Simulating Society Requires Simulating Thought

Chance Jiajie Li[1*], Jiayi Wu[9*], Zhenze Mo[8], Ao Qu[4], Yuhan Tang[5], Kaiya Ivy Zhao[2,3], Yulu Gan[2], Jie Fan[2,7†], Jiangbo Yu[10], Jinhua Zhao[4,5,6], Paul Pu Liang[1,2], Luis Alonso[1], Kent Larson[1]

[1]MIT Media Lab [2]MIT EECS [3]MIT BCS [4]MIT IDSS [5]MIT CEE [6]MIT DUSP
[7]MIT Architecture [8]Northeastern University [9]Brown University [10]McGill University

[*]Equal contribution.

[†]Now at Google.

Correspondence to: `jiajie@mit.edu`

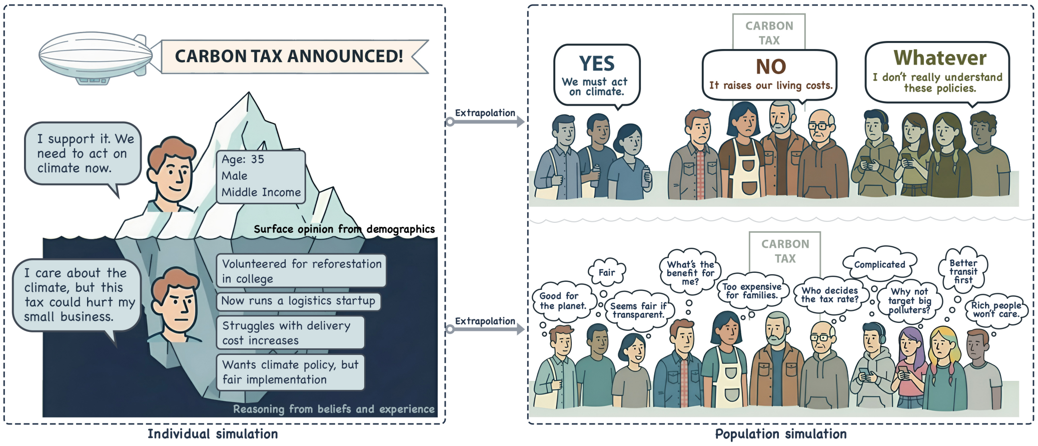
CARBON TAX ANNOUNCED!
I support it. We need to act on climate now.
Age: 35
Male
Middle Income
Surface opinion from demographics
I care about the climate, but this tax could hurt my small business.
Volunteered for reforestation in college
Now runs a logistics startup
Struggles with delivery cost increases
Wants climate policy, but fair implementation
Reasoning from beliefs and experience
Individual simulation
Extrapolation
Extrapolation
CARBON TAX
YES We must act on climate.
NO It raises our living costs.
Whatever I don't really understand these policies.
CARBON TAX
Good for the planet.
Fair
Seems fair if transparent.
What's the benefit for me?
Too expensive for families.
Who decides the tax rate?
Complicated
Why not target big polluters?
Better transit first
Rich people won't care.
Population simulation

Figure 1: **From surface imitation to cognitively grounded social simulation.** Current LLM-based simulations (top left) capture only *surface opinions*, shaped by demographics or language patterns, while the deeper *belief formation processes* remain unmodeled (bottom left, beneath the waterline). This yields population-level simulations that are *flattened and stereotyped*, reflecting aggregated personas rather than genuine diversity. In contrast, cognitively grounded reasoning (bottom right) models the latent belief dynamics behind individual decisions, producing collective patterns that are heterogeneous, interpretable, and causally faithful.

Abstract

Simulating society with large language models (LLMs), we argue, requires more than generating plausible behavior; it demands cognitively grounded reasoning that is structured, revisable, and traceable. LLM-based agents are increasingly used to simulate individual and group behavior, primarily through prompting and supervised fine-tuning. Yet current simulations remain grounded in a behaviorist *"demographics in, behavior out"* paradigm, focusing on surface-level plausibility. As a result, they often lack internal coherence, causal reasoning, and belief traceability, which makes them unreliable for modeling how people reason, deliberate, and respond to interventions.

To address this, we present a **conceptual modeling paradigm**, **Generative Minds (GenMinds)**, which draws from cognitive science to support structured belief representations in generative agents. To evaluate such agents, we introduce the **RECAP** (*REconstructing CAusal Paths*) framework, a benchmark designed to assess reasoning fidelity via causal traceability, demographic grounding, and interven-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Position Paper Track.

tion consistency. These contributions advance a broader shift: from surface-level mimicry to generative agents that simulate thought—not just language—for social simulations.

1 Introduction

The Rise of LLMs in Social Simulation. Over the past two years, LLMs have increasingly become dominant tools for simulating human behavior across language [1, 2], vision [3] and decision-making domains [4].In the field of social simulation, LLMs are now commonly used to emulate public opinions, stakeholder interactions, and policy responses under diverse scenarios [5, 6, 7].

An Oversimplified Paradigm: Demographics In, Behavior Out. Despite the growing use of LLMs in social simulation, most current models rely on simplified input-output mappings, producing behavior based on surface cues rather than simulating the internal belief dynamics behind decisions. This approach mirrors the logic of *behaviorism* in psychology, which models behavior as a function of external stimuli while ignoring internal cognitive states. The limitations of this paradigm echo a broader historical tension between *behaviorism*, *cognitivism*, and *constructivism*[8, 9]: while *cognitivism* emphasized structured internal representations and causal reasoning, and *constructivism* further argued that beliefs are continually shaped by individual and social experience, existing LLM-based agents remain far from either. They typically exhibit shallow reasoning, frequent hallucinations, and limited understanding of causal and contextual dynamics in socially-salient domains such as upzoning, surveillance, or healthcare access, precisely the domains where reasoning fidelity matters most [10, 11].

$$\textit{Behaviorism} \longrightarrow \textit{Cognitivism} \longrightarrow \textit{Constructivism}$$

Structural Failures: Modeling, Evaluation, and Calibration These failures stem directly from the behaviorist paradigm outlined above. By focusing on surface behavior instead of the reasoning behind it, most LLM-based simulations face fundamental limitations in both modeling and evaluation.

In modeling, agents often rely on shallow input-output patterns, without representing how beliefs are formed, updated, or justified. As a result, their internal reasoning is difficult to inspect, especially when context changes. It is hard to determine how a new policy or scenario influences an agent's judgment, or why a particular decision is made. Without access to reasoning traces, agents cannot support diagnostic explanation, causal attribution, or meaningful intervention — all of which are critical for multi-stakeholder policy simulations. Even when models succeed at surface-level generation, they are difficult to adapt to new domains. Fine-tuning LLMs for specific contexts often requires significant compute and high-quality datasets, which are rarely available in real-world policy settings. Yet effective simulation depends on exactly the opposite: the ability to represent evolving stakeholder reasoning grounded in timely, localized information [12, 13].

In evaluation, models are typically judged by output plausibility or alignment with population-level trends[14], but such metrics say little about whether their reasoning is accurate, flexible, or aligned with how people actually think. Post-hoc output analysis is common, but it cannot substitute for reasoning-level evaluation. Aligning agents with real-world stakeholders requires individual-level data and internal benchmarks for reasoning fidelity, both of which are largely missing today.

