Beyond Causal Parrots: The Role of Meta-Causality for Genuine Causal Understanding

Moritz Willig
Computer Science Department
Technical University of Darmstadt
moritz.willig@cs.tu-darmstadt.de

Winter School on Causality and Explainable AI, Paris 2025

Thanks to my collaborators:

Kristian Kersting | Devendra S. Dhami | Matej Zecevic | Tim Woydt | Florian Busch | Jonas Seng | Nicholas Tagliapietra

… and many more!

AIML Lab
TECHNISCHE UNIVERSITÄT DARMSTADT
hessian.AI
dfki ai
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
BOSCH

Causal AI

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

- Judea Pearl, 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, Book of Why.

Do AI Models 'Understand' what they are doing?

"Make it a starlit night."

Do AI Models 'Understand' what they are doing?

"Make it a starlit night."

Instruct Pix2Pix

Do AI Models ‘Understand’ what they are doing?

“Turn on the lights”

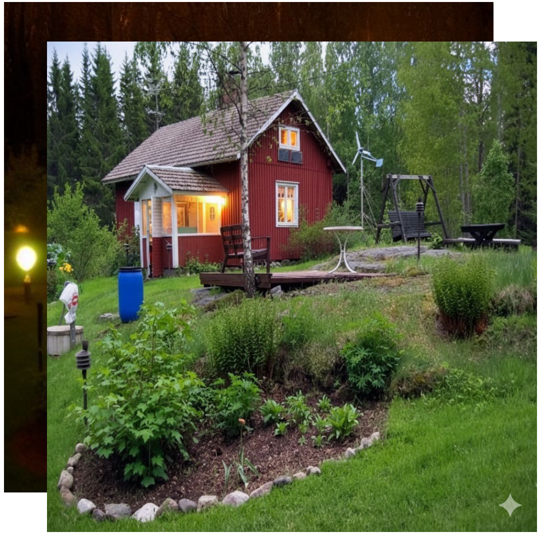
Do AI Models 'Understand' what they are doing?

"Turn on the lights"

Instruct Pix2Pix

Do AI Models ‘Understand’ what they are doing?

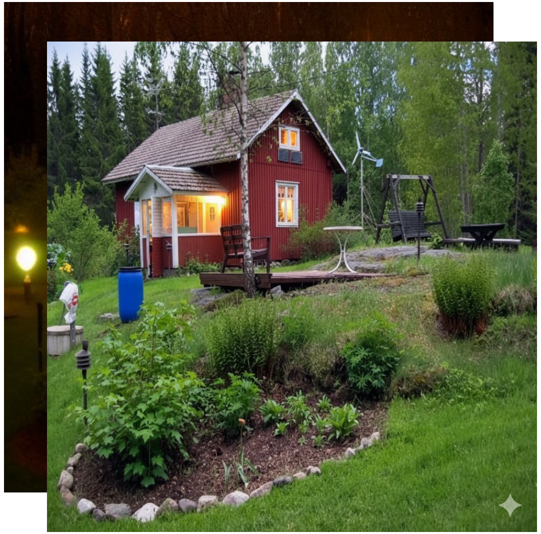
“Turn on the lights”



Gemini-2.5

Do AI Models ‘Understand’ what they are doing?

“Turn on the lights”

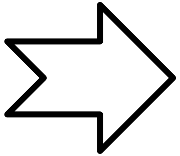


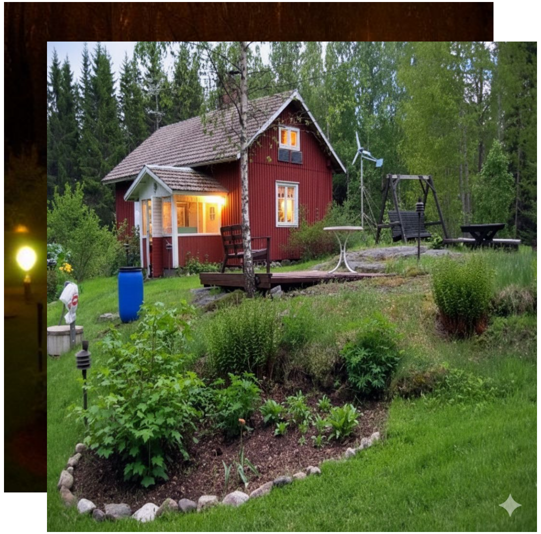
Gemini-2.5

Models unfold according to their inherent structure.

Do AI Models ‘Understand’ what they are doing?

“Turn on the lights”





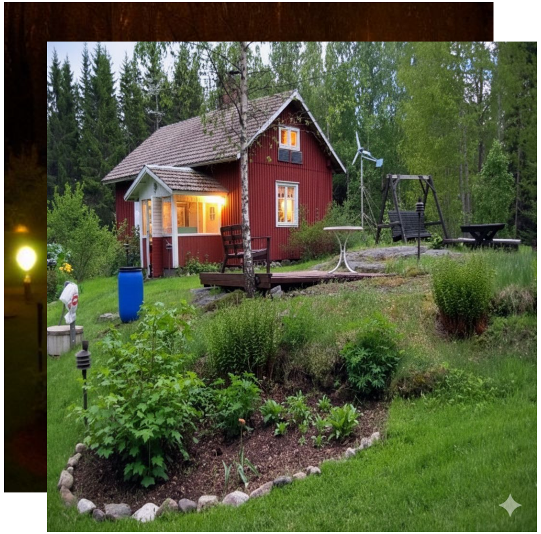
Gemini-2.5

“Does a diffusion model ‘know’ it is causal?”

Models unfold according to their inherent structure.

Do AI Models 'Understand' what they are doing?

"Turn on the lights"



Gemini-2.5

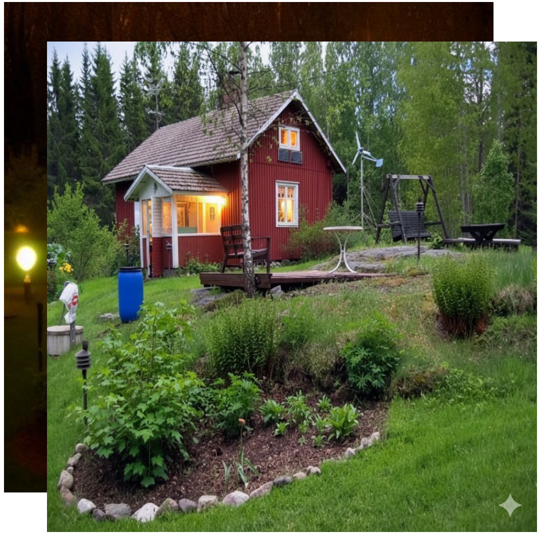
"Does a diffusion model 'know' it is causal?"

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

Models unfold according to their inherent structure.

Do AI Models ‘Understand’ what they are doing?

“Turn on the lights”



Gemini-2.5

Models unfold according to their inherent structure.

“Does a diffusion model ‘know’ it is causal?”

“Does an LLM model ‘know’ it is causal?”

“Does an SCM ‘know’ it is causal?”

Causal Representation Learning

- Learn causal concepts from high-dimensional data.
  - Requires on interventions or sufficient variation in data.
- Guarantees for structuring models according to underlying process.

"**Toward causal representation learning.**"
Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. *Proceedings of the IEEE* 2021

"**Weakly supervised causal representation learning.**"
Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S. Cohen. NeurIPS 2022

"**Learning temporally causal latent processes from general temporal data.**"
Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. ICLR 2022

"**Robust agents learn causal world models.**" Jonathan Richens and Tom Everitt. ICLR 2024

…

Causal Representation Learning

- Learn causal concepts from high-dimensional data.
  - Requires on interventions or sufficient variation in data.
- Guarantees for structuring models according to underlying process.
- …, but no reflection.

“**Toward causal representation learning.**”
Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. *Proceedings of the IEEE* 2021

“**Weakly supervised causal representation learning.**”
Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S. Cohen. NeurIPS 2022

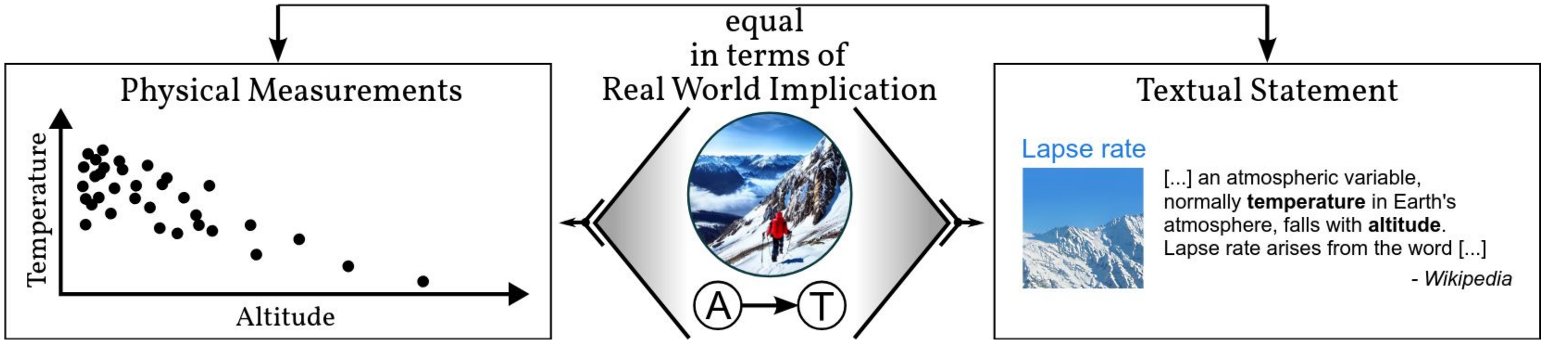
“**Learning temporally causal latent processes from general temporal data.**”
Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. ICLR 2022

“**Robust agents learn causal world models.**” Jonathan Richens and Tom Everitt. ICLR 2024

…

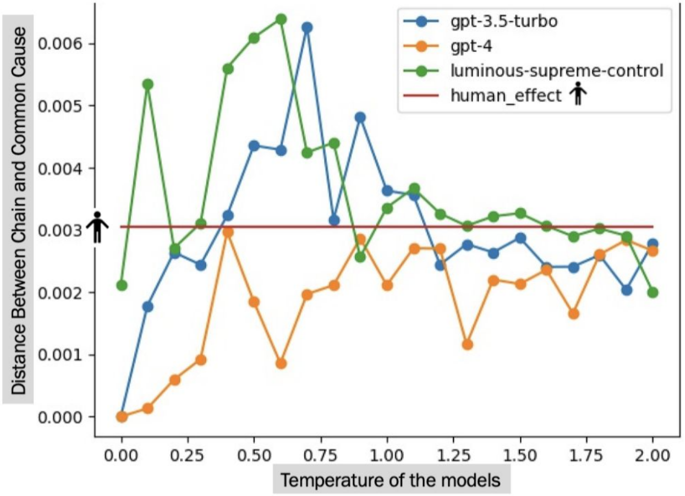
Natural Language Data as an Opportunity

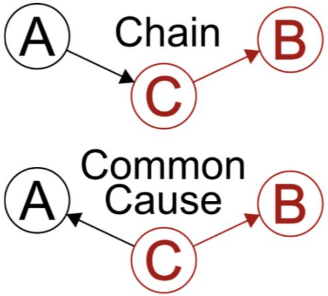
Natural language allows for the explicit representation of causal facts.



