

A Review of the Role of Causality in Developing Trustworthy AI Systems

NILOY GANGULY^{*†}, DREN FAZLIJA^{*}, MARYAM BADAR^{*}, MARCO FISICHELLA^{*}, SANDIPAN SIKDAR^{*}, JOHANNA SCHRADER^{*}, JONAS WALLAT^{*}, KOUSTAV RUDRA^{*‡}, MANOLIS KOUBARAKIS^{*§}, GOURAB K. PATRO^{*}, WADHAH ZAI EL AMRI^{*}, and WOLFGANG NEJDL^{*}, L3S Research Center, Leibniz University of Hannover, Germany

State-of-the-art AI models largely lack an understanding of the cause-effect relationship that governs human understanding of the real world. Consequently, these models do not generalize to unseen data, often produce unfair results, and are difficult to interpret. This has led to efforts to improve the trustworthiness aspects of AI models. Recently, causal modeling and inference methods have emerged as powerful tools. This review aims to provide the reader with an overview of causal methods that have been developed to improve the trustworthiness of AI models. We hope that our contribution will motivate future research on causality-based solutions for trustworthy AI.

CCS Concepts: • **Information systems**; • **Computing methodologies** → **Artificial intelligence**;

Additional Key Words and Phrases: Causality, Counterfactual, Interpretability, Explainability, Robustness, Bias, Discrimination, Fairness, Privacy, Safety, Healthcare

1 INTRODUCTION

The Deep Learning based systems, in recent years, have produced superior results on a wide array of tasks; however, they generally have limited understanding of the relationship between causes and effects in their domain [197]. As a result, they are often brittle and unable to adapt to new domains, can treat individuals or subgroups unfairly, and have limited ability to explain their actions or recommendations [197, 235] reducing the trust of human users [118]. Following this, a new area of research, *trustworthy AI*, has recently received much attention from several policymakers and other regulatory organizations. The resulting guidelines (e.g., [184, 186, 187]), introduced to increase trust in AI systems, make developing trustworthy AI not only a technical (research) and social endeavor but also an organizational and (legal) obligational requirement.

In this paper, we set out to demonstrate, through an extensive survey, that *causal modeling and reasoning* is an emerging and very useful tool for enabling current AI systems to become trustworthy. *Causality* is the science of reasoning about causes and effects. Cause-and-effect relationships are central to how we make sense of the world around us, how we act upon it, and how we respond to changes in our environment. In AI, research in causality was pioneered by the Turing award winner Judea Pearl long back in his 1995 seminal paper [194]. Since then, many researchers have contributed to the development of a solid mathematical basis for causality; see, for example, the books [79, 196, 201], the survey [90] and seminal papers [197, 235].

The TAILOR project [166], an initiative of EU Horizon 2020, with the main objective of integrating learning, reasoning, and optimization in next-generation AI systems, in its first strategic research

^{*}All authors contributed equally, names are randomly ordered

[†]Presently in Indian Institute of Technology, Kharagpur

[‡]Presently in Indian Institute of Technology, Indian School of Mines, Dhanbad

[§]Presently in National and Kapodistrian University of Athens

and innovation roadmap, identifies the following dimensions of AI systems which contribute to trustworthiness: interpretability or explainability, safety and robustness, fairness, equity and justice, accountability and reproducibility, privacy, and sustainability. This is corroborated by several recent surveys and reports [34, 110, 118, 247, 283]. Following the above-mentioned works, we select a number of properties desired for the trustworthiness of AI and survey the role of causality in achieving these properties:

- **Interpretability** Can the AI system’s output be justified with an explanation that is meaningful to the user?
- **Fairness:** How can we ensure that the AI system is not biased, does not discriminate (e.g., against minorities, disabled people, people of a certain gender, etc.), but provides fair prediction and recommendation?
- **Robustness:** How sensitive is the AI system’s output to changes in the input? Does the performance vary significantly in the case of distributional shifts?
- **Privacy:** Is the AI system susceptible to attacks by adversaries? Does the system leak private information? Does the AI system respect user privacy? How can we develop AI systems that ensure user privacy?
- **Safety and Accountability:** What are the risks for humans after an AI system is deployed in the real world? Is the AI system safe? Who or what is responsible for the actions (or failures) of the AI system after deployment? How to audit an AI system and its impact?

There have been several current surveys on trustworthy AI [34, 118, 247, 260, 283] but none has paid significant attention to the role of causality in developing trustworthy AI systems. We provide a detailed comparison of our survey to other related surveys in Section 2.6. We survey the contributions of causality in upholding various trustworthy AI aspects (interpretability, fairness, robustness, privacy, safety, and auditing) in Sections 3 to 7. Considering the recent interest in AI for healthcare, we discuss various trustworthy aspects needed in the health domain and survey how causality has played an important role in ensuring trust in AI-based healthcare applications (Section 8). We also list available datasets, tools and packages relevant to causality-based trustworthy AI research and development (Appendices A to F).

2 PRELIMINARIES ON CAUSALITY

The two most powerful causal frameworks are (i) *structural causal models* [194, 196] developed in AI, and Rubin’s (ii) *potential outcomes framework* [224] developed in Statistics. Here, we give short introductions to structural causal models (Section 2.3), the potential outcomes framework (Section 2.4), and then discuss how they are connected Section 2.5. *Causal graphical models* [253] is another popular framework that we do not cover here. Before discussing the frameworks, we first list some basic notations, and discuss some fundamental concepts on causality: the difference between correlation and causation in Section 2.1, and the ladder of causation through association, intervention, and counterfactuals in Section 2.2.

Notations: We use non-boldface uppercase letters (e.g., X) to denote single random variables and non-boldface lowercase letters (e.g., x) to denote their values. Boldface uppercase letters (e.g., \mathbf{X}) denote a collection of variables, and boldface lowercase letters (e.g., \mathbf{x}) their values. Random variables can have continuous or discrete or categorical values.

2.1 Correlation vs. Causation

Statisticians and others often state that “correlation does not imply causation” and illustrate it with anecdotal examples. Schölkopf and von Kügelgen [237], for example, cite the correlation between chocolate consumption and the number of Nobel prizes per capita from [164], which obviously

networks since this level is characterized by conditional probability sentences e.g., the sentence $P(Y = y|X = x) = 0.8$. For some associations, it might be easy to find causal interpretations, while for others it might not. *But agents on this level of the ladder, cannot differentiate between a cause and an effect.* Therefore, they cannot answer “why” questions.

(2) Intervention: The second level (doing) corresponds to *interventions*. Most tool users are on this level if they plan their actions and not imitating the actions of others. To see the effects of interventions, we can do experiments (e.g., a randomized controlled trial). Pearl and his colleagues have formalized reasoning on this level of the ladder of causation using the *do calculus* [79, 196, 198]. This level of the ladder is characterized by expressions such as $P(Y = y|do(X = x), Z = z)$ which means “the probability of event $Y = y$ given that we intervene and set the value of X to x and subsequently observe event $Z = z$ ” [197].

(3) Counterfactuals: The third level (imagining) corresponds to *counterfactual reasoning* and its associated modal reasoning capabilities (e.g., retrospection, introspection, etc.). On this level, we can imagine worlds that do not exist, including worlds that contradict the world in which we live, and infer why the phenomena we have observed in our domain have taken place. This level of the ladder is characterized by expressions of the form $P(Y = y_{X=x}|X = x', Y = y')$ which stand for “the probability that event $Y = y$ would be observed had X been x , given that we actually observed the events $X = x'$ and $Y = y'$ ” [197]. Such expressions can be computed only if we have a causal formalism such as the structural causal networks presented in Section 2.3 below. Given such a structure, one can then also formalize level 3 reasoning through do calculus.

2.3 Structural Causal Models (SCM)

Structural causal models consist of causal graphs and structural equations. Following the definition by Bareinboim et al. [18], a *structural causal model (SCM)* \mathcal{M} is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$ where: \mathbf{U} is a set of background variables, also called *exogenous* variables, that are determined by factors outside the model; \mathbf{V} is a set of variables, called *endogenous* variables, that are determined by other variables in the model i.e., variables in $\mathbf{U} \cup \mathbf{V}$; \mathcal{F} is a set of functions $\{f_1, f_2, \dots, f_n\}$ such that each f_i is a mapping from the respective domains of $\{U_i\} \cup \mathbf{PA}_i$ to V_i , where $U_i \in \mathbf{U}$, the variables \mathbf{PA}_i are the *parents* of variable V_i and are such that $\mathbf{PA}_i \subseteq \mathbf{V} \setminus \{V_i\}$, and the entire set \mathcal{F} forms a mapping from \mathbf{U} to \mathbf{V} , so, for $i = 1, \dots, n$, each $f_i \in \mathcal{F}$ is such that $V_i \leftarrow f_i(\mathbf{PA}_i, U_i)$, i.e., it assigns a value to V_i that depends on the values of the parents of V_i and the value of the exogenous variable U_i ; $P(\mathbf{U})$ is a probability function defined over the domain of \mathbf{U} . The exogenous ones are determined “outside” of the causal model (therefore assumed to be independent), and their associated probability distribution $P(\mathbf{U})$ gives a summary of the state of the world outside the domain.

The expression $V_i \leftarrow f_i(\mathbf{PA}_i, U_i)$ in the above definition are called *structural equations*¹. The structural equations *induce a causal graph* for the SCM. The nodes of the graph are the variables in the set \mathbf{V} , and there is a directed edge from each variable in \mathbf{PA}_i to V_i for all i . In other words, the structural equations encode the direct causal effects for each edge in the causal graph, and they can be used to determine the value of each endogenous variable V_i in terms of the values of the exogenous variable U_i and the values of the endogenous variables \mathbf{PA}_i that are parents of V_i . The causal processes encoded by structural equations are assumed to be invariant unless we explicitly intervene on them using the do calculus of the second level of the ladder of causality [18].

SCMs can be used for various inference tasks that the forthcoming sections of this paper show to be useful for achieving the trustworthiness of AI systems. Some well-known tasks are listed below.

¹We follow [18] and avoid using an equality sign for structural equations since their interpretation is that of an assignment statement; they should not be interpreted as algebraic equations that can be solved for any variable.

- The first such task is **causal reasoning**, which is the process of deriving conclusions from a causal model e.g., an SCM. SCMs can also be used to study the effects of interventions or distribution changes or carry out counterfactual reasoning [196, 201].
- The opposite task of causal reasoning is **causal discovery** or **causal structure learning**, which is the process of inferring SCMs from data, assumptions, empirical observations, or from data under interventions or distribution changes [196, 201].
- Another interesting task is **causal mediation**, which is the process of looking for the mechanism that explains how a cause is connected with an effect [195]. In causal mediation, we may have an SCM $X \rightarrow Z \rightarrow Y$ where X is the cause, Y is the effect, and Z is the *mediator* i.e., a variable that can be used to answer the question “Why X causes Y ?”. For example, the SCM $Citrus\ Fruit \rightarrow Vitamin\ C \rightarrow Scurvy$, explains why citrus fruits were important in helping sailors avoid scurvy in the 1800s [198]. In such networks, Y is an *indirect* effect of X as opposed to a *direct* effect, which would have been denoted by $X \rightarrow Y$.

2.4 Potential Outcomes (PO) Framework

The *potential outcomes* approach to causality was developed to make statistically valid statements even in cases where randomized controlled studies are not or only partially possible. This section gives a brief introduction to the PO framework [224].

The missing data problem in individualized treatment effect (ITE): We quote an example from Ozer et al. [190]. Consider two possible outcomes for one patient having resectable gallbladder cancer, “survival” and “no survival”. The task is to measure the causal effect of a treatment (e.g., chemotherapy). Here, we have a binary *treatment* variable T with $T = 1$ if the person is getting the chemotherapy and $T = 0$ otherwise (*control*), and its effect on an *outcome* variable Y (typically a measure of health) is to be found. Note that $Y_i(1)$ and $Y_i(0)$ are unobserved outcomes for $T = 1$ and $T = 0$ respectively. Then the *individualized treatment effect* can be captured by the difference between the two potential outcomes, i.e., $\tau_i = Y_i(1) - Y_i(0)$. Since only one of the outcomes will take place, we can not have both $Y_i(1)$ and $Y_i(0)$ for an individual. To overcome this problem, the PO framework takes some assumptions and estimates average causal effects over a population instead of individual effects, which we detail next.

Assumptions to overcome the missing data problem:

The PO framework relies on the following assumptions: (a) **The stable unit treatment value assumption.** The observation on one unit should be unaffected by the assignment of treatments to the other units; it is a reasonable assumption in many situations like controlled studies, since different units can form independent samples from a population. (b) **The consistency assumption.** The potential outcome for an individual remains consistent and converges. Any variation within the exposure group (treatment or control) would result in the same outcome for that individual. Using these assumptions, the PO framework estimates average causal effects over a population, as detailed next.

The average treatment effect (ATE) in PO: Since the ITE $\tau_i (Y_i(1) - Y_i(0))$ under a deterministic treatment model can not be found, an *average treatment effect* on a population is calculated. The difference between the expected outcome in the treatment group and the expected outcome in the control group is expressed as $\mathbb{E}[Y|1] - \mathbb{E}[Y|0]$. So, the ATE can be calculated by first randomizing the treatment and control group assignments of units, followed by a comparison of mean outcomes for treated and untreated units. In real-world settings, some covariates (e.g., smoking and drinking habits) also affect the outcome along with the treatment. Thus, they must be taken into account.

Propensity score matching to take care of covariates: In cases of covariates, one must balance the covariate distribution between the treated and untreated cohorts, *propensity score matching* [222] is one of the most popular methods used for this. Propensity scores are probabilities

of units being assigned to different treatment groups based on the observed covariates, and these are estimated using logistic regression over the covariates. Now the method basically aims to make two groups comparable to each other in terms of covariates, thereby accounting for selection bias [54]. Essentially, each patient from one group is matched with another one for the other group based on the propensity scores. Units (patients) with no matching from the other treatment group are often removed and not used in the ATE calculation. Other measures, such as the *Mahalanobis* metric, were excluded from this review. We refer to [255] for an extensive overview of matching methods, including definitions of prominent matching metrics.

2.5 Connection between SCM and PO

The above discussion has assumed that individuals (or units) are binary quantities. However, this is not a good assumption when dealing with complex units such as people e.g., in the health domain. In these cases, potential outcomes can be defined as random variables and a clear connection to the structural equations with exogenous noise variables in SCMs can be defined. This observation results in the equivalence of the two frameworks, a result which has been originally shown in [196]. [237] explains this equivalence of the two frameworks in the following simple way:

$$Y_i(t) = Y \mid do(T = t) \text{ in an SCM with } \mathbf{U} = \mathbf{u}_i$$

This informally means that an individual i in the PO framework corresponds to a particular value of an exogenous variable U_i in an SCM. The potential outcome is deterministic once we know \mathbf{U} , but since we do not observe \mathbf{u}_i , the counterfactual outcome is treated as a random variable [237].

2.6 Prior Surveys

Some recent survey papers have discussed causality and trustworthy machine learning. Most of them focus either exclusively on causality [76, 234] or trustworthy aspects of machine learning [119], only a few papers tried to cover both causality and different aspects of trustworthiness [117, 235].

Apart from general coverage of important trustworthiness aspects, some of the surveys explicitly covered interpretability and fairness aspects in detail. Guidotti et al. [88] and Linardatos et al. [142] provided a review of machine learning interpretability methods that deal with explaining black box models. Zhou et al. [311] performed a survey on evaluation metrics of explainability. Makhoul et al. [156] pointed out the difficulty in choosing a particular fairness notion in a given domain and scenario. Mehrabi et al. [162] conducted a survey focused on the application of AI for fairness-aware learning in various domains. Moraffah et al. [171] performed a survey on causal inference on model, example-based interpretability, and fairness. In contrast to the above approaches, we provide detailed coverage of causality-based methods over a wide variety of trustworthy aspects.

Kaddour et al. [117] and Schölkopf et al. [235] provide an overview of the application of causality to address trustworthy aspects. Kaddour et al. [117] discussed different causal approaches, such as causal supervised learning and causal generative modeling, and their applications on explainability, fairness, and robustness. They did not cover the application phases of different methods in detail, i.e., whether causal approaches could be applied in pre-processing, in-processing, or post-processing stages, and did not provide more extensive coverage on different aspects of trustworthiness, though. Our work complements these prior works and extends the discussion to robustness, privacy, safety, and accountability. Apart from structural causal models, we also include a discussion on the PO framework which helps in causal analysis from a statistical perspective.

Table 1 provides the coverage statistics of the existing surveys. Note that, this table covers the generic and domain-specific surveys that used causality as a driving force to achieve trustworthiness. It is evident from the table that our survey covers the causality and trustworthiness aspects in more detail.

Table 1. Comparison of our survey with related causality and trustworthy-based survey papers.

Surveys	Trustworthiness Aspects							Domain
	Interpretability	Fairness	Accountability	Robustness	Privacy	Safety	Security	
Kaddour et al. [117]	X	X	-	X	-	-	-	-
Schölkopf and von Kügelgen [237]	-	-	-	X	-	-	-	-
Guidotti et al. [88]	X	-	-	-	-	-	-	-
Linardatos et al. [142]	X	-	-	-	-	-	-	-
Zhou et al. [311]	X	-	-	-	-	-	-	-
Makhlouf et al. [156]	-	X	-	-	-	-	-	-
Mehrabi et al. [162]	-	X	-	-	-	-	-	-
Moraffah et al. [171]	X	X	-	-	-	-	-	-
Zhang et al. [306]	X	X	-	X	-	-	-	Healthcare
Sanchez et al. [231]	X	X	-	X	-	-	-	Healthcare
Vlontzos et al. [275]	X	X	-	X	-	X	-	Healthcare & Image analysis
Gao et al. [76]	-	-	-	-	-	-	-	Recommendation
Our Survey	X	X	X	X	X	X	X	-

In the next sections (Sections 3 to 7), we discuss the listed trustworthy AI properties one by one and discuss how causality is helpful in upholding them in the context of AI.

3 CAUSALITY AND INTERPRETABILITY

Understanding latent models is one of the central challenges in the development and deployment of machine learning applications. Recent trends have shown that interpretability is sacrificed for overparameterization and the promised generalizability [33, 207]. However, to apply these models in high-stakes scenarios such as the legal or medical domain, we will need (and potentially be legally required by the EU AI Act²) to build understandable models. To ensure that the model’s explanation communicates the true reasons behind a prediction, the explanation itself should preferably be causal. This section aims to provide an overview of the early stages of causality in existing interpretability research, its promises, and how it might help build more trustworthy models.

3.1 Preliminaries

Evaluation Criteria. Faithfulness and causability are two important criteria to measure the interpretability of non-causal methods. Faithfulness measures how well the explanation matches the actual underlying processes of the model. To gain human confidence, the accuracy of the explanation is of immense importance in high-stake scenarios, such as medical or health applications. Various metrics have been proposed to measure faithfulness, e.g., feature attribution methods. Causability measures how well explanations depict the causal structure of the problem and can be assessed through a user study using the System Causability Scale (SCS), which uses a Likert scale and is similar to the well-established System Usability Scale (SUS) [100]. A causable explanation helps the recipient build a correct mental model of the problem, which has been shown to greatly impact the user’s trust in the system [246]. For an effective AI system, the target audience must be taken into account when evaluating the causability of the method, that is, causal explanations of a model used in the medical domain should provide more details when dedicated to medical experts than an explanation presented to patients.

Like traditional interpretability methods, causal approaches are either interpretable by their model design or are methods that provide post-hoc explanations for non-interpretable models. In this section, we first provide a brief overview of methods that are causally interpretable by design

²Article 13, <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>

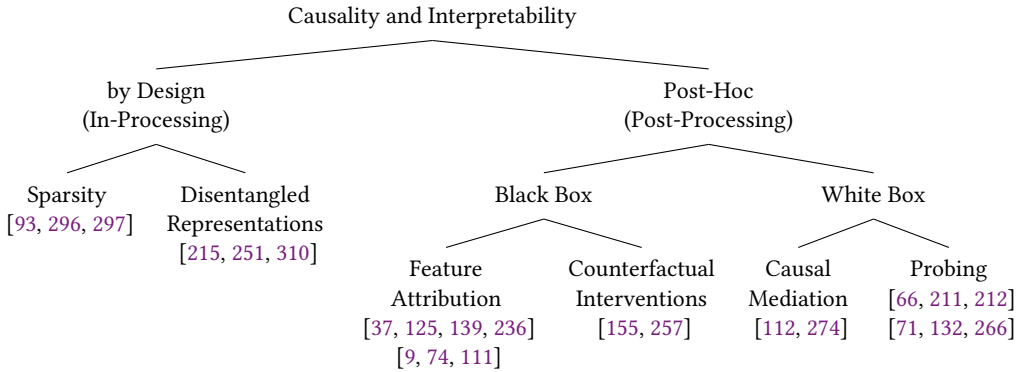


Fig. 2. Structure of approaches introducing causality in interpretability.

and then focus on causal post-hoc interpretability. Figure 2 visualizes an overview of the different approaches and the application of causal interpretability.