Toward Mechanistic and Individual-Level Alignment These limitations call for a shift in how we conceptualize generative social simulation—not as behavior mimicry, but as cognitive modeling. This paper takes up that call. Specifically, we propose leveraging ideas from Theory of Mind (ToM) and cognitive science to extract and simulate reusable, executable reasoning units—what we term **reasoning traces**—rather than simply mimicking human tone or persona [15, 16].

Unlike prompt-driven persona or character approaches that generate "average" group behaviors [17], cognitive models allow agents to represent beliefs, values, and causal assumptions in a compositional manner. This makes it possible to generalize to unseen scenarios, so long as the individual components of the reasoning trace are known. For example, if a stakeholder has previously reasoned about "density" and "transit," then when asked about a novel "transit-oriented development" policy, the agent can reuse those motifs to simulate beliefs without re-training.

Such compositionality is a cornerstone of human cognition, where reasoning emerges from fragments that are reusable, revisable, and structured across contexts [18, 19]. It also improves simulation fidelity: interactions between agents, or between agents and dynamic environments, can be represented through composable and transparent reasoning structures, enabling structured simulation at both micro and macro levels [20]. Moreover, reusing compositional and modular reasoning units reduces the need to regenerate full-context reasoning at every step, improving both interpretability and computational efficiency.

Position and Vision **In this paper, we advocate moving beyond output-level alignment toward aligning the internal reasoning traces of generative agents.** Capturing the causal, compositional, and revisable structure of belief formation, which we refer to as reasoning fidelity, is essential for building cognitively faithful agents that simulate not only what people say but also how they think.

To support this argument, we:

- Illustrate how current approaches fall short by producing outputs that appear coherent but lack internal consistency, adaptability, or traceability;
- Theorize reasoning fidelity as a structural alignment problem grounded in cognitive science;
- Introduce a symbolic-neural framework for simulating belief formation through modular reasoning motifs and causal graphs;
- Present a methodology for extracting and simulating belief structures from natural language, enabling interpretability, counterfactual reasoning, and domain transfer.

In summary, this position paper argues that simulating human society requires more than generating plausible conversations. It requires **simulating the structure of human reasoning**. By grounding agents in modular belief representations and evaluating them on reasoning fidelity, we take a critical step toward building generative minds, not just generative outputs.

2 Social Simulation: Opportunities and Gaps

Social simulation has emerged as a high-impact use case for LLMs [6]. Traditionally, social science relies on surveys, experiments, fieldwork, or game-theoretic models to understand individual and group behavior [21]. While effective, these methods are expensive, hard to scale, and often face ethical and logistical challenges. The rise of LLM-driven agents offers a promising alternative capable of simulating human responses across a wide range of scenarios, roles, and interventions. Therefore, LLM-agent-based social simulation has been increasingly implemented in domains including policy modeling [22], behavior forecasting [23], annotation tasks [24], and opinion surveys[25].

Recent studies demonstrate that LLMs can emulate key aspects of human reasoning and decision-making [26, 27, 28, 29], enabling agents that perceive their environment, make context-sensitive decisions, and articulate motivations. When equipped with role-based prompting or persona conditioning [30, 31], these agents exhibit a property known as *algorithmic fidelity*—the ability to simulate how specific individuals or subgroups might respond in a given situation [25, 32].

Beyond single-agent settings, multi-agent simulations extend this potential by modeling the interactions, conflicts, and consensus dynamics among synthetic populations. However, faithfully simulating social processes introduces several critical requirements:

1. **Fidelity**: Simulated agents should be faithful in respect to individual human reasoning [33, 34].
2. **Individuality**: Simulations should preserve individual-level heterogeneity [33, 35], capturing the positional and contextual diversity of human reasoning.
3. **Extrapolation**: A core challenge is out-of-distribution generalization. Agents must reason in novel scenarios and generalize from skewed population subsets to broader groups [34, 35], requiring belief updates and counterfactual reasoning under uncertainty.

While many recent works aim to create more "human-like" agents, few clearly define what this term means, or how such fidelity is to be evaluated. As a result, although large-scale agent environments now enable rapid, low-cost exploration of collective behavior, due to the lack of internal transparency and explainability of language models today [36], most simulations remain behaviorist at their core, thus simulated societies risk reflecting the biases and homogeneity of the prompts rather than the heterogeneity of human internal belief structures. These methodological and epistemic gaps motivate a shift toward cognitively grounded simulation.

3 Problem Statement: Beyond Behavioral Plausibility in Generative Social Simulation

Most existing efforts to align LLM-based agents focus on model *behavior* [29]: Do agents take stances, express preferences, or engage in natural-sounding conversations as human subjects they aim to imitate? This behavior-centric view is reinforced by popular techniques such as reinforcement learning from human feedback (RLHF) [37, 38], persona prompting [39], and chain-of-thought (CoT) generation [26, 40]. These methods optimize for behavioral *plausibility* rather than structural fidelity of human reasoning.

The behavioral-centric framework overlooks a critical problem: output plausibility is not equivalent to cognitive alignment. In this section, we argue that behavioral fluency fails to serve as a proxy of agents' reasoning fidelity. Without structural representations of belief, sensitivity to counterfactual intervention, or positional diversity of individual human subjects, generative agents risk producing surface-level plausible outputs that are epistemically unaligned. We identify two core challenges of the current agent simulation paradigm: (1) **fidelity**, the agent's capacity for coherent, revisable, and causally grounded belief formation and (2) **individuality**, the agent's ability to model distributed and positional human reasoning, and show how both remain underspecified in existing evaluation **metrics**.

3.1 Fidelity: Coherence, Traceability, and Causal Grounding

One crucial aspect of social simulation is that it must ensure agents think in ways that are causally structured, internally coherent, and dynamically revisable. Current LLM-based agents fail to meet these standards, and this mismatch between surface-level plausibility and internal structural fidelity manifests in a set of persisting reasoning failures.

3.1.1 Traceability and Interpretability

Current LLM-based agents presents two essential gaps in reasoning fidelity: (1) *decoding faithfulness mismatch*, where generated reasoning traces in agent's output generated traces diverge from the model's internal computational path; (2) *cognitive-alignment mismatch*, where the modeled inference path diverges from human belief formation. Subsequently, LLM-based agents face *intervention-invariance mismatch*, where belief updates fail under counterfactual perturbations.

The first critique is largely similar to those observed in CoT-related discussions [41, 42]: agents may produce fluent rationalizations, while their "reasoning traces" are constructed post hoc: assembled from language patterns rather than derived from an underlying belief model [43, 44]. Recent controlled studies confirm that the existing framework of LLM reasoning is largely a data mirage: Zhao et al. show that CoT performance collapses when tasks, the length of reasoning chain, or prompt format deviate moderately from the model's training-set distribution [45]. This finding reinforces that current CoT outputs are not a faithful representation of the model's actual reasoning processes, as they lack stability under distributional shift and therefore fail to reveal the causal decision paths or the underlying dependency assumptions guiding them. While there are some promising approaches beginning to incorporate structured representations, such as knowledge graphs, belief graphs, and additional reasoning layers [46, 47, 48], most of them are domain-specific and disconnected from live generative processes; also, none have yet been operationalized within interactive, generative social simulation pipelines.

The second critique concerns the presumed alignment between model reasoning and human reasoning, which remains largely untested. Although emerging proposals in cognitive science examines LLMs through theory-of-mind and belief-attribution paradigms [49], these studies largely evaluate behavioral correlates (e.g., predicting others' beliefs or heuristics) rather than mapping internal reasoning operators to human causal schemas. Moreover, no standardized benchmark operationalizes measurable correspondences between model-internal belief transitions and human causal inference sequences. Therefore, claims of "human-like reasoning" remain more speculative than empirically grounded, lacking any quantitative measure of cognitive fidelity.

