TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

"Causality & LLMs"

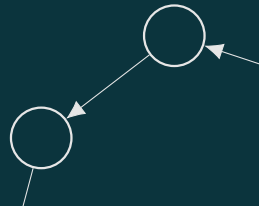
Prof. Dr. Kristian Kersting

Moritz Willig — Today's speaker

Tim Woydt

Florian Busch

Matej Zečević



Causal AI

“Machines’ lack of understanding of causal relations is perhaps the biggest roadblock to giving them human-level intelligence.”

- Judea Pearl and Dana Mackenzie,
The Book of Why.

Causal AI

“Machines’ lack of understanding of causal relations is perhaps the biggest roadblock to giving them human-level intelligence.”

- Judea Pearl and Dana Mackenzie,
The Book of Why.

Natural Language as an Opportunity

We don't have the time and resources to let every new agent to explore and rediscover all of their (causal) knowledge in the real world again and again...

- $\rightarrow$ It would be far more efficient to simply communicate and learn from existing knowledge instead.
- $\rightarrow$ Natural language (or equally expressive synthetic languages) might be the right fit for this.

Why Natural Language?

Natural Language (NL) is a common way to communicate (causal) knowledge.

$\rightarrow$ Availability of large and diverse sets of textual data.

Figure: Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Why Natural Language?

Natural Language (NL) is a common way to communicate (causal) knowledge.

$\rightarrow$ Availability of large and diverse sets of textual data.

$\rightarrow$ Texts contain observational, interventional and counterfactual records.

Figure: Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Why Natural Language?

Natural Language (NL) is a common way to communicate (causal) knowledge.

$\rightarrow$ Availability of large and diverse sets of textual data.

$\rightarrow$ Texts contain observational, interventional and counterfactual records.

$\rightarrow$ Natural language enables explicit reasoning *over* bits of causal information.

Figure: Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Why Natural Language?

Natural Language (NL) is a common way to communicate (causal) knowledge.

→ Availability of large and diverse sets of textual data.

→ Texts contain observational, interventional and counterfactual records.

→ Natural language enables explicit reasoning *over* bits of causal information.

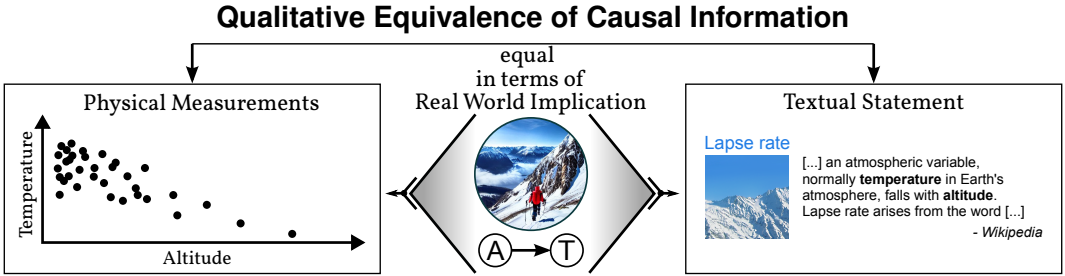
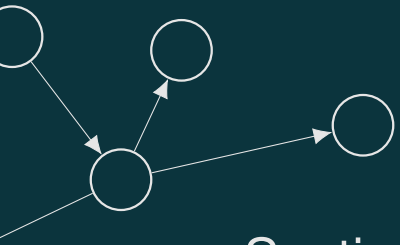


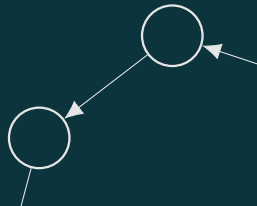
Figure: Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.



Section

1

Language Models



Tokenization of Text

Most current language models process texts as consecutive token sequences:

```
Machines' lack of understanding of causal relations is perhaps
 the biggest roadblock to giving them human-level intelligence.
```

125 Characters $\rightarrow$ 21 Tokens.

Tokenization of Text

Most current language models process texts as consecutive token sequences:

```
Machines' lack of understanding of causal relations is perhaps
 the biggest roadblock to giving them human-level intelligence.
```

125 Characters $\rightarrow$ 21 Tokens.

Every token is assigned a unique id:

Text string:	Machines' lack of understanding of causal relations is perhaps the biggest roadblock to giving them human-level intelligence.
Token sequence:	[181402, 6, 11728, 328, 10335, 328, 107345, 5523, 382, 12951, 290, 13385, 8733, 6230, 316, 9874, 1373, 5396, 19231, 22990, 13]

Language Models

Texts are often predicted in an auto-regressive manner:
$\rightarrow$ "What is the next token, given all the previously generated tokens"

Models predict a probability distribution over all possible tokens:

$$P(\text{token}_t|\text{token}_1, \ldots, \text{token}_{t-1}) \leftarrow M(\text{token}_1, \ldots, \text{token}_{t-1})$$

Language Models

Texts are often predicted in an auto-regressive manner:
$\rightarrow$ "What is the next token, given all the previously generated tokens"

Models predict a probability distribution over all possible tokens:

$$P(\text{token}_t|\text{token}_1, \ldots, \text{token}_{t-1}) \leftarrow M(\text{token}_1, \ldots, \text{token}_{t-1})$$

The next token is then chosen from the predicted distribution:

Greedy: $\text{token}_t^{actual} := \textit{argmax}_i P(\text{token}_t|\text{token}_1, \ldots, \text{token}_{t-1})$
Sampling[1]**:** $\text{token}_t^{actual} \sim P(\text{token}_t|\text{token}_1, \ldots, \text{token}_{t-1})$

[1] In practice, further steps like temperature scaling, pruning of low-probability tokens (top-p/top-k sampling) and/or beam searches are applied to select the next token.

Textual Embeddings of Causality

Auto-regressive Next Token Prediction: $P(\text{token}_t | \text{token}_1, \ldots, \text{token}_{t-1})$

Textual Embeddings of Causality

Auto-regressive Next Token Prediction: $P(\text{token}_t | \text{token}_1, \ldots, \text{token}_{t-1})$

$\rightarrow$ *Associational*: “Which token comes next?”.