3.2 Causal Interpretability By Design

With the recent surge of interest in causality, existing approaches to interpretable ML, such as feature selection and disentanglement representations, have also been extended to include causality in interpretable methods. In the following, we will give an overview of both approaches in causality.

3.2.1 Sparsity (Causal Feature Selection). Feature selection categorizes the data’s attributes into relevant (irrelevant) features, i.e., features that do (not) provide predictive information regarding a target variable, and redundant features, i.e., multiple features with the same predictive information, and removes irrelevant and redundant features from the variable-subset of interest [93]. Causality is introduced into feature selection methods to distinguish between causally relevant and only co-occurring features when choosing the redundant features to keep as relevant attributes [93], which may lead to higher accuracy. To introduce a causal ordering of the features, Bayesian networks (BNs) can be used in causal feature selection to represent causal relationships among features in a DAG by interpreting directed edges as cause-effect relationships. Under the faithfulness assumption, the Markov boundary (MB) of a variable of interest in the BN describes the variable’s local causal relationships, i.e., its parents, children, and spouse [296]. In most cases, a model has only one variable of interest, i.e., the target variable, making it unnecessary to learn a full BN. Instead, these approaches [7, 8] only learn the MB for the variable of interest, using either constraint-based methods based on conditional independence tests or score-based methods that combine greedy search with scoring functions [296]. However, causal feature selection methods generally are less efficient than their non-causal counterparts due to conditioning on the other features and also are less reliable when the data set is small with high dimensionality [297]. For a more detailed overview, the reader is referred to a comprehensive review of causal feature selection by Yu et al. [296]. They also provide a collection of relevant methods in their package CausalFS.

3.2.2 Disentangled Representations. A perfectly disentangled latent representation where each dimension represents a human-understandable concept would naturally be interpretable. However, achieving a fully disentangled representation is not feasible in the general case, as approaches require specific training data [241], supervision [182], or predefined concepts [145]. Nevertheless, there are approaches using SCMs (Section 2.3) to identify confounders and then disentangle representations

to produce more interpretable models. Reddy et al. [215] propose a dataset to investigate the causal effects in disentangled representations. The proposed dataset contains images of geometric shapes with varying properties (e.g., shape, background, color, size) generated according to an underlying SCM. This enables investigations to determine whether learned disentangled representations follow the same causal structure as used to generate the data. Zheng et al. [310] build a structural causal graph and note that the predictions of recommendation systems are using the intertwined information of user interest, as well as the general popularity of items. Therefore, they propose a multitask learning framework to causally disentangle these properties, where individual interest and popularity representations are learned on separate auxiliary datasets and later concatenated for the recommendation. In addition to the added interpretability of the causally disentangled representation, they also report increased robustness in the non-IID setting [310]. Si et al. [251] continue this line of work by combining search and recommendation data to disentangle latent representation into a causally-relevant personalization part and a causally non-relevant part.

3.3 Post-Hoc Causal Interpretability

Post-hoc causal interpretation methods aim at causally explaining existing non-interpretable models after they have been trained, thus supplementing the good performance of complex models with human-understandable causal explanations. Causal methods for explaining non-interpretable models can be divided into two groups: black-box and white-box approaches.

3.3.1 Black-Box. Model agnostic approaches treat the model as a black-box and provide an explanation considering the input and output to make the model’s predictions human-comprehensible without requiring access to the model’s parameters. Approaches for causally explaining black-box models can be separated into two main categories: Feature attribution and counterfactual intervention methods.

Feature Attribution: Feature attribution methods quantify the input feature’s contribution to the prediction, but unlike non-causal approaches, causal approaches aim to retrieve only causally relevant features. In causal filtering, counterfactual input representations are generated and evaluated by masking potentially influential features, e.g., determined by attention and measuring the deterioration in performance.

Kim and Canny [125] apply this method to real-time videos to estimate the features’ causal influence on the prediction. Schwab and Karlen [236] use the importance distribution determined by masking to build a causal explanation model in parallel with the prediction model and minimize the Kullback-Leibler divergence between their importance distributions. Causal methodologies have also been applied in prompt-based interpretability. Cao et al. [37] investigate risk factors and confounders by prompting language models to complete a sentence to elicitate whether the model learned certain information from pre-training. To do so, they built an SCM to identify risk factors by backdoor-paths and propose causal interventions to study the methodologically induced biases. Li et al. [139] further investigates how the input sentence formulation can influence the model predictions by constructing counterfactual model inputs. To do so, they mask potentially relevant context words and study the impact on the amount of factual knowledge retained by the model. Their experiments suggest that co-occurrence, as well as the closeness of the subject (e.g. “Albert Einstein”) with the object (e.g. the birthplace), are highly relevant.

Frye et al. [74] extend the notion of Shapley values to asymmetric Shapley values (ASV) and build up a causal framework to quantify the contribution of a feature to the model’s prediction. However, this notion requires at least some knowledge about the causal ordering of features as it otherwise reduces to classical symmetric Shapley values without ordering. Janzing et al. [111] introduce a causal view to SHAP, considering the model’s inputs as causes of the output. They

argue for the use of interventional conditional distributions in SHAP to quantify the contribution of each observation to the output. This approach allows for determining the causal relevance of the input features without prior knowledge about the features' causal relationships. Alvarez-Melis and Jaakkola [9] use perturbation to explain specific input-output pairs for any input-output structured model by constructing a dense bipartite graph of perturbed inputs and their output tokens to infer a causal model using Bayesian logistic regression. A graph partitioning framework derives an explainable dependency graph by minimizing the k -cut. The perturbations introduce uncertainty information via the frequency with which each token occurs and thus allow one to reveal flaws or biases in a model.

Counterfactual Interventions: Tan et al. [257] use counterfactual interventions to explain decisions of recommendation models. To create counterfactuals, the authors manipulate input items while optimizing for minimally changed items that reverse the recommendation. The minimal changes now serve as explanations for the original recommendations. Mahajan et al. [155] propose a causal proximity regularizer to incorporate knowledge on causal relations from an SCM and thus constrain generated counterfactual explanations to be actionable recourses that are feasible in real-world settings.

3.3.2 *White-Box.* White-box interpretability methods, in contrast to black-box methods, require access to the model parameters and thus are also called model introspective approaches. Causal approaches use causal mediation or probing.

Causal Mediation: To identify the components of a model responsible for a biased prediction, Jeoung and Diesner [112], Vig et al. [274] introduce the method of causal mediation analysis. More formally, causal mediation analysis is a method to examine the intermediate processes where independent variables affect dependent variables. By intervening on the model and measuring direct and indirect effects, Vig et al. [274] investigate gender bias in individual neurons and attention heads and find a small subset of attention heads and neurons in the intermediary layers, which are responsible for biased predictions. Jeoung and Diesner [112] follow a similar causal mediation setup but use it to investigate the effects of debiasing techniques, finding that the debiased representations are robust to fine-tuning.

Probing: Probing usually involves training a small classifier to predict a property of interest from the model's latent embeddings. Elazar et al. [66] construct a counterfactual embedding without the property of interest (POI). They then infer the usage of POI during inference if the performance drops after the removal of POI information. To construct these counterfactual embeddings, several methodologies have been proposed - either by iteratively training classifiers to identify relevant dimensions in the embedding space [211], posing the removal as a minimax [212], or as a gradient-guided search problem [266]. Elazar et al. [66] utilize this *causal probing* approach and study the effects of part-of-speech knowledge on language modeling. Lasri et al. [132] (causally) probe language models on the usage of grammatical number information, finding that this resides in different layers depending on the token type (verbs/nouns). Furthermore, Tucker et al. [266] utilize their gradient-guided counterfactual creation method and find evidence for language models using tree distance-based embeddings to represent syntax. Finally, Feder et al. [71] investigate the effect of concepts like the presence of political figures on model prediction. Using adversarially trained counterfactual representations (without the concept under investigation), they contrast the classification performance of the standard and counterfactual representations.

3.4 Conclusion

The surge in causality papers over the last few years also affected the interpretability field. We believe that either causal - and therefore interpretable by design - models or the usage of causal

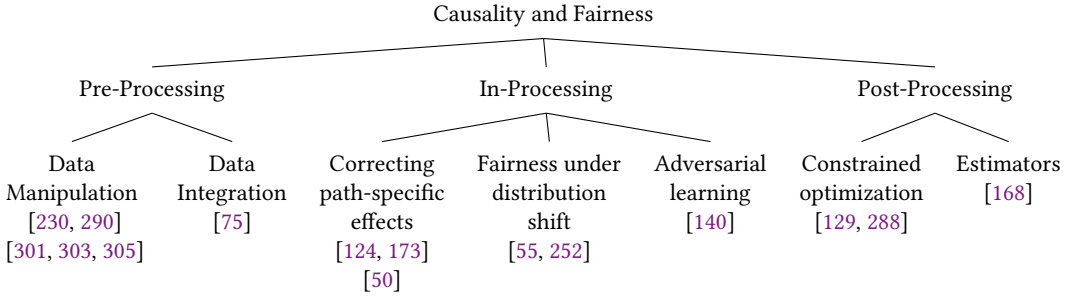


Fig. 3. Segregation of causality-based methods for fairness-aware learning.

(post-hoc) interpretability for standard ML models will be paramount for building trustworthy models. Existing causal methods [112, 274] provide information on the parts of an ML model that causally affect predictions and thus also provide information on biased or erroneous models. However, the underlying processes that lead to biased predictions and, in turn, the best approach to counteract these biases is still not well understood, and methods need to be developed in that direction. We observe structural causal models to be generated and used for many problems but still lack standardized procedures for evaluation. For example, an assessment of the extent to which the provided explanations help humans to causally understand a model’s output would be helpful. The SCS [100] provides a suitable framework for causality evaluation, however, its effectiveness depends upon a wider adoption of user studies in evaluating AI research.

4 CAUSALITY AND FAIRNESS

Addressing issues of fairness is a prerequisite for applying AI-based learning algorithms to support decisions that critically affect people’s lives, such as offender recidivism, loan approval, disease diagnosis, hiring, student admissions, etc. A complete understanding of the causal relationships between the sensitive attributes (e.g. gender, race, marital status, etc.) and the predicted outcome may play a crucial role in analyzing and legitimizing the fair or unfair behavior of a learning algorithm. This section aims to provide an overview of the contributions of causality in existing fairness research, its promises, and how it might be helpful in building fairer models.

The use of causality to describe and quantify fairness is a distinguishing feature of research conducted in the field of causal fairness. Causality has also proven to be effective in mitigating discrimination. We can divide the state-of-the-art causal frameworks for discrimination mitigation into the following categories: (a) Pre-processing methods, (b) In-processing methods, and (c) Post-processing methods as presented in Figure 3.

4.1 Preliminaries

The notions of fairness can be categorized into two groups: *individual* and *group*. Group fairness notions assess the large-scale biased effect of the learning algorithm on a certain legally protected group of the underlying dataset. Individual fairness notions measure the difference in the decisions predicted for similar individuals in a population. Causal fairness notions can be further segregated based on two criteria: (a) *Counterfactual fairness (CF)* and (b) *Interventional Fairness (IF)*.

4.1.1 Counterfactual fairness (CF). CF measures fairness by quantifying the effect of sensitive attributes on the predicted outcome through counterfactuals. If the sensitive attribute (e.g. S =

‘gender’) is binary then it could take two values protected, i.e. the disadvantaged group (p^+ = ‘female’) and non-protected, i.e. favoured (p^- = ‘male’) group.

Counterfactual fairness [130] is achieved by a predictor Y for an individual if the probability of achieving the output $Y = y$ remains the same if the value of sensitive attribute changes from p^- to p^+ as presented in Equation (1), where $X = V \setminus \{S, Y\}$ is the set of all variables except the sensitive and outcome variables.

$$\mathbb{P}(y_{p^+}|X = x, S = p^+) = \mathbb{P}(y_{p^-}|X = x, S = p^-) \quad (1)$$

where $P(y_{p^+})$ is a short notation for $P(Y = y|do(S = p^+))$. This individual fairness notion assumes that the effect of sensitive attributes on the decision along all causal paths is unfair. But this may not be true in some cases, e.g. from Figure 1, the direct effect of race on the admission outcome is unfair; however, the indirect effect of race on the outcome through the qualification variable is fair.

Path specific counterfactual fairness [49], in contrast to the counterfactual fairness notion, attempts to remove the causal effects of sensitive attributes on the outcome along only unfair causal paths. For a set of paths λ , path-specific counterfactual fairness exists if Equation (2) is satisfied, where $\bar{\lambda}$ is the set of remaining paths and $P(y_{p^+})$ is a short notation for $P(Y = y|do(S = p^+))$.

$$\mathbb{P}(y_{p^+}|\lambda, p^-, \bar{\lambda}) = \mathbb{P}(y_{p^-}) \quad (2)$$

PC-fairness [289] is an additional path-specific counterfactual fairness notion for subgroups not just individuals. Given a set of paths (λ) and a factual condition $X = x$ ($X \in V$), a predictor Y attains PC-fairness if it satisfies the following criteria:

$$\mathbb{P}(y_{p^+}|\lambda, p^-, \bar{\lambda}|X) = \mathbb{P}(y_{p^-}|X), \quad (3)$$

where $P(y_{p^+})$ is a short notation for $P(Y = y|do(S = p^+))$.

Counterfactual equalized odds [168] fairness notion is satisfied by a predictor if the respective counterfactual false positive rates (cFPR) and counterfactual false negative rates (cFNR) of the protected group and non-protected group are equal which is not possible practically. Therefore, we can use approximate counterfactual equalized odds. This notion is satisfied by a predictor if the constraints (4) hold, where ε^+ and ε^- are predefined thresholds. $Diff^+$ ($Diff^-$) is the difference between cFPR (cFNR) of the protected and the non-protected group as presented in Equation (5).

$$|Diff^+| \leq \varepsilon^+, \quad \varepsilon^+ \in [0, 1] \text{ and } |Diff^-| \leq \varepsilon^-, \quad \varepsilon^- \in [0, 1] \quad (4)$$

$$Diff^+ = cFPR(p^+) - cFPR(p^-) \text{ and } Diff^- = cFNR(p^+) - cFNR(p^-) \quad (5)$$

Causal Explanation formula [299] is a causal explanation method that helps in dividing the observed discrimination into three counterfactual effects: direct (DE), indirect (IE), and spurious effects (SE) of sensitive attributes on the outcome. The authors designed a causal explanation formula and decomposed the total variation (TV) into DE, SE, and IE, as presented in the following equation.

$$TV_{p^+, p^-}(Y = y) = |SE_{p^+, p^-}(Y = y) + IE_{p^+, p^-}(Y = y|S = p^-) - DE_{p^+, p^-}(Y = y|S = p^-)| \quad (6)$$

This formula demonstrates that the total discrimination experienced by the individuals with $S = p^-$ equals the disparity experienced via SE, plus the advantage lost due to IE, and minus the advantage it would have gained without DE.

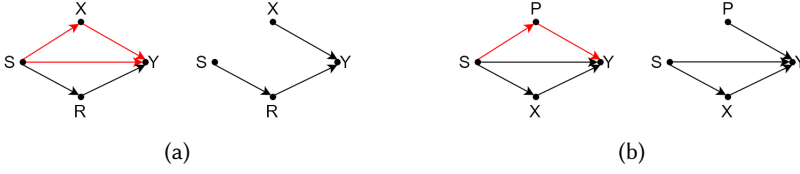


Fig. 4. (a) The left causal graph exhibits unresolved discrimination (along red paths) while the right one is free from unresolved discrimination where S is the sensitive variable, R is the resolving variable, X is the unresolving variable, and Y is the predicted decision. (b) The left causal graph exhibits proxy discrimination (along red paths) while the right one is free from ‘proxy discrimination’ where S is the sensitive variable, P is the proxy variable, and Y is the predicted decision.

4.1.2 Interventional Fairness (IF). IF measures fairness by quantifying the effect of sensitive attributes on the predicted outcome by intervening on the protected and non-protected attributes.

No unresolved discrimination [124] is a group fairness notion that focuses on the direct and indirect causal influence of sensitive attributes on the decision. It is satisfied when there is no direct path between the sensitive attributes and the outcome, except through a resolving/ admissible variable. In a causal graph, a resolving variable is a variable that is influenced by the sensitive attributes in an unbiased manner. The left causal graph in Figure 4 (a) exhibits discrimination along the causal paths: $S \rightarrow Y$ and $S \rightarrow X \rightarrow Y$ while the right one is free from discrimination, where R is the resolving variable, X is unresolving variable, and Y is the predicted outcome.

Proxy discrimination [124] is another indirect causality-based group fairness notion. It is present in a causal graph when the path between the sensitive attributes and the outcome is intercepted by a proxy variable. A predictor Y avoids proxy discrimination if, for a proxy variable P , Equation (7) holds for all potential values of P (p_1, p_2). A proxy variable has the same influence on the outcome as the influence of the sensitive attributes. The left causal graph in Figure 4 (b) exhibits discrimination along the causal paths: $S \rightarrow P \rightarrow Y$; where P is the proxy variable and Y is the predicted outcome.

$$\mathbb{P}(Y = y | do(P = p_1)) = \mathbb{P}(Y = y | do(P = p_2)) \quad \forall p_1, p_2 \in dom(P). \quad (7)$$

Total Effect [196] is the causal version of the statistical parity group fairness notion. It measures the effect of changing sensitive attribute values on the outcome along all causal paths from sensitive attributes to the outcome as presented in Equation (8)

$$TE_{p^+, p^-}(Y = y) = \mathbb{P}(y_{p^+}) - \mathbb{P}(y_{p^-}) \quad (8)$$

Individual direct discrimination [302] identifies direct discrimination at individual level. In any classification task, an individual is compared with n similar individuals sought from the protected group, denoted as D and n similar individuals from the non-protected group, denoted as \bar{D} . The similarity between the two individuals (i, i') is measured using causal inference as presented in Equations (9) and (10), where CE is the causal effect of each selected variable (x_k) on outcome variable, VD is the distance function, and $range$ is the difference between the maximum and minimum of the variable x_k . The individual is deemed not to be discriminated against if the difference between the positive prediction rates for the two groups (D, \bar{D}) is under a predefined threshold.

$$d(i, i') = \sum_{k=1}^{|X|} |CE(x_k, x_{k'}) \cdot VD(x_k, x_{k'})| \quad (9)$$

$$CE(Y = y) = \mathbb{P}(Y = y | do(X)) - \mathbb{P}(Y = y | do(x'_k, X \setminus x_k)) \text{ and } VD(x_k, x_{k'}) = \frac{|x_k - x_{k'}|}{range} \quad (10)$$

Equality of effort [104] assesses discrimination by measuring the amount of effort required by the marginalized individual or group to reach a certain level of the outcome as shown in Equation (11). The minimal effort required to achieve the γ -level of outcome is computed using Equation (12), where G_+ and G_- represent the set of individuals with $S = p^+$ and $S = p^-$ respectively which are similar to the target individual, $E[Y_{G_+}^t]$ is the expected value of outcome under treatment $T = t$ for the set G_+ . This notion of fairness is based on potential outcomes framework.

$$\psi_{G^+}(\gamma) = \psi_{G^-}(\gamma) \quad (11)$$

$$\psi_{G^+}(\gamma) = \operatorname{argmin}_{t \in T} \mathbb{E}[Y_{G_+}^t] \geq \gamma \quad (12)$$

Interventional and justifiable fairness [230] are stronger versions of the Total Effect fairness notion. The total effect intervenes on the sensitive attribute; however, interventional fairness intervenes on all attributes except the sensitive attribute. A classification algorithm is interventionally K -fair if for any assignment of $K = k$ and output $Y = y$ the following equation holds. K is a subset of attributes (V) except the sensitive attribute (S) and the outcome variable ($K \subseteq V \setminus \{S, Y\}$)

$$\mathbb{P}(y_{p^+,k}) = \mathbb{P}(y_{p^-,k}) \quad (13)$$

Justifiable fairness is a special case of interventional fairness, where we only consider those attributes for intervening that are admissible/resolving (E) or a superset of admissible variables:

$$\mathbb{P}(y_{p^+,k}) = \mathbb{P}(y_{p^-,k}), k \supseteq E. \quad (14)$$

Causal fairness [75] identifies a classifier as fair if for any given set of admissible variables E , the following equation holds:

$$\mathbb{P}(y_{p^+,e}) = \mathbb{P}(y_{p^-,e}), e \subseteq E. \quad (15)$$

4.2 Pre-processing Methods

Pre-processing methods are considered to be the most generalizable methods. These methods intend to manipulate the dataset in order to make it bias-free before feeding it to any learning algorithm. **Data Augmentation:** Zhang et al. [303] calculate path-specific effects of sensitive attributes on the predicted outcome and compare them to a predefined threshold τ . If the calculated path-specific effect exceeds τ , this indicates the presence of direct and indirect discrimination. Later they eliminate both direct and indirect discrimination by generating a bias-free dataset through causal network manipulation that guarantees path-specific effects under τ . Zhang et al. [305] further identify and handle the situations in which indirect discrimination cannot be measured because of the non-identifiability of certain path-specific effects. In such cases, the authors suggest setting an upper and lower bound on the effect of indirect discrimination. Another discrimination discovery and prevention causal framework is proposed by Zhang and Wu [301]. In this work, the authors detect direct and indirect system-level discrimination by measuring the path-specific causal effects of sensitive attributes on the outcome. To prevent discrimination, they modified the causal network to generate a new bias-free dataset. Xu et al. [290] has proposed a utility-preserving and fairness-aware causal generative adversarial network (CFGAN) to generate high-quality and bias-free data. Salimi et al. [230] detected discrimination using interventional and justifiable fairness notions. To eliminate discrimination, they used causal dependencies between sensitive attributes and outcome variables to add and remove samples from the training data.