Alternative View: Post-hoc rationalization may be cognitively authentic and functionally sufficient. Drawing from work in social and cognitive psychology, one might argue that people often rationalize decisions or beliefs after the fact [50, 51]. LLMs' tendency to construct rationales retroactively might therefore be seen not as a defect, but as a cognitive parallel to human behaviors.
Response. We do not target post-hoc rationalization per se, but its total detachment from structured belief representation. Human justifications are often imperfect but still rely on internal models of causality, memory, and values and are traceable thereafter [52], whereas LLMs produce rationalizations without structured anchoring. *Form* (i.e. the shape of the rationale) is not the same as *function* (i.e. the structural, belief-guided deliberation). Agents must operate on explicit, traceable structures to simulate human reasoning.

3.1.2 Counterfactual Intervention Sensitivity and Belief Revision

In social simulations, agents are expected to revise their stances when key assumptions or contextual conditions change, which is a hallmark of human reasoning known as counterfactual intervention sensitivity [53, 54]. However, due to the lack of an internal causal structure that anchors beliefs to causes and consequences [11, 55], current LLMs and LLM-based generative agents often respond to such interventions with inertia or token-level paraphrasing. As a result, they cannot explain why a particular belief might hold under some conditions but not others, nor can they simulate the effects of counterfactual changes.

This structural deficiency manifests as inconsistency across prompts or dialogue turns: empirical studies have shown that LLM-based agents may support one policy in one scenario, then oppose in another without any causal reasoning-trace grounding [56, 57]. While there have been research efforts in grounding models and model-based agents with causal memory [58] and knowledge graph [59], most of them focus on graph discovery and construction [60, 61] in specific knowledge domains rather than general human belief systems. Without an explicit model of how beliefs are formed, revised, and connected, agents' utterances are generated in isolation—locally plausible, but globally incoherent.

Alternative View: Human cognition is non-monotonic and contextually fluid, thus demanding coherence is unrealistic. One might argue that humans often hold incoherent or even contradictory beliefs. Demanding that agents simulate perfectly consistent beliefs risks idealizing cognition and misrepresenting actual human messiness [62, 63].
Response. We do not call for rigid logical coherence or monotonic reasoning. Rather, our claim is modest: agents should be able to faithfully simulate belief revision under counterfactual assumptions—not that they maintain perfect consistency across all scenarios. Furthermore, human belief systems can be messy and involve incoherent or even contradictory stances, but it doesn't mean they're structureless. In human cognition, contradictions are often meaningful, reflecting a set of ambivalent conventions and priors in the social system at large [64]; in contrast, LLMs generate contradictions without memory, deliberation, or causal record. We don't require logical perfection but rather grounded incoherence, that agents should be able to simulate how humans arrive at contradictory views and under what conditions those contradictions persist or resolve.

Taken together, these observations define reasoning fidelity as the preservation of structured belief dynamics under decoding transparency, cognitive alignment, and counterfactual intervention sensitivity. Current autoregressive architectures optimize next-token likelihood rather than belief-state transitions, making fidelity an architectural limitation rather than a prompting issue [65]. Until this gap is addressed, generative agents will continue to exhibit the symptoms of alignment while remaining fundamentally unaligned at the level of thought.

3.2 Individuality: Heterogeneity and Positional Reasoning in Social Simulation

Alongside reasoning fidelity, social simulation requires an additional layer of alignment: positional individuality. In particular, we define individuality as the preservation of heterogeneity in agents' latent belief and value representations under shared generative priors. Current LLM-based agents, optimized under shared autoregressive parameters and global priors, collapse toward the mean of the pretraining distribution, erasing structured heterogeneity.

When deployed in real-world contexts, such as civic simulations [66, 67], participatory policy design [68, 69], stakeholder modeling [70], agents that lack internal reasoning fidelity may produce outputs that appear thoughtful, yet encode no coherent decision process beneath the text. This disconnect introduces a set of critical downstream failures:

3.2.1 Illusion of Consensus in Multi-Agent Systems

One critical downstream failure of LLM-based agent social simulation is the illusion of consensus. Recent studies demonstrate that LLMs in multi-agent setups exhibit conformity behavior, and the benchmark shows virtually all models converge in behavior under majority-pressure protocols [71].

This convergence reflects a deeper statistical mechanism: models trained to minimize token-level loss implicitly learn to average across pre-training data distributions, biasing their conditional likelihoods toward high-frequency, socially moderate continuations [72, 73]. When multiple model-based agents interact, this produces a form of *synthetic agreement* aligned with a median perspective that masks underlying conflict, complexity, and epistemic independence [74, 75]. In multi-agent policy or deliberation simulations, this can yield systematically misleading inferences: agents appear to "agree" on a position not because of shared reasoning, but because their generative priors and cross-conditioning push them toward a median narrative that suppresses disagreement and complexity.

3.2.2 Flattened Outputs in Demographic Conditioning

A related but more insidious form of convergence occurs within demographic conditioning. Since LLMs are trained on aggregated corpora and lack explicit conditioning on intersecting social variables, their generative prior implicitly factorizes across dimensions such as race, class,a nd gender. This yields identity flattening, that agents reproduce majority-class correlations that dominate pre-training statistics, producing homogenous or stereotypical portrayals of social groups [76, 77], with resulting responses minimizing token-level loss but erasing intersectional variation. This leads to epistemic harm: the rich, positional knowledge of real-world stakeholders is replaced with monolithic, decontextualized simulations.

Alternative View: Generalization over identity categories is necessary for tractable simulation. One might argue that abstractions over demographic identities are unavoidable and, in fact, desirable when building simulators at scale considering model tractability [78]. Identity flattening may be viewed as a form of necessary regularization.
Response. Our critique is not about the use of abstraction per se, but the fact that LLMs abstract without modeling the joint distribution of beliefs, values, and positionality conditioned on intersecting variables (e.g., age × race × class × institutional exposure) and how abstraction, subsequently, is operationalized without epistemic grounding [76]. In social simulations, such abstractions introduce bias, undermine group heterogeneity, and lack epistemic representativeness, especially when the simulation output is used to inform policy or governance decisions [79].

As agents are increasingly used to test policy options, simulate deliberative processes, or represent groups in synthetic social systems, these reasoning deficiencies risk being institutionalized. Decision-makers may take model outputs at face value, unaware that these outputs do not derive from any concrete belief structure. Ultimately, when agents simulate without reasoning, the outputs they generate can erode trust, misinform policy, and flatten the epistemic diversity of the very populations they are meant to represent [76, 80, 81, 82].

3.3 Evaluation Gaps: Structural Limitations of Current Benchmarks

The deficiencies in reasoning fidelity and individuality are compounded by a third layer: evaluation misalignment. Despite the growing sophistication of generative agents, current benchmarks remain optimized for stylistic fluency, local coherence, and plausibility of individual model behavior and output. Most benchmarks treat language as a proxy of thought, implicitly assuming that coherent expression entails coherent reasoning. As a result, current evaluations risk rewarding surface-level coherence while overlooking causal consistency and belief heterogeneity [71]. This first creates what we term as *traceability gap*: benchmarks assess outputs instead of reasoning trajectories. Stance classification tasks, for example, check whether a model picks a side but remain agnostic about how or why that stance was formed [83, 84]. Dialogue benchmarks reward conversational smoothness, even when agents flip positions over time [85, 86, 87].