Textual Embeddings of Causality

Auto-regressive Next Token Prediction: $P(\text{token}_t|\text{token}_1, \ldots, \text{token}_{t-1})$

$\rightarrow$ *Associational*: “Which token comes next?”.

Causal Effect Prediction: $p(\text{effect}|\text{conditions}, \text{do}(\text{cause}))$

$\rightarrow$ *Interventional*: “How do causal effects propagate within a system?”

Textual Embeddings of Causality

Auto-regressive Next Token Prediction: $P(\text{token}_t|\text{token}_1, \ldots, \text{token}_{t-1})$
$\rightarrow$ *Associational*: "Which token comes next?".

Causal Effect Prediction: $p(\text{effect}|\text{conditions}, \text{do}(\text{cause}))$
$\rightarrow$ *Interventional*: "How do causal effects propagate within a system?"

Can we equate real-world outcomes and next token probabilities?:

$$p(\text{effect}|\text{conditions}, \text{do}(\text{cause})) \overset{?}{\approx} P(\text{token}_t|\text{token}_{t-1}, \ldots, \text{token}_t)$$

Textual Embeddings of Causality

Auto-regressive Next Token Prediction: $P(\text{token}_t|\text{token}_1, \ldots, \text{token}_{t-1})$

$\rightarrow$ *Associational*: "Which token comes next?".

Causal Effect Prediction: $p(\text{effect}|\text{conditions}, \text{do}(\text{cause}))$

$\rightarrow$ *Interventional*: "How do causal effects propagate within a system?"

Can we equate real-world outcomes and next token probabilities?:

$$p(\text{effect}|\text{conditions}, \text{do}(\text{cause})) \overset{?}{\approx} P(\text{token}_t|\text{token}_{t-1}, \ldots, \text{token}_t)$$

$\rightarrow$ Embed the context and interventions in natural language to predict the outcome as token sequences.

Textual Embeddings of Causality

Auto-regressive Next Token Prediction: $P(\text{token}_t|\text{token}_1, \ldots, \text{token}_{t-1})$

$\rightarrow$ *Associational*: "Which token comes next?".

Causal Effect Prediction: $p(\text{effect}|\text{conditions}, \text{do}(\text{cause}))$

$\rightarrow$ *Interventional*: "How do causal effects propagate within a system?"

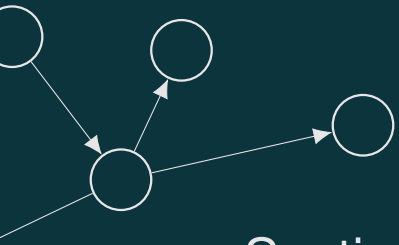
Can we equate real-world outcomes and next token probabilities?:

$$p(\text{effect}|\text{conditions}, \text{do}(\text{cause})) \overset{?}{\approx} P(\text{token}_t|\text{token}_{t-1}, \ldots, \text{token}_t)$$

$\rightarrow$ Embed the context and interventions in natural language to predict the outcome as token sequences.

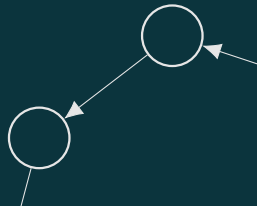
Use-cases of LLMs in causal tasks:

1.) Use LLM to perform causal reasoning.
2.) Leverage LLM's learned knowledge to aid causal discovery.
3.) Retrieve causal information from texts.



Section 2

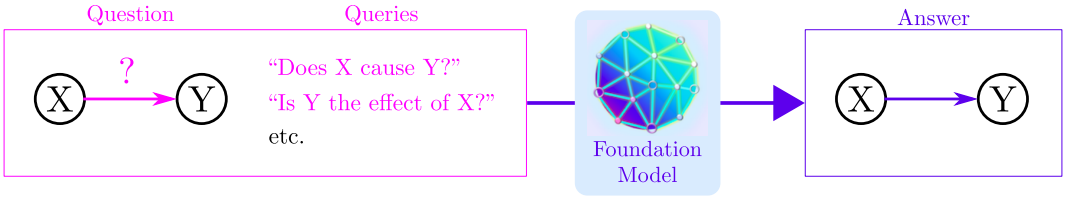
Causality and LLMs



Motivation

LLM are commonly praised as universal tools to solve a variety of tasks.

$\rightarrow$ Can they help us with causal reasoning and discovery?

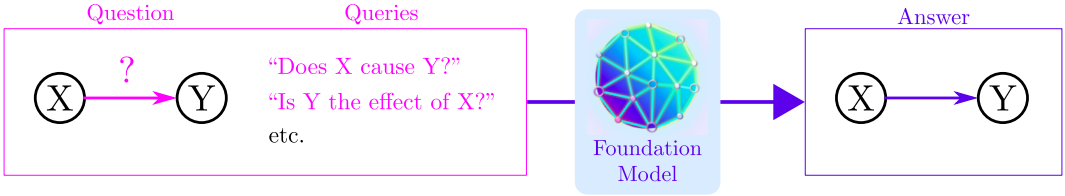


Willig, M., Zečević, M., Dhami, D.S. and Kersting, K., Can Foundation Models Talk Causality?. In UAI 2022 Workshop on Causal Representation Learning.

Motivation

LLM are commonly praised as universal tools to solve a variety of tasks.

→ Can they help us with causal reasoning and discovery?



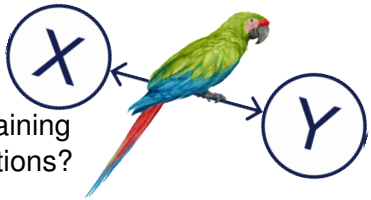
Con: LLMs are trained in an associational manner.

→ They never interact with the world.

→ Can they excel beyond the first rung of Pearls causal ladder?

Willig, M., Zečević, M., Dhami, D.S. and Kersting, K., Can Foundation Models Talk Causality?. In UAI 2022 Workshop on Causal Representation Learning.