Data Integration: Data integration aims to combine data from various sources that capture a comprehensive context and enhance predictive ability. Galhotra et al. [75] modeled the problem of ensuring *causal fairness* in a learning task as a fair data integration problem by combining

additional features with the original dataset. They proposed a conditional testing-based feature selection method that guarantees high predictive performance without adding bias to the dataset.

4.3 In-processing Methods

In-processing methods achieve fairness by altering the learning algorithm.

Correcting path-specific effects: The sensitive attributes can affect the outcome through both fair and unfair causal pathways as explained in Section 2.1. Kilbertus et al. [124] proposed to deal with such a situation by constraining the parameters of the learning algorithm so that the causal effects along both fair and unfair causal paths from sensitive attribute to outcome variable is removed. They used “proxy discrimination” and “unresolved discrimination” fairness notions to detect discrimination. [173] proposed to deal with such situations by constraining the path-specific effect during model training within a certain range. The methods proposed by [124, 173] cater for the causal effects of sensitive attributes on the outcome without distinguishing between fair and unfair causal effects, thus negatively impacting the predictive performance of the learning algorithm. A solution to this problem is proposed by Chiappa [49]. The author presented a causal framework that ensures path-specific counterfactual fairness by correcting the observations of such variables that are descendants of sensitive attributes along only unfair causal paths so that only individual unfair information is eliminated while the individual fair information is retained, hence improving predictive performance of the framework.

Fairness under distribution shift: The problem of learning fair prediction models with covariates distributed differently in the test set than in the training set is studied by Singh et al. [252]. They proposed a method based on feature selection to achieve fairness given the ground truth graph that explains the data. Fairness concerns also surface when AI-based learners deal with dynamically fluctuating environments and produce long-term effects for both individual and protected groups. Creager et al. [55] have proposed that in such dynamical fairness setting when the dynamic parameters are unknown, causal inference can be utilized to estimate the dynamic parameters and improve off policy estimation from historical data.

Adversarial learning: Li et al. [140] proposed an adversarial learning-based approach to achieve the goal of personalized counterfactual fairness for users in recommendation systems. They attempt to remove sensitive features information from the user embeddings by using a filter module and a discriminator module to make the learner’s decisions independent of the sensitive features.

4.4 Post-processing Methods

These methods tailor the outputs of the learner to achieve fair outcomes.

Constrained optimization: Wu et al. [288] proposed a method to bound the unidentifiability of counterfactual quantities and used c -component factorization to identify its source. They proposed a graphical criterion to determine the lower and upper bound on counterfactual fairness in unidentifiable scenarios. Finally, they proposed a post-processing method to reconstruct the decision model to achieve counterfactual fairness. Similarly, Kusner et al. [129] achieved counterfactual fairness by constraining the beneficial effects obtained by an individual under a limit depending on the sensitive attribute of the individual.

Doubly robust estimators: Mishler et al. [168] proposed a post-processed predictor, estimated using doubly robust estimators, to achieve the counterfactual equalized odds fairness notion. Through experiments, they also proved that their method has favorable convergence properties.

4.5 Conclusion

A unique trait of fairness research is the usage of multiple metrics to define/measure it. Consequently, choosing the most appropriate notion of fairness applicable to a particular situation is an important

task. On the other hand, even if a fairness notion is found suitable for a scenario, it may not be applicable due to the problem of identifiability, as Pearl’s SCM framework requires causal quantities, counterfactuals, and interventions, to be identifiable [156]. Most of the discrimination mitigation approaches discussed above to achieve the goal of causal fairness are based on the synthesis of a bias-free dataset. These methodologies are only applicable in a static environment where all data are available in advance. An interesting future direction could be the extension of such methods for online learning, where not all data are available beforehand [75]. Most of the causality-based fairness solutions discussed above rely on the assumption that the underlying data is independent and identically distributed (IID). However, real-world use cases include non-IID data, therefore, another future direction could be to design causality-based decision support systems which relax the assumption of non-IID data and provide non-discriminatory predictions. Finally, all research done in the field of causal fairness is connected to classification tasks [301], it will be interesting to expand it to achieve causal fairness in community detection, word embedding, named entity recognition, representation learning, semantic role labeling, language models, and machine translation [140].

5 CAUSALITY AND ROBUSTNESS

Training of modern machine learning (ML) systems builds upon the assumption that all observations - whether training or test data - are *independent and identically distributed* (IID) under a single distribution. As it is improbable that training data can perfectly represent the sample distribution of real-world data, striving for more robust ML systems is of utmost importance. Robust behavior is not only required for safety-critical applications of ML but also essential for the trustworthiness of AI. We believe that AI systems that are error-prone to small changes in the working environment will not be able to gain the complete trust of humans. Even high-performing ML models often cannot differentiate between *style variables*, which are content-independent and irrelevant information for the task, and the information-rich, relevant, and invariant *content variables* [117, 237]. Causal information about the problem can provide models with a better understanding of the task and eliminate such spurious correlations. Inspired by this perspective, many researchers in recent years investigated the connection between robust machine learning and causality. We summarize related publications and organize them into three broad categories: (i) **pre-processing**, (ii) **in-processing**, and (iii) **post-processing** methods. Figure 5 provides an overview of the techniques discussed in this section.

5.1 Preliminaries

We briefly introduce the most important concepts for this section, including a high-level definition of robustness and the primary sources of non-robust behavior.

Robustness: The notion of robustness boils down to the question of how sensitive the ML model’s output is to changes in the input. Minor changes in the input should not significantly alter the performance of robust AI systems. Instead, performance should degrade *gracefully*: slowly and gradually with the deviation in input distribution. We distinguish between naturally occurring and artificially crafted distributional shifts. The first type of shift is represented by the notion of *out-of-distributional* data, whereas the second type of shift is represented by *adversarial examples*. **Out-Of-Distribution Data:** Out-Of-Distribution (OOD) data represents naturally occurring data with previously unseen characteristics. For instance, computer vision models trained to solve MNIST classification may encounter naturally perturbed images (e.g. numbers written in a different orientation or a different color). Such perturbations represent data points from a different, shifted distribution and, therefore, may lead to poor performance of our computer vision model. These data points are, thus, considered to be **out-of-distribution**. Methods such as *Data Augmentation* (e.g. methods discussed in [250]) can increase robustness towards OOD-data by providing models with

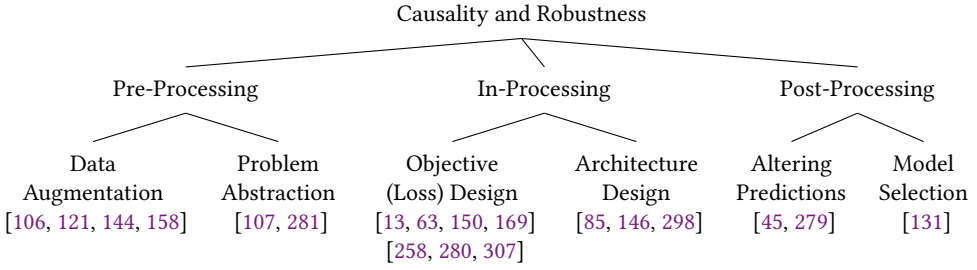


Fig. 5. Structure of approaches introducing causality in robustness.

additional, naturally perturbed data via hand-crafted rules [235, 258]. These methods, however, cannot cover all possible environmental settings [258].

Adversarial Examples: Another hurdle in robust machine learning is the continuous rise of *Adversarial Examples* (AE). AEs represent artificially perturbed input values intending to fool machine learning models. Such examples are especially concerning for the field of Trustworthy AI, as these perturbed input values look benign to humans. Despite the existence of various developed defenses against continuously evolving AEs, research into improved attacks and defenses continues to this day. We refer to [6] for a recent overview of such methods in the computer vision domain.

5.2 Pre-Processing

Pre-processing methods built upon causality are mostly *Data Augmentation* methods, which create causally motivated augmentations. We will also discuss alternative, exciting approaches to pre-processing that we cover under the umbrella term *Problem Abstraction*.

5.2.1 Data Augmentation. Data Augmentation is the most common pre-processing method to induce causality. There are several methods [106, 158] which use the notion of causal graphs to motivate data augmentation. For instance, Ilse et al. [106] try to find and apply transformations that emulate the intervention on high-level domain-specific features (e.g., the orientation of handwritten digits) within data points. Such information is only spuriously correlated to the output label and, as such, should not affect the decision-making of ML models. To find such a transformation without an SCM, the authors propose to train a classifier that can predict the domain of data points (e.g., the given number is rotated by 60°). They then choose the augmentation of a pre-defined set of transformations that leads to the *lowest* accuracy of the domain classifier. Applying the selected augmentation "destroys" the most domain-specific information. Such augmented training data, in turn, reduces the likelihood of ML models overly relying on domain-specific features that are only spuriously correlated to the label. Mao et al. [158] show that it is also possible to generate intervention-simulating data via GANs by identifying *interpretable controls* through GANSpace. Manipulating the data generated through these controls is equivalent to intervening in the underlying SCM. Alternatively, one can generate counterfactual examples by augmenting the data just enough to flip the label. Following this principle, Kaushik et al. [121] developed a human-in-the-loop process that improves robustness on NLP tasks compared to alternative, non-causal methods. As shown by Little and Badawy [144], it is also possible to induce causality through bootstrapping. The authors developed *causal bootstrapping*, which utilizes information provided by a causal graph to sample data whose deducible observations better reflect the domain's causal relationships. Models trained on causally sampled data demonstrate increased robustness against spurious correlations.

5.2.2 Problem Abstraction. Instead of data sampling, some methods try to abstract and simplify the problem. One example is the *Datamodeling* [107] framework, which allows researchers to approximate the behavior of large and complex models on the given data through a set of simple linear functions. Wang et al. [281] simplify the problem for reinforcement learning (RL) agents using information encoded in a causal graph. The authors propose to create state abstractions for RL agents that only contain the relevant one-to-one causal dependencies between variables and actions. RL agents that utilize this state abstraction exhibit higher robustness towards unseen states, cover a wider range of tasks than agents trained with other methods and demonstrate higher sample efficiency.

5.3 In-Processing

It is possible to induce notions of causality as part of the algorithm either through the definition of a *causality-aware optimization objective* or via *architectural design choices*.

5.3.1 Objective (Loss) Design. Most causal in-processing techniques incorporate an optimization objective (e.g. a loss function or a regularization term) that guides ML models to a more causality-aware behavior. We will introduce three possible causal foundations for such objectives.

Content and Style Variables: Zhang et al. [307] use a causal graph to model the generation process of AEs to examine the source of adversarial vulnerability. They conclude that AEs exploit the spurious correlations between style variables and labels to mislead classifiers. However, the *adversarial distribution* is drastically different from the natural one. Consequently, the authors developed a loss function that *aligns* the two distributions. Compared to other robust classifiers, aligned classifiers demonstrate higher accuracy on adversarial data without significantly worse performance on natural data. Wang et al. [280] developed a regularization term for *logistic regression* models allowing researchers to penalize causal and spurious features separately. This form of regularization explicitly requires researchers to categorize features as either causal, spurious, or remaining (not identified as either of the two). Given such information, models optimized with the term showcase improved robustness on lower and higher-dimensional data.

Multiple Environments: One vital insight for causal ML is the connection between the causal relevance of a feature and its invariance across environments [200]. The basic idea is that style variables (e.g. the image background) greatly vary across environments (i.e. unique experimental settings), whereas content variables (e.g. an animal’s anatomy) remain *invariant*. Consequently, guiding models to perform well across environments should lead to models capable of differentiating between content and style variables, which, in turn, makes them more robust. The most prominent example is the loss function *invariant risk minimization* (IRM) [13] that empirically leads to higher robustness. Mitrovic et al. [169] were able to transfer this premise to the Self-Supervised Learning (SSL) setting to improve OOD performance. The resulting SSL objective requires generated data representations to be stable across different interventions simulated by data augmentation. On the assumption that “the prior over the data representation belongs to a general exponential family when conditioning on the target and the environment” [150], the *iCaRL* framework is able to outperform IRM, but causal discovery is needed to identify the causally-relevant latent variables.

Counterfactuals: Researchers also successfully developed counterfactual-based loss functions. Teney et al. [258] designed an auxiliary loss for supervised learning that gives additional attention to pairs of data that are counterfactuals of one another. Raising awareness for counterfactuals can also increase the robustness of already causal methods, such as recommendation via Causal Algorithmic Recourse. Such algorithmic recourse systems try to find minimal-costly actions that result in a counterfactual representing a desirable outcome. Dominguez-Olmedo et al. [63] were able to further enhance this framework by also accounting for uncertainties stemming from adversarially

perturbed features. However, instead of considering all possible perturbations within an ϵ -range, the authors solely considered perturbations within the (SCM-guided) *counterfactual neighborhood* (i.e., only instances in ϵ -range that represent counterfactuals to the given data).

5.3.2 Architecture Design. Goyal et al. [85] try to take advantage of the *independent causal mechanisms* (ICM) principle. It states that the causal dynamics of a domain are built upon “autonomous modules that do not inform or influence each other.” [237]. They achieve this by implementing a sequential architecture of independently acting recurrent subsystems that only communicate sparsely with one another. Each subsystem is designed to emulate a mechanism of the causal generative process in the hope of better capturing the domain’s causal structure. The resulting architecture is more robust to distributional shifts than, e.g. LSTM or Transformers.

An alternative, more feature-focused architecture is *deep Causal Manipulation Augmented Model* (deep CAMA) [298] - a deep generative model whose design is consistent with the causality encoded in a given causal graph. It not only considers the effect of the output label on the input data but also considers manipulable variables (e.g., rotation and color of MNIST digits) and non-manipulable variables (e.g., handwriting style of MNIST digits) influence on the input. The authors achieve this by adding autoencoders for both the variables. This design choice allows the resulting generative model to better distinguish between relevant/causal features and non-relevant ones. Deep CAMA showcased improved robustness against adversarial attacks on MNIST data.

Liu et al. [146] follow a similar approach to designing a robust motion forecasting model. The authors argue that latent variables of the motion forecasting tasks are either (i) *invariant variables* such as laws of physics, which are crucial for correct motion forecasting, (ii) *hidden confounders* like the motion style that can sparsely differ between environments but still need to be considered for optimal performance, or (iii) *non-causal spurious features* that can drastically vary between environments. Based on this categorization, the authors developed an architecture that provides higher robustness toward style shifts.

5.4 Post-Processing

The causality framework also enables researchers to impact the robustness of their ML-pipeline *after the training phase*. Post-processing methods discussed in this section either directly alter the predictions of a given trained model or enable a causality-informed selection between a set of models.

5.4.1 Altering Predictions. *Counterfactual regularization* [117], which tries to remove the confounding effects of unobserved variables, is a common approach to instilling causality through post-processing. One can achieve this by first estimating the effects of unobserved variables, referred to as the *counterfactual prediction* [45]. By then taking the difference between the counterfactual prediction and the *factual prediction* (built upon both causally relevant and spurious information), a deconfounded prediction can be extracted. Using such a counterfactual regularization, Chen et al. [45] successfully improved the quality of trajectory predictions in multi-domain settings. Wang et al. [279] introduce an alternative approach called *Invariant-Feature Subspace Recovery* (ISR). The authors first extract the feature representation of the given data via the model’s hidden layers. They then recover the subspace spanned by invariant features (i.e., content variables), fit a linear predictor in the resulting manifold, and substitute the model’s classification layer with the subspace predictor. Post-processing with ISR improved the performance of trained models on multiple OOD-datasets.

5.4.2 Model Selection. Kyono and van der Schaar [131] introduce a pipeline that utilizes (incomplete) causal information in the form of a DAG in a post-hoc manner. Given a set of trained models and the data these models used, it is possible to score models based on how well the

model’s predictions abide by the “rules” induced by the given causal structure (DAG). The authors propose to choose the model whose predictions follow the given causal relationships the most. Classifiers chosen via this causal model selection technique empirically showcase higher robustness in OOD-learning settings.

5.5 Conclusion

Despite the relative novelty of causal learning, researchers have already applied the notions of causality to domains such as computer vision, NLP, and recommendation. Naturally occurring distributional shifts are the focus of these advancements, though causal learning can also increase robustness towards AEs. The success of causality in fields outside of traditional supervised learning, e.g. reinforcement learning [281] and self-supervised learning [169] further show the legitimacy of improving robustness via causal learning. We see multiple exciting avenues for future studies into robust causal ML. For instance, researchers need to be aware that OOD-datasets may vastly differ in the type of distributional shifts they emulate [292]. Hence, causal solutions for robust ML need to be analyzed and compared to other (both causal and non-causal) state-of-the-art approaches on benchmarks that contain a diverse collection of datasets. It will also be interesting to (further) explore related fields such as *Neurosymbolic AI* (where researchers enhance ML systems through the use of knowledge-based systems) or *Object-Centric Learning* (a special case of Causal Representation Learning [235] where visual scenes are modeled as compositions of objects that interact with one another). We also believe that research into causal solutions for *certified robustness* and *concept drifts* occurring in online learning will lead to interesting new solutions.

6 CAUSALITY AND PRIVACY

In Section 5, we discussed the robustness aspect of AI from the viewpoint of the model and noted out-of-domain generalization as a fundamental challenge. The problem becomes even more challenging in a distributed learning scenario, where the AI model gathers data from different users for training. In such a setting, it must be ensured that the users’ private data is not exposed. Recent studies [261, 295] have demonstrated that weak generalizability of a model could be exploited to design attacks that can expose users’ private information (e.g., whether the user’s data were used in training the model). With their inherent ability to generalize to out-of-distribution (OOD) data, causal models can help in preventing such attacks.

6.1 Preliminaries

Learning paradigm: Machine learning algorithms are often trained on a large amount of training data, which can contain sensitive information. We consider a particular setting where data are collected from one or multiple users to train a machine-learning model. For example, the data could be collected from users’ mobile devices to train a better speech recognition model. The most relevant variant in the context of privacy is **federated learning** (FL) [161]. FL proposes to learn a global model without collecting users’ data into a central server, rather keeping the data in users’ devices. Data are processed locally to update a local model, while intermediate model updates are sent to the central server, which are then aggregated to update the global model. Once updated, the global model is sent to the local devices. In this part, we will specifically look into attacks on federated learning as it involves dealing with users’ data where privacy is of utmost importance.

Privacy attacks: There has been a growing body of work aimed at designing different types of attacks against FL systems (refer to Jere et al. [113] for a comprehensive overview). However, one of the most common types of attack posed to FL systems and where causality-driven models have been deployed more often is **membership inference attack** [249]. In this type of attack, the attacker tries to determine whether a particular data point was used in training the model,

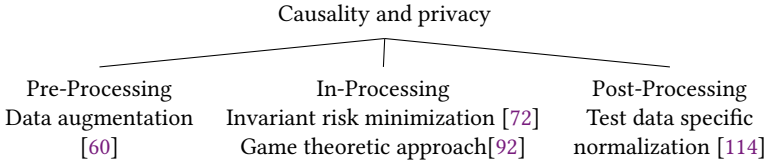


Fig. 6. Causality and privacy approaches.

with only model predictions available to the attacker [176, 295]. The core idea is to exploit the stochastic gradient descent (SGD) algorithm to extract information about a client’s data. The attacks can be initiated both from the server’s side and the client’s side. For a comprehensive review of membership inference attacks on FL, refer to Nasr et al. [176]. Membership inference attack has been linked to model generalizability [176], and causal models are adept at improving the generalizability of models. Hence, the bulk of work deploying causal models against privacy attacks has concentrated on combating membership inference attacks, which we discuss in detail in this section. *Differential privacy* (DP) [64] ensures that the presence or absence of a data point does not significantly influence the output. Hence, a defense mechanism against membership attacks should provide better differential privacy guarantees.