The second limitation concerns *intervention blindness*: most benchmarks assess agents on static inputs but fail to measure their belief revision under counterfactual interventions or other hypothetical perturbations [88, 89, 90, 91, 92]. Finally, current benchmarks leave agents' positional individuality largely unaddressed. Evaluation suites typically aggregates performance across stances, computing mean accuracy or sentiment agreement but rarely quantify inter-agent divergence or distributional variance. Recent work on pluralistic alignment has begun moving in this direction, measuring whether models maintain epistemic diversity or support multiple internally coherent responses [93, 94, 95]; yet these methods remain limited to small-scale reasoning or dialogue tasks and have not been adopted within social simulation pipelines, where evaluation still focuses on output fluency and stance accuracy.

In sum, evaluation protocols for generative agents remain behaviorally calibrated but structurally ungrounded. Bridging this gap requires a new generation of benchmarks that treat reasoning as a structured process rather than a stylistic performance.

4 What Does Human-Like Reasoning Entail?

The failures outlined above stem from a core mismatch between behavioral alignment and structural reasoning alignment. This section turns from diagnosis to design: what structural properties must agents possess to simulate human-like reasoning? We outline potential modeling paradigm shifts in social simulation, theoretical foundations drawn from cognitive science, and definitions of reasoning fidelity.

4.1 Modeling Paradigms in Social Simulation: A Cognitive Turn

While recent efforts in generative agent research focus on improving behavioral plausibility through techniques like persona prompting [30, 31], reinforcement learning from human feedback (RLHF) [37, 38], and chain-of-thought (CoT) generation [26, 96, 42], these methods share a common assumption: that plausible language implies plausible reasoning.

Our position challenges this assumption. These methods remain fundamentally *output-centric*, optimizing for stylistic fluency or stance alignment without simulating how beliefs are causally formed or revised. This often leads to post-hoc rationalizations, identity flattening, and the illusion of consensus.

By contrast, we propose a *cognition-centric* paradigm shift: modeling thought as a structured, revisable, and compositional process. Table 1 outlines this distinction.

Dimension	**Existing Paradigm**	**Our Proposal (GenMinds)**
Reasoning Format	Token-level generation, post-hoc	Structured belief graphs, motifs
Belief Dynamics	Static or reset each prompt	Revisable via causal updates
Evaluation Lens	Output fluency, stance labels	Reasoning fidelity and adaptability
Social Representation	Averaged, flattened views	Divergent, positional cognition

Table 1: Paradigm shift from output mimicry to cognitive modeling in generative agents.

4.2 Theoretical Foundations: Causal, Compositional, Revisable

To move beyond behavioral alignment, we must first define what it means to reason like a human.

Cognitive science offers a well-established answer. Decades of research suggest that human reasoning is not merely reactive output generation, but a process grounded in structured representations, counterfactual simulation, and dynamic belief updating [16, 18, 97]. From these foundations, we identify three defining features of human-like reasoning:

1. **Causal:** Humans reason in terms of causes and consequences. Even young children exhibit Bayesian-like inference over causal relationships and use interventions to test hypotheses about the world [16, 98]. Mental models are structured around "what caused what," emphasizing explanation rather than mere correlation. This causal orientation allows for robust generalization and counterfactual reasoning [97].
2. **Compositional:** Human reasoning is modular and reusable. Cognitive architectures operate by composing shared schemas—what we term *cognitive motifs*—that generalize across domains [18, 20]. These motifs support efficient reasoning by enabling agents to simulate belief structures without re-learning from scratch [19].
3. **Revisable:** Human beliefs evolve dynamically. When presented with new information or contradiction, individuals revise their prior assumptions. This capacity for belief updating has been modeled through probabilistic programming and counterfactual simulation frameworks [19, 99], capturing the adaptive, non-monotonic nature of human thought.

Taken together, the three dimensions of *causal*, *compositional*, and *revisable* reasoning form the foundation of what we call **reasoning fidelity**, defined as the structural integrity of belief formation and revision processes in generative agents.

4.3 Defining Reasoning Fidelity

We define **reasoning fidelity** as an agent's ability to construct, simulate, and revise a structured trace of belief formation that mirrors human causal reasoning patterns. This concept extends the dual-process model proposed by [99], in which language models interact with structured reasoning systems to model inference, belief, and decision-making.

Reasoning fidelity comprises three measurable properties:

1. **Traceability** — the ability to inspect how a belief or stance was formed through intermediate reasoning steps [100, 101];
2. **Counterfactual adaptability** — the capacity to revise beliefs predictably in response to interventions or changes in context [102, 103];
3. **Motif compositionality** — the reuse of modular causal structures (motifs) across different scenarios or domains [99, 104].

These properties define the core evaluation axes in the proposed **RECAP paradigm**, which shifts benchmarking from output plausibility to structural reasoning fidelity (Section 5). For example, traceability is assessed via motif-to-stance inference accuracy, adaptability through belief revision under hypothetical scenarios, and compositionality via motif reuse across unrelated topics.

This framework can be formed through explicit causal belief graphs, as illustrated in our proposed *GenMinds* architecture (Section 5). In such graphs, nodes represent causally relevant concepts (e.g., policy tradeoffs, values, or outcomes), and directed edges encode influence relationships. These graphs are derived from natural language using LLM-guided parsing and persist across interactions, enabling intervention analysis and reasoning trace reconstruction.

Importantly, this architecture is not tied to any particular implementation. While LLMs may serve as one plausible interface for extracting cognitive motifs, the core modeling contribution lies in structuring reasoning as revisable causal graphs. This approach is compatible with both symbolic and neural systems [99], and GenMinds exemplifies one such instantiation of this broader modeling principle.

At the evaluation level, reasoning fidelity fulfills emerging demands for cognitively grounded AI benchmarks [88]. It offers a testable, interpretable standard for assessing agent behavior that goes beyond language mimicry.

Yet current LLM-based agents fall short of this standard. Most optimize for surface alignment by producing plausible stances such as "I support policy X," without modeling the underlying belief process. They lack persistent belief states, causal coherence, and principled revision under counterfactuals. This results in brittle or contradictory responses, agreement bias between agents, and an absence of traceable justification.

5 Toward Cognitively Grounded Simulation: Modeling and Evaluation Principles

After outlining the cognitive foundations necessary for human-like reasoning, we translate these principles into a modeling and evaluation framework for cognitively grounded simulation. This includes *GenMinds*, which models structured belief formation, and *RECAP*, which evaluates reasoning fidelity in generative agents.

5.1 GenMinds: A Framework for Modeling Human-Like Reasoning

Structured Thought Capture: From Semi-Structured Interviews to Causal Graphs. To build generative agents that simulate human reasoning rather than merely output plausible stances, we propose modeling individuals' internal logic through **semi-structured interviews**, adaptively conducted by large language models (LLMs). These interviews elicit causal explanations in everyday language (e.g., "why do you support X?" "what does Y influence?"), which are then parsed into directed acyclic graphs representing the participant's belief structure [3]. Each node encodes a concept (e.g., fairness, safety, family needs), and each edge encodes a directional causal relation, with confidence and polarity scores.

Shared Knowledge. We introduce *cognitive motifs* as minimal causal reasoning units extracted from natural language. These motifs—e.g., "Surveillance → Crime Rate → Public Safety"—capture widely shared conceptual dependencies across individuals. When aggregated across interviews, they form a topology of commonly held belief structures.