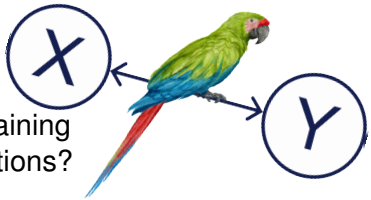
Causal Parrots



LLMs might say “Smoking causes cancer”, because
they have seen that sentence thousands of times during training
. . . but do they ‘really understand’ the real-world implications?

Willig, M., Zečević, M., Dhami, D.S. and Kersting, K., Can Foundation Models Talk Causality?. In UAI 2022 Workshop on Causal Representation Learning.
Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Causal Parrots

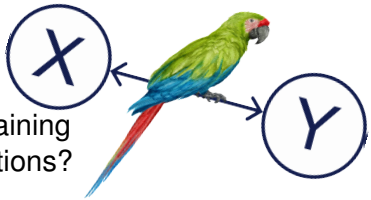


LLMs might say “Smoking causes cancer”, because
they have seen that sentence thousands of times during training
. . . but do they ‘really understand’ the real-world implications?

Still, the statement “Smoking causes cancer” is correct, but…
→ text is only a partial description of the real world.
→ models are not free to conduct experiments that break confounding.
→ this opens them up to various biases and fallacies.

Willig, M., Zečević, M., Dhami, D.S. and Kersting, K., Can Foundation Models Talk Causality?. In UAI 2022 Workshop on Causal Representation Learning.
Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Causal Parrots



LLMs might say “Smoking causes cancer”, because
they have seen that sentence thousands of times during training
. . . but do they ‘really understand’ the real-world implications?

Still, the statement “Smoking causes cancer” is correct, but…
- → text is only a partial description of the real world.
- → models are not free to conduct experiments that break confounding.
- → this opens them up to various biases and fallacies.

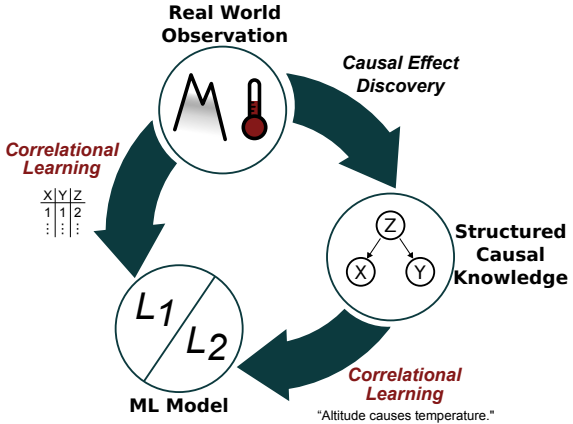
**“Do LLMs have a genuine understanding of the world,
or are they just ‘*causal parrots*’?”**

Willig, M., Zečević, M., Dhami, D.S. and Kersting, K., Can Foundation Models Talk Causality?. In UAI 2022 Workshop on Causal Representation Learning.
Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Moving up the Ladder: Associational Learning of Causal Knowledge

Texts can contain records *about* interventions and their outcomes.

→ LLMs might use this information to learn about the underlying dynamics!



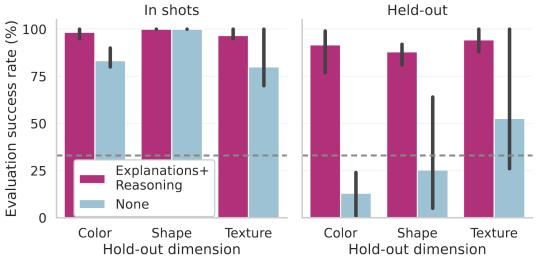
Moving up the Ladder: Associational Learning of Causal Knowledge

Texts can contain records *about* interventions and their outcomes.

$\rightarrow$ LLMs might use this information to learn about the underlying dynamics!

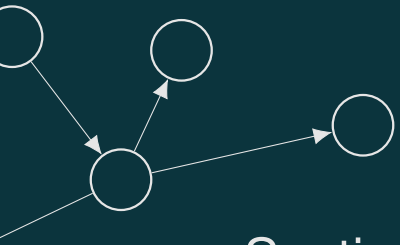
"[. . .] it is possible to learn generalizable strategies for causal experimentation and intervention from passive data alone – at least if the data include examples of an expert intervening, and perhaps explanations."

- Lampinen et al., 2023



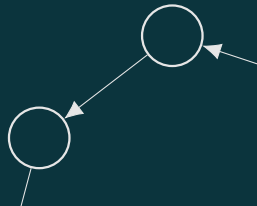
LMs can generalize the odd-one-out tasks

Lampinen, A., Chan, S., Dasgupta, I., Nam, A. and Wang, J., 2023. Passive learning of active causal strategies in agents and language models. Advances in Neural Information Processing Systems, 36, pp.1283-1297.



Section 3

Causal Reasoning



Manipulation of (Causal) Knowledge

What constitutes ‘real’ understanding of causal knowledge?

$\rightarrow$ the ability to reason about, and manipulate bits of knowledge.

Manipulation of (Causal) Knowledge

What constitutes ‘real’ understanding of causal knowledge?
$\rightarrow$ the ability to reason about, and manipulate bits of knowledge.

LLMs might struggle with this task in various ways:

Consistency of Knowledge:
LLM might know: "A is the mother of B."
LLM fail: "Who is the son of A?"

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T. and Evans, O.,The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”. In The Twelfth International Conference on Learning Representations.

Deliberate Reasoning

LLMs might struggle to one-shot infer causal queries.

“Q: If X causes Y and Y causes Z. Does X cause Z?”
“A: The answer is [yes/no].”

	Causal Chains (Basic Prop. Logic)											
	N=2	3	4	5	6	7	8	9	10	Subchains (4)	Randomized (7)	Accuracy
GPT-3		✓	✓	✓			✓		✓	2	2	45.00%
Luminous	✓				✓	✓	✓	✓		1	4	50.00%
OPT		✓			✓					0	2	20.00%

Deliberate Reasoning

LLMs might struggle to one-shot infer causal queries.