Evaluation: The success of a membership inference attack is measured in terms of accuracy on the binary classification task of determining membership [249]. Similarly, Yeom et al. [295] introduce *advantage*, which is measured as the difference in true and false positive rates in membership prediction. Hence, the goal of any defense mechanism against membership attacks is to reduce the attack’s accuracy or advantage.

Causal models for improving privacy: Yeom et al. [295] demonstrate that model overfitting significantly contributes to the leakage of membership information. For example, consider a model A , a data point z , and a loss function $l \leq B$ where B is a constant. An attack strategy is to first query the model (i.e., computing $A(z)$) and then output that z as a member of the training set with probability $1 - l(A, z)/B$. For such an attack, the performance in determining membership is shown to be proportional to the generalization error (i.e., the extent of overfitting). Hence, the defense mechanisms propose to deploy model generalization techniques such as learning rate decay, dropout [229] or adversarial regularization [177]. However, these mechanisms assume that the train and the test datasets are sampled from the same distribution, which is not always the case, particularly for FL. In fact, Tople et al. [261] argue that vulnerability to such attacks is exacerbated when a model is deployed on unseen data. They further utilize the generalization property of causal models and establish a theoretical link between causality and privacy. It is further argued that the models learned through causal features generalize better to distribution shifts and provide better privacy guarantees than equivalent association models. This has resulted in a growing body of work exploring causal models against privacy attacks [42, 60, 72, 92, 114, 261]. We segregate these methods into **pre-**, **in-** and **post-**processing methods (refer to figure 6).

6.2 Pre-processing

The key idea here is to manipulate the training data at each client, which would result in better generalization. de Luca et al. [60] introduce a causality-based data augmentation to mitigate the problem of domain generalization in FL. The authors argue that data augmentation can reduce the heterogeneity across user data distribution, thereby making them more similar for the server model. The proposed data augmentation method is based on the notion of structural causal model (SCM) as described in section 2.3. The authors posit that a data point X_i is generated by a common cause Z

and some random variation ϵ_i . Following SCM, this could be represented as $X_i := g(Z, \epsilon_i)$ where g is the causal mechanism that remains invariant. An augmented data point \hat{X}_i could then be generated by defining a transformation τ over ϵ_i and then following the SCM $\hat{X}_i := g(Z, \tau(\epsilon_i))$.

6.3 In-processing

This class of methods aims at learning domain invariant features while training. Francis et al. [72] propose to collaboratively learn causal features common to all collaborating users/clients through invariant risk minimization [13] (introduced in Section 5.3). The proposed method is able to defend better against membership inference and property inference attacks in comparison to the vanilla FL algorithms. Gupta et al. [92] propose FL games, a game theoretic method, to solve the problem of generalization in FL. The key idea is to learn causal representations that are invariant across users. Each client serves as a player that competes to optimize its local objective, while the server guides the optimization to a global objective. An equilibrium is reached when all the local models across users become equivalent, thereby achieving generalization across users' data distributions as well as finding invariant representations. Although the authors do not explicitly demonstrate the effectiveness of their model against specific attacks, it can be argued that better generalization should lead to improved robustness against inference attacks. Given these methods aim to discover a set of invariant features that directly influence the outcome (hence speculated as causal), they can be loosely placed under causal discovery. However, it is difficult to conclude whether the discovered features are indeed causal.

6.4 Post-processing

Jiang et al. [114] argue that the FL training could be represented as a structural causal model with four variables, input data (X), raw extracted features (R), normalized features (F) (obtained by applying batch normalization [108]) and the output (Y), following a causal structure $X \rightarrow R \rightarrow F \rightarrow Y$. Although heterogeneity of data can lead to each client fitting its individual feature distribution as opposed to the global objective, batch normalization (BN) layers normalize the training data to a uniform distribution, thereby allowing the global model to converge. However, when dealing with unseen test data (D^u), BN layers with the estimated training statistics run the risk of normalizing it improperly, thereby introducing an edge $D^u \rightarrow F$ and making D^u a confounder. Performing a causal intervention by introducing a surrogate variable S , which is test-specific statistics of raw features R (i.e., BN normalization parameters are now computed from the test set instead of the training set), leads to blocking the path between D^u and F and getting rid of the confounder. This achieves better generalization and hence better robustness against attacks.

6.5 Conclusion

In this section, we presented an overview of various attempts at improving defenses against privacy attacks on FL systems through causality-driven methods. Most of them have exploited the OOD generalization capabilities of causal methods as this successfully defends against inference attacks, which calls for similar investigations with respect to other types of attacks. Except for Tschantz et al. [263], who propose to apply results from causality theory while studying differential privacy (DP) [64], this has largely remained unexplored and might be another interesting line of future research.

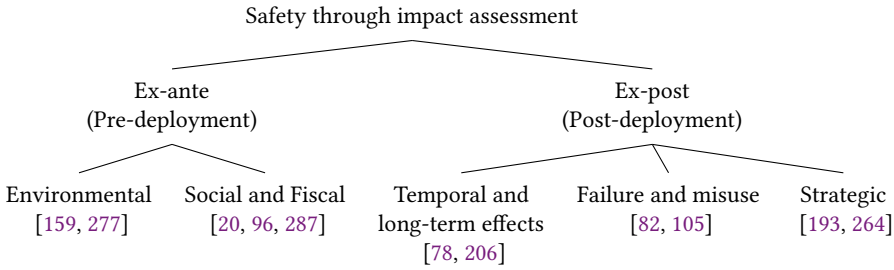


Fig. 7. Safety through (causal) impact assessment.

7 CAUSALITY AND AUDITING (SAFETY AND ACCOUNTABILITY)

Automated systems often enabled by AI, when deployed at scale, interact with people, each other, and also with other ecosystem or environmental parameters, thereby having a potential for widespread unforeseen (desirable or undesirable) effects. Some undesirable effects can cause significant societal harm, such as the creation of online filter bubbles and polarization due to personalized consumption-centric information retrieval systems powered by various ML models [192]. Following such effects of AI, there has been a lot of interest in finding potential negative effects before deployment or on the run and make necessary changes to avoid or neutralize such effects. Selbst [238] describes Algorithmic Impact Assessment (AIA) as a regulatory strategy for addressing and correcting algorithmic harms. EU regulatory guidelines also emphasize the need for environmental, social, and economic impact assessments before deployment. The phrase ‘impact assessment’ itself suggests a causal relationship between system design and impact. Thus, causal inference has long been used for impact assessment of algorithmic and non-algorithmic systems. Depending on when an assessment occurs, it can be categorized into (i) ex-ante (before deployment) and (ii) ex-post (after deployment) impact assessment.

7.1 Ex-ante impact assessment

When impacts are assessed before deployment of a particular system or policy (often using prior knowledge or use cases, empirical data from system testing, and system design details), we talk of ex-ante impact assessment. Ex-ante impact assessment often predicts the risks and impacts of the proposed system. Such assessments are used as tools to assess potential environmental, financial, social, and human rights ramifications of systems (both algorithmic and non-algorithmic), projects, and policies, and grant some measure of control and voice to designers or developers, affected population, and authorities to make, induce or enforce changes accordingly.

Environmental impact assessment (EIA): Both deductive and inductive causal inference have long been used in EIA [159, 277]. While with a deductive approach, a hypothesis about a causal relationship is formed and tested, using an inductive approach, data are collected from observations, and a causal relationship is inferred from the instances. Causal networks have long been used for EIA [159, 277] since they bring both network (multiple interaction pathways between environment and various activities) and cause-effect (various activities affecting the elements of the environment) logic to the analysis. Causal networks allow an analysis of impacts through sequences of interactions [159]—also referred to as *sequence diagrams* [36]. Causal networks (structural causal models) have helped find indirect impacts on multiple levels [36]. While initial EIA studies have helped find forms and parameters of relationships between various kinds of activities/developments and environmental elements, they are also utilized to (ex-ante) estimate the environmental impacts

of planned new or upcoming projects, including complex AI systems in the near future [254].

Social and Fiscal impact assessment (SIA and FIA): Becker [20] defines SIA as *the process of identifying the future consequences of a current or proposed action which are related to individuals, organizations and social macro-systems*. Similar to the works in EIA, SIA also uses both deductive and inductive methods to discover the causal relationships between actions (e.g. inventory optimization, logistic changes) and consequences (e.g., increase in sales, popularity, overall perception of the product) [58, 287]. Note that most of the works here use the potential outcomes framework. Moreover, economists have long been looking for ways to construct counterfactuals and answer causal and policy evaluation questions [94]. The majority of works on FIA or economic impact assessment try to understand the effects of introducing new economic policies or changing existing ones [96]. Such analyses are done in a multi-agent setting with modeling of the preferences and choices of agents along with their ability to infer evaluations and outcomes. Causal and counterfactual economic assessment has been a big part of FIA of many automated systems [96].

7.2 Ex-post impact assessment

Since ex-ante impact assessments are limited to the available prior knowledge and use cases, it is often not possible to identify all possible risks and impacts of a system or policy, which motivates ex-post impact assessment. When impacts are assessed after deployment (often in real-time using the running record of the system, real-time audits, and evaluations), we call this ex-post assessment. Ex-post impact assessment often detects the risks and impacts on the go once the system is introduced. While ex-ante impact assessments often have clear guidelines or metrics for specific types of impacts (e.g., average CO₂ emission for environmental impact, total financial cost, opportunity cost and return-on-investment for economic impact, overall positive or negative opinions of the population for social impact), ex-post impact assessments are generally broader since they need to define what constitutes an impact in real-time. Causal inference is used to tackle various categories of risks and impacts of a system or policy in real time.

Temporal and long-term effects: Many kinds of temporal and long-term effects are seen in real-world systems which operate inside an eco-system of stakeholders. For example, the creation of online filter bubbles and polarization due to personalized consumption-centric information retrieval systems powered by various ML models [78, 206], the content homogenization effects observed in online marketplaces as a response to popularity bias in recommender systems [43]. Causality, along with behavioral modeling, has helped assess such systems in real-time and find out the elements responsible [239].

Effects from system failure and system misuse: Systems are often designed with some desired criteria (e.g., accuracy, fairness, robustness, etc.). If a real-world model deviates from the desired criteria and produces unwanted outputs, *accountability* helps identify the reason behind that failure and take required actions. For example Gillespie et al. [82] studies how the inclusion of a robot as a team member in surgeries increases the complexity and errors that it was supposed to reduce. Making systems and stakeholders accountable for such failures of the system is an integral part of safety. Ibrahim et al. [105] propose a bottom-up causal approach using goal-specific accountability mechanisms. Their mechanisms can help identify the root cause of specific type(s) of events or failures, which can then be used to eliminate the underlying (technical) problem and also to assign blame. Causality has also been very helpful in analyzing (networked or other) system attacks and finding out the root cause and the source host or process from where the attack has originated [51, 148].

Strategic risks and effects: While interpretability and explainability are essential for trustworthiness, one must also consider various risks of using interpretable and explainable decision-making systems in the real world. For example Shokri et al. [248] analyze connections between model

explanations and the leakage of sensitive information in the model's training set even if a model is used as a black-box; They show that back-propagation-based explanations can leak a significant amount of information about individual training data points, exhibiting a potential conflict between privacy and interpretability. Similarly Tsirtsis and Gomez Rodriguez [264] show that counterfactual explanations can reveal various details of decision-making systems, thereby making them vulnerable towards strategic behaviors and, therefore, non-robust. Since in real-world settings, individuals (either using the decision-making system or being affected by its decisions) often try to optimize their outcome, their rational strategic behavior can cause issues of privacy and robustness. Because of the above strategic risks, recent works [193, 264] suggest modeling real-world strategic settings as games (from applied game theory) and then designing decision-making systems that can mitigate the effects of strategic behavior. Essentially, these works [193, 264] try to design decision-making systems that incentivize individuals only to improve a desired quality (to improve individual outcomes) and not do any strategic manipulations.

7.3 Conclusion

We discussed the importance of impact assessment and auditing for the safety and accountability of AI systems and the use of causality in different types of ex-ante and ex-post assessment. However, as Moss et al. [172] outline, there is no pre-existing or universal definition of impact that can be used in safety assessment of algorithmic systems. Instead, the idea of an impact can vary depending on the context, scale, and application domain. Adding to that, Irving and Asbell [109] justifiably argues that figuring out AI safety would need social scientists since it depends on human values and expectations. Given clear definitions of undesirable effects, causality can be very helpful, particularly in studying the effects of different design elements of algorithmic systems in simulated environments, and also in figuring out which design element(s) are responsible for certain undesirable effects observed in either real or simulated environments.

8 CAUSALITY IN HEALTHCARE

In the previous sections, we mostly demonstrated the effectiveness of causality-based methods for improving different trustworthy aspects of AI. However, causal methods have proved to be reliable tools in many application domains as well. In this section, we look into *healthcare*, where causal methods have been demonstrated to be particularly advantageous. Additionally, we also point out how causal methods could be effective in enhancing the trustworthiness aspects of such applications.

Several existing papers highlight the application of causality approaches in the healthcare domain. Through SCMs, discussed in **Section 2.3**, we can better analyze diseases (by revealing the causal relations of the diseases' features), improve the accuracy of automatic diagnosis, or even discover new drugs. Moreover, causal reasoning can also be used to investigate drugs' outcome through the PO framework, discussed in **Section 2.4**. Figure 8 depicts these different use cases. In this section, we will survey existing research based on the two broad categories of causal inference, SCMs and PO, while focusing on interpretability, robustness, fairness, and privacy.

8.1 Causality in Healthcare through SCM framework

The literature provides many reviews which focus on the usage of SCMs in healthcare and personalized medicine [231, 275, 306]. Zhang et al. [306] provided an introduction to causality using different medical examples like lung cancer causal graphs and shed light on the different challenges and issues encountered when dealing with causal inference, such as missing data, biased data, and transferability of models. Vlontzos et al. [275] present the benefits of introducing causality and its use in the field of medical imaging. Their survey reviews several applications that incorporate

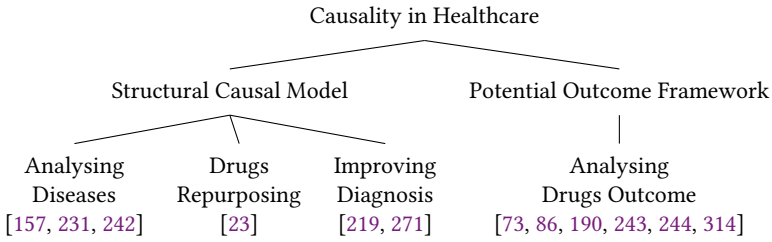


Fig. 8. Structure of different healthcare applications used with causality.

causal discovery and causal inference in medical imaging. Sanchez et al. [231] specifically focused on Alzheimer’s disease (AD), a progressive neurodegenerative disorder, to highlight the utility of causal machine learning in precision medicine. Shen et al. [242] used two causal discovery methods to discover the causal relationship utilizing observational data of AD. The authors used the *fast causal inference* (FCI) [253], a constraint-based algorithm, and the *fast greedy equivalence search* (FGES) [208], a score based algorithm. These approaches are then compared and benchmarked with a well-established causal graph. Mani and Cooper [157] tried to identify causal factors of clinical conditions for intensive care unit (ICU) patients using medical discharge reports. In the above references, we observe that causality methods are mostly used to interpret and explain the outcomes of medical models. Such approaches and ideas make the AI system interpretable by design, which is desirable in healthcare use cases. By providing an explanation and finding causal relations between different factors of a particular disease, different insights are gained, which in turn contribute to a better, trustworthy AI system.

Drug discovery is another research area that benefits from causality. Belyaeva et al. [23] show that causal models can be used to repurpose drugs for new diseases like SARS-CoV-2. The research team integrated transcriptomic, proteomic, and structural data for different diseases. They first used autoencoders to match a drug’s signature with a reverse disease signature in the latent space. Using the augmented Steiner tree, the disease interactome is then identified. Lately, they have verified the causal interaction of the drugs with genes by using a causal structure discovery algorithm and building a causal network.

Apart from interpretability, few studies addressed fairness and robustness issues. In precision medicine, we aim for a fair system that provides personalized and equitable treatment to each individual without any bias [47]. Chen et al. [47] gave the example of biased systems in healthcare. An algorithm trained only on USA cancer pathology data may lead to wrong classification, when deployed on data from Turkish cancer patients, due to protocol variations or population shifts (imbalanced data). The authors argued that causality is one of the technologies (others being fairness-aware federated learning, features disentanglement, etc.) that contribute towards a fair algorithm in healthcare by performing causal analysis to identify the bias factors. We should therefore include causality and analyze the causal structures, which could be discovered or provided by the clinicians to make biased AI algorithms fair in real-world scenarios and healthcare applications.

Robustness is another aspect of trustworthy AI that is addressed in various causality papers related to healthcare applications. It is important to design a system that is robust to change in the distribution and provides reliable outputs and accurate results. Van Amsterdam et al. [271] improved their lung cancer image prediction algorithm by eliminating bias signals. By predicting the collider variable (tumor size) and the prognostic factor (tumor heterogeneity), it was then possible to unbiased the estimation and make the system more robust. Richens et al. [219] improved

the accuracy of medical diagnosis through the use of causal machine learning. Their counterfactual algorithm helps them to improve the decision-making process and classify different vignettes correctly based on Bayesian networks that model known relationships between multiple diseases and integrate the causal relationship between different variables. Such algorithms incorporate the collider variables and also different causal relations in the network architecture design, which makes them in-processing methods that enhance robustness.

8.2 Causality in Healthcare through the PO framework

The PO framework is commonly used in the medical field. It provides methods to conduct causal analysis from a statistical perspective, as already introduced in **Section 2.4**. In real-world scenarios, observational data are biased (e.g. biased labeling, under-representation, etc.). PO framework methods provide a way to remove the selection bias in the historical data, thereby leading to a fair system (e.g. the *propensity score matching* method: **Section 2.4**).

Shi and Norgeot [243] review different research works which focused on learning causal effects from observational data in the medical domain. These survey articles summarize different methods used to estimate treatment effects. In the medical field, it is common to use the PO framework to test whether a particular drug is beneficial or harmful. For instance, Graham et al. [86] used propensity score matching [222] to examine whether Dabigatran or Warfarin increase the risk of death in elderly patients from nonvalvular atrial fibrillation. Similarly, Ozer et al. [190] investigated the benefits of chemotherapy in comparison to only undergoing surgery for patients with resectable gallbladder cancer. By performing propensity score matching [222] analysis, it was possible to find out that chemotherapy increases the survival rate. Friedrich and Friede [73] tried several propensity score-based methods utilizing the PO framework, in addition to other approaches such as g-computation and doubly robust estimators [153]. They designed a simulation to compare these different methods. It mimics data from a small non-randomized study on the efficacy of hydroxychloroquine for COVID-19 patients Ziff et al. [314] used PO framework-based analysis (propensity score matching [222]) to evaluate the safety and efficacy of the drug digoxin for patients with heart failure.

As discussed in **Section 6**, privacy is one of the main foundations required for a trustworthy AI system. An AI system should not allow the identification of specific patients based on the available training datasets. This aspect was one of the major concerns covered in the paper [244]. This paper states that the publicly available dataset helps assess a single treatment's efficacy. However, to investigate multiple treatments and correctly estimate the causal effect through observational data, the authors generated a new large synthetic dataset that imitates real-world data distributions and preserves individual patients' privacy. They reported the ϵ -*identifiability* metric that estimates the probability that an individual is identifiable and ensures that this value remains low after the data generation process.

8.3 Conclusion

Work in the healthcare domain encompassing causality mainly considers the aspects of interpretability and robustness and touch upon fairness and privacy to some extent. The other tenets of trustworthy AI, like safety and accountability, must be explored further. Integrating causality into precision medicine is still facing several challenges that need to be addressed. When discovering causal relations (i.e. causal discovery) among different features, ground truth causal graphs are not always available to validate the results [101]. This implies over-trusting the data, and the discovered relations can be overcome by working closely with experts and clinicians to integrate their knowledge. Available data in the healthcare domain are commonly unstructured, highly complex, and multimodal [231]. Thus, there is an urge to develop better decision-making algorithms that

not only find correlations in these data but also understand causal relations and perform causal reasoning.

9 CONCLUSION

In this article, we surveyed causal modeling and reasoning tools for enhancing the trustworthy aspects of AI models, which include interpretability, fairness, robustness, privacy, safety, and accountability. While in recent years, the community has witnessed an unprecedented surge of research in this context, important facets still remain unexplored. We expect significant advancements in the coming years and hope this survey will act as an important resource to the community and at the same time guide future research connecting trustworthy AI and causality.