We represent these motifs in a symbolic causal graph (CBN), enabling alignment of diverse opinions while maintaining transparency of reasoning. By grounding this structure in semi-structured interviews, we connect population-level reasoning to individual narratives.

Inference via Symbolic–Neural Hybrid Graph Simulation. We define reasoning as a form of forward inference over belief graphs: given a causal structure and an intervention (e.g., "increasing housing near transit"), the agent uses probabilistic updates (e.g., do-calculus) to simulate belief shifts and final stances. A language model selects relevant interventions and assembles motifs into a causal Bayesian network. This hybrid method ensures both interpretability and expressive power, enabling agents to trace "why" a conclusion was reached and what would change it.

Be Aware of Unknown. While causal motifs help model explicit reasoning patterns, real-world beliefs are often incomplete or contradictory. Our framework is designed to highlight missing links or uncertain dependencies by visualizing weakly supported or isolated nodes in the graph.

We encourage future systems to maintain uncertainty visualization and prompt-based elicitation to expand motif coverage, rather than overfitting to known paths. This allows belief modeling to remain adaptive and open-ended, rather than overly deterministic.

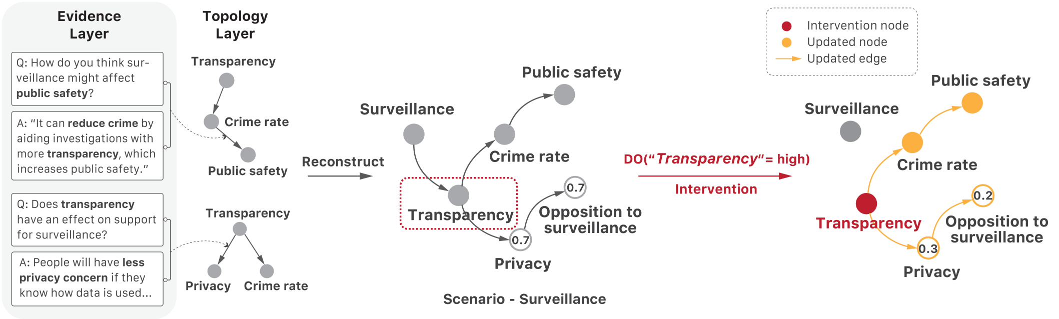
Evidence Layer
Topology Layer
Q: How do you think surveillance might affect public safety?
A: "It can reduce crime by aiding investigations with more transparency, which increases public safety."
Q: Does transparency have an effect on support for surveillance?
A: People will have less privacy concern if they know how data is used...
Transparency
Crime rate
Public safety
Transparency
Privacy
Crime rate
Reconstruct
Surveillance
Public safety
Crime rate
Transparency
0.7
Opposition to surveillance
0.7
Privacy
Scenario - Surveillance
DO("Transparency"= high)
Intervention
Intervention node
Updated node
Updated edge
Surveillance
Public safety
Crime rate
Transparency
0.2
Opposition to surveillance
0.3
Privacy

Figure 2: **Motif-based belief graph and intervention.** Natural language responses are parsed into motif-level causal links, forming a personalized belief graph. A simulated intervention on `Transparency` propagates downstream updates, shown as highlighted nodes and edges.

Illustrative Example: From Interviews to Reasoning Agent Structures

To concretize how motif-based causal reasoning operates in our framework, we present a real scenario from our semi-structured interviews on urban surveillance.

Step 1: Extracting causal motifs from QA responses. We start with Q and A responses annotated with concept nodes and directional relations. For instance:

- **QA#1:** *Q: How do you think surveillance might affect **public safety**? A: "It can **reduce crime** by aiding investigations with more **transparency**, which increases public safety."* ⇒ Motif: `Transparency` → `Crime rate` → `Public safety`
- **QA#2:** *Q: Does **transparency** have an effect on support for surveillance? A: "People will have **less privacy concern** if they know how data is used..."* ⇒ Motif: `Privacy` ← `Transparency` → `Crime rate`

Step 2: Composing a Causal Belief Network. These motifs are compiled into a belief graph representing the participant's reasoning. Nodes are concepts; edges indicate directional influence. Confidence scores are derived from motif density or respondent emphasis.

Step 3: Simulating belief change via intervention. We apply a hypothetical intervention:

```
do (Transparency = high)
```

This reflects a policy shift such as increasing camera accountability. Using belief propagation over the CBN, the downstream posteriors update as follows:

$$\begin{aligned} P(\texttt{Privacy Concern}) &: 0.7 \rightarrow 0.3 \\ P(\texttt{Opposition to Surveillance}) &: 0.7 \rightarrow 0.2 \end{aligned}$$

This chain demonstrates the potential of motif-based causal modeling to simulate how real individuals update their beliefs in response to policy changes, thereby moving beyond static opinion snapshots.

5.2 RECAP: Principles for Evaluating Reasoning Fidelity

To advance cognitively aligned simulation, we propose a benchmark framework called **RECAP**, that shifts evaluation from surface-level correctness to the internal structure and coherence of reasoning.

Design Principles.

- **Traceability:** Can the agent construct a transparent chain of intermediate beliefs?
- **Demographic Sensitivity:** Can it represent diverse reasoning paths across identities or contexts?

- **Intervention Coherence:** Does it revise beliefs in response to hypothetical changes in a consistent, causally grounded way?

Structure and Inputs.

- Situated prompt in a morally or socially complex domain;
- Human-annotated responses capturing causal motifs and belief chains;
- A task such as graph reconstruction, stance explanation, or counterfactual reasoning that requires structured inference.

Metrics.

- *Motif Alignment:* Structural similarity between human and model belief graphs;
- *Belief Coherence:* Internal consistency of the model's reasoning trace;
- *Counterfactual Robustness:* Sensible belief updates under interventions.

Grounding in Human Reasoning. All items originate from real-world, semi-structured interviews, capturing how people explain and revise their beliefs. This grounding ensures the benchmark reflects the complexity and causal depth of actual human reasoning.

Toward a Shared Format. RECAP is not a static dataset but a replicable schema for structured reasoning evaluation. Grounded in human-derived motifs, it aims to promote interpretability, adaptability, and socially responsible agent design.

6 A Call for Cognitively Grounded Simulation

As large language models become embedded in social simulations and policy tools, we face a pivotal choice: whether to pursue agents that merely sound human, or agents that can reason in structured, human-like ways. This paper argues for the latter. We call for a shift from behavior-level mimicry to cognitively grounded reasoning, where agents represent beliefs, simulate causal relationships, revise assumptions, and reveal their internal logic.

We introduced *Generative Minds* and *RECAP* as conceptual scaffolds to support this shift—prioritizing reasoning fidelity, traceability, and epistemic diversity over surface plausibility. These are not fixed systems, but a framework for developing agents that simulate how people think, not just what they say.

This paradigm enables more transparent diagnostics, pluralistic modeling of public reasoning, and structured evaluations that align with the complexity of real-world decisions.

Implications of Adopting Reasoning Fidelity as a Core Standard. Adopting reasoning fidelity as a core standard would shift generative agent research from stylistic fluency to structural interpretability. It reshapes alignment evaluation, promotes modular and revisable architectures, and incentivizes cognitively grounded benchmarks. In high-stakes applications such as civic simulation, participatory policy, and AI governance, agents with causal transparency and revisable beliefs are essential for trust, auditability, and fairness. Without this shift, we risk institutionalizing brittle models that obscure bias and flatten the diversity of public thought.

