Reporting on Real-World Datasets and Packages for Causal AI Research*

Dren Fazlija[0000−0002−9133−3526]

L3S Research Center, Appelstr. 4, 30167 Hannover, Germany
dren.fazlija@l3s.de

Abstract. Causal reasoning has garnered much attention in the AI research community, resulting in an influx of causality-based AI methods in recent years. We believe that this sudden rise of Causal AI has led to many publications that primarily evaluate their proposed algorithms in specifically designed experimental setups. Hence, comparisons between different causal methods, as well as existing state-of-the-art non-causal approaches, become increasingly more difficult. To make Causal AI more accessible and to facilitate comparisons to non-causal methods, we analyze the use of real-world datasets and existing causal inference tools within relevant publications. Furthermore, we support our hypothesis by outlining well-established tools for benchmarking different trustworthy aspects of AI models (interpretability, fairness, robustness, privacy, and safety) healthcare tools and how these systems are not prevalent in respective Causal AI publications.

Keywords: Causality · Datasets · Benchmarks · Artificial Intelligence.

1 Introduction

Modern Artificial Intelligence (AI) systems tend to demonstrate a limited understanding of the relationship between causes and effects intrinsic to their working environment [5]. Understanding such relations is quintessential to our ability to effortlessly adapt and interact with our world [2]. Hence, it is reasonable to assume that AI systems would benefit from the ability to reason about causal effects within their domain. The idea of utilizing concepts of *Causality* – the science of reasoning about causes and effects – to enhance AI systems has its origins in Judea Pearl's seminal paper [4] back from 1995. Ever since then, continuous contributions by prominent AI researchers like Bernhard Schölkopf or Yoshua Bengio (see, e.g., [7]) have significantly increased the popularity of this research area. The recent rise of Causal AI is well-documented by several surveys (e.g., [1, 3]), including its utility within the healthcare domain (e.g., [1, 6, 8]). The authors of [3] – a pervasive review of existing Causal Machine Learning (ML) methods – observed several issues that complicate the evaluation of proposed causal methods. They state that (*i*) there are relatively few open-sourced

* Supported by the Center for Digital Innovations of Lower Saxony (ZDIN).

software packages related to Causal ML, (*ii*) some Causal ML publications do not thoroughly compare their approach to non-causal techniques, and (*iii*) it is generally difficult to obtain ground-truth evaluation data due to the lack of public benchmarks for causal model training. To analyze and explain these shortcomings of current Causal AI research, we decided to investigate the use of real-world datasets and existing packages and benchmarks within recent Causal AI literature. We directly build upon the curated list of [1]. It contains the used real-world datasets of each Causal AI method discussed in the survey and some packages related to different aspects of Trustworthy AI (interpretability, fairness, robustness, privacy, and safety) and the healthcare domain.

1.1 Contributions

1. We analyze the use of existing real-world datasets in Causal AI publications discussed in [1]. We report on the most popular dataset types, how influential the datasets are, and the machine learning tasks for which current causality-based techniques are used. Additionally, we provide potential explanations for the observed use of real-world datasets.
2. For each aspect of trustworthy AI (and for the healthcare domain), we highlight the most common causality tasks (e.g., causal discovery or causal effect estimation) within the publications of [1]. To promote using existing tools for causal reasoning, we provide a short overview of the most prominent tools and packages for relevant tasks. Whenever possible, we will refer to suitable datasets that provide causal encodings, such as predefined counterfactuals, causal graphs, or the ability to simulate well-defined interventions in the data domain.
3. We review relevant benchmarks of each facet of trustworthy AI and investigate whether they are currently used to evaluate existing causal approaches. Similar to the causality-related tools, we will also provide potential reasons why some benchmarks are not fully utilized yet.

Acknowledgments. This work has received funding from the Lower Saxony Ministry of Science and Culture under grant number ZN3492 within the Lower Saxony "Vorab" funding program of the Volkswagen Foundation.

References

1. Ganguly, N., Fazlija, D., Badar, M., Fisichella, M., Sikdar, S., Schrader, J., Wallat, J., Rudra, K., Koubarakis, M., Patro, G.K., Zai El Amri, W., Nejdl, W.: A Review of the Role of Causality in Developing Trustworthy AI Systems (Feb 2023). https://doi.org/10.48550/ARXIV.2302.06975, https://arxiv.org/abs/2302.06975
2. Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., Danks, D.: A theory of causal learning in children: causal maps and bayes nets. Psychological review **111**(1), 3 (2004)
3. Kaddour, J., Lynch, A., Liu, Q., Kusner, M.J., Silva, R.: Causal machine learning: A survey and open problems. arXiv preprint arXiv:2206.15475 (2022), https://arxiv.org/abs/2206.15475

4. Pearl, J.: Causal diagrams for empirical research. Biometrika **82**(4), 669–688 (1995)
5. Pearl, J.: The seven tools of causal inference, with reflections on machine learning. Communications of the ACM **62**(3), 54–60 (2019)
6. Sanchez, P., Voisey, J.P., Xia, T., Watson, H.I., O'Neil, A.Q., Tsaftaris, S.A.: Causal machine learning for healthcare and precision medicine. Royal Society Open Science **9**(8), 220638 (2022)
7. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
8. Vlontzos, A., Rueckert, D., Kainz, B.: A review of causality for learning algorithms in medical image analysis. arXiv preprint arXiv:2206.05498 (2022)