"Causal Parrots: Large Language Models May Talk Causality But Are Not Causal."
Matej Zečević*, Moritz Willig*, Devendra Singh Dhami and Kristian Kersting. Transactions on Machine Learning Research. 2023

LLMs adopt Human Biases in Causal Perception





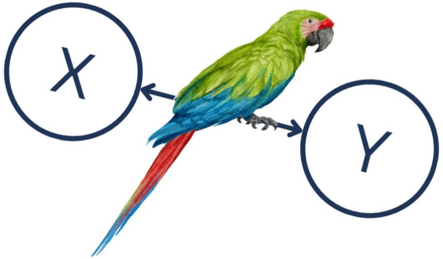
$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} | \mathbf{C}$$

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

"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."

"**Chain versus common cause: Biased causal strength judgments in humans and large language models**"
Anita Keshmirian, Moritz Willig, Babak Hemmatian, Kristian Kersting, Ulrike Hahn and Tobias Gerstenberg. CogSci 2024

Genuinely Causal or Causal Parrots?



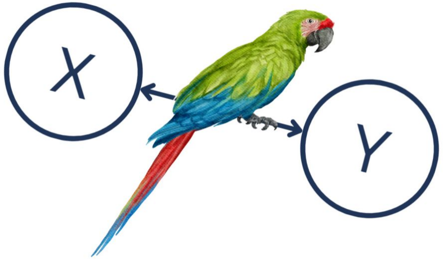
LLMs have no real-world interactions during training.

Can they can excel beyond the first rung of the causal ladder?

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

"**Causal Parrots: Large Language Models May Talk Causality But Are Not Causal.**"
Matej Zečević*, Moritz Willig*, Devendra Singh Dhami and Kristian Kersting. Transactions on Machine Learning Research. 2023

Genuinely Causal or Causal Parrots?



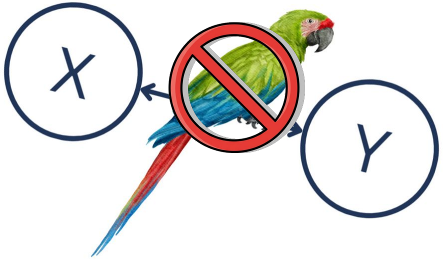
LLMs have no real-world interactions during training.
Can they can excel beyond the first rung of the causal ladder?

…they can free themselves through deliberate reasoning.

	Causal Chains (Basic Prop. Logic)									Subchains (4)	Randomized (7)	Accuracy
	N=2	3	4	5	6	7	8	9	10			
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% *

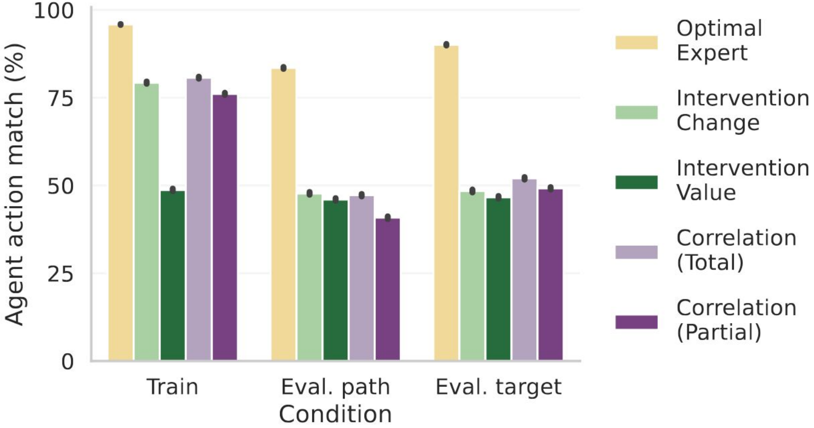
"Causal Parrots: Large Language Models May Talk Causality But Are Not Causal."
Matej Zečević*, Moritz Willig*, Devendra Singh Dhami and Kristian Kersting. Transactions on Machine Learning Research. 2023

Reasoning beyond the first Rung



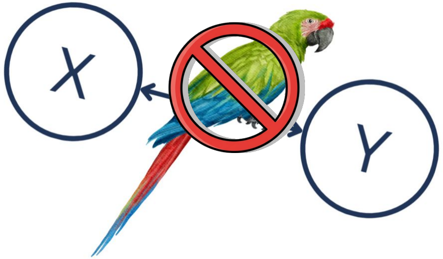
Natural Language contains information *about* interventions.

Lampinen et al. showed that observing experts' interventions plus explanations can suffice to acquire generalizable strategies.



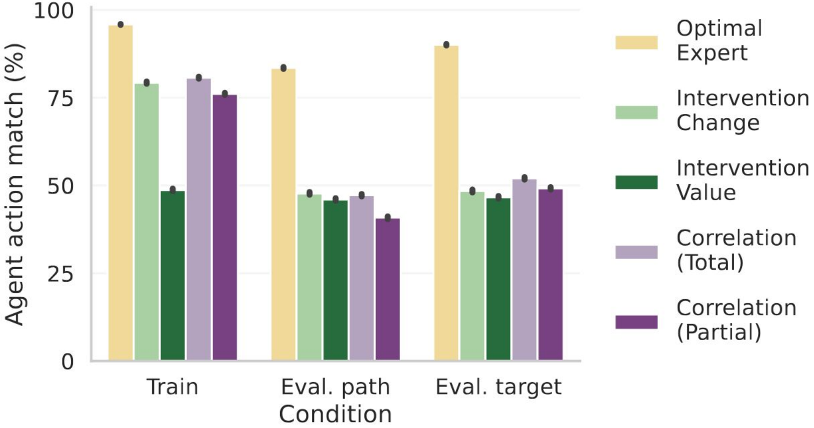
"Passive learning of active causal strategies in agents and language models"
Andrew Lampinen, Stephanie Chan, Ishita Dasgupta, Andrew Nam and Jane Wang. NeurIPS 2023.

Reasoning beyond the first Rung



Natural Language contains information *about* interventions.

Lampinen et al. showed that observing experts' interventions plus explanations can suffice to acquire generalizable strategies.

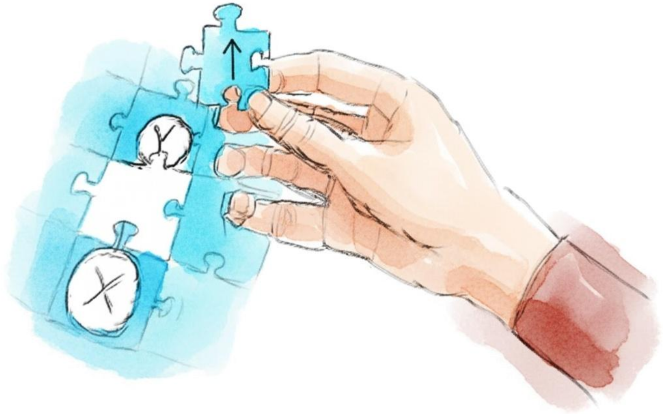


Language models can adapt to reason *over* causal relations.

"Passive learning of active causal strategies in agents and language models"
Andrew Lampinen, Stephanie Chan, Ishita Dasgupta, Andrew Nam and Jane Wang. NeurIPS 2023.

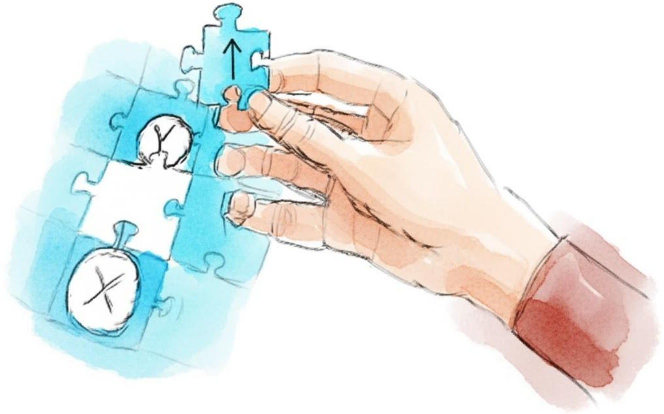
Meta-Causality

We would like to have a framework that allows general AI/ML models to piece together and manipulate causal relations.



Meta-Causality

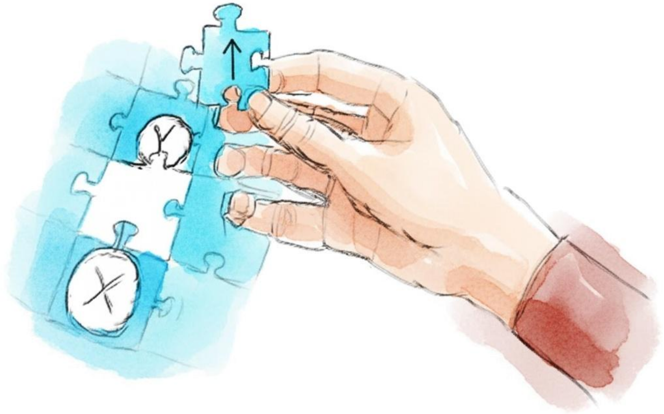
We would like to have a framework that allows general AI/ML models to piece together and manipulate causal relations.



- Predict under which conditions causal edges emerge and vanish.

Meta-Causality

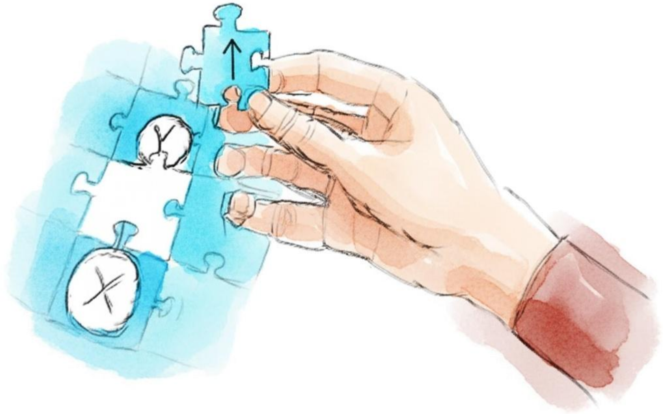
We would like to have a framework that allows general AI/ML models to piece together and manipulate causal relations.



- Predict under which conditions causal edges emerge and vanish.
- Reason over system dynamics.

Meta-Causality

We would like to have a framework that allows general AI/ML models to piece together and manipulate causal relations.



- Predict under which conditions causal edges emerge and vanish.
- Reason over system dynamics.
- Attribution beyond static root-causes, but for the existence of causal links themselves.

