Grounding Before Generalizing: How AI Differs from Humans in Causal Transfer

Liangru Xiang[1,4,*], **Yuxi Ma**[2,3,4,5,*], **Zhihao Cao**[1,4,*], **Yixin Zhu**[3,2,4,5], **Song-Chun Zhu**[1,2,4]

[*]Equal contributors | [1]Department of Automation, Tsinghua University [2]Institue for Artificial Intelligence, Peking University [3]School of Psychological and Cognitive Sciences, Peking University [4]State Key Laboratory of General Artificial Intelligence [5]Beijing Key Laboratory of Behavior and Mental Health, Peking University

Abstract

Extracting abstract causal structures and applying them to novel situations is a hallmark of human intelligence.

- Using the OpenLock paradigm requiring sequential discovery of CC and CE structures, here we show that models exhibit fundamentally delayed or absent transfer.
- Even successful models require initial environmental-specific mapping before efficiency gains emerge, whereas humans leverage prior structural knowledge from the very first solution attempt.

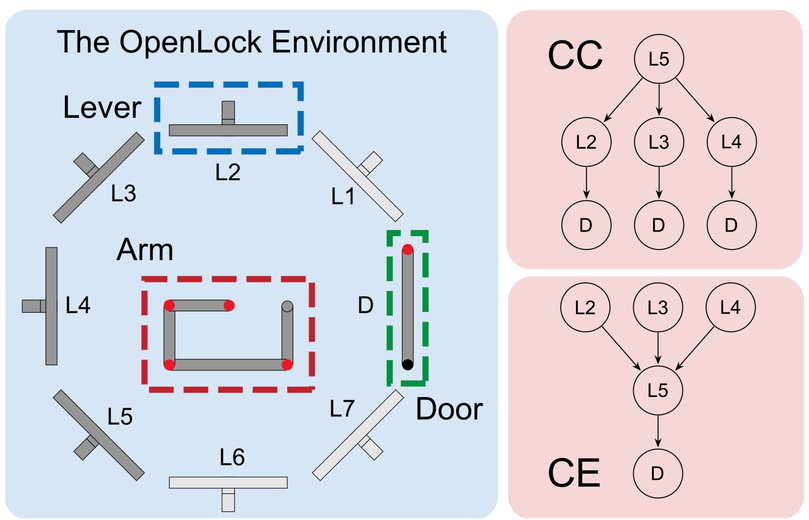
Research Question

causal learning | active exploration | transfer

Whether state-of-the-art AI models possess human-like mechanisms for abstract causal structure transfer is an open question.

We address this gap by adapting the OpenLock paradigm to systematically probe four state-of-the-art AI models.

OpenLock Paradigm



OpenLock environment and causal structure schematics. The virtual environment contains seven levers and one door; agents discover sequential causal dependencies through active exploration.

- Each environment contains eight interactive components: seven levers and one door.
- Success requires discovering three unique solutions within a budget of 30 attempts, where each attempt is strictly limited to a three-action sequence: two lever manipulations followed by a door-opening attempt.
- The two experimental variants instantiate distinct causal graph topologies over four active components (three levers and the door); the remaining four levers are inactive and serve as distractors.