ACKNOWLEDGMENTS

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Action “NoBIAS - Artificial Intelligence without Bias” (grant agreement number 860630) and Network of Excellence “TAILOR - A Network for Trustworthy Artificial Intelligence” (grant agreement number 952215), the Lower Saxony Ministry of Science and Culture under grant number ZN3492 within the Lower Saxony “Vorab” of the Volkswagen Foundation and supported by the Center for Digital Innovations (ZDIN), and the Federal Ministry of Education and Research (BMBF), Germany under the project “LeibnizKILabor” with grant No. 01DD20003 and from Volkswagen Foundation and the Ministry for Science and Culture of Lower Saxony, Germany (MWK) under the "Understanding Cochlear Implant Outcome Variability using Big Data and Machine Learning Approaches" (grant no. ZN3429) project.

REFERENCES

- [1] E. D. Abraham, K. D’Oosterlinck, A. Feder, Y. O. Gat, A. Geiger, et al. 2022. CEbaB: Estimating the Causal Effects of Real-World Concepts on NLP Model Behavior. *arXiv preprint arXiv:2205.14140* (2022).
- [2] A. Abyaneh, N. Scherrer, P. Schwab, S. Bauer, B. Schölkopf, et al. 2022. FED-CD: Federated Causal Discovery from Interventional and Observational Data. *arXiv preprint arXiv:2211.03846* (2022).
- [3] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi. 2018. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4971–4980.
- [4] O. Ahmed, F. Träuble, A. Goyal, A. Neitz, M. Wüthrich, et al. 2021. CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning. In *International Conference on Learning Representations*.
- [5] K. Ahuja, K. Shanmugam, K. Varshney, and A. Dhurandhar. 2020. Invariant risk minimization games. In *International Conference on Machine Learning*. PMLR, 145–155.
- [6] N. Akhtar, A. Mian, N. Kardan, and M. Shah. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access* 9 (2021), 155161–155196.
- [7] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. 2010. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research* (2010).
- [8] C. F. Aliferis, A. R. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. 2010. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II: Analysis and Extensions. *Journal of Machine Learning Research* (2010).
- [9] D. Alvarez-Melis and T. S. Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint* (2017).
- [10] American Psychological Association. 2022. PsycINFO. Retrieved December 30, 2022 from <https://www.apa.org/pubs/databases/psycinfo/index>
- [11] D. Anguita, A. Ghio, L. Oneto, X. Parra Perez, and J. L. Reyes Ortiz. 2013. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*. 437–442.
- [12] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.
- [13] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. 2019. Invariant risk minimization. *arXiv* (2019).

- [14] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics* 38, 2 (2011), 915–931.
- [15] S. Avram, C. G. Bologna, J. Holmes, G. Bocci, T. B. Wilson, et al. 2021. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic acids research* 49, D1 (2021), D1160–D1169.
- [16] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, et al. 2018. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging* 38, 2 (2018), 550–560.
- [17] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, et al. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems* 32 (2019).
- [18] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. [n. d.].
- [19] A. L. Beam et al. 2016. Medical Data for Machine Learning. <https://github.com/beamandrew/medical-data>.
- [20] H. A. Becker. 2001. Social impact assessment. *European Journal of Operational Research* 128, 2 (2001), 311–321.
- [21] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. <https://arxiv.org/abs/1810.01943>
- [22] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. 2013. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research* 47 (jun 2013), 253–279.
- [23] A. Belyaeva, L. Cammarata, A. Radhakrishnan, C. Squires, K. D. Yang, et al. 2021. Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing. *Nature Communications* 12, 1 (2021).
- [24] J. Bennett, S. Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. New York, NY, USA., 35.
- [25] P. J. Bickel, E. A. Hammel, and J. W. O’Connell. 1975. Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science* 187, 4175 (1975), 398–404.
- [26] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, et al. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Technical Report MSR-TR-2020-32. Microsoft. <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>
- [27] D. Blanco-Melo, B. E. Nilsson-Payant, W.-C. Liu, S. Uhl, D. Hoagland, et al. 2020. Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell* 181, 5 (2020), 1036–1045.
- [28] S. L. Blodgett, L. Green, and B. O’Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868* (2016).
- [29] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*. 12–58.
- [30] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).
- [31] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).
- [32] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* (2015), 247–274.
- [33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.
- [34] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, et al. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *CoRR* abs/2004.07213 (2020). arXiv:2004.07213 <https://arxiv.org/abs/2004.07213>
- [35] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, et al. 2018. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097* (2018).
- [36] L. W. Canter. 1982. Environmental impact assessment. *Impact Assessment* 1, 2 (1982), 6–40.
- [37] B. Cao, H. Lin, X. Han, F. Liu, and L. Sun. 2022. Can Prompt Probe Pretrained Language Models? Understanding the Invisible Risks from a Causal View. In *ACL*.
- [38] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, et al. 2022. MONAI: An open-source framework for deep learning in healthcare. (11 2022). <https://doi.org/10.48550/arXiv.2211.02701>
- [39] L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, et al. 2015. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreservation and biobanking* 13, 5 (2015), 311–319.
- [40] A. D. Center. 2016. The Medical Expenditure Panel Survey (MEPS). Retrieved November 25, 2022 from <https://meps.ahrq.gov/mepsweb/>
- [41] D. Chai, L. Wang, K. Chen, and Q. Yang. 2020. Fedeval: A benchmark system with a comprehensive evaluation model for federated learning. *arXiv preprint arXiv:2011.09655* (2020).

- [42] V. Chandrasekaran, D. Edge, S. Jha, A. Sharma, C. Zhang, et al. 2021. Causally Constrained Data Synthesis for Private Data Release. *arXiv preprint arXiv:2105.13144* (2021).
- [43] A. J. Chaney, B. M. Stewart, and B. E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM conference on recommender systems*. 224–232.
- [44] C. Chen, K. Lin, C. Rudin, Y. Shaposhnik, S. Wang, et al. 2018. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615* (2018).
- [45] G. Chen, J. Li, J. Lu, and J. Zhou. 2021. Human trajectory prediction via counterfactual analysis. *IEEE/CVF*.
- [46] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao. 2020. Causalm: Python package for causal machine learning. *arXiv preprint arXiv:2002.11631* (2020).
- [47] R. J. Chen, T. Y. Chen, J. Lipkova, J. J. Wang, D. F. Williamson, et al. 2021. Algorithm fairness in ai for medicine and healthcare. *arXiv preprint arXiv:2110.00603* (2021).
- [48] M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, et al. 2019. BabyAI: First Steps Towards Grounded Language Learning With a Human In the Loop. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=rJeXC0cYX>
- [49] S. Chiappa. 2019. Path-specific counterfactual fairness. In *AAAI*, Vol. 33. 7801–7808.
- [50] S. Chiappa and W. S. Isaac. 2018. A causal Bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*. Springer, 3–20.
- [51] J. Chow, B. Pfaff, T. Garfinkel, K. Christopher, and M. Rosenblum. 2004. Understanding data lifetime via whole system simulation. In *USENIX Security Symposium*. 321–336.
- [52] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. 2018. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6172–6180.
- [53] Cochrane. 2022. Cochrane Library. Retrieved December 30, 2022 from <https://www.cochranelibrary.com/central>
- [54] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. 2008. Sample selection bias correction theory. In *International conference on algorithmic learning theory*. Springer, 38–53.
- [55] E. Creager, D. Madras, T. Pitassi, and R. Zemel. 2020. Causal modeling for fairness in dynamical systems. In *ICML*. PMLR, 2185–2195.
- [56] F. Croce, M. Andriushchenko, V. Schwag, E. Debenedetti, N. Flammarion, et al. 2020. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670* (2020).
- [57] DataCanvas. 2022. YLearn. <https://github.com/DataCanvasIO/YLearn>.
- [58] J. B. de Araujo and C. M. L. Ugaya. 2018. Development of S-LCIA models: a review of multivariate data analysis methods. *Social LCA* (2018), 67.
- [59] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, et al. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.
- [60] A. B. de Luca, G. Zhang, X. Chen, and Y. Yu. 2022. Mitigating Data Heterogeneity in Federated Learning with Data Augmentation. *arXiv preprint arXiv:2206.09979* (2022).
- [61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, et al. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
- [62] D. Dheeru and E. K. Taniskidou. 2017. UCI machine learning repository. (2017).
- [63] R. Dominguez-Olmedo, A. H. Karimi, and B. Schölkopf. 2022. On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*. PMLR, 5324–5342.
- [64] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*. Springer, 486–503.
- [65] A. D’Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, et al. 2020. Fairness is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT ’20)*. Association for Computing Machinery, New York, NY, USA, 525–534. <https://doi.org/10.1145/3351095.3372878>
- [66] Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg. 2021. Amnesic Probing: Behavioral Explanation With Amnesic Counterfactuals. *Trans. Assoc. Comput. Linguistics* (2021).
- [67] H. Elsahar, P. Vougiouklis, A. Remaci, C. Gravier, J. Hare, et al. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- [68] Elsevier. 2023. Science Direct. Retrieved February 5, 2023 from <https://www.sciencedirect.com/>
- [69] A. Ess, B. Leibe, and L. Van Gool. 2007. Depth and appearance for mobile scene analysis. In *2007 IEEE 11th international conference on computer vision*. IEEE, 1–8.

- [70] H. Fanaee-T and J. Gama. 2014. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* 2, 2 (2014), 113–127.
- [71] A. Feder, N. Oved, U. Shalit, and R. Reichart. 2021. CausaLM: Causal Model Explanation Through Counterfactual Language Models. *Comput. Linguistics* (2021).
- [72] S. Francis, I. Tenison, and I. Rish. 2021. Towards causal federated learning for enhanced robustness and privacy. *arXiv preprint arXiv:2104.06557* (2021).
- [73] S. Friedrich and T. Friede. 2020. Causal inference methods for small non-randomized studies: Methods and an application in COVID-19. *Contemporary Clinical Trials* 99 (2020).
- [74] C. Frye, C. Rowat, and I. Feige. 2020. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. In *NeurIPS*.
- [75] S. Galhotra, K. Shanmugam, P. Sattigeri, K. R. Varshney, R. Bellamy, et al. 2022. Causal Feature Selection for Algorithmic Fairness. (2022).
- [76] C. Gao, Y. Zheng, W. Wang, F. Feng, X. He, et al. 2022. Causal Inference in Recommender Systems: A Survey and Future Directions. *arXiv preprint arXiv:2208.12397* (2022).
- [77] E. Gao, J. Chen, L. Shen, T. Liu, M. Gong, et al. 2021. Federated causal discovery. *arXiv preprint arXiv:2112.03555* (2021).
- [78] V. R. K. Garimella and I. Weber. 2017. A long-term analysis of polarization on Twitter. In *Eleventh international AAAI conference on web and social media*.
- [79] H. Geffner, R. Dechter, and J. Halpern (Eds.). 2022. *Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM Books.
- [80] N. George. 2018. All Lending Club loan data. Retrieved February 6, 2023 from <https://www.kaggle.com/datasets/wordsforthewise/lending-club>
- [81] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi. 2015. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*. 2551–2559.
- [82] B. M. Gillespie, J. Gillespie, R. J. Boorman, K. Granqvist, J. Stranne, et al. 2021. The impact of robotic-assisted surgery on team performance: a systematic mixed studies review. *Human factors* 63, 8 (2021), 1352–1379.
- [83] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 7816 (2020), 459–468.
- [84] Government of Canada. [n. d.]. Algorithmic Impact Assessment tool. Retrieved November 30, 2022 from <https://open.canada.ca/aia-eia-js/?lang=en>
- [85] A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, et al. 2021. Recurrent Independent Mechanisms. ICLR.
- [86] D. J. Graham, M. E. Reichman, M. Wernecke, R. Zhang, M. R. Southworth, et al. 2015. Cardiovascular, Bleeding, and Mortality Risks in Elderly Medicare Patients Treated With Dabigatran or Warfarin for Nonvalvular Atrial Fibrillation. *Circulation* 131, 2 (2015).
- [87] U. Groemping. 2019. South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep* 4 (2019), 2019.
- [88] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, et al. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
- [89] I. Gulrajani and D. Lopez-Paz. 2020. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434* (2020).
- [90] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. 2020. A Survey of Learning Causality with Data: Problems and Methods. *ACM Comput. Surv.* 53, 4 (2020), 75:1–75:37. <https://doi.org/10.1145/3397269>
- [91] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–37.
- [92] S. Gupta, K. Ahuja, M. Havaei, N. Chatterjee, and Y. Bengio. 2022. FL Games: A federated learning framework for distribution shifts. *arXiv preprint arXiv:2205.11101* (2022).
- [93] I. Guyon, C. Aliferis, et al. 2007. Causal feature selection. In *Computational methods of feature selection*.
- [94] T. Haavelmo. 1943. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society* (1943), 1–12.
- [95] F. M. Harper and J. A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (dec 2015), 19 pages. <https://doi.org/10.1145/2827872>
- [96] M. Haseeb, L. W. Miharjo, A. R. Gill, K. Jermisittiparsert, et al. 2019. Economic impact of artificial intelligence: new look for the macroeconomic assessment in Asia-Pacific region. *International Journal of Computational Intelligence Systems* 12, 2 (2019), 1295.
- [97] C. He, K. Balasubramanian, E. Ceyani, C. Yang, H. Xie, et al. 2021. Fedgraphnn: A federated learning system and benchmark for graph neural networks. *arXiv preprint arXiv:2104.07145* (2021).
- [98] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

- 8340–8349.
- [99] D. Hendrycks and T. Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).
- [100] A. Holzinger, A. Carrington, and H. Müller. 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations. *KI - Kunstliche Intelligenz* (2020).
- [101] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9 (2019).
- [102] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, et al. (Eds.). Curran Associates, Inc., 9737–9748. <http://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks.pdf>
- [103] S. Hu, Y. Li, X. Liu, Q. Li, Z. Wu, et al. 2022. The oarf benchmark suite: Characterization and implications for federated learning systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 4 (2022), 1–32.
- [104] W. Huan, Y. Wu, L. Zhang, and X. Wu. 2020. Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*. 743–751.
- [105] A. Ibrahim, S. Kyriakopoulos, and A. Pretschner. 2021. Causality-based accountability mechanisms for socio-technical systems. *Journal of Responsible Technology* 7 (2021), 100016.
- [106] M. Ilse, J. M. Tomczak, and P. Forré. 2021. Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*. PMLR, 4555–4562.
- [107] A. Ilyas, S. M. Park, L. Engstrom, G. Leclerc, and A. Madry. 2022. Datamodels: Understanding Predictions with Data and Data with Predictions. In *International Conference on Machine Learning*. PMLR, 9525–9587.
- [108] S. Ioffe and C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.
- [109] G. Irving and A. Askill. 2019. AI safety needs social scientists. *Distill* 4, 2 (2019), e14.
- [110] A. Jacovi, A. Marasovic, T. Miller, and Y. Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *FACCT ’21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 624–635. <https://doi.org/10.1145/3442188.3445923>
- [111] D. Janzing, L. Miniorics, and P. Blöbaum. 2020. Feature relevance quantification in explainable AI: A causal problem. *AISTATS* (2020).
- [112] S. Jeoung and J. Diesner. 2022. What Changed? Investigating Debiasing Methods using Causal Mediation Analysis. *CoRR* (2022).
- [113] M. S. Jere, T. Farnan, and F. Koushanfar. 2020. A taxonomy of attacks on federated learning. *IEEE Security & Privacy* 19, 2 (2020), 20–28.
- [114] M. Jiang, X. Zhang, M. Kamp, X. Li, and Q. Dou. 2021. TsmoBN: Interventional Generalization for Unseen Clients in Federated Learning. *arXiv preprint arXiv:2110.09974* (2021).
- [115] Jigsaw. 2019. Jigsaw Unintended Bias in Toxicity Classification. Retrieved November 23, 2022 from <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification>
- [116] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.
- [117] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva. 2022. Causal Machine Learning: A Survey and Open Problems. *arXiv preprint arXiv:2206.15475* (2022).
- [118] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi. 2023. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* 55, 2 (2023), 39:1–39:38. <https://doi.org/10.1145/3491209>
- [119] J. N. Kaur, E. Kiciman, and A. Sharma. 2022. Modeling the Data-Generating Process is Necessary for Out-of-Distribution Generalization. *arXiv preprint arXiv:2206.07837* (2022).
- [120] D. Kaushik, E. Hovy, and Z. C. Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434* (2019).
- [121] D. Kaushik, A. Setlur, E. Hovy, and Z. C. Lipton. 2020. Explaining the efficacy of counterfactually augmented data. *arXiv preprint arXiv:2010.02114* (2020).
- [122] N. R. Ke, A. Didolkar, S. Mittal, A. Goyal, G. Lajoie, et al. 2021. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848* (2021).
- [123] M. H. G. L. P. O. M. O. V. S. Keith Battocchi, Eleanor Dillon. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. <https://github.com/microsoft/EconML>. Version 0.x.
- [124] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, et al. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems* 30 (2017).
- [125] J. Kim and J. Canny. 2017. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. *ICCV*.

- [126] R. Kohavi et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.. In *Kdd*, Vol. 96. 202–207.
- [127] A. Krizhevsky, G. Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [128] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, et al. 2016. Zoneout: Regularizing rnns by randomly preserving hidden activations. *arXiv preprint arXiv:1606.01305* (2016).
- [129] M. Kusner, C. Russell, J. Loftus, and R. Silva. 2019. Making decisions that reduce discriminatory impacts. In *International Conference on Machine Learning*. PMLR, 3591–3600.
- [130] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).
- [131] T. Kyono and M. van der Schaar. 2019. Improving model robustness using causal knowledge. *arXiv preprint arXiv:1911.12441* (2019).
- [132] K. Lasri, T. Pimentel, A. Lenci, T. Poibeau, and R. Cotterell. 2022. Probing for the Usage of Grammatical Number. In *ACL*.
- [133] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsis. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2022), e1452.
- [134] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- [135] A. Lerner, Y. Chrysanthou, and D. Lischinski. 2007. Crowds by example. In *Computer graphics forum*, Vol. 26. Wiley Online Library, 655–664.
- [136] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales. 2017. Deeper, broader and artier domain generalization. In *ICCV*. 5542–5550.
- [137] H. Li, E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, et al. 2017. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature genetics* 49, 5 (2017), 708–718.
- [138] L. Li, X. Qi, T. Xie, and B. Li. 2020. Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131* (2020).
- [139] S. Li, X. Li, L. Shang, Z. Dong, C. Sun, et al. 2022. How Pre-trained Language Models Capture Factual Knowledge? A Causal-Inspired Analysis. In *ACL*.
- [140] Y. Li, H. Chen, S. Xu, Y. Ge, and Y. Zhang. 2021. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1054–1063.
- [141] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, et al. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [142] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.
- [143] P. Lison and J. Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. (2016).
- [144] M. A. Little and R. Badawy. 2019. Causal bootstrapping. *arXiv preprint arXiv:1910.09648* (2019).
- [145] K. Liu, G. Cao, F. Zhou, B. Liu, J. Duan, et al. 2022. Towards Disentangling Latent Space for Unsupervised Semantic Face Editing. *IEEE Trans. Image Process.* (2022).
- [146] Y. Liu, R. Cadei, J. Schweizer, S. Bahmani, and A. Alahi. 2022. Towards Robust and Adaptive Motion Forecasting: A Causal Representation Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17081–17092.
- [147] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, et al. 2022. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 4525–4542. <https://www.usenix.org/conference/usenixsecurity22/presentation/liu-yugeng>
- [148] Y. Liu, M. Zhang, D. Li, K. Jee, Z. Li, et al. 2018. Towards a Timely Causality Analysis for Enterprise Security.. In *NDSS*.
- [149] Z. Liu, P. Luo, X. Wang, and X. Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- [150] C. Lu, Y. Wu, J. M. Hernández-Lobato, and B. Schölkopf. 2021. Invariant Causal Representation Learning for Out-of-Distribution Generalization. In *International Conference on Learning Representations*.
- [151] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*. Springer, 189–202.
- [152] S. M. Lundberg and S.-I. Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [153] H. MA and R. JM. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- [154] A. Magrini, S. Di Blasi, F. M. Stefanini, et al. 2017. A conditional linear Gaussian network to assess the impact of several agronomic settings on the quality of Tuscan Sangiovese grapes. *Biometrical Letters* 54, 1 (2017), 25–42.