Open Challenges and Next Steps We are actively developing:

- Agent architectures for modular belief reasoning and counterfactual revision;
- Tools for causal motif extraction and belief graph construction;
- Datasets across domains such as housing, surveillance, and healthcare.

Alongside these efforts, we identify several open challenges:

- Constructing causal belief networks from natural language transcripts remains challenging, due to ambiguity in concept identification, causal direction, polarity, and conceptual granularity;
- Causality alone cannot capture the full range of human reasoning. People also rely on associative, analogical, and emotional processes that resist strict symbolic modeling. Our initial focus on casuality is a strategic and computationally tractable starting point, not an endpoint.

We invite the community to co-develop evaluation protocols, agent designs, and data pipelines that advance cognitively aligned simulation.

To simulate society faithfully, we must simulate thought.

Acknowledgments

This work builds on research originally developed at MIT. We thank Kent Larson, Jinhua Zhao, Paul Liang, and Deb Roy for their valuable guidance and discussions throughout this work. We are especially grateful to Kent Larson for his continuous mentorship, to Deb Roy for his conceptual insights and high standards that shaped the framing of this work, to Jinhua Zhao for his thoughtful resonance and encouragement, and to Paul Liang for providing key technical perspectives.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] C. Xie, C. Chen, F. Jia, Z. Ye, S. Lai, K. Shu, J. Gu, A. Bibi, Z. Hu, D. Jurgens *et al.*, "Can large language model agents simulate human trust behavior?" in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[3] J. Yang, R. Ding, E. Brown, X. Qi, and S. Xie, "V-IRL: Grounding Virtual Intelligence in Real Life," in *European conference on computer vision*, 2024.

[4] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg *et al.*, "A generalist agent," *arXiv preprint arXiv:2205.06175*, 2022.

[5] J. Piao, Y. Yan, J. Zhang, N. Li, J. Yan, X. Lan, Z. Lu, Z. Zheng, J. Y. Wang, D. Zhou *et al.*, "Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society," *arXiv preprint arXiv:2502.08691*, 2025.

[6] Y. Huang, Z. Yuan, Y. Zhou, K. Guo, X. Wang, H. Zhuang, W. Sun, L. Sun, J. Wang, Y. Ye *et al.*, "Social science meets llms: How reliable are large language models in social simulations?" *arXiv preprint arXiv:2410.23426*, 2024.

[7] M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, M. Botvinick, and C. Summerfield, "Ai can help humans find common ground in democratic deliberation," *Science*, vol. 384, no. 6696, pp. 620–626, 2024.

[8] P. H. Miller, *Theories of developmental psychology*. Macmillan, 2003.

[9] H. Gardner, *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books, 1985.

[10] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, "How language model hallucinations can snowball," *arXiv preprint arXiv:2305.13534*, 2023.

[11] H. Chi, H. Li, W. Yang, F. Liu, L. Lan, X. Ren, T. Liu, and B. Han, "Unveiling causal reasoning in large language models: Reality or mirage?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 96 640–96 670, 2024.

[12] V. Cedeno-Mieles, Z. Hu, X. Deng, Y. Ren, A. Adiga, C. Barrett, S. Ekanayake, G. Korkmaz, C. J. Kuhlman, D. Machi *et al.*, "Mechanistic and data-driven agent-based models to explain human behavior in online networked group anagram games," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 357–364.

[13] D. Anzola and C. García-Díaz, "What kind of prediction? evaluating different facets of prediction in agent-based social simulation," *International Journal of Social Research Methodology*, vol. 26, no. 2, pp. 171–191, 2023.

[14] A. Pachot and T. Petit, "Can large language models accurately predict public opinion? a review," —, 2024, ffhal-04688498f. [Online]. Available: https://hal.archives-ouvertes.fr/hal-04688498f

[15] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum, "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing," *Nature Human Behaviour*, vol. 1, no. 4, p. 0064, 2017.

[16] A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks, “A theory of causal learning in children: Causal maps and bayes nets,” *Psychological Review*, vol. 111, no. 1, pp. 3–32, 2004.

[17] J. S. Park, J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative agents: Interactive simulacra of human behavior,” in *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023, pp. 1–22.

[18] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, “How to grow a mind: Statistics, structure, and abstraction,” *science*, vol. 331, no. 6022, pp. 1279–1285, 2011.

[19] N. D. Goodman, J. B. Tenenbaum, and T. Gerstenberg, “Concepts in a probabilistic language of thought,” 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:16858487

[20] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, “Building machines that learn and think like people,” *Behavioral and Brain Sciences*, vol. 40, p. e253, 2017.

[21] K. Sigmund, H. De Silva, A. Traulsen, and C. Hauert, “Social learning promotes institutions for governing the commons,” *Nature*, vol. 466, pp. 861–3, 08 2010.

[22] K. Sreedhar, A. Cai, J. Ma, J. V. Nickerson, and L. B. Chilton, “Simulating cooperative prosocial behavior with multi-agent llms: Evidence and mechanisms for ai agents to inform policy decisions,” in *Proceedings of the 30th International Conference on Intelligent User Interfaces*, ser. IUI ’25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1272–1286. [Online]. Available: https://doi.org/10.1145/3708359.3712149

[23] R. M. del Rio-Chanona, M. Pangallo, and C. Hommes, “Can generative ai agents behave like humans? evidence from laboratory market experiments,” 2025. [Online]. Available: https://arxiv.org/abs/2505.07457

[24] F. Gilardi, M. Alizadeh, and M. Kubli, “Chatgpt outperforms crowd workers for text-annotation tasks,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, Jul. 2023. [Online]. Available: http://dx.doi.org/10.1073/pnas.2305016120

[25] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, “Out of one, many: Using language models to simulate human samples,” *Political Analysis*, vol. 31, no. 3, p. 337–351, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1017/pan.2023.2

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[27] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[28] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, “Tree of thoughts: Deliberate problem solving with large language models,” *Advances in neural information processing systems*, vol. 36, pp. 11 809–11 822, 2023.

[29] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin *et al.*, “A survey on large language model based autonomous agents,” *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[30] Y. Shao, L. Li, J. Dai, and X. Qiu, “Character-llm: A trainable agent for role-playing,” *arXiv preprint arXiv:2310.10158*, 2023.

[31] J. Chen, X. Wang, R. Xu, S. Yuan, Y. Zhang, W. Shi, J. Xie, S. Li, R. Yang, T. Zhu *et al.*, “From persona to personalization: A survey on role-playing language agents,” *arXiv preprint arXiv:2404.18231*, 2024.

[32] Y. Chaudhary and J. Penn, “Large language models as instruments of power: New regimes of autonomous manipulation and control,” *arXiv preprint arXiv:2405.03813*, 2024.

[33] X. Zhang, J. Lin, X. Mou, S. Yang, X. Liu, L. Sun, H. Lyu, Y. Yang, W. Qi, Y. Chen *et al.*, “Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users,” *arXiv preprint arXiv:2504.10157*, 2025.

[34] Q. Mi, M. Yang, X. Yu, Z. Zhao, C. Deng, B. An, H. Zhang, X. Chen, and J. Wang, "Mf-llm: Simulating collective decision dynamics via a mean-field large language model framework," *arXiv preprint arXiv:2504.21582*, 2025.

[35] J. R. Anthis, R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, E. Brynjolfsson, J. Evans, and M. S. Bernstein, "Position: Llm social simulations are a promising research method," in *Forty-second International Conference on Machine Learning Position Paper Track*.

