Moritz Willig

1 Motivation

Current foundation models, despite their scale, primarily rely on associational learning methods. Consequently, they are prone to reasoning shortcuts and fallacies, indicating that while they can *simulate* reasoning on the surface, they lack a genuine understanding of the mechanisms driving the processes they reason about. My research goal is to bridge this gap, equipping AI systems with the ability to perform **genuine causal reasoning**. To achieve this, I argue we must move beyond static causal graphs:

(1) Meta-Causal Reasoning. Most real-world processes are dynamic. A genuine (causal) understanding, therefore, not only requires inference *within* causal graphs, but reasoning *about* the emergence and evolution of causal relations from their underlying processes.
(2) Reflection and Adaptation. Agents must not only predict based on pretrained weights, but adapt decisions based on emerging evidence. Specifically, agents must understand how their own actions can alter the causal structure of a system. I propose that *meta-causality* –reflection on the formation of causal links– is a prerequisite for agents to adapt and foresee the structural consequences of their own actions.

2 Contributions

My research progresses from evaluating the limitations of current LLMs to establishing a fundamental theory of meta-causality. My work has been published at major venues including **NeurIPS** [1,2], **ICLR (Spotlight)** [3], and **TMLR** [4], with contributions at smaller, targeted conferences (UAI, CogSci, PGM; [5–7]).

I. Evaluating Causal Understanding: Can Foundation Models Talk Causality?
To understand the capabilities of modern AI systems, my colleagues and I conducted some of the first formal examinations of causal reasoning in large language models (LLMs). In our work on '*Causal Parrots*' [4,8], we demonstrated that while reasoning skills of LLMs are steadily improving, enabling them to recall, combine and reason over causal facts, they fail to understand the underlying processes that give rise to these structures. We introduced the theory of 'correlation of causal facts' to explain how models can answer interventional and counterfactual queries, although never interacting with the real world during training. This work highlighted that scale and correlational back-propagation might be insufficient to achieve genuine causal understanding, motivating the need for fundamental changes in how we model causal reasoning within modern AI systems.

II. Meta-Causal Models: Reasoning About System Dynamics.
Moving beyond critique of LLMs' capabilities, I developed the framework of *Meta-Causal Models*. Standard causal inference assumes static graphs; however, I argue that causal relations emerge and vanish as the underlying systems progress. Even if exact structural equations remain unknown, the qualitative causal configuration of a system at a point in time constitutes a particular *meta-causal state*. In my work [2,3], we formalized meta-causal variables as factors that are informative of the presence and dynamics of causal relations in a system. These interactions allow agents to perform 'reflective' reasoning *about* causal dynamics, and predict how causal links shift. This formalization enables agents to be active reasoners, capable of adapting strategies through meta-aware interventions in ways that static models cannot.

3 Future Research Directions

My future agenda aims to advance the theoretical understanding of causal emergence, operationalize meta-causal reasoning, and move from theoretical frameworks to data-driven simulation and application. I therefore propose the following three research directions:

I. Causal Emergence. Current formal causal study commonly relies on graphical structures. While this approach enables pure theoretical study on idealized causal graphs, it also detaches them from the underlying processes that give rise to them in the first place. I argue that the study of causal structures must always be thought together with their underlying real-world processes. I aim to *advance the theoretical connection between low-level associational processes and their emergent causal features*.

II. Meta-Causal Understanding. Causal understanding is critical not just for making predictions, but to understand how interventions couple back into a system, potentially destabilizing the very causal relations agents may rely upon. Stressing a physical link may yield diminishing returns and eventually reach a breaking point where the *type* of the relation gets inherently altered. Genuine agents must be aware of such boundary conditions that preserve or destroy causal links. Conversely, an agent might strategically build or break relations to achieve a goal. Building on my prior work, I aim to *enable AI agents to reason about these boundary conditions*, allowing them to leverage the plasticity of causal structures through deliberate planning.

III. Process-First Causal Simulation. There is a lack of datasets that capture the complexity of dynamic causal emergence, and most evaluations rely on data that is simulated according to some predefined causal graph. Those graphs, however, rarely promote complex interactions that lead to higher-level causal effects. I propose a *process-first simulation* approach. By employing *LLMs as stochastic world engines in combination with per-instance flow models*, I aim to generate sufficiently complex synthetic worlds where the generative process comes first, and only then causal effects are analyzed. This will provide the necessary ground truth to train and benchmark a new generation of causally aware agents.

4 Conclusion

My research philosophy is rooted in the belief that AI must rise from statistical imitation to algorithmic reflection. By advancing Meta-Causal Models, I aim to build systems that do not merely talk about the world, but understand and adapt to the mechanisms that drive it.

References

[1] **Moritz Willig**, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. "Do Not Marginalize Mechanisms, Rather Consolidate!" In: *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*. 2023.

[2] **Moritz Willig**, Woydt Tim, Devendra Singh Dhami, and Kristian Kersting. "When Causal Dynamics Matter: Adapting Causal Strategies through Meta-Aware Interventions". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*. 2025.

[3] **Moritz Willig**, Tim Nelson Tobiasch, Florian Peter Busch, Jonas Seng, Devendra Singh Dhami, and Kristian Kersting. "Systems with Switching Causal Relations: A Meta-Causal Perspective". In: *The Thirteenth International Conference on Learning Representations (ICLR 2025)*. 2025.

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

[5] Harsh Poonia, **Moritz Willig**, Zhongjie Yu, Matej Zečević, Kristian Kersting, and Devendra Singh Dhami. "χSPN: Characteristic Interventional Sum-Product Networks for Causal Inference in Hybrid Domains". In: *The 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024)*. 2024.

[6] Anita Keshmirian, **Moritz Willig**, Babak Hemmatian, Kristian Kersting, Ulrike Hahn, and Tobias Gerstenberg. "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 (CogSci 2024)*. Vol. 46. 2024.

[7] Florian Peter Busch, **Moritz Willig**, Jonas Seng, Kristian Kersting, and Devendra Singh Dhami. "Ψnet: Efficient Causal Modeling at Scale". In: *International Conference on Probabilistic Graphical Models (PGM 2024)*. PMLR. 2024, pp. 452–469.

[8] **Moritz Willig***, Matej Zečević*, Devendra Singh Dhami, and Kristian Kersting. "Can Foundation Models Talk Causality?" In: *UAI 2022 Workshop on Causal Representation Learning*. 2022.