- [155] D. Mahajan, C. Tan, and A. Sharma. 2019. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. (12 2019). <http://arxiv.org/abs/1912.03277>
- [156] K. Makhlof, S. Zhoua, and C. Palamidessi. 2020. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553* (2020).
- [157] S. Mani and G. F. Cooper. 2000. Causal discovery from medical textual data. *AMIA Symp.* (2000).
- [158] C. Mao, A. Cha, A. Gupta, H. Wang, J. Yang, et al. 2021. Generative interventions for causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3947–3956.
- [159] A. R. D. Mareddy, A. Shah, and N. Davergave. 2017. *Environmental impact assessment: theory and practice*. Butterworth-Heinemann.
- [160] R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 92–97.
- [161] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.
- [162] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
- [163] S. Merity, C. Xiong, J. Bradbury, and R. Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843* (2016).
- [164] F. H. Messerli. 2012. Chocolate consumption, cognitive function, and nobel laureates. *The New England Journal of Medicine* 367, 10 (2012), 1562–1564.
- [165] Microsoft. 2022. Responsible AI Toolbox. <https://github.com/microsoft/responsible-ai-toolbox>. Version 0.23.
- [166] M. Milano and M. Schoenauer. 2022. The TAILOR Strategic Research and Innovation Roadmap. Available at <https://tailor-network.eu/research-overview/strategic-research-and-innovation-roadmap/>.
- [167] J. Miller, C. Hsu, J. Troutman, J. Perdomo, T. Zrníc, et al. 2020. WhyNot. <https://doi.org/10.5281/zenodo.3875775>
- [168] A. Mishler, E. H. Kennedy, and A. Chouldechova. 2021. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *FAccT*. 386–400.
- [169] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell. 2020. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922* (2020).
- [170] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. 2016. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research* 17, 1 (2016), 1103–1204.
- [171] R. Moraffah, M. Karami, R. Guo, A. Raglin, and H. Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.
- [172] E. Moss, E. A. Watkins, J. Metcalf, and M. C. Elish. 2020. Governing with algorithmic impact assessments: six observations. In *Watkins, Elizabeth and Moss, Emanuel and Metcalf, Jacob and Singh, Ranjit and Elish, Madeleine Clare, Governing Algorithmic Systems with Impact Assessments: Six Observations (May 14, 2021)*. AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES).
- [173] R. Nabi, D. Malinsky, and I. Shpitser. 2019. Learning optimal fair policies. In *ICML*. PMLR, 4674–4682.
- [174] R. Nabi and I. Shpitser. 2018. Fair inference on outcomes. In *AAAI*, Vol. 32.
- [175] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133* (2020).
- [176] M. Nasr, R. Shokri, and A. Houmansadr. 2018. Comprehensive privacy analysis of deep learning. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*. 1–15.
- [177] M. Nasr, R. Shokri, and A. Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 634–646.
- [178] National Center for Biotechnology Information. 2022. PubMed. Retrieved December 30, 2022 from <https://pubmed.ncbi.nlm.nih.gov/>
- [179] J. Ni, J. Li, and J. McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.
- [180] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, et al. 2018. Adversarial Robustness Toolbox v1.2.0. *CoRR* 1807.01069 (2018). <https://arxiv.org/pdf/1807.01069>
- [181] K. Niemelä et al. 2016. Awesome Health. <https://github.com/kakoni/awesome-healthcare>.
- [182] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or. 2020. Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics* (2020).
- [183] H. Nori, S. Jenkins, P. Koch, and R. Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).
- [184] OECD. 2021. Tools for trustworthy AI. 312 (2021). <https://doi.org/https://doi.org/10.1787/008232ec-en>

- [185] B. of Governors of the Federal Reserve System (US). 2007. *Report to the congress on credit scoring and its effects on the availability and affordability of credit*. Board of Governors of the Federal Reserve System.
- [186] U. G. A. Office. [n. d.]. Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. GAO-21-519SP Report. <https://www.gao.gov/products/GAO-21-519SP>
- [187] E. C. I. H.-L. E. G. on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. Retrieved August 10, 2021 from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419
- [188] OpenPowerlifting. 2019. Powerlifting Database. Retrieved December 28, 2022 from <https://www.kaggle.com/datasets/open-powerlifting/powerlifting-database>
- [189] Organisation for Economic Co-operation and Development. 2023. OECD Statistics. Retrieved February 6, 2023 from <https://stats.oecd.org/>
- [190] M. Ozer, S. Y. Goksu, N. N. Sanford, M. Porembka, H. Khurshid, et al. 2022. A Propensity Score Analysis of Chemotherapy Use in Patients With Resectable Gallbladder Cancer. *JAMA Network Open* 5, 2 (2022).
- [191] W. Pan, S. Cui, J. Bian, C. Zhang, and F. Wang. 2021. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *SIGKDD*. 1287–1297.
- [192] E. Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. penguin UK.
- [193] G. K. Patro, L. Porcaro, L. Mitchell, Q. Zhang, M. Zehlike, et al. 2022. Fair Ranking: A Critical Review, Challenges, and Future Directions. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAcT '22)*. Association for Computing Machinery, New York, NY, USA, 1929–1942. <https://doi.org/10.1145/3531146.3533238>
- [194] J. Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.
- [195] J. Pearl. 2001. Direct and Indirect Effects. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, Jack S. Breese and Daphne Koller (Eds.), Morgan Kaufmann, 411–420. https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=126&proceeding_id=17
- [196] J. Pearl. 2009. *Causality: Models, reasoning and inference* (second ed.). Cambridge University Press.
- [197] J. Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* 62, 3 (2019), 54–60.
- [198] J. Pearl and D. Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.
- [199] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, et al. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1406–1415.
- [200] J. Peters, P. Bühlmann, and N. Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78, 5 (2016), 947–1012.
- [201] J. Peters, D. Janzing, and B. Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.
- [202] R. C. Petersen, P. Aisen, L. A. Beckett, M. Donohue, A. Gamst, et al. 2010. Alzheimer’s disease neuroimaging initiative (ADNI): clinical characterization. *Neurology* 74, 3 (2010), 201–209.
- [203] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, et al. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066* (2019).
- [204] Project SEAPHE. 2007. Project SEAPHE: Databases. Retrieved December 28, 2022 from <http://www.seaphe.org/databases.php>
- [205] ProQuest. 2022. ProQuest. Retrieved December 30, 2022 from <https://www.proquest.com/>
- [206] P. Ramaciotti Morales and J.-P. Cointet. 2021. Auditing the effect of social network recommendations on polarization in geometrical ideological spaces. In *Fifteenth ACM Conference on Recommender Systems*. 627–632.
- [207] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR* (2022).
- [208] J. D. Ramsey. 2015. Scaling up Greedy Causal Search for Continuous Variables. *arXiv preprint arXiv:1507.07749* (2015).
- [209] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, et al. 2019. *Deep Learning for the Life Sciences*. O’Reilly Media. <https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837>.
- [210] J. Rauber, R. Zimmermann, M. Bethge, and W. Brendel. 2020. Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX. *Journal of Open Source Software* 5, 53 (2020), 2607. <https://doi.org/10.21105/joss.02607>
- [211] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *ACL*.
- [212] S. Ravfogel, M. Twiton, Y. Goldberg, and R. Cotterell. 2022. Linear Adversarial Concept Erasure. In *ICML*.
- [213] S. Razick, G. Magklaras, and I. M. Donaldson. 2008. iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics* 9, 1 (2008), 1–19.
- [214] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. 2019. Do imagenet classifiers generalize to imagenet?. In *International Conference on Machine Learning*. PMLR, 5389–5400.

- [215] A. G. Reddy, B. G. L., and V. N. Balasubramanian. 2022. On Causally Disentangled Representations. In *AAAI*.
- [216] H. Reichenbach. 1956. *The direction of time*. Vol. 65. Univ of California Press.
- [217] P. A. Reyfman, J. M. Walter, N. Joshi, K. R. Anekalla, A. C. McQuattie-Pimentel, et al. 2019. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *American journal of respiratory and critical care medicine* 199, 12 (2019), 1517–1536.
- [218] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [219] J. G. Richens, C. M. Lee, and S. Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications* 11, 1 (2020).
- [220] F. L. Rios, G. Moffa, and J. Kuipers. 2021. Benchpress: a scalable and platform-independent workflow for benchmarking structure learning algorithms for graphical models. arXiv:stat.ML/2107.03863
- [221] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*. Springer, 549–565.
- [222] P. R. Rosenbaum and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983).
- [223] S. Rosenthal, N. Farra, and P. Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 502–518. <https://doi.org/10.18653/v1/S17-2088>
- [224] D. B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 5 (1974).
- [225] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301* (2018).
- [226] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.
- [227] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* (2019).
- [228] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, et al. 2018. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577* (2018).
- [229] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, et al. 2018. MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246* (2018).
- [230] B. Salimi, L. Rodriguez, B. Howe, and D. Suci. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *MOD*. 793–810.
- [231] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O’Neil, et al. 2022. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science* 9 (2022).
- [232] E. Santana and G. Hotz. 2016. Learning a driving simulator. *arXiv preprint arXiv:1608.01230* (2016).
- [233] P. Schmidt and A. D. Witte. 1988. *Predicting recidivism in north carolina, 1978 and 1980*.
- [234] B. Schölkopf. 2022. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. 765–804.
- [235] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, et al. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.
- [236] P. Schwab and W. Karlen. 2019. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. *NeurIPS* (2019).
- [237] B. Schölkopf and J. von Kügelgen. 2022. From Statistical to Causal Learning. arXiv. <https://doi.org/10.48550/ARXIV.2204.00607>
- [238] A. D. Selbst. 2021. AN INSTITUTIONAL VIEW OF ALGORITHMIC IMPACT. *Harvard Journal of Law & Technology* 35, 1 (2021).
- [239] A. Sharma, J. M. Hofman, and D. J. Watts. 2015. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 453–470.
- [240] A. Sharma, E. Kiciman, et al. 2019. DoWhy: A Python package for causal inference. <https://github.com/microsoft/dowhy>.
- [241] W. Shen and R. Liu. 2017. Learning Residual Images for Face Attribute Manipulation. In *CVPR*.
- [242] X. Shen, S. Ma, P. Vemuri, G. Simon, M. W. Weiner, et al. 2020. Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer’s Pathophysiology. *Scientific Reports* 10, 1 (2020).
- [243] J. Shi and B. Norgeot. 2022. Learning Causal Effects From Observational Data in Healthcare: A Review and Summary. *Frontiers in Medicine* 9 (2022).
- [244] J. Shi, D. Wang, G. Tesei, and B. Norgeot. 2022. Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments. *Frontiers in Artificial Intelligence* 5 (2022).

- [245] Y. Shimoni, E. Karavani, S. Ravid, P. Bak, T. H. Ng, et al. 2019. An evaluation toolkit to guide model selection and cohort definition in causal inference. *arXiv preprint arXiv:1906.00442* (2019).
- [246] D. Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human Computer Studies* (2021).
- [247] B. Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *Int. J. Hum. Comput. Interact.* 36, 6 (2020), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- [248] R. Shokri, M. Strobel, and Y. Zick. 2021. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 231–241.
- [249] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.
- [250] C. Shorten and T. M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.
- [251] Z. Si, X. Han, X. Zhang, J. Xu, Y. Yin, et al. 2022. A Model-Agnostic Causal Learning Framework for Recommendation using Search Data. In *WWW*.
- [252] H. Singh, R. Singh, V. Mhasawade, and R. Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3–13.
- [253] P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, Prediction, and Search, Second Edition*. MIT Press.
- [254] E. Strubell, A. Ganesh, and A. McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3645–3650.
- [255] E. A. Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.
- [256] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, et al. 2017. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171, 6 (2017), 1437–1452.
- [257] J. Tan, S. Xu, Y. Ge, Y. Li, X. Chen, et al. 2021. Counterfactual Explainable Recommendation. *International Conference on Information and Knowledge Management, Proceedings*.
- [258] D. Teney, E. Abbasnedjad, and A. v. d. Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. In *European Conference on Computer Vision*. Springer, 580–599.
- [259] The World Bank Group. 2022. World Development Indicators. Retrieved November 29, 2022 from <https://data.worldbank.org/indicator>
- [260] S. Thiebess, S. Lins, and A. Sunyaev. 2021. Trustworthy artificial intelligence. *Electron. Mark.* 31, 2 (2021), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- [261] S. Tople, A. Sharma, and A. Nori. 2020. Alleviating privacy attacks via causal learning. In *International Conference on Machine Learning*. PMLR, 9537–9547.
- [262] A. Torralba and A. A. Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*. IEEE, 1521–1528.
- [263] M. C. Tschantz, S. Sen, and A. Datta. 2020. SoK: Differential privacy as a causal property. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 354–371.
- [264] S. Tsirtsis and M. Gomez Rodriguez. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems* 33 (2020), 16749–16760.
- [265] G. Tsitsiridis, R. Steinkamp, M. Giurgiu, B. Brauner, G. Fobo, et al. 2023. CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Research* 51, D1 (2023), D539–D545.
- [266] M. Tucker, P. Qian, and R. Levy. 2021. What if This Modified That? Syntactic Interventions with Counterfactual Embeddings. In *ACL-IJCNLP*.
- [267] Udacity. 2016. Self-Driving Car. <https://github.com/udacity/self-driving-car>.
- [268] U.S. Department of Education’s Office for Civil Rights (OCR). 2023. Civil Rights Data Collection. Retrieved January 30, 2023 from <https://ocrdata.ed.gov/>
- [269] U.S. National Library of Medicine. 2023. MEDLINE. Retrieved January 23, 2023 from https://www.nlm.nih.gov/medline/medline_overview.html
- [270] B. Ustun, A. Spangher, and Y. Liu. 2019. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*. 10–19.
- [271] W. A. C. Van Amsterdam, J. J. C. Verhoeff, P. A. de Jong, T. Leiner, and M. J. C. Eijkemans. 2019. Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning. *npj Digital Medicine* 2, 1 (2019).
- [272] S. Van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber. 2018. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. *arXiv preprint arXiv:1802.10353* (2018).
- [273] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5018–5027.
- [274] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, et al. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *NeurIPS*.

- [275] A. Vlontzos, D. Rueckert, and B. Kainz. 2022. A Review of Causality for Learning Algorithms in Medical Image Analysis. *arXiv preprint arXiv:2206.05498* (2022).
- [276] T. V. Vo, Y. Lee, T. N. Hoang, and T.-Y. Leong. 2022. Bayesian Federated Estimation of Causal Effects from Observational Data. In *The 38th Conference on Uncertainty in Artificial Intelligence*.
- [277] G. Voegeli, W. Hediger, and F. Romerio. 2019. Sustainability assessment of hydropower: Using causal diagram to seize the importance of impact pathways. *Environmental Impact Assessment Review* 77 (2019), 69–84.
- [278] R. Voigt, D. Jurgens, V. Prabhakaran, D. Jurafsky, and Y. Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- [279] H. Wang, H. Si, B. Li, and H. Zhao. 2022. Provable Domain Generalization via Invariant-Feature Subspace Recovery. *arXiv preprint arXiv:2201.12919* (2022).
- [280] Z. Wang, K. Shu, and A. Culotta. 2021. Enhancing Model Robustness and Fairness with Causality: A Regularization Approach. *arXiv preprint arXiv:2110.00911* (2021).
- [281] Z. Wang, X. Xiao, Z. Xu, Y. Zhu, and P. Stone. 2022. Causal dynamics learning for task-independent state abstraction. *arXiv preprint arXiv:2206.13452* (2022).
- [282] A. Williams, N. Nangia, and S. R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).
- [283] J. M. Wing. 2021. Trustworthy AI. *Commun. ACM* 64, 10 (2021), 64–71. <https://doi.org/10.1145/3448248>
- [284] World Economic Forum. 2023. World Economic Forum. Retrieved February 6, 2023 from <https://www.weforum.org/reports/>
- [285] World Intellectual Property Organization. 2023. Global Brand Database. Retrieved February 6, 2023 from <https://branddb.wipo.int/en/>
- [286] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3597–3606.
- [287] S. R. Wu, J. Chen, D. Apul, P. Fan, Y. Yan, et al. 2015. Causality in social life cycle impact assessment (SLCIA). *The International Journal of Life Cycle Assessment* 20, 9 (2015), 1312–1323.
- [288] Y. Wu, L. Zhang, and X. Wu. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.
- [289] Y. Wu, L. Zhang, X. Wu, and H. Tong. 2019. Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems* 32 (2019).
- [290] D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu. 2019. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.
- [291] J. N. Yan, Z. Gu, H. Lin, and J. M. Rzeszutowski. 2020. Silva: Interactively assessing machine learning fairness using causality. In *CHI*. 1–13.
- [292] N. Ye, K. Li, L. Hong, H. Bai, Y. Chen, et al. 2021. OoD-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721* (2021).
- [293] I.-C. Yeh and C.-h. Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications* 36, 2 (2009), 2473–2480.
- [294] Yelp. [n. d.]. Yelp Open Dataset. <https://www.yelp.com/dataset>. Accessed: 2022-11-23.
- [295] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 268–282.
- [296] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, et al. 2020. Causality-based Feature Selection: Methods and Evaluations. *Comput. Surveys* (2020).
- [297] K. Yu, L. Liu, and J. Li. 2021. A Unified View of Causal and Non-causal Feature Selection. *ACM Transactions on Knowledge Discovery from Data* 15 (2021).
- [298] C. Zhang, K. Zhang, and Y. Li. 2020. A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems* 33 (2020), 289–301.
- [299] J. Zhang and E. Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [300] K. Zhang, S. Zhu, M. Kalander, I. Ng, J. Ye, et al. 2021. gCastle: A Python Toolbox for Causal Discovery. *arXiv preprint arXiv:2111.15155* (2021).
- [301] L. Zhang and X. Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 4, 1 (2017), 1–16.
- [302] L. Zhang, Y. Wu, and X. Wu. 2016. Situation Testing-Based Discrimination Discovery: A Causal Inference Approach.. In *IJCAI*, Vol. 16. 2718–2724.
- [303] L. Zhang, Y. Wu, and X. Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *IJCAI*. AAAI Press, 3929–3935.

- [304] L. Zhang, Y. Wu, and X. Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.
- [305] L. Zhang, Y. Wu, and X. Wu. 2018. Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering* 31, 11 (2018), 2035–2050.
- [306] W. Zhang, R. Ramezani, and A. Naeim. 2021. Causal Inference in medicine and in health policy, a summary. *arXiv preprint arXiv:2105.04655* (2021).
- [307] Y. Zhang, M. Gong, T. Liu, G. Niu, X. Tian, et al. 2021. CausalAdv: Adversarial Robustness through the Lens of Causality. *arXiv* (2021).
- [308] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).
- [309] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang. 2018. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496* (2018).
- [310] Y. Zheng, C. Gao, X. Li, X. He, Y. Li, et al. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *WWW*.
- [311] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593.
- [312] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang. 2020. Learning to generate novel domains for domain generalization. In *European conference on computer vision*. Springer, 561–578.
- [313] Y. Zhu, J. Wong, A. Mandlekar, and R. Martin-Martin. 2020. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293* (2020).
- [314] O. J. Ziff, D. A. Lane, M. Samra, M. Griffith, P. Kirchhof, et al. 2015. Safety and efficacy of digoxin: systematic review and meta-analysis of observational and controlled trial data. *BMJ* 351 (2015).
- [315] Zimnat. 2020. Zimnat Insurance Recommendation Challenge. Retrieved November 25, 2022 from <https://zindi.africa/competitions/zimnat-insurance-recommendation-challenge>

A REVIEW OF THE ROLE OF CAUSALITY IN DEVELOPING TRUSTWORTHY AI SYSTEMS – DATASETS AND PACKAGES

As a result of our literature review on causality-based solutions for Trustworthy AI, a need for an extensive overview of relevant datasets and packages was observed. To make causal machine learning (ML) more accessible and to facilitate comparisons to non-causal methods, we created a curated list of datasets used for recent Causal ML publications. This appendix also includes an overview of useful causal and non-causal tools and packages to assess different trustworthy aspects of ML models (interpretability, fairness, robustness, privacy, and safety). We also provide a similar overview for the healthcare domain. Each aspect has its dedicated section that is structured as follows:

- (1) An overview of **publicly available real-world datasets** used in cited publications of this survey
- (2) Some **benchmarks and packages for Causal Machine Learning** that researchers could utilize
- (3) A number of **well-established tools**, that allow for a better comparison to non-causal machine learning

We want to clarify that this section does not (and cannot) aim for completeness. Instead, we want to provide researchers interested in working on a selection of aspects of Trustworthy AI with a concise overview of exciting avenues for experimenting with causal machine learning. The resources are hyperlinked and sorted based on when their associated causal papers first appear in the corresponding subsections (e.g., datasets used in pre-processing papers will appear first). We highly encourage readers to seek additional reading material, such as Chapter 9 of [117] or the two Github repositories for datasets³ and algorithms⁴ resulting from [91].