[36] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," 2019. [Online]. Available: https://arxiv.org/abs/1811.10154

[37] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell, "Open problems and fundamental limitations of reinforcement learning from human feedback," 2023. [Online]. Available: https://arxiv.org/abs/2307.15217

[38] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, "A survey of reinforcement learning from human feedback," 2024. [Online]. Available: https://arxiv.org/abs/2312.14925

[39] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić, "Personality traits in large language models," 2025. [Online]. Available: https://arxiv.org/abs/2307.00184

[40] Z. Yu, L. He, Z. Wu, X. Dai, and J. Chen, "Towards better chain-of-thought prompting strategies: A survey," 2023. [Online]. Available: https://arxiv.org/abs/2310.04959

[41] S. H. Tanneru, D. Ley, C. Agarwal, and H. Lakkaraju, "On the hardness of faithful chain-of-thought reasoning in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2406.10625

[42] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, and C. Callison-Burch, "Faithful chain-of-thought reasoning," in *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*, 2023.

[43] N. Kroeger, D. Ley, S. Krishna, C. Agarwal, and H. Lakkaraju, "Are large language models post hoc explainers?" 2024. [Online]. Available: https://openreview.net/forum?id=MOtZlKkvdz

[44] N. Joshi, A. Saparov, Y. Wang, and H. He, "Llms are prone to fallacies in causal inference," 2024. [Online]. Available: https://arxiv.org/abs/2406.12158

[45] C. Zhao, Z. Tan, P. Ma, D. Li, B. Jiang, Y. Wang, Y. Yang, and H. Liu, "Is chain-of-thought reasoning of llms a mirage? a data distribution lens," 2025. [Online]. Available: https://arxiv.org/abs/2508.01191

[46] N. Kassner, O. Tafjord, A. Sabharwal, K. Richardson, H. Schuetze, and P. Clark, "Language models with rationality," 2023. [Online]. Available: https://arxiv.org/abs/2305.14250

[47] A. Creswell, M. Shanahan, and I. Higgins, "Selection-inference: Exploiting large language models for interpretable logical reasoning," 2022. [Online]. Available: https://arxiv.org/abs/2205.09712

[48] L. LUO, Y.-F. Li, G. Haffari, and S. Pan, "Reasoning on graphs: Faithful and interpretable large language model reasoning," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=ZGNWW7xZ6Q

[49] T. Ullman, "Large language models fail on trivial alterations to theory-of-mind tasks," 2023. [Online]. Available: https://arxiv.org/abs/2302.08399

[50] F. Cushman, "Rationalization is rational," *Behavioral and Brain Sciences*, vol. 43, p. e28, 2020.

[51] E. Eyster, S. Li, and S. Ridout, “A theory of ex post rationalization,” 2022. [Online]. Available: https://arxiv.org/abs/2107.07491

[52] T. Felin and M. Holweg, “Theory is all you need: Ai, human cognition, and causal reasoning,” *Strategy Science*, vol. 9, no. 4, pp. 346–371, 2024.

[53] K. Epstude and N. J. Roese, “The functional theory of counterfactual thinking,” *Personality and Social Psychology Review*, vol. 12, no. 2, pp. 168–192, 2008, pMID: 18453477. [Online]. Available: https://doi.org/10.1177/1088868308316091

[54] N. Roese, “Counterfactual thinking,” *Psychological Bulletin*, vol. 121, pp. 133 – 148, 01 1997.

[55] G. Gendron, J. M. Rožanec, M. Witbrock, and G. Dobbie, “Counterfactual causal inference in natural language with large language models,” 2024. [Online]. Available: https://arxiv.org/abs/2410.06392

[56] R. U. Sosa, K. N. Ramamurthy, M. Chang, and M. Singh, “Reasoning about concepts with LLMs: Inconsistencies abound,” in *First Conference on Language Modeling*, 2024. [Online]. Available: https://openreview.net/forum?id=oSG6qGkt1I

[57] Y. Saxena, S. Chopra, and A. M. Tripathi, “Evaluating consistency and reasoning capabilities of large language models,” in *2024 Second International Conference on Data Science and Information System (ICDSIS)*, 2024, pp. 1–5.

[58] K. Han, K. Kuang, Z. Zhao, J. Ye, and F. Wu, “Causal agent based on large language model,” 2024. [Online]. Available: https://arxiv.org/abs/2408.06849

[59] J. Jiang, K. Zhou, W. X. Zhao, Y. Song, C. Zhu, H. Zhu, and J.-R. Wen, “Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph,” 2024. [Online]. Available: https://arxiv.org/abs/2402.11163

[60] B. Zhang and H. Soh, “Extract, define, canonicalize: An llm-based framework for knowledge graph construction,” 2024. [Online]. Available: https://arxiv.org/abs/2404.03868

[61] Y. Zhang, Y. Zhang, Y. Gan, L. Yao, and C. Wang, “Causal graph discovery with retrieval-augmented generation based large language models,” 2024. [Online]. Available: https://arxiv.org/abs/2402.15301

[62] E. Smith and P. Hancox, “Representation, coherence and inference,” *Artif. Intell. Rev.*, vol. 15, pp. 295–323, 06 2001.

[63] N. Pfeifer and G. Kleiter, “Coherence and nonmonotonicity in human reasoning,” *Synthese*, vol. 146, pp. 93–109, 08 2005.

[64] D. Bostick, “The emergent nature of knowledge–structured resonance, coherence, and the collapse of probability in human cognition,” 2025.

[65] G. Gui and O. Toubia, “The challenge of using llms to simulate human behavior: A causal inference perspective,” *SSRN Electronic Journal*, 2023. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4650172

[66] D. Jarrett, M. Pîslar, M. A. Bakker, M. H. Tessler, R. Köster, J. Balaguer, R. Elie, C. Summerfield, and A. Tacchetti, “Language agents as digital representatives in collective decision-making,” 2025. [Online]. Available: https://arxiv.org/abs/2502.09369

[67] J. Burton, E. Lopez-Lopez, S. Hechtlinger, Z. Rahwan, S. Aeschbach, M. Bakker, J. Becker, A. Berditchevskaia, J. Berger, L. Brinkmann, L. Flek, S. Herzog, S. Huang, S. Kapoor, A. Narayanan, A.-M. Nussberger, T. Yasseri, P. Nickl, A. Almaatouq, and R. Hertwig, “How large language models can reshape collective intelligence,” *Nature human behaviour*, vol. 8, 09 2024.

[68] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, and A. D. Procaccia, “Webuildai: Participatory framework for algorithmic governance,” *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Nov. 2019. [Online]. Available: https://doi.org/10.1145/3359283

[69] A. Kormilitzin, N. Tomasev, K. McKee, and D. Joyce, “A participatory initiative to include lgbt+ voices in ai for mental health,” *Nature Medicine*, vol. 29, pp. 1–2, 01 2023.

[70] S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, and M. Fritz, “Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 83 548–83 599, 2024.

[71] Z. Weng, G. Chen, and W. Wang, “Do as we do, not as you think: the conformity of large language models,” in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=st77ShxP1K

[72] R. Baltaji, B. Hemmatian, and L. R. Varshney, “Persona inconstancy in multi-agent llm collaboration: Conformity, confabulation, and impersonation,” 2024. [Online]. Available: https://arxiv.org/abs/2405.03862

[73] J. C. Yang, D. Dailisan, M. Korecki, C. I. Hausladen, and D. Helbing, “Llm voting: Human choices and ai collective decision-making,” *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, p. 1696–1708, Oct. 2024. [Online]. Available: http://dx.doi.org/10.1609/aies.v7i1.31758

[74] A. Proskurnikov, M. Cao *et al.*, “Consensus in multi-agent systems,” *Wiley encyclopedia of electrical and electronics engineering, Wiley & Sons*, vol. 2, p. 14, 2016.