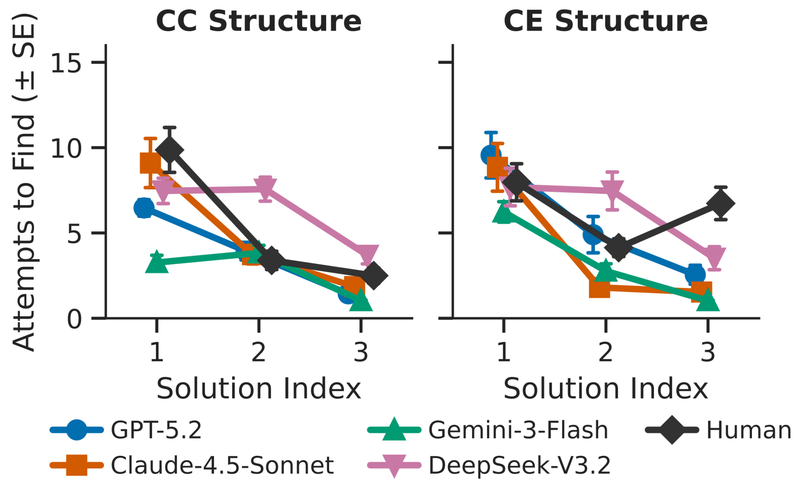
Experiment 1: Design

4 AI models	**3** input modalities
30 attempt budget	**80** human baseline N

- Following the protocol of Edmonds et al. (2018), each model was given 30 attempts to find all solutions within a single OpenLock environment.
- To ensure performance stability across random environment instantiations, we tested 30 independent agents per model on each of the two causal structures.
- GPT-5.2, Claude-4.5-Sonnet, and Gemini-3-Flash were evaluated across all three conditions; DeepSeek-V3.2 was evaluated under the text-only condition only.

Experiment 1 Result: Sequential Discovery Patterns

- Human learners exhibited non-linear acceleration: discovery cost dropped sharply from the first solution to the second.
- Claude-4.5-Sonnet closely mirrored this pattern; GPT-5.2 and Gemini-3-Flash showed only gradual, incremental improvement.
- Discussion: this abrupt efficiency gain in humans is consistent with representational change; the smooth improvement curves of GPT-5.2 and Gemini suggest iterative statistical refinement.



Sequential discovery efficiency across causal structures. Marginal discovery cost (attempts required to find each successive solution) within a single environment, shown separately for CC and CE structures.

Experiment 1 Result: Overall Performance and Causal Structure Asymmetry

- In the text-only condition, models matched or exceeded human discovery efficiency.
- Gemini-3-Flash outperformed humans in both accuracy and efficiency.
- Humans showed consistent success rates across both structures; in contrast, all models exhibited systematic asymmetries between structures.
- Discussion: the CC/CE asymmetries indicate reliance on heuristic biases rather than direction-neutral causal representations.

Model	Cond.	Success CC	Success CE	Attempts CC	Attempts CE
Human	---	65.0	65.0	19.4	22.0
GPT	T	100.0	66.7	11.8	19.7
Claude	T	67.7	86.7	16.9	14.6
Gemini	T	100.0	100.0	8.1	10.0
DeepSeek	T	96.7	86.2	18.6	20.1

Experiment 1 Result: Impact of Modality on Causal Discovery

Adding visual information degraded performance for most models.

- For GPT-5.2, the TI condition required significantly more attempts than the T condition (M=24.10 vs M=15.77).
- Gemini-3-Flash showed a smaller but significant efficiency drop from T to TI (M=9.08 vs M=10.41).
- Claude-4.5-Sonnet was the exception, showing no significant difference between T and TI.
- Discussion: current VLMs appear to lack the hierarchical control necessary to filter low-level visual features when abstract symbolic reasoning is required.

Model	Cond.	Success CC	Success CE	Attempts CC	Attempts CE
GPT	I	38.7	10.3	26.1	29.1
GPT	TI	66.7	50.0	22.8	25.4
Claude	I	45.2	64.5	22.9	19.8
Claude	TI	86.7	93.3	17.9	10.2
Gemini	I	100.0	100.0	10.0	12.3
Gemini	TI	100.0	100.0	8.7	11.8