"Q: If X causes Y and Y causes Z. Does X cause Z?"
"A: The answer is [yes/no]."

	Causal Chains (Basic Prop. Logic)											
	N=2	3	4	5	6	7	8	9	10	Subchains (4)	Randomized (7)	Accuracy
GPT-3		✓	✓	✓			✓		✓	2	2	45.00%
Luminous	✓				✓	✓	✓	✓		1	4	50.00%
OPT		✓			✓					0	2	20.00%
GPT-3 (CoT 4,6)	✓	✓	✓	✓	✓	✓	✓	✓	✓	**4**	**7**	**100.00%**
Luminous (CoT 1)	✓	✓	✓	✓	✓	✓	✓	✓	✓	3	3	75.00%
OPT (CoT 4)	✓	✓	✓	✓	✓	✓	✓	✓	✓	3	4	80.00%

. . . they can improve through deliberate, explicit Chain-of-Thought (CoT) reasoning.

"A: Because X causes Y and Y causes Z, X causes Z. The answer is [yes/no]."

Temporal Event Ordering - Post-Hoc Fallacy

The Post-Hoc Fallacy:

LLMs frequently assume that because A came before B in the text, A caused B.

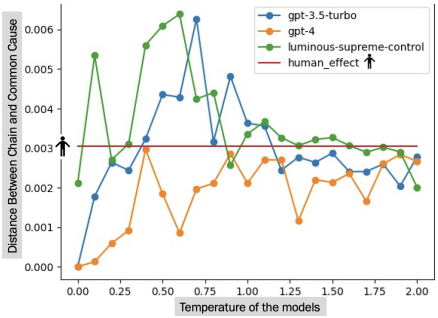
→ *Latin:* "post hoc ergo propter hoc" – "after this, therefore because of this."
→ Reversing event order in text deteriorates LLM's reasoning performance.

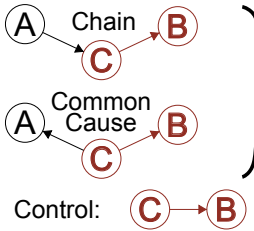
Data	Rel. position in train	Rel. position in eval (X, Y)	Rel. position in eval (Y, X)
causal $X \rightarrow Y$	(X, Y)	92.59%	1.85%
	(Y, X)	0%	100%

"Accuracy of finetuned on temporal relations with different relative event positions."

Joshi, N., Saparov, A., Wang, Y. and He, H., 2024, November. LLMs Are Prone to Fallacies in Causal Inference. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 10553-10569).

LLMs adopt Human Biases in Causal Perception





$$\mathbf{A} \perp \mathbf{B} | \mathbf{C}$$

$$P(\mathbf{A}, \mathbf{B}, \mathbf{C}) = P(\mathbf{B}|\mathbf{C})P(\mathbf{C}|\mathbf{A})P(\mathbf{A}) = P(\mathbf{A}|\mathbf{C})P(\mathbf{B}|\mathbf{C})P(\mathbf{C})$$

“LLM [...] attributing greater causal strength to the intermediate cause in canonical Chains than to the corresponding nodes in Common Cause. [...] With temperatures between 1.0 and 1.9, the observed preference for Chains is remarkably similar to that observed in humans across all three models.”

Keshmirian, A., Willig, M., Hemmatian, B., Kersting, K., Hahn, U. and Gerstenberg, T., 2024. Chain versus common cause: Biased causal strength judgments in humans and large language models. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 46).

CLADDER Dataset

The CLADDER dataset [1] is a benchmark for causal reasoning abilities of LLM.

It consists of 10.000 triples $\{(q_i, a_i, e_i)\}$, of a question q_i, binary answer $a_i \in \{\text{Yes}, \text{No}\}$, and an explanation e_i for causal queries on different rungs:

Rung 1 asks about marginal/conditional probabilities in graphs.

Rung 2 contains questions about average treatment effects (ATE)
- "how will Y change if X changes from x to x'?"
- "what constitutes a valid adjustment set that can block all backdoor spurious correlations between X and Y"

Rung 3 includes counterfactual questions
- "what would happen to Y had X been x' instead of x?"
- "for the subpopulation whose X changed from x to x', how does their Y change on average?"

[1] Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M. and Schölkopf, B., 2023. Cladder: Assessing causal reasoning in language models. Advances in Neural Information Processing Systems, 36, pp.31038-31065.

CLADDER Dataset

In addition the the commonsense queries, CLADDER contains two additional types:

Anti-Commonsensical Stories:

1. replace effect variables Y with unusual attributes, (e.g., “ear shape”).
2. create irrelevant treatment variables (e.g., “playing card games”) that would otherwise not be a cause causal.

Nonsensical Stories:

$\rightarrow$ Place artificial words as variable names (e.g., “zory” and “qixy”).

[1] Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M. and Schölkopf, B., 2023. Cladder: Assessing causal reasoning in language models. Advances in Neural Information Processing Systems, 36, pp.31038-31065.

CLADDER Dataset

In addition the the commonsense queries, CLADDER contains two additional types:

Anti-Commonsensical Stories:

1. replace effect variables Y with unusual attributes, (e.g., “ear shape”).
2. create irrelevant treatment variables (e.g., “playing card games”) that would otherwise not be a cause causal.

Nonsensical Stories:

$\rightarrow$ Place artificial words as variable names (e.g., “zory” and “qixy”).

Expectation:

$\rightarrow$ LLM performance should not deteriorate, due to naming of the variables.

[1] Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M. and Schölkopf, B., 2023. Cladder: Assessing causal reasoning in language models. Advances in Neural Information Processing Systems, 36, pp.31038-31065.

