S Schanke, G Burtch, Y Hong, and et al. Digital lyrebirds: Experimental evidence that deepfakes of political candidates increase online political incivility. *Management Science*, 2024.

A Schecter and et al. Ai assistance and human decision quality: Evidence from high-stakes tasks. *Management Science*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. NeurIPS 2023.

Weiyan Shi, Xeuwei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. Effects of persuasive dialogues: Testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI)*, 2020. CHI 2020.

D Shin and et al. Personality matters: How agent personality shapes human reliance and performance. *Information Systems Research*, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. NeurIPS 2023.

D Siemon, E Elshan, and et al. Beyond anthropomorphism: Social presence in human–ai interaction. *Journal of Management Studies*, 2025.

YW Song, TT Yan, F Jia, LJ Chen, and H Li. Developing generative ai for value co-creation: An intervention-based randomized field experiment in a healthcare context. *Journal of Operations Management*, 2025.

Marta Stelmaszak, Mareike Möhlmann, and Carsten Sørensen. When algorithms delegate to humans. *MIS Quarterly*, 2025.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. NeurIPS 2023.

D Treiman and et al. The consequences of iterative human–ai collaboration. *Management Science*, 2024.

M Von Zahn, S Liebich, and et al. Knowing (not) to know: Explainable artificial intelligence and metacognition. *Information Systems Research*, 2025.

PX Wan, ZG Huang, and et al. Outpatient reception via collaboration between humans and ai: Evidence from a field deployment. *Nature Medicine*, 2024.

Lingli Wang, Ni Huang, Yumei He, De Liu, Xunhu Guo, and et al. Artificial intelligence (ai) assistant in online shopping: A randomized field experiment on a livestream selling platform. *Information Systems Research*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023. OpenReview: 1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. NeurIPS 2022.

Jason Xianghua Wu, Yan Diana Wu, Kay-Yut Chen, and Lei Hua. Building socially intelligent ai systems: Evidence from the trust game using artificial agents with deep learning. *Management Science*, 2023. doi: 10.1287/mnsc.2023.4782.

Yuqian Xu, Hongyan Dai, and Wanfeng Yan. Identity disclosure and anthropomorphism in voice chatbot design: A field experiment. *Management Science*, 2024. doi: 10.1287/mnsc.2023.4781. Articles in Advance.

Cathy Yang, Kevin Bauer, Xitong Li, and Oliver Hinz. My advisor, her ai, and me: evidence from a field experiment on human–ai collaboration and investment decisions. *Management Science*, 72 (1):242–264, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. NeurIPS 2023.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI)*, 2019. CHI 2019.

Sangseok You, Cathy Liu Yang, and Xitong Li. Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems*, 2022. doi: 10.1080/07421222.2022.2063553.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020. FAT* 2020.