Structural Causal Models as Boundary Objects in AI System Development

Hans-Martin Heyn*
Eric Knauss*
hans-martin.heyn@gu.se
eric.knauss@cse.gu.se
Department of Computer Science and Engineering
Chalmers | University of Gothenburg, Sweden

ABSTRACT

Artificial Intelligence (AI), and especially machine learning can be used to find statistical patterns in datasets with thousands of variables with ease. But an understanding of causality is difficult to learn for a machine. For humans however, realising causal relations is often not a difficult process, as we can refer to experience or scientific knowledge. Here we propose the use of structural causal models, represented through direct acyclic graphs, to design, determine, and communicate causal relations hidden beyond the statistical models of an AI. The idea is to make human insight in causal relations explicit and use this knowledge during AI system development. In a joint-industry project we discovered that structural causal models can serve as living boundary objects that facilitate coordination of domain experts, data scientists, systems engineers, and AI experts in AI system development.

KEYWORDS

artificial intelligence, causal reasoning, machine learning, requirement engineering, safety

ACM Reference Format:
Hans-Martin Heyn and Eric Knauss. 2022. Structural Causal Models as Boundary Objects in AI System Development. In *1st Conference on AI Engineering - Software Engineering for AI (CAIN'22), May 16–24, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3522664.3528615

1 INTRODUCTION

Modern machine learning, and especially deep learning, are opaque learning machines [5] because they rely on statistical learning to find "dependence among random variables from observational data" [7]. For successful training of machine learning models the provided data must be suitable to infer the underlying mathematical structure. Advances in connectivity and data storage lead to "big data" for providing huge amounts of data. However, there is no guarantee that the data provided for training are actually correctly representing the problem we want to solve with our AI system. Judea Pearl describes it as:

> *"To achieve human-level intelligence, learning machines need the guidance of a blueprint of reality, a model - similar to a road map that guides us in driving through an unfamiliar city"*

Informed machine learning tries to give some "blueprint of reality" by including prior knowledge into the machine learning pipeline as described in the taxonomy by von Rueden et al. [9]. While conducting the joint industrial-academic research project VEDLIoT[1] we observed that prior knowledge in the form of causal models can help in deciding which data is required for training and testing, and which runtime conditions need to be fulfilled. Previously, causal models at design time of a system were mostly use to define safety argumentation of a system, such as described in the work of Ibrahim et al. [2]. In this extended abstract we propose to use causal models for planning acquisition of training data and for ensuring the desired run-time behaviour of the AI system.

2 CAUSAL MODELS AS BOUNDARY OBJECTS

Why Causal Models? Many causal relationships are learnt by humans through experience or formalised through scientific methods. For a machine learning model, causal relations, and especially the direction of cause and effect are not directly learnable. Although research into causal learning is progressing (see all the work from Pearl et al., e.g., [4, 6] and Peters et al. [7]), inferring causal structure from data is an "ill-posed" problem because "even complete knowledge of the probabilistic model[2] usually does not determine the underlying causal model" [7].

Causal models in practise. Causal relations can be represented through directed acyclic graphs [6]. During the research project VEDLIoT, we started to draw directed graphs to capture the desired causal behaviour of the AI system. For one of the industry use cases, which encompasses a machine learning based pedestrian detection [3], the starting point is depicted in Figure 1a. Given a target on the road, the pixel values of the camera image change, which should (through a trained deep neural network) lead to the event "Detection of target".

We found causal models were helpful to plan the acquisition of training data and to find requirements for run-time monitoring.

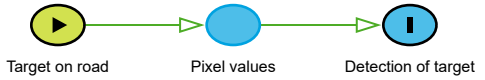
A causal model assists the planning of (training) datasets. If one includes the environmental aspect of weather, our knowledge will

*Both authors contributed equally to this research.

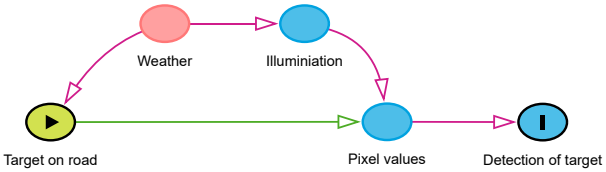
CAIN'22, May 16–24, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9275-4/22/05.
https://doi.org/10.1145/3522664.3528615

[1] http://www.vedliot.eu
[2] which probably never exists in machine learning, because we train models with sampled, non-infinite datasets



(a) Basic functionality of the Pedestrian detection.



(b) Considering the influence of weather. Green edges: Causal paths Red edges: Indirect, biasing paths.

Figure 1: Directed acyclic graphs representing causal models

tell us that weather changes the illumination of the camera images (sunny day vs. cloudy day). The illumination in turn effects the pixel values. But furthermore, weather also effects the probability of a target *being* on the road. It is more probable to encounter a pedestrian walking on the road on a nice sunny day compared to a cloudy, rainy day. Therefore, there is not only a causal relation between weather and pixel values (via illumination), but also between weather and the probability that a target is encountered on the road. Weather is therefore a so called confounder, which can introduce a bias. In a directed graph this causal relation can easily be depicted, as shown in Figure 1b.