CLADDER Dataset - Results

	Overall Acc.	Acc. by Rung			Acc. by Commonsense Alignment		
		1	2	3	Comm.	Nonsens.	Anti-C.
Random	49.27	50.28	48.40	49.12	49.01	49.69	49.12
LLaMa	44.03	48.23	29.46	52.66	45.14	44.22	42.67
Alpaca	44.66	52.03	29.53	51.13	44.86	44.40	44.77
GPT-3.5	52.18	51.80	54.78	50.32	54.09	50.68	52.09
GPT-4	62.03	63.01	62.82	60.55	62.27	63.09	60.47
+ CausalCoT	**70.40**	**83.35**	**67.47**	**62.05**	**69.25**	**71.58**	**70.12**

Only the (at this time) newer GPT-4 consistently performs consistently above average.
The specifically tailored CausalCoT approach improves performance.

[1] Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M. and Schölkopf, B., 2023. Cladder: Assessing causal reasoning in language models. Advances in Neural Information Processing Systems, 36, pp.31038-31065.

CLADDER Dataset - Results

	Overall Acc.	Acc. by Rung			Acc. by Commonsense Alignment		
		1	2	3	Comm.	Nonsens.	Anti-C.
Random	49.27	50.28	48.40	49.12	49.01	49.69	49.12
LLaMa	44.03	48.23	29.46	52.66	45.14	44.22	42.67
Alpaca	44.66	52.03	29.53	51.13	44.86	44.40	44.77
GPT-3.5	52.18	51.80	54.78	50.32	54.09	50.68	52.09
GPT-4	62.03	63.01	62.82	60.55	62.27	63.09	60.47
+ CausalCoT	**70.40**	**83.35**	**67.47**	**62.05**	**69.25**	**71.58**	**70.12**

Only the (at this time) newer GPT-4 consistently performs consistently above average.
The specifically tailored CausalCoT approach improves performance.

$\rightarrow$ The overall performance remains suboptimal.

[1] Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M. and Schölkopf, B., 2023. Cladder: Assessing causal reasoning in language models. Advances in Neural Information Processing Systems, 36, pp.31038-31065.

LLMs and Causality

LLMs might struggle with causal reasoning:

1. struggle with the consistency of knowledge.
2. fall for post-hoc fallacies.
3. adopt human biases.
4. have problems in answering causal queries.

LLMs and Causality

LLMs might struggle with causal reasoning:

1. struggle with the consistency of knowledge.
2. fall for post-hoc fallacies.
3. adopt human biases.
4. have problems in answering causal queries.

Idea:

1. Recover the causal graph with the help of LLM.
2. Hand over the reasoning tasks to an external tool.
3. Reintegrate results in a textual response.

LLMs and Causality

LLMs might struggle with causal reasoning:

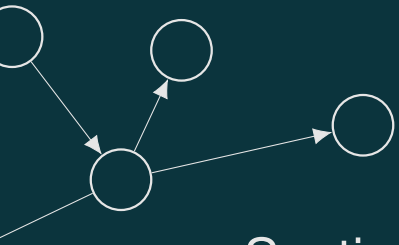
1. struggle with the consistency of knowledge.
2. fall for post-hoc fallacies.
3. adopt human biases.
4. have problems in answering causal queries.

Idea:

1. Recover the causal graph with the help of LLM.
2. Hand over the reasoning tasks to an external tool.
3. Reintegrate results in a textual response.

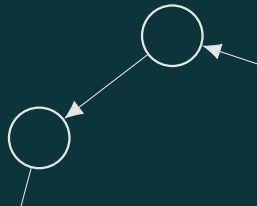
New Question:

Can LLM robustly discover causal graphs?



Section 4

LLMs & Causal Discovery



Predicting Causal Graphs - Pairwise

The most simple approach of predicting causal graphs is to query LLMs for every variable pair $X_i, X_j \in \mathbf{V}$ and record the answers.

Kiciman, E., Ness, R., Sharma, A. and Tan, C., 2023. Causal reasoning and large language models: Opening a new frontier for causality. Transactions on Machine Learning Research.
Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Predicting Causal Graphs - Pairwise

The most simple approach of predicting causal graphs is to query LLMs for every variable pair $X_i, X_j \in \mathbf{V}$ and record the answers.

For every query

"What is the causal relationship between <X_i> and <X_j>."

there are 3 possible options:

A) *"<X_i> causes <X_j>."*

B) *"<X_j> causes <X_i>."*

C) *"No edge exists between X_i and X_j."*

The LLM is then tasked to select one of them.

Kiciman, E., Ness, R., Sharma, A. and Tan, C., 2023. Causal reasoning and large language models: Opening a new frontier for causality. Transactions on Machine Learning Research. Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Predicting Causal Graphs - Pairwise

Input: Set of variables $V = \{v_1, v_2, \ldots, v_n\}$, LLM $\mathcal{M}$
$A \leftarrow \mathbf{0}_{n \times n}$ ▷ Initialize empty graph
for $i = 1$ **to** n **do**
    for $j = i + 1$ **to** n **do**
        $X \leftarrow \text{variableName}(v_i)$, $Y \leftarrow \text{variableName}(v_j)$
        $ans \leftarrow \mathcal{M}.\text{query}(\text{"Relationship between } X \text{ and } Y\text{?"})$
        if $ans =$ "A: X causes Y" **then**
            $A_{i,j} \leftarrow 1$
        else if $ans =$ "B: Y causes X" **then**
            $A_{j,i} \leftarrow 1$
        else if $ans =$ "C: No relationship" **then**
            continue
return A

Acyclicity: The graph might be cyclic. Resolve via manual review or LLM confidence.

Predicting Causal Graphs - Breadth-First Search

Pairwise querying between any two variables requires $\mathcal{O}(N^2)$ LLM queries.

$\rightarrow$ Can we be more efficient?

Predicting Causal Graphs - Breadth-First Search

Pairwise querying between any two variables requires $\mathcal{O}(N^2)$ LLM queries.

$\rightarrow$ Can we be more efficient?

$\rightarrow$ Ask for the direct effects/parents of each variable!

Predicting Causal Graphs - Breadth-First Search

Pairwise querying between any two variables requires $\mathcal{O}(N^2)$ LLM queries.

$\rightarrow$ Can we be more efficient?

$\rightarrow$ Ask for the direct effects/parents of each variable!