A INTERPRETABILITY

A.1 Datasets Used by Cited Publications

- **CANDLE** [215]: A dataset of realistic images of objects in a specific scene generated based on observed and unobserved confounders (object, size, color, rotation, light, and scene). As each of the 12546 images is annotated with the ground-truth information of the six generating factors, it is possible to emulate interventions on image features. → **Used by:** [215]
- **MIND** [286]: A news recommendation dataset built upon user click logs of Microsoft News. It contains 15 million impression logs describing the click behavior of more than 1 Million users across over 160k English news articles. Each news article entry contains its title, category, abstract, and body. Each log entry is made up of the users' click events, non-clicked events, and historical news click behaviors prior to this impression. → **Used by:** [251]
- **MovieLens** [95]: A group of datasets containing movie ratings between 0 and 5 (with 0.5 increments) collected from the MovieLens website. Movies are described through their title, genre, and relevance scores of tags (e.g., romantic or funny). GroupLens Research constantly releases new up-to-date MovieLens databases in different sizes. → **Used by:** [310]
- **Netflix Prize** [24]: A movie rating dataset consisting of about 100 Million ratings for 17,770 movies given by 480,189 users. Ratings consists of four entries: user, movie title, date of grade, and a grade ranging from 1 to 5. Users and movies are represented with integer IDs. → **Used by:** [310]

³<https://github.com/rguo12/awesome-causality-data>

⁴<https://github.com/rguo12/awesome-causality-algorithms>

- **WMT 14** [29]: WMT⁵ is a yearly workshop in which researchers develop machine translation models for several different tasks. WMT14 was created for the event in 2014 and included a translation, a quality estimation, a metrics, and a medical translation task. Each category comprises different subtasks (e.g., translating between two specific languages). → **Used by:** [9]
- **OpenSubtitles** [143]: A text corpus comprising over 2.6 billion sentences from movie dialogues. The data stem from pre-processing 3,735,070 files from the online database *OpenSubtitles.org*⁶. This corpus covers dialogues from ca. 2.8 million movies in 62 languages. → **Used by:** [9]
- **LAMA** [203]: A probe designed to examine the factual and commonsense knowledge in pretrained language models. It is built upon four different, prominent corpora of facts that cover a wide range of knowledge types. → **Used by:** [37]
- **Comma.ai Driving Dataset** [232]: A video dataset made up of 11 video clips of variable size capturing the windshield view of an Acura ILX 2016. The driving data contains 7.25 hours of footage, which was mostly recorded on highways. Each video is accompanied by measurements such as the car's speed, acceleration, or steering angle. → **Used by:** [125]
- **Udacity Driving Dataset** [267]: A driving video dataset developed for the Udacity *Self-Driving Car Nanodegree Program*⁷. The GitHub repository contains two annotated datasets in which computer vision systems have to label objects, such as cars or pedestrians, within driving footage. → **Used by:** [125]
- **T-REx** [67]: A dataset of large-scale alignments between Wikipedia abstracts and Wikidata triples. Such triples encode semantic information in the form of subject-predicate-object relationships. T-REx consists of 11 million triples with 3.09 million Wikipedia abstracts (6.2 million sentences). → **Used by:** [139]
- **MNIST** [134]: An extraordinarily well-known and widely used image dataset comprising 28 × 28 grayscale images of handwritten digits. It contains 60,000 training and 10,000 test samples. → **Used by:** [236]
- **ImageNet** [61]: Another well-known, more sophisticated image dataset containing more than 14 million images. The images depict more than 20,000 *synsets* (i.e., concepts "possibly described by multiple words or word phrases"⁸). → **Used by:** [236]
- **Adult (Census Income)** [62, 126]: A tabular dataset containing anonymized data from the 1994 Census bureau database.⁹ Classifiers try to predict whether a given person will earn over or under 50,000 USD worth of salary. Each person is described via 15 features (including their id), e.g., gender, education, and occupation. → **Used by:** [74, 155]
- **Human Activity Recognition** [11]: This dataset contains smartphone-recorded sensor data from 30 subjects performing *Activities of Daily Living*. The database differentiates between the activities walking (upstairs, downstairs, or on the same level), sitting, standing, and laying. → **Used by:** [111]
- **Yelp** [294]: A dataset of almost 7 million Yelp user reviews of around 150k businesses across 11 cities in the US and Canada. Review entries contain not only their associated text and an integer star rating between 1 and 5 but also additional information like the amount of *useful*, *funny*, and *cool* votes for the review. → **Used by:** [257]

⁵<https://machinetranslate.org/wmt>

⁶<https://www.opensubtitles.org/>

⁷<https://udacity.com/self-driving-car>

⁸<https://www.image-net.org/about.php>

⁹<http://www.census.gov/en.html>

- **Amazon (Product) Data** [179]: An extensive dataset of 233.1 million Amazon reviews between May 1996 and October 2018. The data include not only information about the review itself and product metadata (e.g., descriptions, price, product size, or package type) but also *also bought* and *also viewed* links. → **Used by:** [257]
- **Sangiovese Grapes** [154]: A conditional linear Bayesian network that captures the effects of different canopy management techniques on the quality of Sangiovese grapes. Based on a two-year study of Tuscan Sangiovese grapes, the authors created a network with 14 features (13 of which are continuous variables). The data used for experiments in [155] are linked in their repository (see the link behind the term “Sangiovese Grapes”). → **Used by:** [155]
- **WikiText-2** [163]: An NLP benchmark containing over 100 million tokens extracted from verified Good and Featured articles on Wikipedia. Contrary to previous token collections, however, WikiText-2 is more extensive and comprises more realistic tokens (e.g., lower-case tokens). → **Used by:** [112]
- **Jigsaw Toxicity Detection** [115]: A dataset of comments made across around 50 English-language news sites built to analyze unintended bias in toxicity classification within a Kaggle competition organized by Jigsaw and Google. Each comment in the training set comes with a human-annotated toxicity label (e.g., obscene or threat) and labels for mentioned identities (e.g., gender, ethnicity, sexuality, or religion) in the comment. → **Used by:** [112]
- **RTGender** [278]: A collection of comments made on online content across different platforms such as Facebook or Reddit. Each post and comment is annotated with the gender of the author in order to analyze gender bias in social media. → **Used by:** [112]
- **CrowS-Pairs** [175]: A benchmark designed to investigate the social bias of NLP models. Each entry consists of two sentences: one representing a stereotypical statement for a given bias type (e.g., religion or nationality) and an anti-stereotypical version of the statement, where the described group/identity was substituted. → **Used by:** [112]
- **Professions** [274]: A set of templates (originating from [151]) that were augmented with professions from [30]. Each sentence template follows the pattern “The [occupation] [verb] because”, and each profession has a crowdsourced rating that describes its definitionality and stereotypicality. → **Used by:** [274]
- **WinoBias** [308]: A collection of 3,160 WinoCoRef style sentences created to estimate gender bias within NLP models. Sentences come in pairs that only differ by the gender of one pronoun, with each sentence describing an interaction between two people with different occupations. → **Used by:** [274]
- **Winogender Schemas** [225]: A Winograd-style collection of templates that generate pairs of sentences that only differ by the gender of one pronoun. Researchers can generate 720 different sentences by defining the building blocks *occupation*, *participant*, and *pronoun* as a benchmark for gender bias detection. → **Used by:** [274]
- **English UD Treebank** [160]: The English UD Treebanks represents a subset of a data collection containing uniformly analyzed sentences across six different languages. The English treebank consists of 43,948 sentences and 1,046,829 tokens. → **Used by:** [66]
- **Gender-Neutral GloVe Word Embeddings** [309]: This variant of GloVe produces gender-neutral word embeddings by maintaining all gender-related information exclusively in specific dimensions of word vectors. The resulting word embeddings can be a starting point for more unbiased NLP. → **Used by:** [211, 212]
- **Biographies** [59]: A collection of 397,340 online biographies covering 28 occupations (e.g., professors, physicians, or rappers). Each biography is stored as a dictionary containing the title, the (binary) gender, the length of the first sentence, and the entire text of the biography. → **Used by:** [211, 212]

- **TwitterAAE corpus** [28]: A collection of 59.2 million tweets sent out by 2.8 million users from the US in 2013. Each tweet is annotated with a vector describing the “likely demographics of the author and the neighborhood they live in.” [28] These demographic approximations of users were built upon US census data. → **Used by:** [211]
- **CelebA** [149]: A face image dataset containing 202,599 images of size 178×218 from 10,177 unique celebrities. Each image is annotated with 40 binary facial attributes (e.g., *Is this person smiling?*) and five landmark positions describing the 2D position of the eyes, the nose, and the mouth (split into *left* and *right* side of the mouth). → **Used by:** [212]

A.2 Interesting Causal Tools

- **CausalFS** [296]: An open-source package for C++ that contains 28 local causal structure learning methods for feature selection. It is specifically designed to facilitate the development and benchmarking of new causal feature selection techniques.
- **CEBaB** [1]: A recently designed benchmark to estimate and compare the quality of concept-based explanation for NLP. CEBaB includes a set of restaurant reviews accompanied by human-generated counterfactuals, which enables researchers to investigate the model’s ability to assess causal concept effects.
- **CausaLM Datasets** [71]: As part of the analysis of *CausaLM*, the authors developed four NLP datasets for evaluating causal explanations. These datasets represent real-world applications of ML that come with ground-truth information.
- **Competing with Causal Toolboxes:** Several causal tools like *YLearn* [57], *DoWhy* [240], *CausalML* [46], or *EconML* [123] introduce an entire causal inference pipeline with their own interpreter module. Comparing newly developed interpretation techniques with such packages could be very insightful.

A.3 Prominent Non-Causal Tools

- **LIME** [218]: A very prominent Python package that allows researchers to explain individual predictions of image, text, and tabular data classifiers. Applicable to any black-box classifier that implements a function that outputs class probabilities given raw text or a NumPy array.
- **ROAR** [102]: A benchmark method that evaluates interpretability approaches based on how well they quantify feature importance. The technique was used to assess model explanations of image classifiers over multiple datasets.
- **SHAP** [152]: Another well-known interpretability package which is based on game theory. Although compatible with any ML model, SHAP comes with a C++-based algorithm for tree ensemble algorithms such as XGBoost.
- **InterpretML** [183]: An open-source package developed by Microsoft that includes multiple state-of-the-art methods for model interpretability. It also allows users to train an *Explainable Boosting Machine* (EBM) - a model that provides exact explanations and performs as well as random forests and gradient-boosted trees.

B FAIRNESS

B.1 Datasets Used by Cited Publications

- **Adult (Census Income)** [62, 126]: A tabular dataset containing anonymized data from the 1994 Census bureau database.¹⁰ Classifiers try to predict whether a given person will earn over or under 50,000 USD worth of salary. Each person is described via 15 features (including

¹⁰<http://www.census.gov/en.html>

their id), e.g., gender, education, and occupation. → **Used by:** [75, 174, 191, 230, 288, 290, 291, 304, 305]

- **COMPAS Recidivism Risk** [12]: A set of criminological datasets published by ProPublica to evaluate the bias of COMPAS - an algorithm used to assess the likelihood of criminal defendants reoffending. All COMPAS-related datasets include data from over 10,000 defendants, each being described via 52 features (e.g., age, sex, race) and with a label indicating whether they were rearrested within two years. → **Used by:** [50, 75, 168, 173, 174, 230]
- **FICO Credit Risk** [185]: In this dataset, ML models have to predict whether or not credit applicants will at least once be more than 90 days due with their payment within a two-year timespan. It includes anonymized information about HELOC applicants described through 23 features (e.g., months since the most recent delinquency or number of inquiries in last 6 months) [44] → **Used by:** [55]
- **German Credit Risk** [62]: A collection of data from 1,000 anonymized German bank account holders that applied for a credit. Based on the 20 features of the applicant and their application (e.g., credit history, purpose of credit, or employment status), models need to estimate the risk of giving the person a credit and categorize them as either good or bad credit recipients.. → **Used by:** [75]
- **Medical Expenditure (MEPS)** [40]: A collection of large-scale surveys of US citizens, their medical providers, and employers. It includes information like race, gender, and the ICD-10 code of the diagnosis of a patient. The given information can be used to predict the total number of patients' hospital visits. → **Used by:** [75]
- **MIMIC III** [116]: A dataset of anonymized clinical records of the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Records contain information like ICD-9 codes for diagnoses and medical procedures, vital signs, medication, or even imaging data. The dataset includes records from 38,597 distinct adult patients. → **Used by:** [252]
- **MovieLens** [95]: A group of datasets containing movie ratings between 0 and 5 (with 0.5 increments) collected from the MovieLens website. Movies are described through their title, genre, and relevance scores of tags (e.g., romantic or funny). GroupLens Research constantly releases new up-to-date MovieLens databases in different sizes. → **Used by:** [140]
- **Zimnat Insurance Recommendation** [315]: A data collection of almost 40,000 Zimnat (a Zimbabwean insurance provider) customers. The data contain personal information (e.g., marital status or occupation) and the insurance products that the customers own. In inference time, models must predict which product was artificially removed based on customer information. → **Used by:** [140]
- **Civil Rights Data Collection (CRDC)** [268]: This is an online collection of education-related data. Since 1968, the U.S. Department of Education's Office for Civil Rights (OCR) biennially collects data from U.S. public primary and secondary schools. The dataset includes information such as race distribution, the percentage of students who take college entrance exams, or whether specific courses (e.g., Calculus) are offered. → **Used by:** [129]
- **Berkeley** [25]: A simple gender bias dataset published back in 1975 containing information on all 12,763 applicants to the University of California, Berkeley graduate programs in Fall 1973. Each candidate entry consists of the candidate's major, gender, year of application (always 1973), and whether they were accepted. → **Used by:** [291]

B.2 Interesting Causal Tools

- **Collection of Annotated Datasets** [133]: As part of a survey that provides a thorough overview of commonly used datasets for evaluating the fairness of ML, Le Quy et al. generated

Bayesian Networks encompassing the relationships of attributes for each dataset. This information could be used as a reference point for potential causal annotations of fairness-related datasets.

- **WhyNot** [167]: A Python package that provides researchers with many simulation environments for analyzing causal inference and decision-making in a dynamic setting. It allows benchmarking of multiple decision-making systems on 13 different simulators. Crucially for this section, WhyNot also enables comparisons based on other evaluation criteria, such as the fairness of the decision-making.
- **gCastle** [300]: An end-to-end causal structure learning toolbox that is equipped with 19 techniques for Causal Discovery. It also assists users in data generation and evaluating learned structures. Having a firm understanding of the causal structure is crucial for fairness-related research.
- **Benchmark** [220]: A benchmark for causal structure learning allowing users to compare their causal discovery methods with over 40 variations of state-of-the-art algorithms. The plethora of available techniques in this single tool could facilitate research into fair ML through causality.
- **CausalML** [46]: The Python package enables users to analyze the Conditional Average Treatment Effect (CATE) or Individual Treatment Effect (ITE) observable in experimental data. The package includes tree-based algorithms, meta-learner algorithms, instrumental variable algorithms, and neural-network-based algorithms. Fair-ML researchers could use the provided methods to investigate the causal effect of sensitive attributes on the predicted outcome.

B.3 Prominent Non-Causal Tools

- **AI Fairness 360** [21]: An open-source library (compatible with both Python and R) that allows researchers to measure and mitigate possible bias within their models/algorithms. It includes six real-world datasets, five fairness metrics, and 15 bias mitigation algorithms.
- **Fairlearn** [26]: A Python package developed by Microsoft, which is part of the Responsible AI toolbox¹¹. It contains various fairness metrics, six unfairness-mitigating algorithms, and five datasets.
- **Aequitas** [228]: An open-source auditing tool designed to assess the bias of algorithmic decision-making systems. It provides utility for evaluating the bias of decision-making outcomes and enables users to assess the bias of actions taken directly.
- **ML-Fairness-Gym** [65]: A third-party extension of the OpenAI gym designed to analyze bias within RL agents. Although not built upon real-world data, the simulations developed for this benchmark can lead to insights applicable to the real world. It comes with four simulation environments.

C ROBUSTNESS

C.1 Datasets Used by Cited Publications

- **Rotated MNIST** [81]: The dataset consists of MNIST images with each domain containing images rotated by a particular angle $\{0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ\}$ → **Used by:** [106]
- **ColoredMNIST** [13]: The dataset consists of input images with digits 0-4 colored red and labelled 0 while digits 5-9 are colored green representing the two domains. → **Used by:** [13, 106, 150]

¹¹<https://github.com/microsoft/responsible-ai-toolbox>

- **PACS** [136]: An image classification dataset categorized into 10 classes that are scattered across four different domains, each having a distinct trait: photograph, art, cartoon and sketch. → **Used by**: [106]
- **Amazon (Product) Data** [179]: An extensive dataset of 233.1 million Amazon reviews between May 1996 and October 2018. The data include not only information about the review itself and product metadata (e.g., descriptions, price, product size, or package type) but also *also bought* and *also viewed* links. → **Used by**¹²: [121, 280]
- **SemEval-2017 Task 4** [223]: SemEval¹³ is a yearly NLP workshop where participants compete on different sentiment analysis tasks. Each workshop comes with its own set of tasks to solve. In SemEval-2017 Task 4, NLP models compete on sentiment analysis tasks on English and Arabic Twitter data. → **Used by**: [121]
- **Yelp** [294]: A dataset of almost 7 million Yelp user reviews of around 150k businesses across 11 cities in the US and Canada. Review entries contain not only their associated text and an integer star rating between 1 and 5 but also additional information like the amount of *useful*, *funny*, and *cool* votes for the review. → **Used by**: [121]
- **IMDb extension** [120]: A set of 2440 IMDb reviews, where a human-annotated counterfactual example accompanies each review. The human annotators were found through Amazon’s crowdsourcing platform *Mechanical Turk*¹⁴. The dataset is designed to assess the performance of sentiment analysis and natural language inference models. → **Used by**: [121, 258, 280]
- **SNLI extension** [120]: The original SNLI dataset [31] is a text dataset developed to evaluate natural language inference (NLI) models. Models must decide whether a given hypothesis is contradictory to, entailed by, or neutral to the given premise. Kaushik et al. [120] extended this dataset via humanly-manufactured counterfactual examples. → **Used by**: [121, 258]
- **Parkinson’s voice data** [144]: A set of extracted features from audio samples (i.e., sustained phonations) of patients with Parkinson’s disease and people from healthy control groups. This dataset combines data from three different and independent labs from the **US**, **Turkey**, and **Spain**. The classification task is to detect patients with Parkinson’s disease. → **Used by**: [144]
- **DomainNet** [199]: An unsupervised domain adaptation image dataset containing six domains (referring to the “style” of the image, e.g., sketch, quick drawing or real image) and about ca. 600k images distributed among 345 categories.
- **ImageNet** [61]: Another well-known, more sophisticated image dataset containing more than 14 million images. The images depict more than 20,000 *synsets* (i.e., concepts “possibly described by multiple words or word phrases”¹⁵). → **Used by**: [158, 169]
- **ImageNet-C** [99]: This dataset tests the model’s robustness by applying corruptions to validation images of ImageNet. Each of the 15 corruption types (e.g., gaussian noise, snow, motion blur, or contrast) comes with five levels of corruption intensity. → **Used by**: [158]
- **ImageNet-V2** [214]: A new test set for ImageNet designed to assess the model’s generalization ability. Despite closely following the original dataset creation process, models trained on the original ImageNet demonstrate worse performance on ImageNet-V2. ImageNet models with better generalization should perform stably on both variants. → **Used by**: [158]
- **ObjectNet** [17]: An image dataset designed to demonstrate the transfer learning ability of ImageNet models. Due to this, ObjectNet provides no training set. Instead, all 50,000

¹²[280] used a subset of Kindle Book reviews from an older version of this dataset

¹³<https://semeval.github.io/>

¹⁴<https://www.mturk.com/>

¹⁵<https://www.image-net.org/about.php>

images of ObjectNet combine into a single test set. Each image depicts an object with random backgrounds, viewpoints, and rotations of the object. → **Used by:** [158]