Meta-Causal Models

Meta-Causal Models are a novel framework designed to explicitly model and reason about the emergence and change of causal relationships.

abstract away from structural equations

Meta-Causal Models capture <u>qualitative changes</u> in cause-effect relations.

reason over the presence of causal relations themselves.

Inherently reflective w.r.t. the underlying SCM.

Meta-Causal Models

For an **underlying process** with state transitions $\sigma : \mathcal{S} \rightarrow \mathcal{S}$, we have a causal abstraction $\varphi : \mathcal{S} \rightarrow \boldsymbol{\mathcal{X}}$.

Meta-Causal Models consider the **functional type** of structural equations:

$$T_{s,ij} := \tau_{ij}(\varphi(s), \varphi \circ \sigma)$$

Meta-Causal States (MCS) are type matrices: $T \in \mathcal{T}^{N \times N}$

Meta-Causal Models (MCM) model transitions between states:

$$\delta : \mathcal{T}^{N \times N} \times \mathcal{S} \rightarrow \mathcal{T}^{N \times N}$$

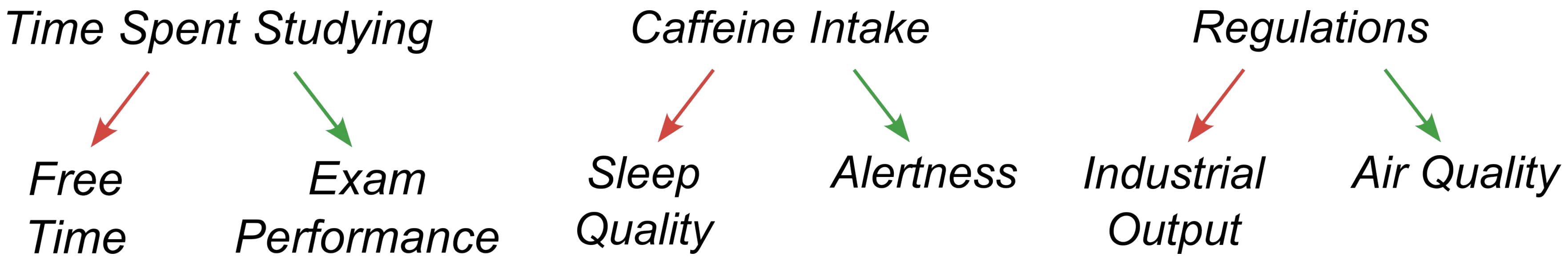
"**Systems with Switching Causal Relations: A Meta-Causal Perspective**", Moritz Willig, Tim Nelson Tobiasch, Florian Peter Busch, Jonas Seng, Devendra Singh Dhami, Kristian Kersting. ICLR 2025

Functional Types

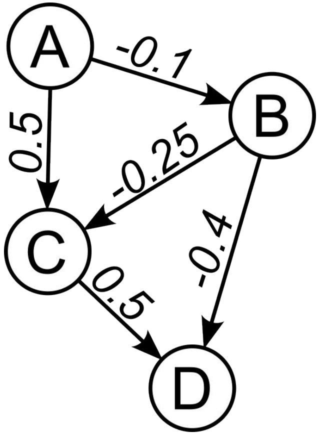
- Abstract away from specific structural equations.
- Consider qualitative edge types. E.g. ‘*suppressing*’, ‘*reinforcing*’, …

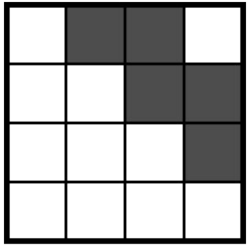
Functional Types

- Abstract away from specific structural equations.
- Consider qualitative edge types. E.g. ‘*suppressing*’, ‘*reinforcing*’, …

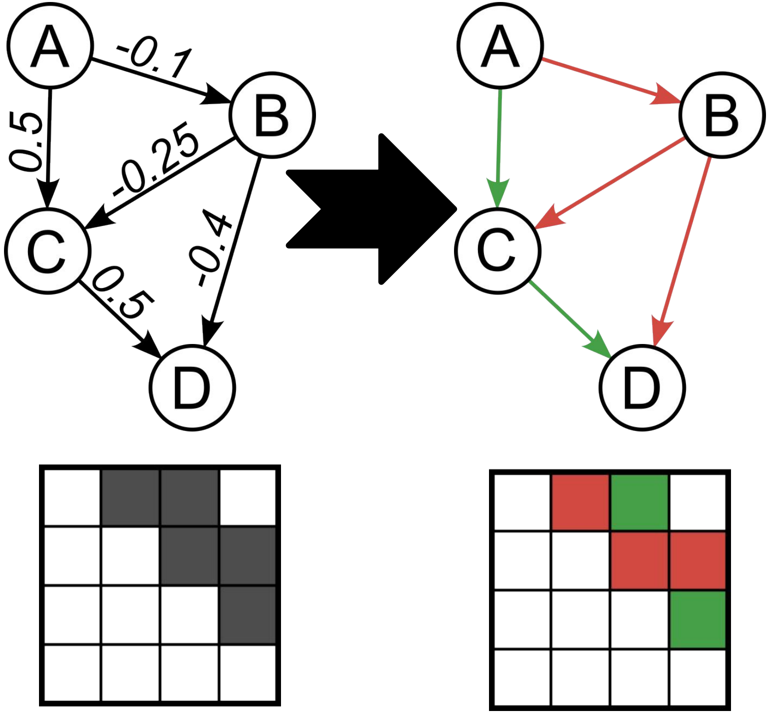


Functional Types

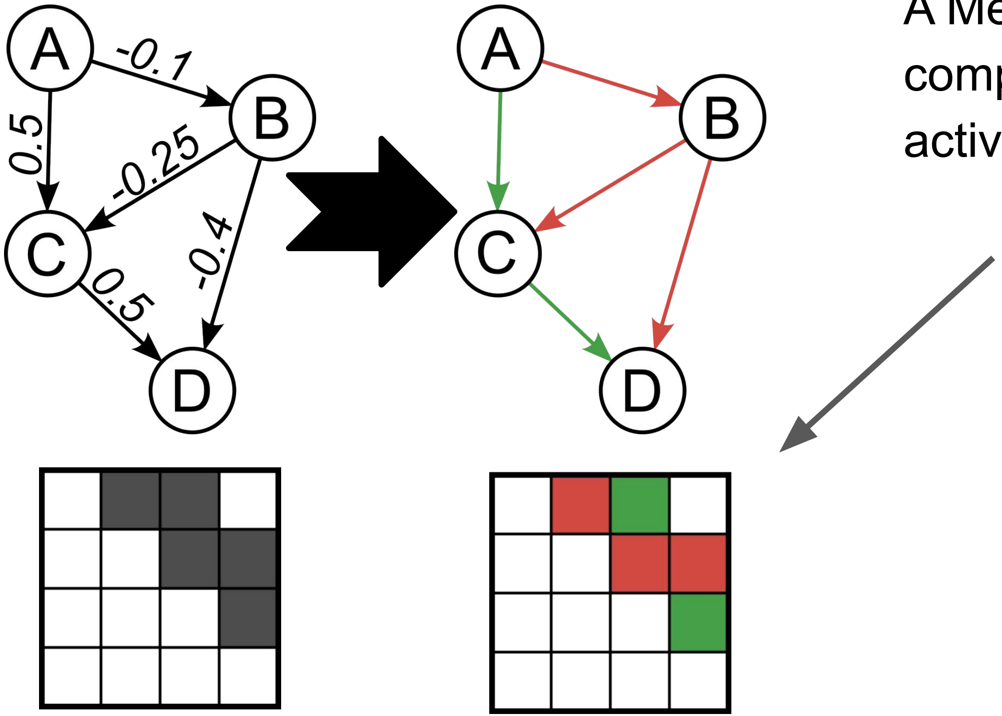




Functional Types



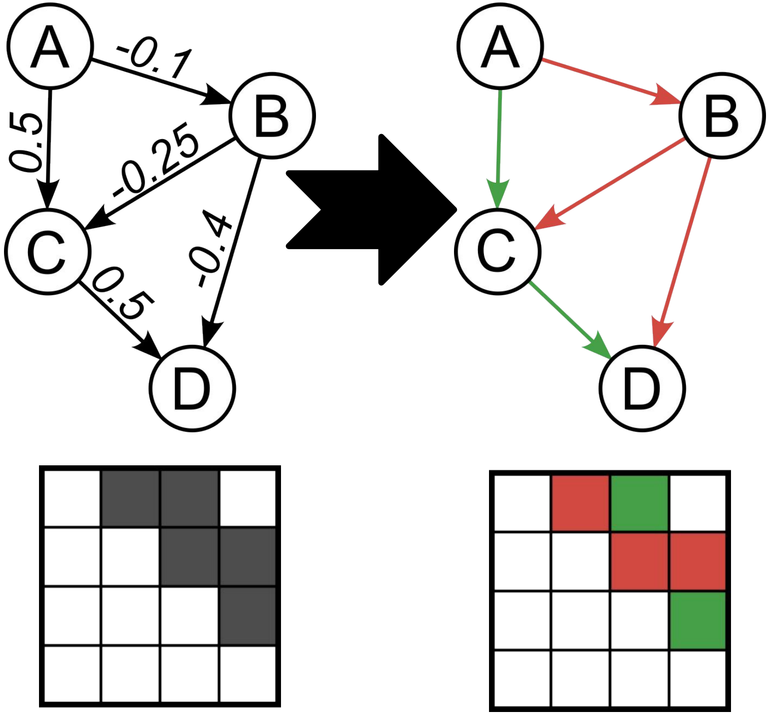
Functional Types



A Meta-Causal State is composed of the currently active types of all edges:

$$T \in \mathcal{T}^{N \times N}$$

Functional Types



A Meta-Causal State is composed of the currently active types of all edges:

$$T \in \mathcal{T}^{N \times N}$$

“But why do we need all of this formalism for the type encoder?”:

structural equations

$$T_{s,ij} := \tau_{ij}(\varphi(s), \varphi \circ \sigma)$$

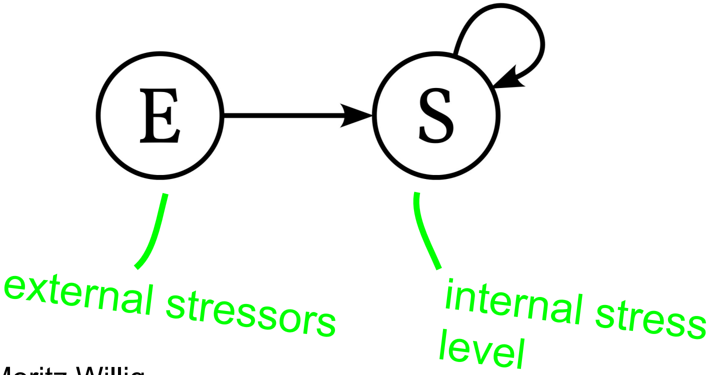
type encoder

system state

Dynamic Switching of Types

So far, we considered static graphs…

Self-reinforcing Stress Example

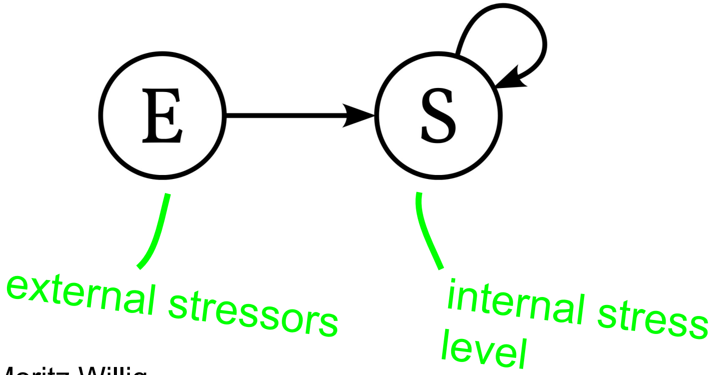


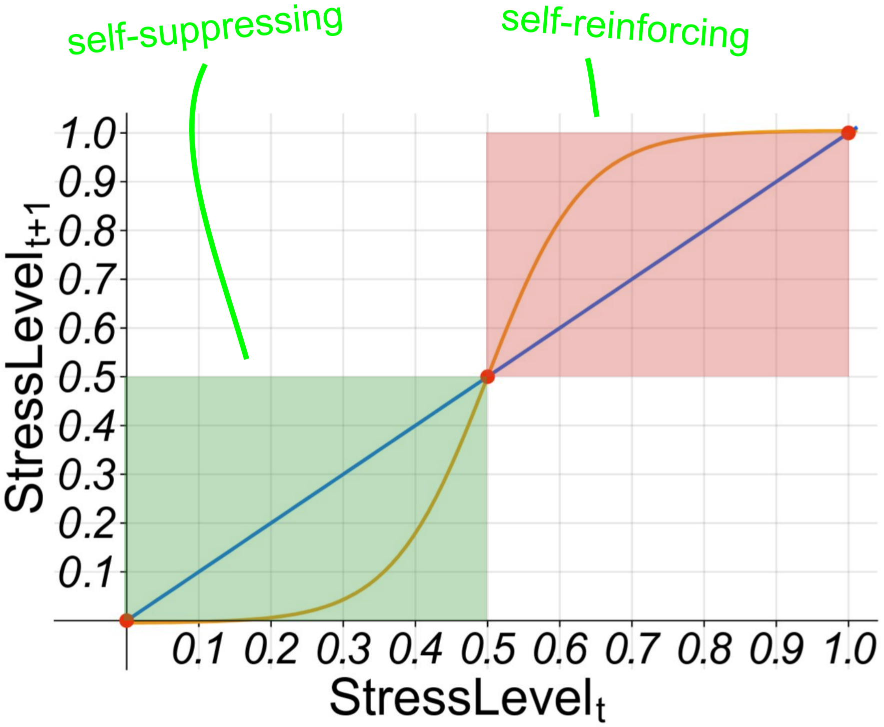
The influence of Internal Stress on itself across two consecutive days.

Dynamic Switching of Types

So far, we considered static graphs…

Self-reinforcing Stress Example





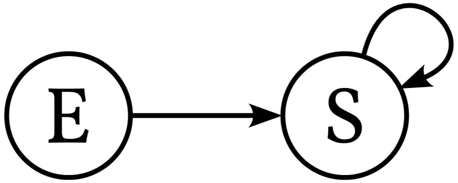
The influence of Internal Stress on itself across two consecutive days.

Dynamic Switching of Types

So far, we considered static graphs…

Self-reinforcing Stress Example

Same structural equation, but changing relation type

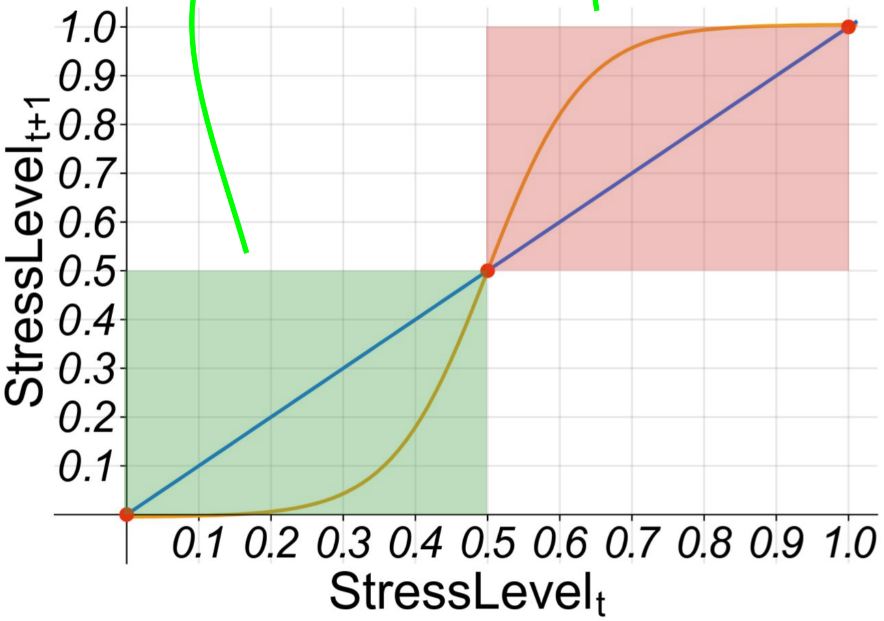


external stressors

internal stress level

self-suppressing

self-reinforcing



The influence of Internal Stress on itself across two consecutive days.

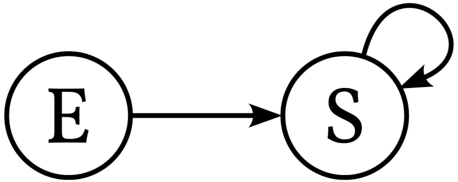
Dynamic Switching of Types

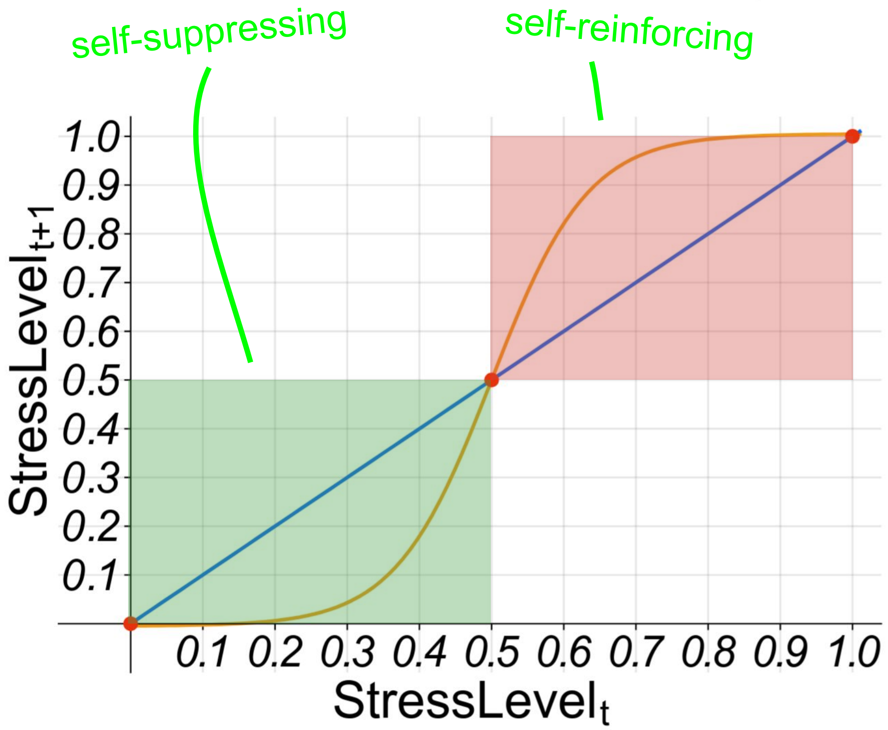
So far, we considered static graphs…

Self-reinforcing Stress

Example

Same structural equation, but changing relation type





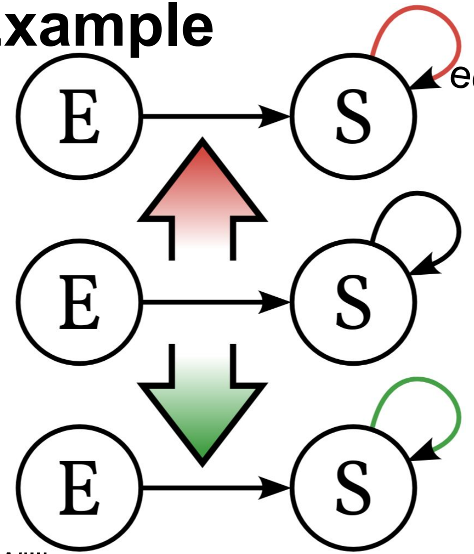
$$T_{\mathbf{x}} := \begin{bmatrix} 0 & 1 \\ 0 & \alpha \end{bmatrix} \text{ with } \alpha := \operatorname{sign}(s - 0.5)$$