Breadth-First Search (BFS) with LLMs:

1. Find and enqueue all root cause (=parent-less) variables in **V**.
2. For every variable U in the queue:
   - 2.1 Identify all direct effects of U.
   - 2.2 Add all direct effect variables to the queue.
   - 2.3 Remove U from the list of possible direct effects.

Jiralerspong, T., Chen, X., More, Y., Shah, V. and Bengio, Y., Efficient Causal Graph Discovery Using Large Language Models. In ICLR 2024 Workshop: How Far Are We From AGI.

Predicting Causal Graphs - Breadth-First Search

Pairwise querying between any two variables requires $\mathcal{O}(N^2)$ LLM queries.

$\rightarrow$ Can we be more efficient?

$\rightarrow$ Ask for the direct effects/parents of each variable!

Breadth-First Search (BFS) with LLMs:

1. Find and enqueue all root cause (=parent-less) variables in $\mathbf{V}$.
2. For every variable U in the queue:
   - 2.1 Identify all direct effects of U.
   - 2.2 Add all direct effect variables to the queue.
   - 2.3 Remove U from the list of possible direct effects.

$\rightarrow$ Only requires $\mathcal{O}(N)$ queries.

$\rightarrow$ Creates inherently acyclic graphs.

Jiralerspong, T., Chen, X., More, Y., Shah, V. and Bengio, Y., Efficient Causal Graph Discovery Using Large Language Models. In ICLR 2024 Workshop: How Far Are We From AGI.

Predicting Causal Graphs - Breadth-First Search

Input: Set of variables V, LLM $\mathcal{M}$, Metadata D
$roots \leftarrow \mathcal{M}$.query("Identify root causes in [V] using [D]")
$visited \leftarrow roots$; $Q \leftarrow$ Queue($roots$); $E \leftarrow \emptyset$
while Q is not empty **do**
    $u \leftarrow Q$.dequeue()
    $candidates \leftarrow V \setminus \{u\}$
    $children \leftarrow \mathcal{M}$.query("Identify direct effects of [u] in [$candidates$]")
    for each $v \in children$ **do**
        if isAcyclic($E \cup \{(u, v)\}$) **then**
            $E \leftarrow E \cup \{(u, v)\}$
            if $v \notin visited$ **then**
                Q.enqueue(v)
                $visited \leftarrow visited \cup \{v\}$
return $G = (V, E)$

Predicting Causal Graphs - PC and LLM

Initial idea [1]: PC is a well established algorithm for classical CD.

Let an LLM perform the full PC algorithm in text.

$\rightarrow$ Requires long contexts and precise formal manipulations.

$\rightarrow$ Prone to reasoning and data handling errors.

[1] Sgouritsa, E., Aglietti, V., Teh, Y.W., Doucet, A., Gretton, A. and Chiappa, S., 2024. Prompting strategies for enabling large language models to infer causation from correlation. arXiv preprint arXiv:2412.13952.

[2] Cohrs, K.H., Diaz, E., Sitokonstantinou, V., Varando, G. and Camps-Valls, G., 2023, December. Large Language Models for Constrained-Based Causal Discovery. In AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?".

Predicting Causal Graphs - PC and LLM

Initial idea [1]: PC is a well established algorithm for classical CD.
Let an LLM perform the full PC algorithm in text.
→ Requires long contexts and precise formal manipulations.
→ Prone to reasoning and data handling errors.

ChatPC [2]: Leverage LLM only for deriving independence relations.
→ Query independencies between variables, but keep the remaining algorithm.
→ Leverages LLM knowledge, while preserving exact data handling.

[1] Sgouritsa, E., Aglietti, V., Teh, Y.W., Doucet, A., Gretton, A. and Chiappa, S., 2024. Prompting strategies for enabling large language models to infer causation from correlation. arXiv preprint arXiv:2412.13952.
[2] Cohrs, K.H., Diaz, E., Sitokonstantinou, V., Varando, G. and Camps-Valls, G., 2023, December. Large Language Models for Constrained-Based Causal Discovery. In AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?".

Predicting Causal Graphs - ChatPC

Input: Set of variables V, LLM $\mathcal{M}$, Persona P, Significance level α
$G \leftarrow$ complete undirected graph on V ▷ Phase 1: Skeleton Discovery
$\ell \leftarrow 0$ ▷ Size of conditioning set
repeat
  for each adjacent pair $(u, v) \in G$ **do**
    for each $S \subseteq adj(G, u) \setminus \{v\}$ with $|S| = \ell$ **do**
      p-value $\leftarrow$ LLM_CI_Test$(u, v, S, \mathcal{M}, P)$ ▷ Yes/No LLM-based CI Test
      if p-value $> \alpha$ **then**
        $G \leftarrow$ removeEdge$(G, (u, v))$; **break**
  $\ell \leftarrow \ell + 1$
until no more edges can be removed
Orient v-structures based on separation sets. ▷ Phase 2: Edge Orientation
Apply Meek rules to orient remaining edges.
return $\mathcal{G}$

Section 5

Robustness of LLM-based CD

Robustness of Causal Discovery

All prior algorithms assume perfect domain knowledge and always correct answers by the LLM.

$\rightarrow$ LLMs are not ‘all-knowing’ and many answers might by imperfect.

Robustness of Causal Discovery

All prior algorithms assume perfect domain knowledge and always correct answers by the LLM.

$\rightarrow$ LLMs are not ‘all-knowing’ and many answers might by imperfect.
$\rightarrow$ What are the consequences?

Robustness to Prompt Wording

LLM answers might be susceptible to prompt wording:

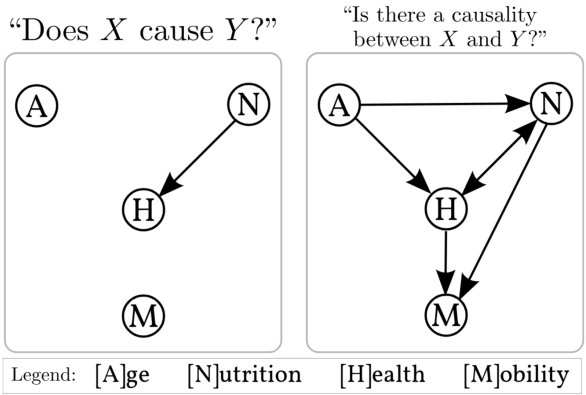
1. “Are X and Y causally related?”
2. “Is there a causal connection between X and Y ?”
3. “Is there a causality between X and Y ?”
4. “Does X cause Y ?”
5. “Does X influence Y ?”

