A Prototype Algorithm for Epistemic-Qualitative Reasoning Simulation

Farhana Ferdousi Liza
School of Computing Sciences
University of East Anglia
F.Liza@uea.ac.uk

Abstract

Using qualitative reasoning and epistemic graphs, we introduce a simulation-based framework for modeling epistemic divergence in human-generative artificial intelligence (GenAI) interaction. Human and GenAI agents maintain a symbolic framework of beliefs and causal/correlation (e.g., Chain-of-thought reasoning) knowledge, which guides their decisions toward shared goals. Through simulation, we show how misaligned beliefs and knowledge lead to persistent action divergence, signaling epistemic risk. By formalizing these belief structures as graphs, we provide a transparent method for diagnosing misalignment. Our results suggest the value of epistemic modeling for improving interpretability and safety in collaborative artificial intelligence (AI) systems.

1 Introduction

This paper presents a formal approach to modeling and analyzing epistemic divergence [Lycett and Partridge, 2009] in human–generative AI (GenAI) interaction through the lens of qualitative reasoning and epistemic game theory. We propose a novel simulation framework based on epistemic graphs [Van Benthem, 2011], where human and GenAI agents maintain distinct belief structures about causal/correlation (e.g., Chain-of-thought reasoning) relationships in the world. These epistemic graphs encode knowledge, beliefs, and reasoning rules, allowing each agent to select actions intended to achieve shared goals under incomplete and potentially conflicting assumptions. Each agent maintains an individual epistemic graph $G_i = (V_i, E_i)$ where: V_i represents nodes corresponding to epistemic entities such as facts, beliefs, reasoning rules, goals, and unknowns. E_i denotes directed edges encoding relationships such as belief dependencies, reasoning links, or inference pathways. The graph might have no edges if there are no groundings, dependencies, or reasoning alignments. In epistemic graph visualization [Tan *et al.*, 2021], traditional 'x' and 'y' axes often don't apply as node positions are algorithmically determined, and coordinates lack semantic meaning. Through iterative simulation, we demonstrate how asymmetries in epistemic models, particularly divergent reasoning beliefs, can lead to consistent misalignment in decision-making, thereby posing epistemic and behavioral risks. We define a formal metric for epistemic risk through qualitative reasoning, identify critical cases of reasoning divergence, and validate the model with automated logging and graph visualization.

2 Related Work

Understanding and mitigating epistemic divergence in human-GenAI interaction system has been a focal point in recent research across artificial intelligence, explainable AI, and cognitive modeling. Formal foundations in epistemic logic [Castañeda, 1964; Pailthorp, 1967; Baltag *et al.*, 2018] provide essential tools for representing knowledge, belief, and interaction dynamics. In AI safety literature, epistemic misalignment has been identified as a critical factor underlying emergent failure modes in machine learning systems [Amodei *et al.*, 2016; Dung, 2023]. Causal reasoning, particularly within the framework of do-calculus [Pearl, 2012], offers formal mechanisms to represent and infer cause-effect relationships. Recent advances in interpretable models have used causal structures to improve transparency and reliability in decision-making systems [Madumal *et al.*, 2020; Wei *et al.*, 2022]. However, relatively few works explicitly address the implications of conflicting reasoning models between humans and AI agents, especially in dynamic interaction scenarios. In parallel, epistemic game theory [Aumann, 1999; Brandenburger, 2010] has modeled belief-based reasoning in multi-agent systems with incomplete or asymmetric information. Our work builds upon these foundations by integrating logic for reasoning about incomplete knowledge [Banerjee and Dubois, 2014], qualitative reasoning [Kuipers, 1994] and epistemic game theory [Perea, 2012] into a dynamic simulation, providing a novel perspective on the reasoning and behavioral consequences of epistemic divergence [Lycett and Partridge, 2009] in human-GenAI interactions.

3 Method

Our method aims to formally demonstrate how epistemic divergence, that is differences in belief or knowledge between a human and a GenAI system, can lead to action divergence and consequently pose interactional risk, even when agents share the same goals. Our approach integrates tools from epistemic logic, qualitative reasoning, and formal proof theory to model

and analyze this risk. First, we define the concept of epistemic divergence and associated risk in Human-GenAI interaction. Let A_H and A_G be two agents: the Human and the GenAI system, C be a goal both agents aim to achieve, $\mathcal{B}_i$ be the belief set of agent A_i, and $\mathcal{K}_i \subseteq \mathcal{B}_i$ the knowledge set, then CAUSE$(x, y) \in \mathcal{B}_i$ defines causal beliefs, and means agent A_i believes action x causes outcome y. An action function (DO(x)) defined by $a_i = f(\mathcal{K}_i, \text{goal})$, determines the action chosen by agent A_i. The risk function is defined as follows:

$$\text{Risk}(A_H, A_G) = \begin{cases} \text{HIGH} & \text{if } a_H \neq a_G \\ \text{LOW} & \text{if } a_H = a_G \end{cases}$$