Dynamic Switching of Types

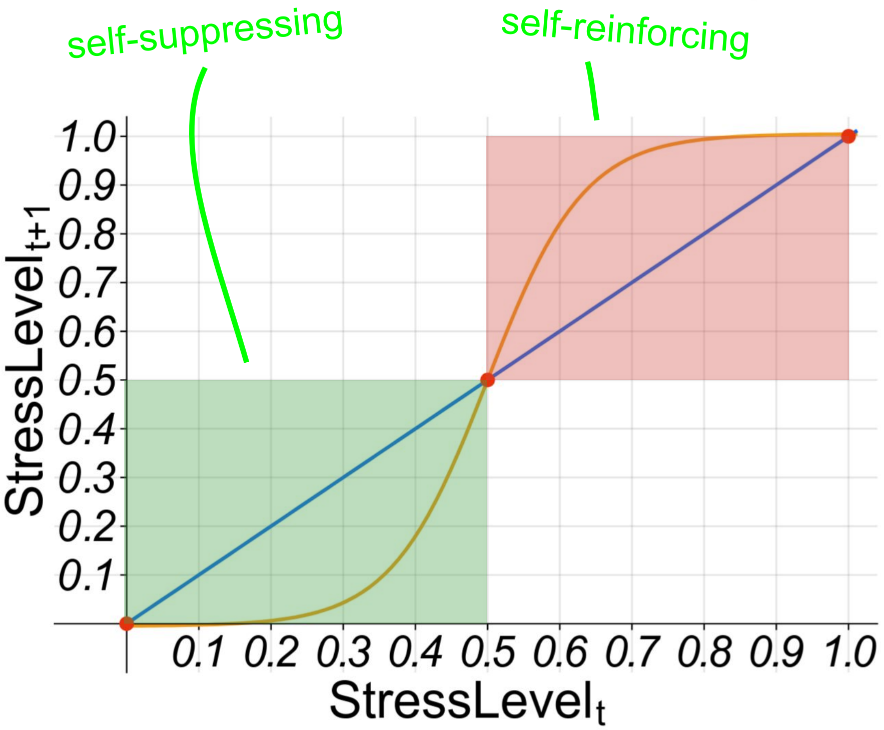
So far, we considered static graphs…

Self-reinforcing Stress

Example

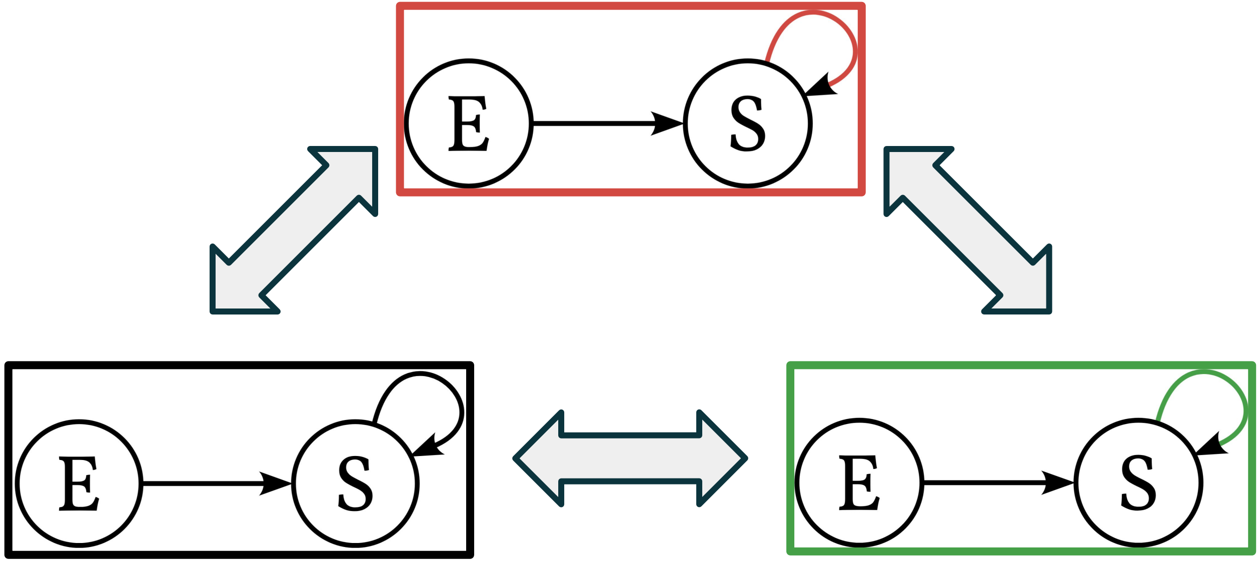


Same structural equation, but changing relation type



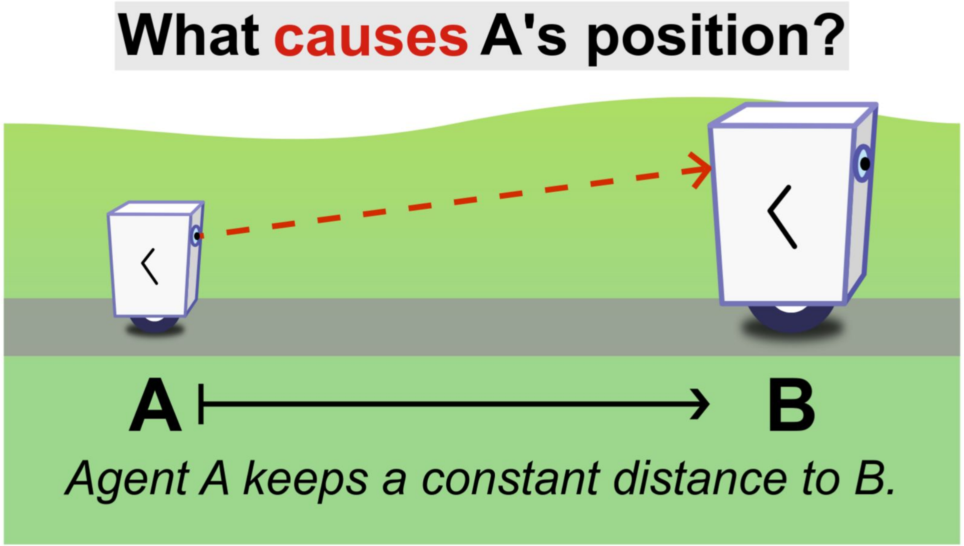
$$T_{\mathbf{x}} := \begin{bmatrix} 0 & 1 \\ 0 & \alpha \end{bmatrix} \text{ with } \alpha := \operatorname{sign}(s - 0.5)$$

Meta-Causal Models

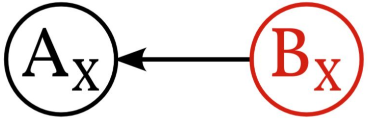


MCM model state transitions: $\delta : \mathcal{T}^{N \times N} \times \mathcal{S} \rightarrow \mathcal{T}^{N \times N}$

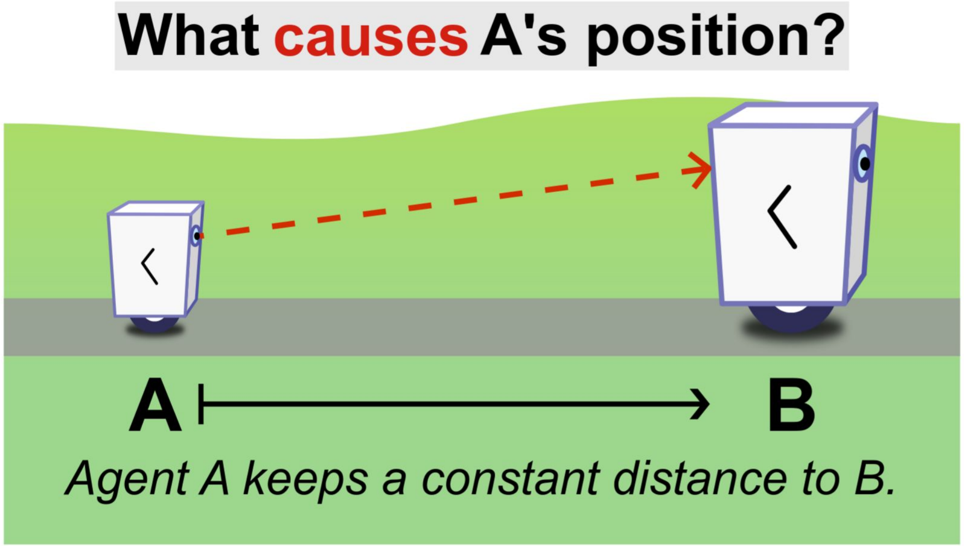
Meta-Causal Attribution



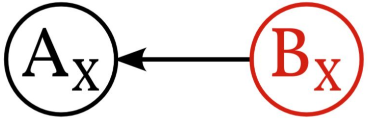
Meta-Causal Attribution



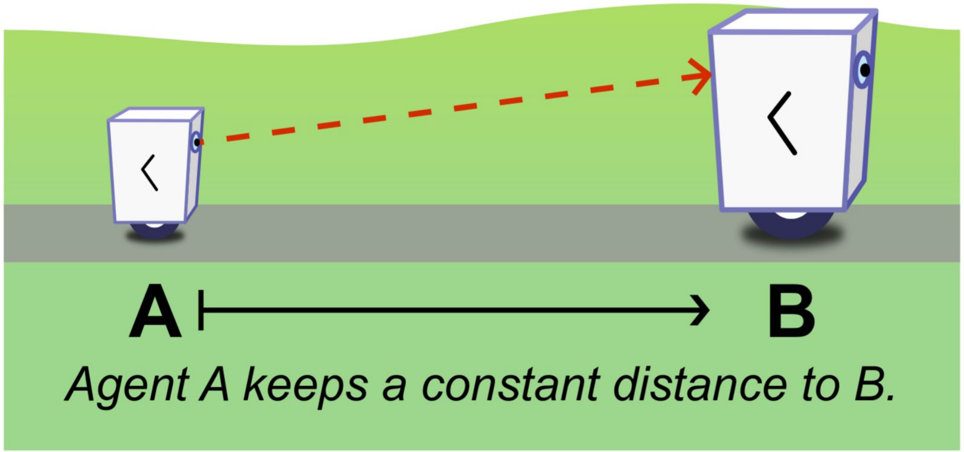
Classical Attribution
A_X is caused by the structural equation $A_X := f(B_X)$.



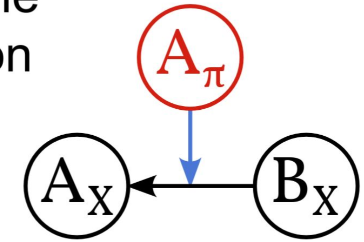
Meta-Causal Attribution



What causes A's position?

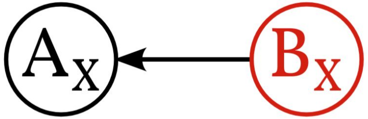


Classical Attribution
A_X is caused by the structural equation $A_X := f(B_X)$.

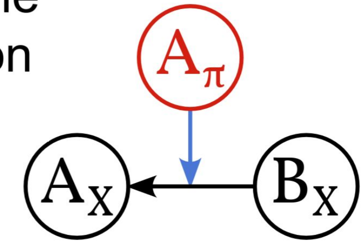


Meta-Causal Attribution
But the relation $B_X \rightarrow A_X$ only *exists* due to A's policy A_π.

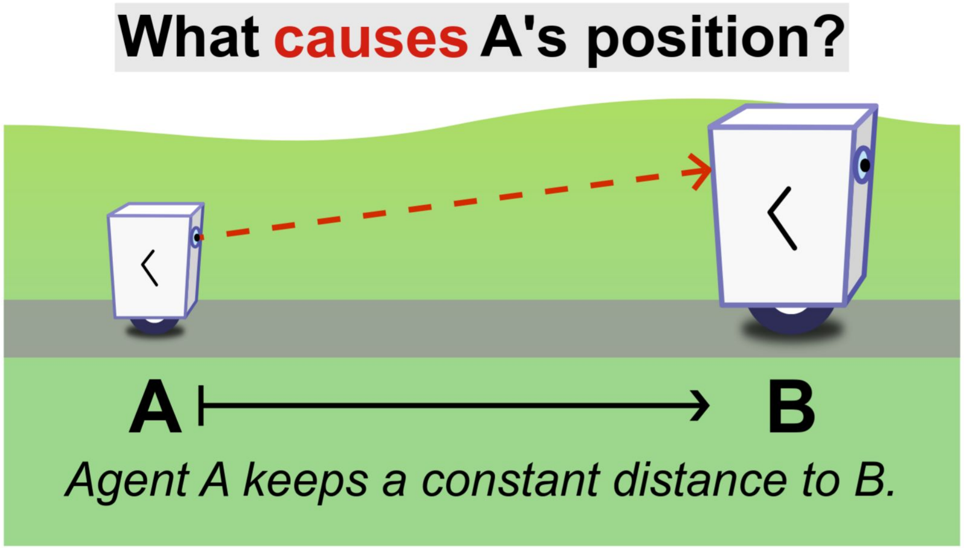
Meta-Causal Attribution



Classical Attribution
A_X is caused by the structural equation $A_X := f(B_X)$.