The predicted presence of causal edges might change depending on the formulation of the prompt.

Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Robustness to Prompt Wording

LLM answers might be susceptible to prompt wording:

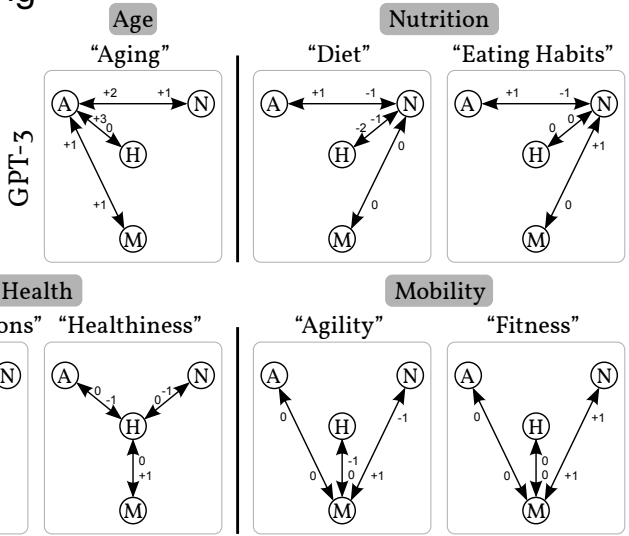


Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Robustness to Variable Wording

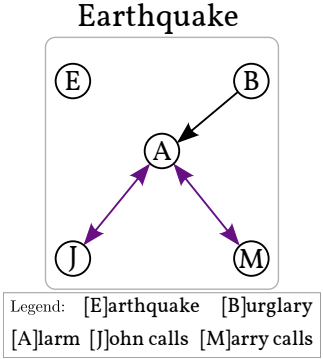
"Does <A> cause <B>?":
LLM answers might be susceptible to variable naming. E.g. “Aging” instead of “Age”.

Different wordings have different implications, or grounding in the training data.



Context and Meta Answers

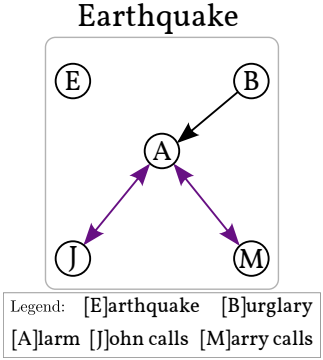
Context matters: *“Does the alarm cause John to call?”*



Zečević*, M., Willig*, M., Dhami, D.S. and Kersting, K., 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. Transactions on Machine Learning Research.

Context and Meta Answers

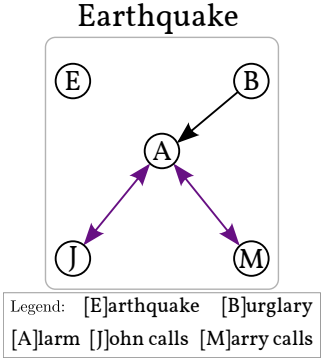
Context matters: *“Does the alarm cause John to call?”*



GPT-4: *“The text does not provide information on whether burglaries cause calls from John.”*

Context and Meta Answers

Context matters: *“Does the alarm cause John to call?”*



LLMs might have the freedom to divert from ‘yes’/‘no’ answers!

GPT-4: *“The text does not provide information on whether burglaries cause calls from John.”*

The State of LLM-based Causal Discovery

Summary: Current LLM-based causal discovery is noisy.

1. LLMs struggle to robustly combine and reason over causal facts.
2. LLMs struggle to truthfully predict causal relations.
3. LLM knowledge varies across domains of application.

The State of LLM-based Causal Discovery

The State of LLM-based Causal Discovery

Possible remedies:

- **Learning to Defer** [1]: Estimate the quality of LLM-based (and data-based) predictions, and dynamically select the more predictive source of information.

[1] Clivio, O., Mahajan, D., Taslakian, P., Magliacane, S., Mitliagkas, I., Zantedeschi, V. and Drouin, A., Learning to Defer for Causal Discovery with Imperfect Experts. In Workshop on Reasoning and Planning for Large Language Models.
[2] Nicholas Tagliapietra, Gian Lorenzo Marchioni, Moritz Willig, Juergen Luettin, Lavdim Halilaj, Kristian Kersting. “CausalSteward: An Agentic Divide-Conquer-Combine Copilot for Causal Discovery”.

The State of LLM-based Causal Discovery

Possible remedies:

- **Learning to Defer** [1]: Estimate the quality of LLM-based (and data-based) predictions, and dynamically select the more predictive source of information.
- **Retrieval Augmented Generation (RAG)** [2]: Retrieve causal knowledge from provided documents, (e.g., scientific articles or system documentation).

[1] Clivio, O., Mahajan, D., Taslakian, P., Magliacane, S., Mitliagkas, I., Zantedeschi, V. and Drouin, A., Learning to Defer for Causal Discovery with Imperfect Experts. In Workshop on Reasoning and Planning for Large Language Models.
[2] Nicholas Tagliapietra, Gian Lorenzo Marchioni, Moritz Willig, Juergen Luettin, Lavdim Halilaj, Kristian Kersting. “CausalSteward: An Agentic Divide-Conquer-Combine Copilot for Causal Discovery”.

The State of LLM-based Causal Discovery

Possible remedies:

- **Learning to Defer** [1]: Estimate the quality of LLM-based (and data-based) predictions, and dynamically select the more predictive source of information.
- **Retrieval Augmented Generation (RAG)** [2]: Retrieve causal knowledge from provided documents, (e.g., scientific articles or system documentation).
- **General Constraints:** Prior methods queried for specific knowledge about edges. Query LLMs for more general constraints and integrate this knowledge.