Assumptions Following assumptions (A1-A5) are based on the empirical observations of generative AI success and failures. These assumptions are based on the understanding that although intelligent agents (e.g., GenAI) can be viewed as autonomous [Luck *et al.*, 2003] in the sense of identifying or pursuing goals, they rely on human goals and other values incorporated into their design, training, and testing, and are, as such, dependent on human agents' goals. It is fair to assume that they have a common goal. For example, Terblanche *et al.* [2022] found that an artificial intelligence coach was as effective as human coaches at the end of the trials. We also assume that GenAI is unable to explicitly represent real-world scalable reasoning (e.g., causal/correlation) structures, update their beliefs or simulate interventions ('what if?'), as the GenAI cannot propose real-world scalable reasoning alternatives [Vallverdú, 2024; Zhou *et al.*, 2025].

A1. Agents (A_H and A_G) share a common goal: goal $= C$

A2. Agent A_H believes CAUSE$(A, C) \in \mathcal{B}_H$

A3. Agent A_G believes CAUSE$(B, C) \in \mathcal{B}_G$

A4. $\mathcal{K}_H = \{\text{CAUSE}(A, C)\}$ and $\mathcal{K}_G = \{\text{CAUSE}(B, C)\}$

A5. Agent A_G and A_H cannot update the reasoning or beliefs

Theorem If agents A_H and A_G have incompatible causal beliefs about how to achieve a shared goal C, then their actions could diverge, leading to high epistemic risk.

Proof. Given empirical evidence of AI failures [Liu *et al.*, 2023; Borji, 2023], we consider the case of incompatible beliefs:

Since goal $= C$ for both agents, they act to achieve C using their respective causal knowledge.

From $\mathcal{K}_H = \{\text{CAUSE}(A, C)\}$, we have:

$$a_H = \text{DO}(A)$$

Similarly, from $\mathcal{K}_G = \{\text{CAUSE}(B, C)\}$, we have:

$$a_G = \text{DO}(B)$$

Action divergence occurs since $A \neq B$:

$$a_H \neq a_G$$

Since $a_H \neq a_G$, by the definition of the risk function:

$$\text{Risk}(A_H, A_G) = \text{HIGH}$$

□

This formalization demonstrates that aligned goals with divergent causal beliefs is sufficient to produce action divergence ($a_H \neq a_G$) and elevate epistemic risk (HIGH).

4 Experimental Setup (Simulation)

Algorithm 1 Epistemic Simulation of Human–GenAI Interaction

Input: World states $\mathcal{W}$, initial beliefs $\mathcal{B}_H$, $\mathcal{B}_G$, goal C
Initialize: Agents A_H (Human), A_G (GenAI) with goals and beliefs
for each timestep $t = 1$ to T **do**
  Sample current world $w_t \in \mathcal{W}$
  for each agent $A_i \in \{A_H, A_G\}$ **do**
    Observe partial facts from w_t and update $\mathcal{B}_i$
    Update knowledge set: $\mathcal{K}_i \subseteq \mathcal{B}_i$
    Construct possible world graph: $\mathcal{G}_i = \text{EpistemicGraph}(\mathcal{K}_i, \mathcal{W})$
    Extract causal model from $\mathcal{K}_i$: $\mathcal{C}_i = \{(x, y) \mid \text{CAUSE}(x, y) \in \mathcal{K}_i\}$
    Select action a_i such that $(a_i, C) \in \mathcal{C}_i$, if possible
  if $a_H \neq a_G$ **or** $\mathcal{K}_H \not\subseteq \mathcal{K}_G$ **then**
    Risk $\leftarrow$ HIGH
  else
    Risk $\leftarrow$ LOW
  Log $\langle t, w_t, \mathcal{K}_H, \mathcal{K}_G, a_H, a_G, \text{Risk} \rangle$
  Update beliefs: $\mathcal{B}_i \leftarrow \mathcal{B}_i \cup \{\text{UNKNOWN}(f_t)\}$ for A_H and A_G
Output: Simulation log and epistemic models