Meta-Causal Attribution
But the relation $B_X \rightarrow A_X$ only *exists* due to A's policy A_π.

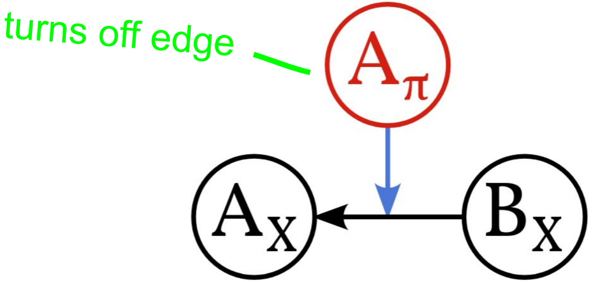


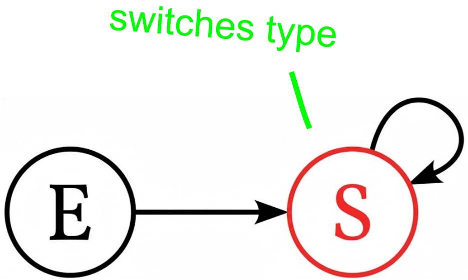
Meta-Causality consider factors that lead to the emergence of edges.

Meta-Causal Variables (MCVs)

MCVs are the factors that lead to switching type relations:

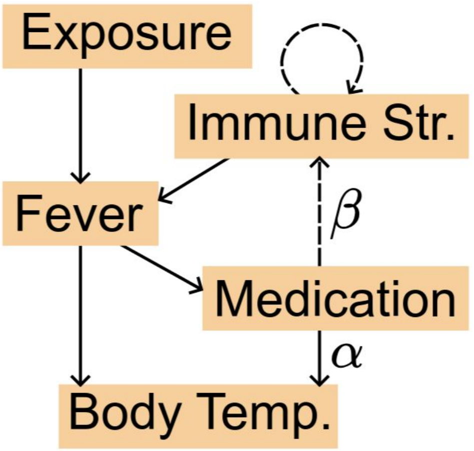
$$\mathbf{C} := \{\, \mathrm{X}_k \in \mathbf{X} \mid \exists\, \mathrm{X}_i, \mathrm{X}_j \in \mathbf{X} \,.\exists\, \mathbf{x}, \mathbf{x}' \in \boldsymbol{\mathcal{X}} \text{ s.t.}$$
$$(\mathbf{x}_{\bar{k}} = \mathbf{x}'_{\bar{k}}) \wedge (x_k \neq x'_k) \wedge (\mathcal{I}(\mathbf{x}, \mathrm{X}_i, \mathrm{X}_j) \neq \mathcal{I}(\mathbf{x}', \mathrm{X}_i, \mathrm{X}_j))\}$$





“**When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions**”,
Moritz Willig, Tim Woydt, Devendra Singh Dhami, Kristian Kersting. NeurIPS 2025

Medication MCA



Compare two medications

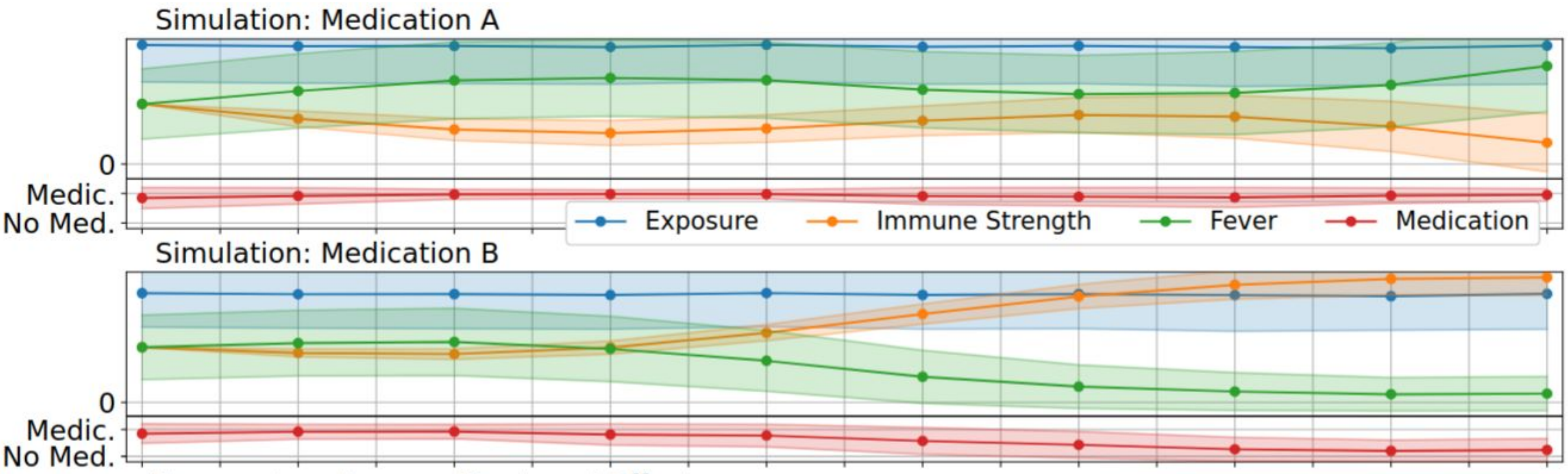
A: High direct impact, suppresses immune development.

B: Lower direct impact, lower immune suppression.

Disclaimer: highly simplified. Assumption: Both drugs are assumed to be equally suited to treat fever.

“**When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions**”, Moritz Willig, Tim Woydt, Devendra Singh Dhami, Kristian Kersting. NeurIPS 2025

Medication MCA



“**When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions**”,
Moritz Willig, Tim Woydt, Devendra Singh Dhami, Kristian Kersting. NeurIPS 2025

Meta-Causal Analysis

Similar to how causal effects quantify the influence between variables, **Meta-Causal Effects** quantify changes in the state transitions.

Questions answered by MCA:

- What is the **probability of** a system to **adapt** a desired MCS?
- How **stable** is a particular MCS?
- Which **transition pathways** can be taken to obtain a particular MCS?

“**When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions**”, Moritz Willig, Tim Woydt, Devendra Singh Dhami, Kristian Kersting. NeurIPS 2025

LMCD

1) Start with some data

Algorithm 1 Linearized Meta-Causal Dynamics (LMCD) Algorithm

Input: SCM: $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, P_{\mathbf{U}})$, data: $\mathbf{x}^{\mathbf{I}} = (\mathbf{x}^i)_{i=1}^N \in \mathbf{X}^N$, id. func.: $\mathcal{I} : \mathbf{X} \to \mathrm{T}$

for each $\mathbf{x}^i$ in $\mathbf{x}^{\mathbf{I}}$ **do**

  $\mathbf{x}^{i,t+1} \leftarrow \mathbf{F}((\mathbf{x}^i\,|_{\mathbf{V}}) \cup (\mathbf{u}^{t+1} \sim P_{\mathbf{U}}))$ ▷ Advance the system.

  $(\mathrm{T}^{i,t}, \mathrm{T}^{i,t+1}) \leftarrow (\mathcal{I}(\mathbf{x}^i), \mathcal{I}(\mathbf{x}^{i,t+1}))$ ▷ Identify MCS transition pair.

$U \leftarrow (\bigcup_i l(\mathrm{T}^{i,t})) \cup (\bigcup_i l(\mathrm{T}^{i,t+1}))$ ▷ Determine set of unique MCS.

for each (u, v) in $\{1, \dots, |U|\}^2$ **do** ▷ Approximate transition dynamics, $P \in \mathbb{R}^{|U| \times |U|}$.

  $P_{u,v} \leftarrow \sum_{i \in [1..N]} (\mathbf{1}(((l(\mathrm{T}^{i,t}) = u) \wedge (l(\mathrm{T}^{i,t+1}) = v)))/ \sum_{i \in [1..N]} \mathbf{1}(l(\mathrm{T}^{i,t} = v))$

$[Q \leftarrow e^{P-I}]$ ▷ Optional: Compute continuous time rate matrix. (I is the identity matrix.)

return $P, [Q]$

"**When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions**",
Moritz Willig, Tim Woydt, Devendra Singh Dhami, Kristian Kersting. NeurIPS 2025

LMCD

2) Identify the state of the system

Algorithm 1 Linearized Meta-Causal Dynamics (LMCD) Algorithm

Input: SCM: $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, P_{\mathbf{U}})$, data: $\underline{\mathbf{x}^{\mathbf{I}} = (\mathbf{x}^i)_{i=1}^N} \in \mathbf{X}^N$, id. func.: $\mathcal{I} : \mathbf{X} \to \mathrm{T}$
for each $\mathbf{x}^i$ in $\mathbf{x}^{\mathbf{I}}$ **do**
  $\mathbf{x}^{i,t+1} \leftarrow \mathbf{F}((\mathbf{x}^i\,|_{\mathbf{V}}) \cup (\mathbf{u}^{t+1} \sim P_{\mathbf{U}}))$ ▷ Advance the system.
  $(\mathrm{T}^{i,t}, \mathrm{T}^{i,t+1}) \leftarrow (\underline{\mathcal{I}(\mathbf{x}^i)}, \mathcal{I}(\mathbf{x}^{i,t+1}))$ ▷ Identify MCS transition pair.
$U \leftarrow (\bigcup_i l(\mathrm{T}^{i,t})) \cup (\bigcup_i l(\mathrm{T}^{i,t+1}))$ ▷ Determine set of unique MCS.
for each (u, v) in $\{1, \ldots, |U|\}^2$ **do** ▷ Approximate transition dynamics, $P \in \mathbb{R}^{|U| \times |U|}$.
  $P_{u,v} \leftarrow \sum_{i \in [1..N]} (\mathbf{1}(((l(\mathrm{T}^{i,t}) = u) \wedge (l(\mathrm{T}^{i,t+1}) = v))) / \sum_{i \in [1..N]} \mathbf{1}(l(\mathrm{T}^{i,t} = v))$
$[Q \leftarrow e^{P-I}]$ ▷ Optional: Compute continuous time rate matrix. (I is the identity matrix.)
return $P, [Q]$

“**When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions**”,
Moritz Willig, Tim Woydt, Devendra Singh Dhami, Kristian Kersting. NeurIPS 2025

LMCD

3) Advance the system and identify MCS, again.