[75] Z. Wu and T. Ito, “The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems,” 2025. [Online]. Available: https://arxiv.org/abs/2502.16565

[76] A. Wang, J. Morgenstern, and J. P. Dickerson, “Large language models that replace human participants can harmfully misportray and flatten identity groups,” *Nature Machine Intelligence*, pp. 1–12, 2025.

[77] K. Lee, S. H. Kim, S. Lee, J. Eun, Y. Ko, H. Jeon, E. H. Kim, S. Cho, S. Yang, E. mee Kim, and H. Lim, “Spectrum: A grounded framework for multidimensional identity representation in llm-based agent,” 2025. [Online]. Available: https://arxiv.org/abs/2502.08599

[78] I. van Rooij and T. Wareham, “Parameterized complexity in cognitive modeling,” *Comput. J.*, vol. 51, no. 3, p. 385–404, May 2008. [Online]. Available: https://doi.org/10.1093/comjnl/bxm034

[79] X. Mou, X. Ding, Q. He, L. Wang, J. Liang, X. Zhang, L. Sun, J. Lin, J. Zhou, X. Huang, and Z. Wei, “From individual to society: A survey on social simulation driven by large language model-based agents,” 2024. [Online]. Available: https://arxiv.org/abs/2412.03563

[80] P. Davidsson, “Multi agent based simulation: Beyond social simulation,” in *Multi-Agent-Based Simulation*, S. Moss and P. Davidsson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 97–107.

[81] C. Castelfranchi, “The theory of social functions: Challenges for multi-agent-based social simulation and multi-agent learning,” *Journal of Cognitive Systems Research*, vol. 2, pp. 5–38, 2001.

[82] J. Harding, W. D’Alessandro, N. Laskowski, and R. Long, “Ai language models cannot replace human research participants,” *AI & SOCIETY*, vol. 39, 07 2023.

[83] F. Alqasemi, H. Al-Baadani, and M. A. Al-Hagery, “Stance detection using two popular benchmarks: A survey,” in *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, 2022, pp. 1–6.

[84] B. Schiller, J. Daxenberger, and I. Gurevych, “Stance detection benchmark: How robust is your stance detection?” 2020. [Online]. Available: https://arxiv.org/abs/2001.01565

[85] S. Mehri, M. Eric, and D. Hakkani-Tur, “Dialoglue: A natural language understanding benchmark for task-oriented dialogue,” 2020. [Online]. Available: https://arxiv.org/abs/2009.13570

[86] N. Dziri, H. Rashkin, T. Linzen, and D. Reitter, “Evaluating attribution in dialogue systems: The begin benchmark,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1066–1083, 2022.

[87] H. Zhan, Z. Li, Y. Wang, L. Luo, T. Feng, X. Kang, Y. Hua, L. Qu, L.-K. Soon, S. Sharma, I. Zukerman, Z. Semnani-Azad, and G. Haffari, “Socialdial: A benchmark for socially-aware dialogue systems,” in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2712–2722. [Online]. Available: https://doi.org/10.1145/3539618.3591877

[88] L. Ying, K. M. Collins, L. Wong, I. Sucholutsky, R. Liu, A. Weller, T. Shu, T. L. Griffiths, and J. B. Tenenbaum, “On benchmarking human-like intelligence in machines,” 2025. [Online]. Available: https://arxiv.org/abs/2502.20502

[89] Z. Jin, Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. Lyu, K. Blin, F. Gonzalez Adauto, M. Kleiman-Weiner, M. Sachan *et al.*, “Cladder: Assessing causal reasoning in language models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 31 038–31 065, 2023.

[90] J. Ma, “Causal inference with large language model: A survey,” 2025. [Online]. Available: https://arxiv.org/abs/2409.09822

[91] E. Kiciman, R. Ness, A. Sharma, and C. Tan, “Causal reasoning and large language models: Opening a new frontier for causality,” *Transactions on Machine Learning Research*, 2023.

[92] Y. Chen, V. K. Singh, J. Ma, and R. Tang, “Counterbench: A benchmark for counterfactuals reasoning in large language models,” 2025. [Online]. Available: https://arxiv.org/abs/2502.11008

[93] T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff, and Y. Choi, “A roadmap to pluralistic alignment,” 2024. [Online]. Available: https://arxiv.org/abs/2402.05070

[94] L. H. Zhang, S. Milli, K. Jusko, J. Smith, B. Amos, Wassim, Bouaziz, M. Revel, J. Kussman, L. Titus, B. Radharapu, J. Yu, V. Sarma, K. Rose, and M. Nickel, “Cultivating pluralism in algorithmic monoculture: The community alignment dataset,” 2025. [Online]. Available: https://arxiv.org/abs/2507.09650

[95] T. Sorensen, L. Jiang, J. D. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, M. Sap, J. Tasioulas, and Y. Choi, “Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 18, p. 19937–19947, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1609/aaai.v38i18.29970

[96] T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, K. Lukošiūtė, K. Nguyen, N. Cheng, N. Joseph, N. Schiefer, O. Rausch, R. Larson, S. McCandlish, S. Kundu, S. Kadavath, S. Yang, T. Henighan, T. Maxwell, T. Telleen-Lawton, T. Hume, Z. Hatfield-Dodds, J. Kaplan, J. Brauner, S. R. Bowman, and E. Perez, “Measuring faithfulness in chain-of-thought reasoning,” 2023. [Online]. Available: https://arxiv.org/abs/2307.13702

[97] T. Gerstenberg and S. Stephan, “A counterfactual simulation model of causation by omission,” *Cognition*, vol. 216, p. 104842, 2021.

[98] J. Jara-Ettinger, L. E. Schulz, and J. B. Tenenbaum, “The naive utility calculus as a unified, quantitative framework for action understanding,” *Cognitive Psychology*, vol. 123, p. 101334, 2020.

[99] L. Wong, G. Grand, A. K. Lew, N. D. Goodman, V. K. Mansinghka, J. Andreas, and J. B. Tenenbaum, “From word models to world models: Translating from natural language to the probabilistic language of thought,” *arXiv preprint arXiv:2306.12672*, 2023.

[100] K. Stenning and M. Van Lambalgen, *Human reasoning and cognitive science*. MIT Press, 2012.

[101] T. Bosse, C. M. Jonker, and J. Treur, "Reasoning by assumption: formalisation and analysis of human reasoning traces," in *Mechanisms, Symbols, and Models Underlying Cognition: First International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2005, Las Palmas, Canary Islands, Spain, June 15-18, 2005, Proceedings, Part I 1*. Springer, 2005, pp. 427–436.

[102] T. Gerstenberg, "Counterfactual simulation in causal cognition," *Trends in Cognitive Sciences*, 2024.

[103] N. Van Hoeck, P. D. Watson, and A. K. Barbey, "Cognitive neuroscience of human counterfactual reasoning," *Frontiers in human neuroscience*, vol. 9, p. 420, 2015.

[104] J. Russin, S. W. McGrath, D. J. Williams, and L. Elber-Dorozko, "From frege to chatgpt: Compositionality in language, cognition, and deep neural networks," *arXiv preprint arXiv:2405.15164*, 2024.