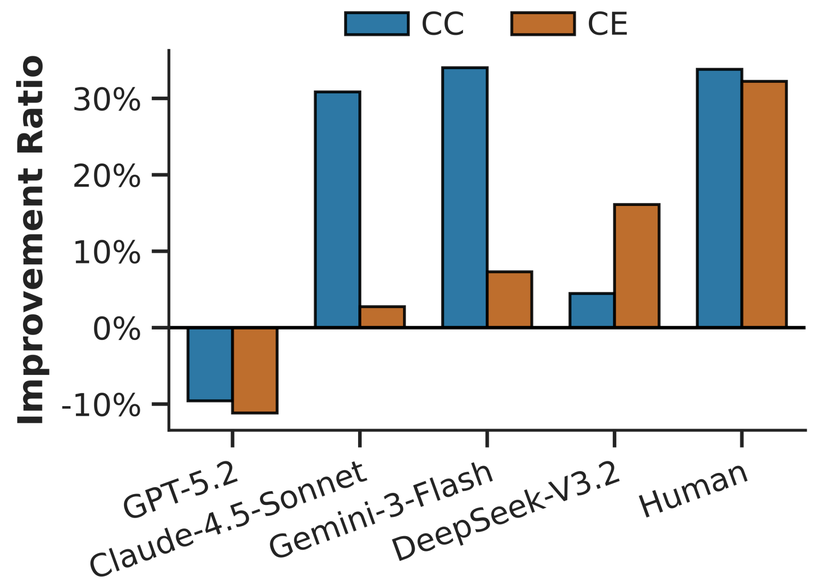
Experiment 2: Design

We modified the prompts for all models to include explicit textual descriptions of all three solutions from a previously completed environment with the same underlying causal structure.

- The previous environment had a different spatial configuration of levers and potentially different color assignments.
- We tested 30 agents per model per structure.

Experiment 2 Result: Overall Transfer Effects

- Human participants demonstrated robust structural transfer, significantly reducing average attempts from baseline to transfer.
- Gemini-3-Flash was the only model to achieve statistically significant overall transfer.
- Claude-4.5-Sonnet and DeepSeek-V3.2 showed numerical trends toward improvement that did not reach statistical significance.
- GPT-5.2 showed no positive transfer.



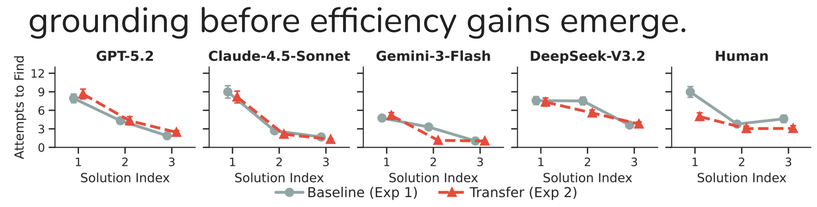
Efficiency gain from causal structure transfer. Improvement ratio in average attempt counts between Experiment 1 (Baseline) and Experiment 2 (Transfer), calculated as (Attempts_base - Attempts_trans)/Attempts_base.

Model	Baseline M (SD)	Transfer M (SD)	Improv.
Human	20.66 (9.09)	13.85 (10.00)	**+33.0*** **
GPT-5.2	15.77 (7.95)	17.43 (8.61)	-10.5
Claude-4.5-Sonnet	15.74 (9.53)	12.92 (8.47)	+17.9
Gemini-3-Flash	9.08 (2.76)	7.33 (3.60)	**+19.3** **
DeepSeek-V3.2	19.35 (6.63)	17.32 (5.89)	+10.5

Experiment 2 Result: Delayed Transfer Effects in Sequential Discovery

Humans exhibit immediate transfer, with significantly lower first-solution cost under transfer. All models exhibit delayed transfer, with performance gains emerging only at the second solution (if at all).

- None of the four models showed a significant reduction in first-solution discovery cost.
- Models benefiting from prior structure did so only after independently discovering an initial solution in the new environment.
- Discussion: this pattern suggests that humans construct decontextualized causal schemas, while models require initial environmental grounding before efficiency gains emerge.



Impact of causal structure transfer on sequential discovery dynamics. Marginal discovery cost (attempts required to find each successive solution) for Baseline (Experiment 1, gray) and Transfer (Experiment 2, red) conditions, shown for each model and for human participants.