Algorithm 1 Linearized Meta-Causal Dynamics (LMCD) Algorithm

Input: SCM: $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, P_{\mathbf{U}})$, data: $\mathbf{x}^{\mathbf{I}} = (\mathbf{x}^i)_{i=1}^N \in \mathbf{X}^N$, id. func.: $\mathcal{I} : \mathbf{X} \rightarrow \mathrm{T}$

for each $\mathbf{x}^i$ in $\mathbf{x}^{\mathbf{I}}$ **do**

  $\mathbf{x}^{i,t+1} \leftarrow \mathbf{F}((\mathbf{x}^i|_{\mathbf{V}}) \cup (\mathbf{u}^{t+1} \sim P_{\mathbf{U}}))$ ▷ Advance the system.

  $(\mathrm{T}^{i,t}, \mathrm{T}^{i,t+1}) \leftarrow (\mathcal{I}(\mathbf{x}^i), \mathcal{I}(\mathbf{x}^{i,t+1}))$ ▷ Identify MCS transition pair.

$U \leftarrow (\bigcup_i l(\mathrm{T}^{i,t})) \cup (\bigcup_i l(\mathrm{T}^{i,t+1}))$ ▷ Determine set of unique MCS.

for each (u, v) in $\{1, \ldots, |U|\}^2$ **do** ▷ Approximate transition dynamics, $P \in \mathbb{R}^{|U|\times|U|}$.

  $P_{u,v} \leftarrow \sum_{i\in[1..N]}(\mathbf{1}(((l(\mathrm{T}^{i,t}) = u) \wedge (l(\mathrm{T}^{i,t+1}) = v)))/\sum_{i\in[1..N]} \mathbf{1}(l(\mathrm{T}^{i,t} = v))$

$[Q \leftarrow e^{P-I}]$ ▷ Optional: Compute continuous time rate matrix. (I is the identity matrix.)

return $P, [Q]$

"**When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions**",
Moritz Willig, Tim Woydt, Devendra Singh Dhami, Kristian Kersting. NeurIPS 2025

LMCD

3) Compute transition dynamics.

Algorithm 1 Linearized Meta-Causal Dynamics (LMCD) Algorithm

Input: SCM: $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, P_{\mathbf{U}})$, data: $\mathbf{x}^{\mathbf{I}} = (\mathbf{x}^i)_{i=1}^N \in \mathbf{X}^N$, id. func.: $\mathcal{I} : \mathbf{X} \to \mathrm{T}$

for each $\mathbf{x}^i$ in $\mathbf{x}^{\mathbf{I}}$ **do**

    $\mathbf{x}^{i,t+1} \leftarrow \mathbf{F}((\mathbf{x}^i|_{\mathbf{V}}) \cup (\mathbf{u}^{t+1} \sim P_{\mathbf{U}}))$ ▷ Advance the system.

    $(\mathrm{T}^{i,t}, \mathrm{T}^{i,t+1}) \leftarrow (\mathcal{I}(\mathbf{x}^i), \mathcal{I}(\mathbf{x}^{i,t+1}))$ ▷ Identify MCS transition pair.

$U \leftarrow (\bigcup_i l(\mathrm{T}^{i,t})) \cup (\bigcup_i l(\mathrm{T}^{i,t+1}))$ ▷ Determine set of unique MCS.

for each (u, v) in $\{1, \ldots, |U|\}^2$ **do** ▷ Approximate transition dynamics, $P \in \mathbb{R}^{|U|\times|U|}$.

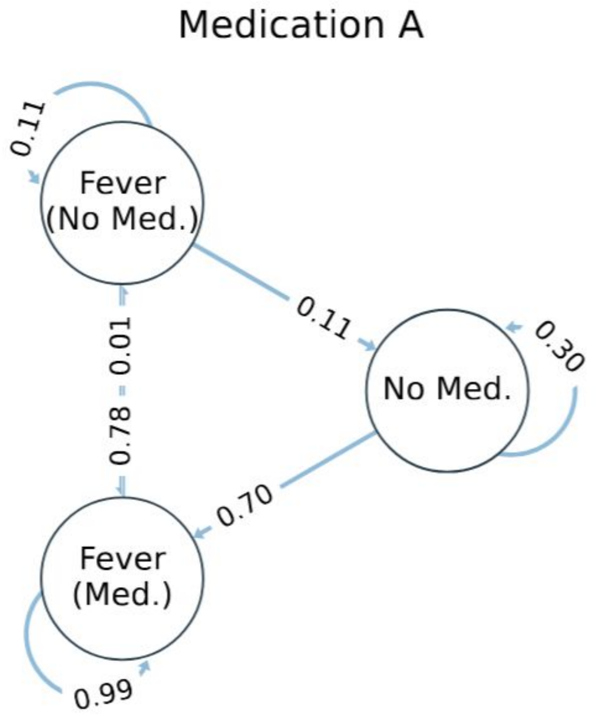
    $P_{u,v} \leftarrow \sum_{i\in[1..N]}(\mathbf{1}(((l(\mathrm{T}^{i,t}) = u) \wedge (l(\mathrm{T}^{i,t+1}) = v)))/\sum_{i\in[1..N]} \mathbf{1}(l(\mathrm{T}^{i,t} = v))$

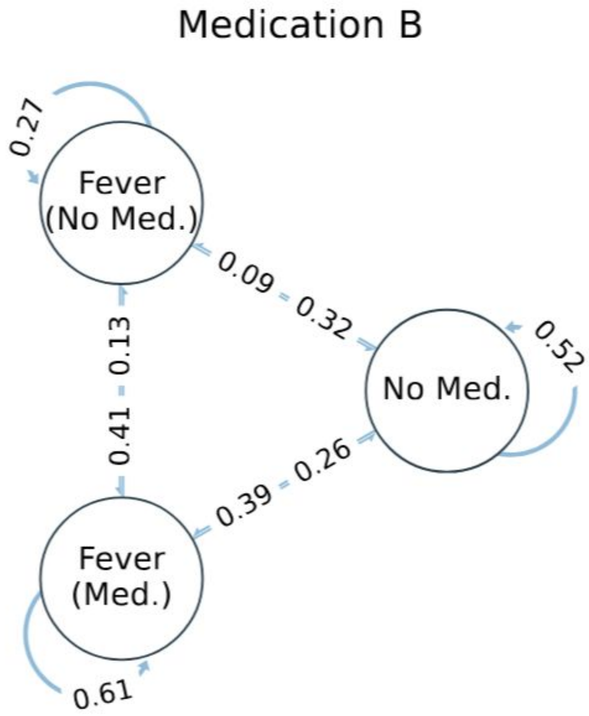
$[Q \leftarrow e^{P-I}]$ ▷ Optional: Compute continuous time rate matrix. (I is the identity matrix.)

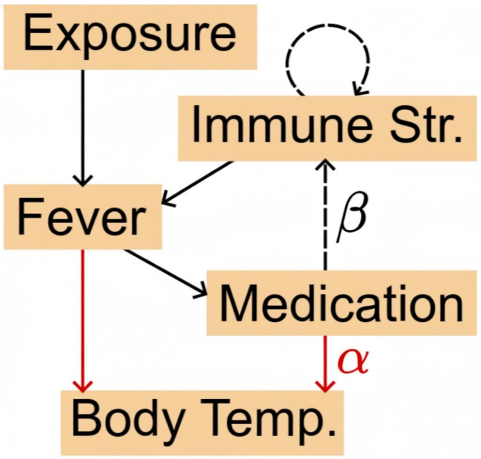
return P, $[Q]$

"**When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions**",
Moritz Willig, Tim Woydt, Devendra Singh Dhami, Kristian Kersting. NeurIPS 2025