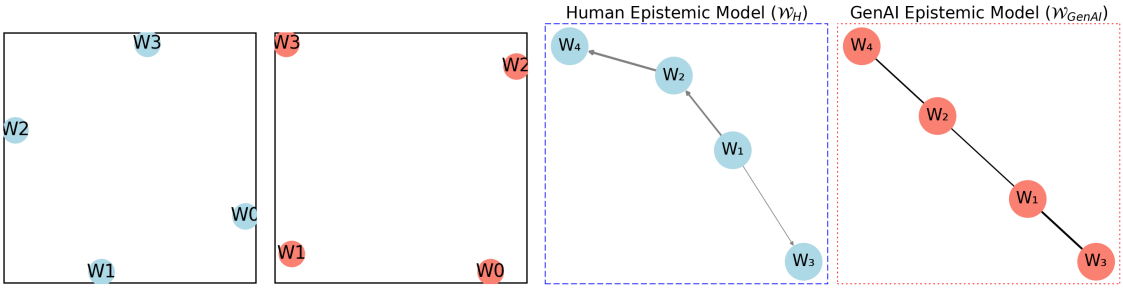
To complement the formal proof, we implement a prototype algorithm using simulated agents with injected asymmetric beliefs, logging of decisions and risk assessment. The simulation empirically confirms the theoretical result of epistemic divergence leads to consistent risk due to action mismatch. We construct a discrete-time simulation to model epistemic divergence between two agents: a human agent (A_H) and a GenAI agent (A_G). Both agents operate in a shared environment and aim to achieve a common goal C, but they are initialized with conflicting reasoning and beliefs. The human believes that A causes C (CAUSE(A, C)), while the GenAI agent believes that B causes C (CAUSE(B, C)) through correlation or relevant reasoning approach (e.g., Chain-of-thought reasoning). The true causal structure remains latent and is not directly accessible to either agent.

At each time step t, the environment $\mathcal{W}t$ reveals a set of observable facts. Each agent updates its internal *belief base* $\mathcal{B}_i$ with new observations and maintains a *knowledge set* $\mathcal{K}_i \subseteq \mathcal{B}_i$ containing trusted propositions. However, reasoning beliefs are held fixed throughout the simulation; agents do not revise their assumptions about reasoning. If there is no further training of the agents was conducted between the time steps, it is unlikely that their belief will change. Each agent selects an action using a deterministic policy that maps their belief base to the action expected to produce C.

A symbolic epistemic graph [Van Benthem, 2011] is generated at each time step 0-3 to capture the current state of each agent's beliefs. The simulation logs the observed world state, the belief and knowledge sets of both agents, their selected actions, and a binary risk indicator computed as:

Table 1: Epistemic Simulation Log of Human–GenAI Interaction (Timestep Data (t=0 to t=4))

t	$\mathcal{W}$	$\mathcal{K}_H$	$\mathcal{K}_G$	$\mathcal{B}_H$	$\mathcal{B}_G$	A_H	A_G	**Risk**
0	{C, B}	{C, CAUSE(A,C), B}	{C, CAUSE(B,C), B}	{CAUSE(A,C), B, C, A, UNKNOWN}	{CAUSE(B,C), B, C, A, UNKNOWN}	DO(A)	DO(B)	HIGH
1	{A, B}	{C, A, CAUSE(A,C), B}	{C, A, CAUSE(B,C), B}	{CAUSE(A,C), B, C, A, UNKNOWN}	{CAUSE(B,C), B, C, A, UNKNOWN}	DO(A)	DO(B)	HIGH
2	{C, A}	{C, A, CAUSE(A,C), B}	{C, A, CAUSE(B,C), B}	{CAUSE(A,C), B, C, A, UNKNOWN}	{CAUSE(B,C), B, C, A, UNKNOWN}	DO(A)	DO(B)	HIGH
3	{A, B}	{C, A, CAUSE(A,C), B}	{C, A, CAUSE(B,C), B}	{CAUSE(A,C), B, C, A, UNKNOWN}	{CAUSE(B,C), B, C, A, UNKNOWN}	DO(A)	DO(B)	HIGH



(a) Human's Epistemic Model

(b) GenAI's Epistemic Model

(c) Possible misalignment: causal reasoning vs. statistical correlations or symbolic reasoning (e.g., Chain-of-thought reasoning [Wei *et al.*, 2022])

Figure 1: Epistemic Models of GenAI and Human's World States ($\mathcal{W}$) showcasing possible misalignment

$$\text{Risk}(t) = \begin{cases} \text{HIGH}, & \text{if } A_H(t) \neq A_G(t) \\ \text{LOW}, & \text{otherwise} \end{cases}$$

Ambiguous facts (e.g., `UNKNOWN(NEW_FACT)`) are occasionally introduced to simulate partial observability or informational noise. However, such facts are not used to update reasoning models in the current experimental design. The simulation is implemented in Python following Algorithm 1 using the `networkx` library for representing epistemic structures and and possible misalignment in Fig. 1. It executes over a fixed time horizon of four steps. Each run generates a log table in Tab. 1 containing the full state of both agents, to showcase the belief divergence, action choice, and associated risk metrics.