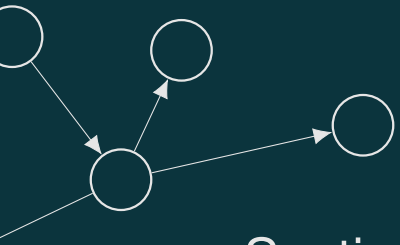
[1] Clivio, O., Mahajan, D., Taslakian, P., Magliacane, S., Mitliagkas, I., Zantedeschi, V. and Drouin, A., Learning to Defer for Causal Discovery with Imperfect Experts. In Workshop on Reasoning and Planning for Large Language Models.
[2] Nicholas Tagliapietra, Gian Lorenzo Marchioni, Moritz Willig, Juergen Luettin, Lavdim Halilaj, Kristian Kersting. “CausalSteward: An Agentic Divide-Conquer-Combine Copilot for Causal Discovery”.

LLMs for improved Causal Discovery

LLMs can be used to place more general priors on relations between variables:

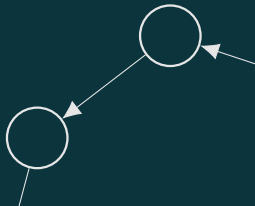
1. **Edge existence**, $X \to Y$: There must be a directed edge (x, y);
2. **Edge forbidden**, $X \not\to Y$: There must not be a directed edge (x, y);
3. **Order constraint**, $x < y$: The variable x occurs before the variable y;
4. **Ancestral constraint**, $x \rightsquigarrow y$: There must be a directed path from the variable x to the variable y.

Ban, T., Chen, L., Lyu, D., Wang, X., Zhu, Q., Tu, Q. and Chen, H., 2025. Integrating large language model for improved causal discovery. IEEE Transactions on Artificial Intelligence.

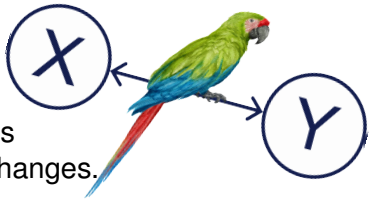


Section 6

Language Enables Reflective Reasoning

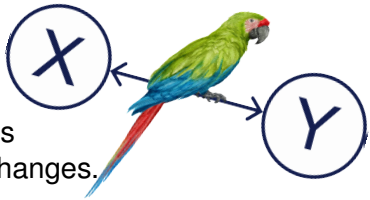


Reflection and Meta Causality



Understanding: Genuine understanding isn't just about knowing that 'A causes B', but understanding the conditions under which that relationship holds, and to adapt when it changes.

Reflection and Meta Causality

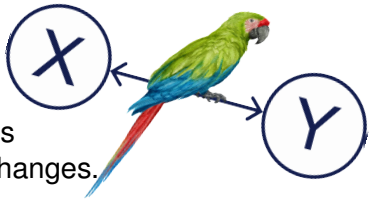


Understanding: Genuine understanding isn't just about knowing that 'A causes B', but understanding the conditions under which that relationship holds, and to adapt when it changes.

Reflection & Adaptation: Genuine AI systems should not just produce due to their intrinsic weights, but deliberately think about the causal mechanisms at play.

→ *"Does a VAE 'know' it is causal?"*

Reflection and Meta Causality

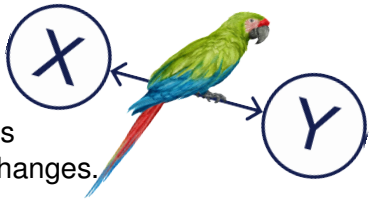


Understanding: Genuine understanding isn't just about knowing that 'A causes B', but understanding the conditions under which that relationship holds, and to adapt when it changes.

Reflection & Adaptation: Genuine AI systems should not just produce due to their intrinsic weights, but deliberately think about the causal mechanisms at play.

- → *"Does a VAE 'know' it is causal?"*
- → *"Does an LLM model 'know' it is causal?"*

Reflection and Meta Causality

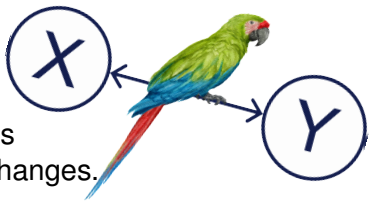


Understanding: Genuine understanding isn't just about knowing that 'A causes B', but understanding the conditions under which that relationship holds, and to adapt when it changes.

Reflection & Adaptation: Genuine AI systems should not just produce due to their intrinsic weights, but deliberately think about the causal mechanisms at play.

- → *"Does a VAE 'know' it is causal?"*
- → *"Does an LLM model 'know' it is causal?"*
- → *"Does an SCM 'know' it is causal?"*

Reflection and Meta Causality



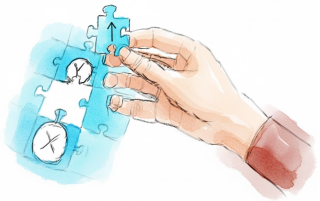
Understanding: Genuine understanding isn't just about knowing that 'A causes B', but understanding the conditions under which that relationship holds, and to adapt when it changes.

Reflection & Adaptation: Genuine AI systems should not just produce due to their intrinsic weights, but deliberately think about the causal mechanisms at play.

→ *"Does a VAE 'know' it is causal?"*

→ *"Does an LLM model 'know' it is causal?"*

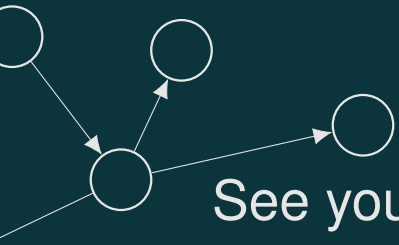
→ *"Does an SCM 'know' it is causal?"*



Reasoning *over* causal facts might be considered a meta-task to classical causal reasoning.

→ Meta-Causal Models (MCM)

(More on that in a separate lecture…)



See you next week!