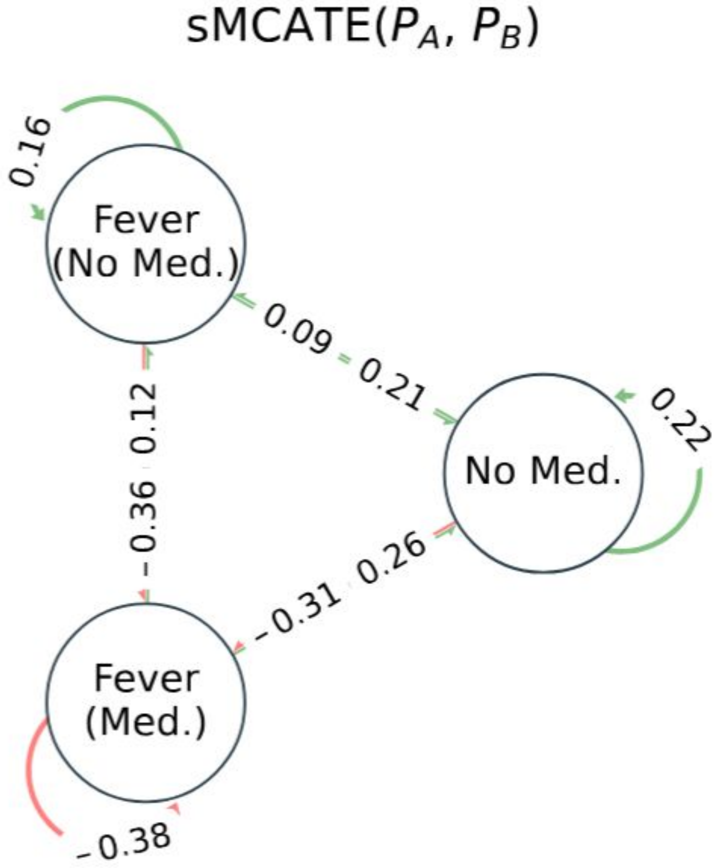
Medication MCM





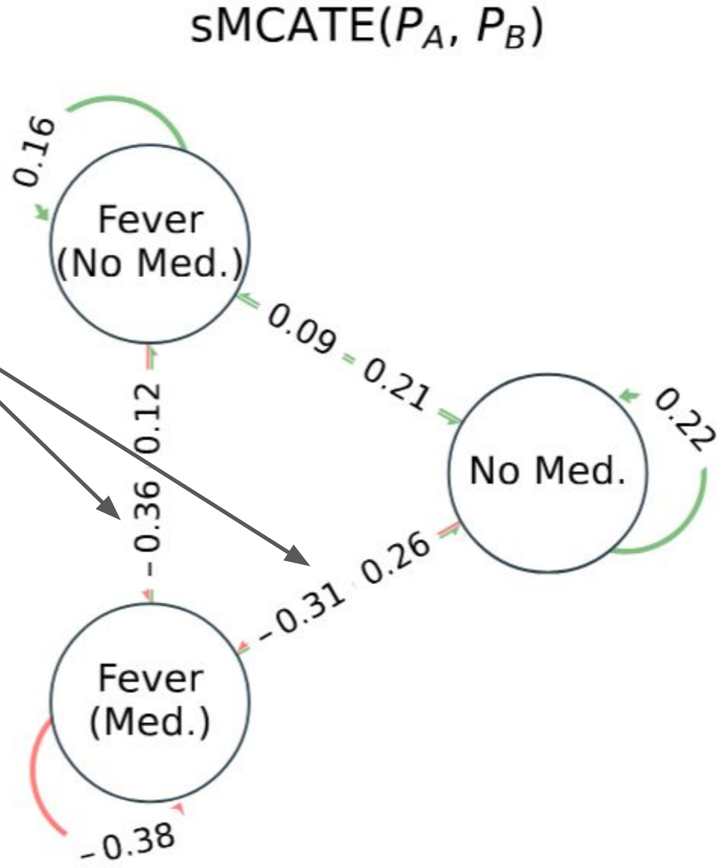


Medication MCA



Medication MCA

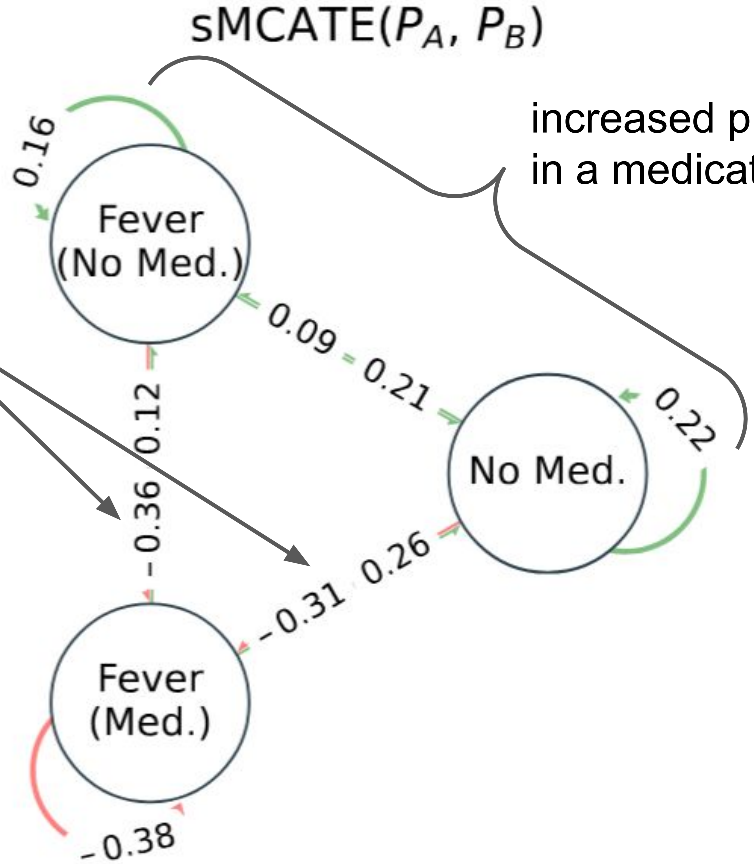
reduced transition
probabilities into fever state!



Medication MCA

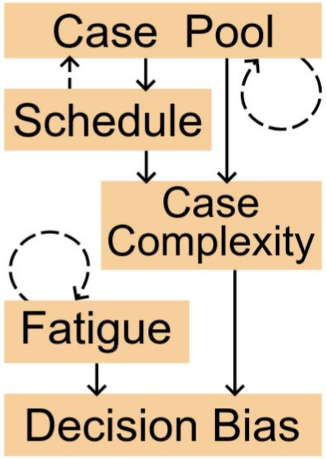
reduced transition probabilities into fever state!

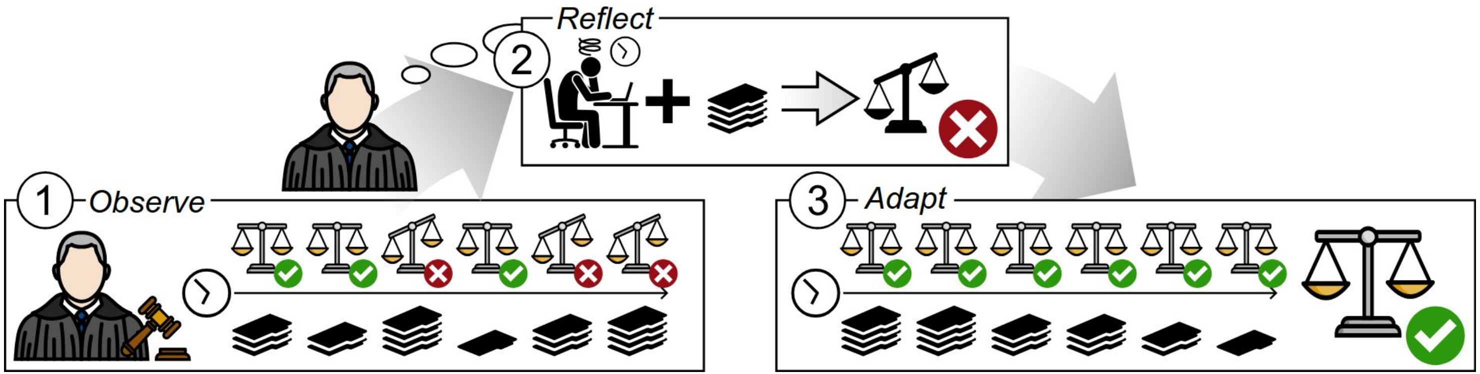
increased probability of staying in a medication free state!



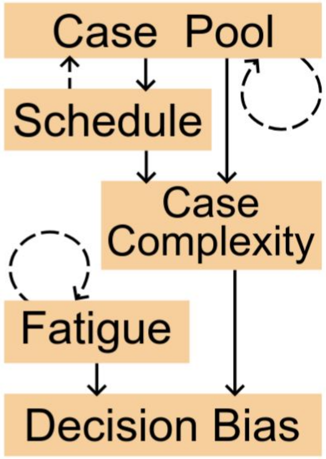
Judicial Decision-Making

Throughout the day a judge picks cases from a case pool and makes decision. Upon reflecting, the judge notices that biased decisions are due to high fatigue and case complexity.



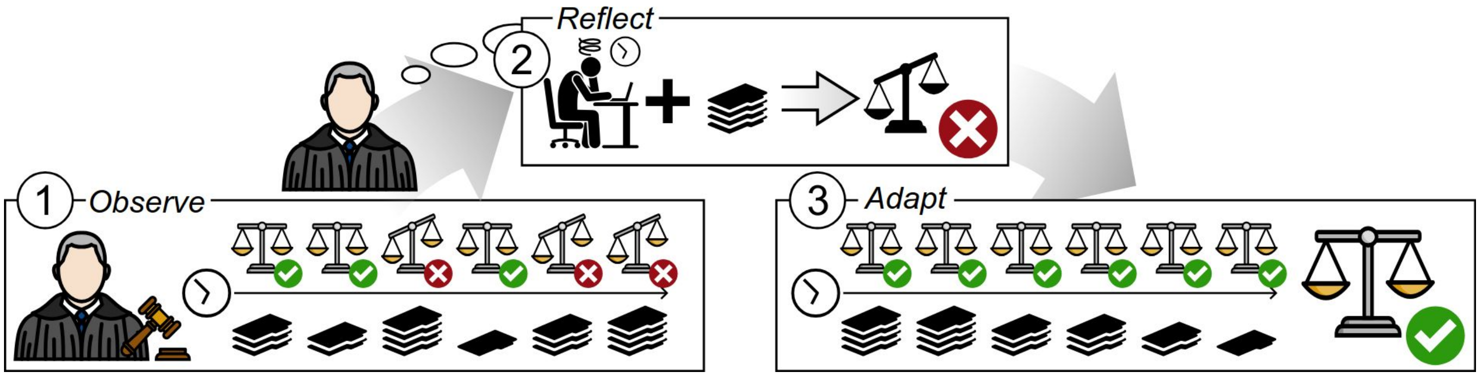


Judicial Decision-Making



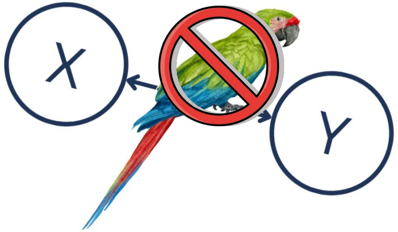
Throughout the day a judge picks cases from a case pool and makes decision. Upon reflecting, the judge notices that biased decisions are due to high fatigue and case complexity.

The key insight here is not just that fatigue causes bias, but under what conditions this causal link *becomes active*.



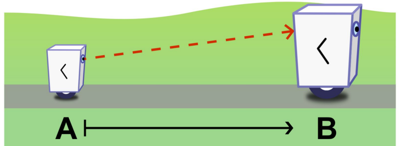
From Parroting to Understanding: A Meta-Causal Path

- **Reflection & Adaptation:** Intelligence isn't just about knowing that A causes B, but understanding the conditions under which that relationship holds, and to adapt when it changes.



From Parroting to Understanding: A Meta-Causal Path

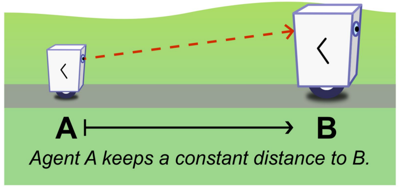
- **Reflection & Adaptation:** Intelligence isn't just about knowing that A causes B, but understanding the conditions under which that relationship holds, and to adapt when it changes.
- **Meta-Causal Models** allow to explicitly reason about *how* and *why* causal relationships change.

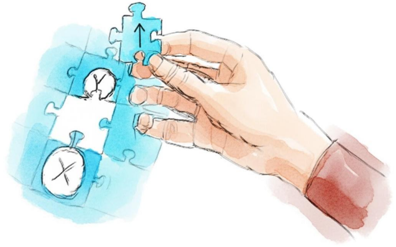


Agent A keeps a constant distance to B.

From Parroting to Understanding: A Meta-Causal Path

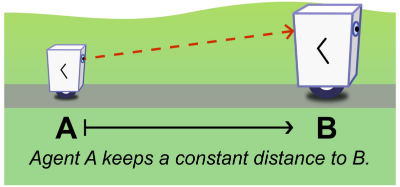
- **Reflection & Adaptation:** Intelligence isn't just about knowing that A causes B, but understanding the conditions under which that relationship holds, and to adapt when it changes.
- **Meta-Causal Models** allow to explicitly reason about *how* and *why* causal relationships change.
- **Genuine AI systems** should not just produce due to their intrinsic weights, but deliberately think about the causal mechanisms at play.

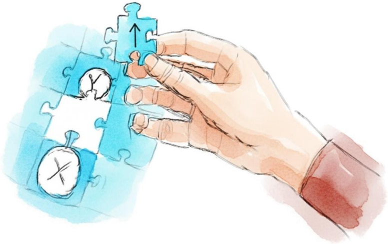




From Parroting to Understanding: A Meta-Causal Path

- **Reflection & Adaptation:** Intelligence isn't just about knowing that A causes B, but understanding the conditions under which that relationship holds, and to adapt when it changes.
- **Meta-Causal Models** allow to explicitly reason about *how* and *why* causal relationships change.
- **Genuine AI systems** should not just produce due to their intrinsic weights, but deliberately think about the causal mechanisms at play.





"Meta-Causality may be the dividing line between systems that merely describe the world from those that truly understand it."