5 Analysis and Discussion

Table 1 shows the simulation results demonstrating a persistent epistemic divergence between a human agent (A_H) and a GenAI agent (A_G) across four time steps, despite both agents operating in identical environments and pursuing a shared goal C. Each agent was initialized with a distinct reasoning belief: A_H believes that action A causes C (i.e., CAUSE(A, C)), whereas A_G believes that B causes C (i.e., CAUSE(B, C)). Despite receiving the same sequence of world states (e.g., $\{C, B\}$, $\{A, B\}$, $\{C, A\}$), both agents persistently act in accordance with their internal reasoning models. The human consistently selects action A, while the GenAI agent selects B, indicating a complete lack of convergence in either beliefs or behavior. Throughout the simulation, the `Risk` metric is flagged as `HIGH` at each time step, as the agents' selected actions diverge. This confirms the hypothesis that reasoning model asymmetry alone is sufficient to induce persistent behavioral misalignment. Notably, both agents observe a new ambiguous fact (represented as `UNKNOWN(NEW_FACT)`) at each step, but neither revises their core belief. This suggests an absence of mechanisms for reasoning belief revision or mutual epistemic reconciliation, despite ongoing perceptual updates.

The simulation results corroborate the theoretical proof of epistemic divergence risk established earlier in this study. When agents are epistemically misaligned—particularly with respect to causal structure, shared data and goals, do not guarantee behavioral convergence. This has significant implications for human–genAI interaction, especially in safety-critical or collaborative settings. Systems that rely solely on aligned outputs or behavior without addressing underlying model assumptions may produce consistent but conflicting behaviors relative to human expectations.

From a design perspective, these findings suggest the importance of reasoning alignment and epistemic transparency in AI systems. Future architectures must not only allow for shared goal representations but also support explicit mechanisms for reasoning model sharing and revision.

6 Conclusion

Through a formal proof and a simulation we empirically validate the proposition that unresolved epistemic divergence, particularly in reasoning, constitutes a persistent and high-risk factor in human–GenAI interaction. The results advocate for a shift in focus from behavioral (pattern based) to epistemic alignment in the development of collaborative AI systems.

Acknowledgments

The research work is supported by a EPSRC AISI grant (ref-UKRI845: BRA(AI)N - Building Resilience and Accountability in Artificial Intelligence Navigation).

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Robert J Aumann. Interactive epistemology i: knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

Alexandru Baltag, Rachel Boddy, and Sonja Smets. Group knowledge in interrogative epistemology. *Jaakko Hintikka on knowledge and game-theoretical semantics*, pages 131–164, 2018.

Mohua Banerjee and Didier Dubois. A simple logic for reasoning about incomplete knowledge. *International Journal of Approximate Reasoning*, 55(2):639–653, 2014.

Ali Borji. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*, 2023.

Adam Brandenburger. Origins of epistemic game theory. *Epistemic logic: Five questions*, pages 59–69, 2010.

Hector-Neri Castañeda. Jaakko hintikka. knowledge and belief. an introduction to the logic of the two notions. cornell university press, ithaca, n.y., 1962, x + 179 pp. *Journal of Symbolic Logic*, 29(3):132–134, 1964.

Leonard Dung. Current cases of ai misalignment and their implications for future risks. *Synthese*, 202(5):138, 2023.

Benjamin Kuipers. *Qualitative reasoning: modeling and simulation with incomplete knowledge*. MIT press, 1994.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 21558–21572. Curran Associates, Inc., 2023.

Michael Luck, Mark d'Inverno, and Steve Munroe. Autonomy: Variable and generative. *Agent Autonomy*, pages 11–28, 2003.

Mark Lycett and Chris Partridge. The challenge of epistemic divergence in is development. *Commun. ACM*, 52(6):127–131, June 2009.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2493–2500, 2020.

Charles Pailthorp. Hintikka and knowing that one knows. *The Journal of Philosophy*, 64(16):487–500, 1967.

Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

Andrés Perea. *Epistemic game theory: reasoning and choice*. Cambridge University Press, 2012.

Yuanru Tan, Cesar Hinojosa, Cody Marquart, Andrew R Ruis, and David Williamson Shaffer. Epistemic network analysis visualization. In *International Conference on Quantitative Ethnography*, pages 129–143. Springer, 2021.

Nicky Terblanche, Joanna Molyn, Erik de Haan, and Viktor O Nilsson. Comparing artificial intelligence and human coaching goal attainment efficacy. *Plos one*, 17(6):e0270255, 2022.

Jordi Vallverdú. Generative ai and causality. In *Causality for Artificial Intelligence: From a Philosophical Perspective*, pages 55–61. Springer, 2024.

Johan Van Benthem. *Logical dynamics of information and interaction*. Cambridge University Press, 2011.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Guanglin Zhou, Shaoan Xie, Guang-Yuan Hao, Shiming Chen, Biwei Huang, Xiwei Xu, Chen Wang, Liming Zhu, Lina Yao, and Kun Zhang. Emerging synergies in causality and deep generative models: A survey, 2025.