The training data set must account for the spurious association caused by the variable "weather" to avoid a bias in the inference results of the final AI model. In practise, the data collection needs to be planed such as to collect more data during bad weather than during sunny weather. This allows for the data to have an equal proportion of pedestrians visible in both bad and sunny weather[3].

A causal model provides rules for runtime monitoring. Monitoring the behaviour and performance of AI models at runtime can be important especially for safety critical applications. However, it is still challenging to derive the necessary requirements for such runtime monitors because test conditions are difficult to define [1] or the deployment environment is unknown at design time [8].

By using causal models, assumptions about the deployment environment can be made explicit and test condition can be derived directly from the causal models. In the model depicted in Figure1b we assumed that weather influences the probability of encountering pedestrians on the road and adjusted our training dataset accordingly. This causal assumptions provides us with a testable probability condition: Weather and the probability of encountering a pedestrian are statistically depended, and the dependence can be inferred at runtime, given access to weather data and detection rate of pedestrians. By providing a clear description of the assumed causal relations through a causal model at design time, validation and runtime testing criteria based on runtime data can be derived

[3]If we were to ignore the spurious association caused by weather, and collect 50% of the data during bad weather and the remaining 50% of the data during sunny weather, the data set would be biased in the sense that more pedestrians are present during sunny weather. This in turn would cause the model to perform better in detecting pedestrians on the road during sunny conditions.

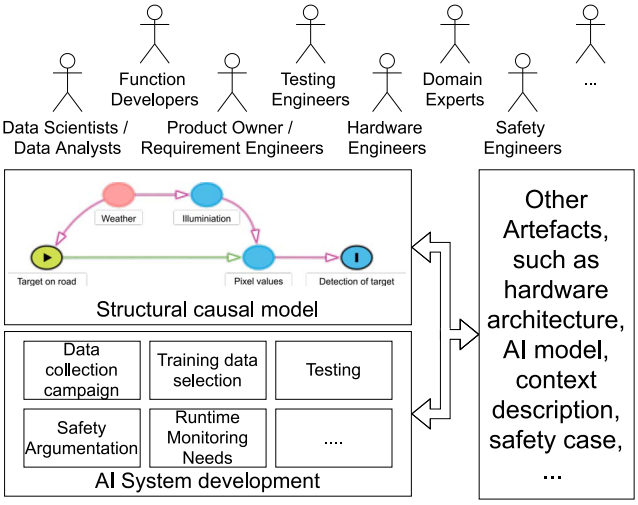


Figure 2: Structural causal models as boundary object in an AI system development environment.

to check the validity of the made causal assumptions (e.g., about the environment):

> *"[...] a researcher who has scientific knowledge in the form of a structural equation model is able to predict patterns of in dependencies in the data, based solely on the structure of the model's graph. [...] Conversely, it means that observing patterns of in dependencies in the data enables us to say something about whether a hypothesised model is correct"* [6, Chapter 2.1].

3 CONCLUSION

"Boundary objects are objects which are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites" [10]. Based on our experience in VEDLIoT, causal models can serve as boundary objects to capture and communicate assumptions about the deployment environment of AI systems between different stakeholders and development artefacts, as illustrated in Figure 2.

ACKNOWLEDGMENTS

This project has received funding from the EU Horizon 2020 research and innovation program under grant agreement No 957197.

REFERENCES

[1] Markus Borg, Cristofer Englund, et al. 2018. Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry. *arXiv preprint arXiv:1812.05389* (2018).
[2] Amjad Ibrahim, Severin Kacianka, et al. 2019. Practical causal models for cyber-physical systems. In *NASA Formal Methods Symposium*. Springer, 211–227.
[3] Franz Meierhöfer, Roland Weiss, et al. 2021. *Specification for selected pilots / use cases.* Technical Report 957197. Horizon 2020 Research Framework. https://vedliot.eu/deliverable/deliverable-d23/
[4] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
[5] Judea Pearl. 2019. The Limitations of Opaque Learning Machines. In *Possible Minds: 25 Ways of Looking at AI*, Johnm Brockman (Ed.). Penguin Press, London, Chapter 2.
[6] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer.* John Wiley & Sons.

[7] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.
[8] Quazi Marufur Rahman, Peter Corke, and Feras Dayoub. 2021. Run-time monitoring of machine learning for robotic perception: A survey of emerging trends. *IEEE Access* 9 (2021), 20067–20075.
[9] Laura von Rueden, Sebastian Mayer, et al. 2021. Informed Machine Learning-A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering* (2021).
[10] Rebekka Wohlrab, Patrizio Pelliccione, et al. 2019. Boundary objects and their use in agile systems engineering. *Journal of Software: Evolution and Process* 31, 5 (2019), e2166.