- **ImageNet ILSVRC-2012** [226]: The dataset used for the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). The 1.5 million images depict objects from 1,000 different synsets. → **Used by:** [169]
- **ImageNet-R** [98]: A variation of ImageNet designed to evaluate the susceptibility to spurious correlations of ImageNet models. It includes 30,000 artistic renditions (e.g., paintings, origami, or sculptures) of 200 ImageNet object classes. The images were primarily collected from Flickr¹⁶. → **Used by:** [169]
- **The Arcade Learning Environment (ALE)** [22]: A suite of Atari 2600 games that allows researchers to develop AI agents (mostly RL agents) for more than 100 games. ALE supports OpenAI gym, Python, and C++ and provides researchers with a plethora of features to evaluate different agents. → **Used by:** [85, 169]
- **CIFAR** [127]: The two CIFAR datasets, CIFAR-10 and CIFAR-100, are labeled images stemming from the now withdrawn Tiny Images dataset¹⁷. The more prominent set, CIFAR-10, contains 60000 32×32 color images separated into ten mutually exclusive classes, with 6000 images per class. CIFAR-100 is simply a 100-class version of CIFAR-10. → **Used by:** [107, 307]
- **Functional Map of the World (FMoW)** [52]: A collection of over 1 million satellite images depicting more than 200 countries. Each satellite contains at least one of 63 box annotations categorizing visible landmarks, such as *flooded road* or *airport*. → **Used by:** [107]
- **Chemical Environment** [122]: This synthetic environment was designed to evaluate causal reinforcement learning (RL) agents exhaustively. In this task, agents must change the colors of a given set of objects. However, altering one object influences the color of other objects. The causal dynamics are set by either a user-defined causal graph or a randomly generated DAG. → **Used by:** [281]
- **robosuite** [313]: A simulation framework built upon the MuJoCo physics engine allowing researchers to simulate contact dynamics for robot learning tasks. Given a set of cubes, RL agents must maneuver a robotic arm to solve different tasks (e.g., stacking the cubes or lifting one to a specified height). → **Used by:** [281]
- **COMPAS Recidivism Risk** [12]: A set of criminological datasets published by ProPublica to evaluate the bias of COMPAS - an algorithm used to assess the likelihood of criminal defendants reoffending. All COMPAS-related datasets include data from over 10,000 defendants, each being described via 52 features (e.g., age, sex, race) and with a label indicating whether they were rearrested within two years. → **Used by:** [63]
- **Adult (Census Income)** [62, 126]: A tabular dataset containing anonymized data from the 1994 Census bureau database.¹⁸ Classifiers try to predict whether a given person will earn over or under 50,000 USD worth of salary. Each person is described via 15 features (including their id), e.g., gender, education, and occupation. → **Used by:** [63]
- **South German Credit** [87]: Designed as a successor to the German Credit dataset, this dataset contains 1000 credit scoring entries from a south german bank between 1973 and 1975. Each row contains 20 columns (e.g., savings, job, and credit history) based on which models must assess the risk of granting credit. → **Used by:** [63]
- **Bail (DATA 1978)** [233]: A collection of criminal records from 9,327 individuals that were released from a North Carolina prison between 1977 and 1978. This dataset was created

¹⁶<https://www.flickr.com/>

¹⁷<http://groups.csail.mit.edu/vision/TinyImages/>

¹⁸<http://www.census.gov/en.html>

to investigate factors that influence the likelihood of recidivism. Each record contains 19 variables, including a binary ethnicity variable (black or not black) and a variable indicating previous use of hard drugs. → **Used by:** [63]

- **Colored FashionMNIST** [5]: This dataset was inspired by Arjovsky et al. [13] Colored MNIST dataset. Ahuja et al. use the same coloring approach to induce spurious correlations into FashionMNIST data (greyscaled Zalando articles). → **Used by:** [150]
- **VLCS** [262]: A collection of 10,729 images from four standard datasets designed to evaluate the OOD performance of image classifiers. Each image depicts a bird, car, chair, dog, or person. → **Used by:** [150]
- **VQA-CP** [3]: A dataset for Visual Question Answering (VQA) models that actively punishes the use of spurious correlations. This is achieved by rearranging the VQA v1 and VQA v2 data splits. The resulting training and test data differ in the "distribution of answers per question type". → **Used by:** [258]
- **COCO** [141]: An object detection dataset containing 328k images that depict 91 different types of objects. Each object within an image has its unique annotation, leading to more than 2.5 million labels across the entire dataset. → **Used by:** [258]
- **Law School Admission Data** [204]: A tabular dataset of admission data from 25 US law schools between 2005 and 2007. This dataset contains information from more than 100,000 applicants (e.g., gender, ethnic group, LSAT score), with each entry having a binary admission status variable. → **Used by:** [280]
- **MNIST** [134]: An extraordinarily well-known and widely used image dataset comprising 28×28 grayscale images of handwritten digits. It contains 60,000 training and 10,000 test samples. → **Used by:** [144, 298, 307]
- **Sequential MNIST Resolution Task** [128]: A sequential version of MNIST, where pixels of an handwritten digit are shown one at a time. → **Used by:** [85]
- **Bouncing Ball** [272]: A simulation environment where multiple balls of different sizes and weights independently move according to Newtonian physics. This environment is used to assess the model's physical reasoning capabilities under different conditions (e.g., different amounts of balls). → **Used by:** [85]
- **BabyAI** [48]: A RL framework that supports the development of agents that can understand language instructions. For this purpose, the authors developed agents that simulate human experts capable of communicating with task-solving agents using synthetic natural language. The platform provides 19 levels to alter the difficulty of the task. → **Used by:** [85]
- **ETH and UCY** [69, 135]: Both *ETH* [69] and *UCY* [135] are datasets containing real-world pedestrian trajectories. More novel papers combine both datasets to simulate multiple training and testing environments. Together, they contain trajectories of 1536 detected pedestrians collected from five locations. → **Used by:** [45, 146]
- **Stanford Drone dataset** [221]: A video dataset containing over 100 top-view scenes of the Stanford University campus that were shot with a quadcopter. The videos depict 20,000 manually annotated targets (e.g., pedestrians, bicyclists, or cars). → **Used by:** [45, 146]
- **Waterbirds** [227]: A binary image classification task where models must decide whether the depicted bird is a waterbird or a landbird. Good-performing models must rely on something other than the intrinsic spurious correlation between the background and the label (e.g., only 56 out of 4795 training images depict a waterbird with a land background). → **Used by:** [279]
- **CelebA** [149]: A face image dataset containing 202,599 images of size 178×218 from 10,177 unique celebrities. Each image is annotated with 40 binary facial attributes (e.g., *Is this person smiling?*) and five landmark positions describing the 2D position of the eyes, the nose, and the mouth (split into *left* and *right* side of the mouth). → **Used by:** [279]

- **MultiNLI** [282]: A text dataset developed to evaluate natural language inference (NLI) models. Models must decide whether a given hypothesis is contradictory to, entailed by, or neutral to the given premise. Contrary to other NLI datasets, MultiNLI includes text from 10 written and spoken English domains. → **Used by:** [279]
- **Abalone** [62]: In this task, ML models need to predict the number of rings an *abalone* (a shellfish) has based on the given features *sex*, *width*, *height*, and *shell diameter*. The dataset contains 4177 entries. → **Used by:** [131]
- **Bike Sharing in Washington D.C.** [70]: This dataset contains the hourly and daily count of rental bikes used in Washington D.C. between 2011 and 2012 (17,379 entries). Given weather and seasonal information, models need to predict the count of total rental bikes. → **Used by:** [131]
- **OpenPowerlifting** [188]: This powerlifting competition dataset includes more than 22,000 competitions and more than 412,000 competitors as of April 2019. The data stem from OpenPowerlifting¹⁹, with each entry containing information about the lifter, the equipment used, weight class, and their performance across different powerlifting disciplines. → **Used by:** [131]

C.2 Interesting Causal Tools

- **CANDLE** [215]: A dataset of realistic images of objects in a specific scene generated based on observed and unobserved confounders (object, size, color, rotation, light, and scene). As each of the 12546 images is annotated with the ground-truth information of the six generating factors, it is possible to emulate interventions on image features.
- **CausalWorld** [4]: A simulation framework and benchmark that provides RL agents different learning tasks in a robotic manipulation environment. The environment comes with a causal structure on which users and agents can intervene on variables such as object masses, colors or sizes.
- **gCastle** [300]: An end-to-end causal structure learning toolbox that is equipped with 19 techniques for Causal Discovery. It also assists users in data generation and evaluating learned structures. Having a firm understanding of the causal structure allows models to deduce the content and style variables of the domain.
- **Benchmark** [220]: A benchmark for causal structure learning allowing users to compare their causal discovery methods with over 40 variations of state-of-the-art algorithms. The plethora of available techniques in this single tool could facilitate research into robustness of ML systems through causality.

C.3 Prominent Non-Causal Tools

- **DomainBed and OOD-Bench** [89, 292]: **DomainBed** is a benchmark for OOD-learning that enables performance comparisons with more than 20 OOD-algorithms on 10 different, popular OOD-datasets. **OOD-Bench** is built upon DomainBed and introduces a measurement to quantify the Diversity shift and Correlation shift inherit to OOD-datasets. The resulting categorization allows researchers to pinpoint strengths and weaknesses of OOD-learning algorithms.
- **RobustBench** [56]: A standardized adversarial robustness benchmark capable of emulating a variety of adversarial attacks for image classification through *AutoAttack*. It also provides multiple continuously updated leaderboards of the most robust models, which allows for direct comparisons between causal and non-causal methods.

¹⁹<https://www.openpowerlifting.org/>

- **Foolbox** [210]: A popular Python library that allows researchers to test their adversarial defenses against state-of-the-art adversarial attacks. Foolbox is very compatible, natively supporting Pytorch, Tensorflow and JAX models.
- **VeriGauge** [138]: A Python toolbox that allows users to verify the robustness of their adversarial defense approach for deep neural networks. It not only covers a multitude of verification techniques but also comes with an up-to-date leaderboard.
- **Adversarial Robustness Toolbox (ART)** [180]: An extensive Python library for Adversarial Machine Learning. It not only equips researchers with various attacks and defenses across four different attack threats (evasion, extraction, poisoning, and inference) but also provides the means to assess the performance of such algorithms thoroughly. ART is compatible with many popular frameworks and supports various data types and learning tasks.

D PRIVACY

D.1 Datasets Used by Cited Publications

- **Rotated MNIST** [81]: The dataset consists of MNIST images with each domain containing images rotated by a particular angle $\{0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ\}$ → **Used by:** [60, 72]
- **PACS** [136]: An image classification dataset categorized into 10 classes that are scattered across four different domains, each having a distinct trait: photograph, art, cartoon and sketch. → **Used by:** [60]
- **Office-Home** [273]: Image classification dataset analogous to PACS, having four distinct image domains: Art, ClipArt, Product and Real-World. → **Used by:** [60]
- **ColoredMNIST** [13]: The dataset consists of input images with digits 0-4 colored red and labelled 0 while digits 5-9 are colored green representing the two domains. → **Used by:** [72, 92]
- **Colored FashionMNIST** [5]: This dataset was inspired by Arjovsky et al. [13] Colored MNIST dataset. Ahuja et al. use the same coloring approach to induce spurious correlations into FashionMNIST data (greyscaled Zalando articles). → **Used by:** [92]
- **CIFAR** [127]: The two CIFAR datasets, CIFAR-10 and CIFAR-100, are labeled images stemming from the now withdrawn Tiny Images dataset²⁰. The more prominent set, CIFAR-10, contains 60000 32×32 color images separated into ten mutually exclusive classes, with 6000 images per class. CIFAR-100 is simply a 100-class version of CIFAR-10. → **Used by:** [92, 114]
- **Digits-DG** [312]: An image dataset specifically designed to evaluate the performance of models on OOD data. It includes images from four different handwritten digits databases. Each dataset represents a unique domain as images from different datasets significantly differ in terms of, e.g., handwriting style or background color. → **Used by:** [114]
- **Camelyon17** [16]: A publicly available medical dataset containing 1000 histology images from five Dutch hospitals. Given an image, classification models need to detect breast cancer metastases. → **Used by:** [114]

D.2 Interesting Causal Tools For Federated Learning

The publications reviewed in Section 6 are largely causal approaches to Federated Learning (FL). As such, we mainly provide an overview of causal and non-causal tools for FL.

- **Federated Causal Discovery** [2, 77]: Until this point, we suggested general causal discovery tools like *gCastle* [300] or *benchpress* [220]. However, the provided methods translate poorly into the federated setting due to the decentralized data. As such, we would like to refer readers to recently developed **Federated Causal Discovery** techniques (e.g., [2, 77]). These

²⁰<http://groups.csail.mit.edu/vision/TinyImages/>

methods are specifically designed to conduct causal discovery on decentralized data in a privacy-preserving manner.

- **CANDLE** [215]: A dataset of realistic images of objects in a specific scene generated based on observed and unobserved confounders (object, size, color, rotation, light, and scene). As each of the 12546 images is annotated with the ground-truth information of the six generating factors, it is possible to emulate interventions on image features. Users/Devices could be simulated by altering the scenery.
- **Federated Causal Effect Estimation** [276]: Similar to causal discovery, standard causal effect estimation methods were not designed for decentralized data. Only very recently, **Vo et al.** developed a causal effect estimation framework compatible with federated learning. Despite this line of work’s infancy, we believe that this publication is important for more privacy-preserving causal learning.

D.3 Prominent Non-Causal Federated Learning Tools

- **LEAF** [35]: A benchmark containing datasets explicitly designed to analyze FL algorithms. The six datasets include existing re-designed databases such as *CelebA* [149] to emulate different devices/users and newly created datasets. LEAF also provides evaluation methods and baseline reference implementations for each dataset.
- **FedEval** [41]: A publicly available evaluation platform for FL. It allows researchers to compare their FL methods with existing state-of-the-art algorithms on seven datasets based on five FL-relevant metrics (Accuracy, Communication, Time efficiency, Privacy, and Robustness). The benchmark utilizes Docker container technology to simulate the server and clients and socket IO for simulating communication between the two.
- **OARF** [103]: An extensive benchmark suite designed to assess state-of-the-art FL algorithms for both horizontal and vertical FL. It includes 22 datasets that cover different domains for both FL variants. Additionally, OARF provides several metrics to evaluate FL algorithms, and its modular design enables researchers to test their own methods.
- **FedGraphNN** [97]: An FL benchmark for Graph Neural Networks (GNN). In order to provide a unified platform for the development of graph-based FL solutions, FedGraphNN supplies users with 36 graph datasets across seven different domains. Researchers can also employ and compare their own *PyTorch (Geometric)* models with different GNNs.
- **ML-Doctor** [147]: A codebase initially used to compare and evaluate different inference attacks (membership inference, model stealing, model inversion, and attribute inference). Its modular structure enables researchers to assess the effectiveness of their privacy-preserving algorithms against SOTA privacy attacks.

E SAFETY

E.1 Datasets Used by Cited Publications

- **ScienceDirect** [68]: A bibliographic database that hosts over 18 million publications from more than 4,000 journals and more than 30,000 e-books from the publisher Elsevier. Launched back in 1997, ScienceDirect includes papers from engineering and medical research areas and social sciences and humanities. → **Used by:** [277]
- **World Bank** [259]: A publicly available collection of datasets that facilitate the analysis of global development. Researchers can use this data to compare countries under different developmental aspects, including agricultural progress, poverty, population dynamics, and economic growth. → **Used by:** [287]

- **World Economic Forum (WEF)** [284]: The WEF is an international non-governmental based in Switzerland that publishes economic reports such as the Global Competitiveness Report. The reports are available online, with some of the data being easily accessible through websites like [Knoema](#). → **Used by:** [96]
- **OECD.Stat** [189]: This webpage includes data and metadata for OECD countries and selected non-member economies. The online platform allows researchers to traverse the collected data through given data themes or via search-engine queries. → **Used by:** [96]
- **Global Brand Database** [285]: An online database hosted by the World Intellectual Property Organization (WIPO) that contains information about Trademark applications (e.g., owner of the trademark, its status, or the designation country). It currently contains almost 53 million records from 73 data sources. → **Used by:** [96]
- **PubMed** [178]: A widely-known, free-to-access search engine for biomedical and life science literature developed and maintained by the National Center for Biotechnology Information (NCBI). Researchers can find more than 34 million citations and abstracts of articles. PubMed does not host the articles themselves but frequently provides a link to the full-text articles. → **Used by:** [82]
- **ProQuest Central** [205]: A database containing dissertations and theses in a multitude of disciplines. It currently contains more than 5 million graduate works. → **Used by:** [82]
- **Cochrane Central Register of Controlled Trials (CENTRAL)** [53]: A database of reports for randomized and quasi-randomized controlled trials collected from different online databases. Although it does not contain full-text articles, the CENTRAL includes bibliographic details and often an abstract of the report. → **Used by:** [82]
- **PsycINFO** [10]: A database hosted and developed by American Psychological Association containing abstracts for more than five million articles in the field of psychology. → **Used by:** [82]
- **Lending Club** [80]: A dataset that contains information about all accepted and rejected peer-to-peer loan applications of LendingClub. Currently, the data are only available through the referenced Kaggle entry, as the company no longer provides peer-to-peer loan services²¹. → **Used by:** [264]
- **Taiwanese Credit Data** [270, 293]: A real-world dataset containing payment data collected in October 2005 from a Taiwanese bank. The commonly used pre-processed version²² [270] contains data from 30,000 individuals described through 16 features (e.g., marital status, age, or payment history). → **Used by:** [264]

E.2 Interesting Causal Tools

- **CausalImpact** [32]: This R package allows users to conduct causal impact assessment for planned interventions on serial data given a response time series and an assortment of control time series. For this purpose, CausalImpact enables the construction of a Bayesian structural time-series model that can be used to predict the resulting counterfactual of an intervention.
- **Causal Inference 360** [245]: A Python package developed by IBM to infer causal effects from given data. Causal Inference 360 includes multiple estimation methods, a medical dataset, and multiple simulation sets. The provided methods can be used for any complex ML model through a scikit-learn-inspired API.

²¹<https://www.lendingclub.com/investing/peer-to-peer>

²²Available at <https://github.com/ustunb/actionable-recourse/tree/master/examples/paper/data> under the name "credit_processed.csv"

- **gCastle** [300]: An end-to-end causal structure learning toolbox that is equipped with 19 techniques for Causal Discovery. It also assists users in data generation and evaluating learned structures. Having a firm understanding of the causal structure is crucial for safety-related research.
- **Benchpress** [220]: A benchmark for causal structure learning allowing users to compare their causal discovery methods with over 40 variations of state-of-the-art algorithms. The plethora of available techniques in this single tool could facilitate research into safety and accountability of ML systems through causality.
- **CauseEffectPairs** [170]: A collection of more than 100 databases, each annotated with a two-variable cause-effect relationship (e.g., access to drinking water affects infant mortality). Given a database, models need to distinguish between the cause and effect variables.

E.3 Prominent Non-Causal Tools

- **Government of Canada’s AIA tool** [84]: The Algorithmic Impact Assessment (AIA) tool is a questionnaire developed in the wake of Canada’s Directive on Automated Decision Making²³. Employees of the Canadian Government wishing to employ automatic decision-making systems in their projects first need to assess the impact of such systems via this tool. Based on answers given to ca. 80 questions revolving around different aspects of the projects, AIA will output two scores: one indicating the risks that automation would bring and one that quantifies the quality of the risk management.
- **Aequitas** [228]: An open-source auditing tool designed to assess the bias of algorithmic decision-making systems. It provides utility for evaluating the bias of decision-making outcomes and enables users to assess the bias of actions taken directly.
- **Error Analysis (Responsible AI)** [165]: As part of the Responsible AI toolbox, Error Analysis is a model assessment tool capable of identifying subsets of data in which the model performs poorly (e.g., black citizens being more frequently misclassified as potential re-offenders). It also enables users to diagnose the root cause of such poor performance.
- **ML-Doctor** [147]: A codebase initially used to compare and evaluate different inference attacks (membership inference, model stealing, model inversion, and attribute inference). Due to its modular structure, it can also be used as a Risk Assessment tool for analyzing the susceptibility against SOTA privacy attacks.

F HEALTHCARE

F.1 Datasets Used by Cited Publications

- **Alzheimer’s Disease Neuroimaging Initiative (ADNI)** [202]: A medical dataset containing multi-modal information of over 5,000 volunteering subjects. ADNI includes clinical and genomic data, biospecimens, MRI, and PET images. Researchers need to apply for data access. → **Used by:** [231, 242]
- **SARS-CoV-2 infected cells (Series GSE147507)** [27]: A genomic dataset that contains RNA-seq data from SARS-CoV-2-infected cells of both humans and ferrets. The data are publicly available on the NCBI Gene Expression Omnibus (GEO) server under the accession number GSE147507. → **Used by:** [23]
- **The Genotype-Tissue Expression (GTEx) project** [39]: An online medical platform that provides researchers with tissue data. Data samples stem from 54 non-diseased tissue sites across nearly 1000 individuals whose genomes were processed via sequencing methods such as WGS, WES, and RNA-Seq. → **Used by:** [23]

²³<http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592>

- **L1000 Connectivity Map (Series GSE92742)** [256]: A connectivity map (CMap) connects genes, drugs, and disease states based on their gene-expression signatures. The CMap provided by [Subramanian et al.](#) includes over 1.3 million L1000 profiles of 25,200 unique perturbagens²⁴. The data are publicly available on the NCBI Gene Expression Omnibus (GEO) server under the accession number GSE92742. → **Used by:** [23]
- **iRefIndex** [213]: This protein-protein interaction (PPI) network is a graph-based database of molecular interactions between proteins from over ten organisms. The current version of iRefIndex (version 19) contains over 1.6 million PPIs. → **Used by:** [23]
- **DrugCentral** [15]: An online platform that provides up-to-date drug information. Users can traverse the database online through the corresponding website or via an API. The platform currently contains information on almost 5,000 active ingredients. → **Used by:** [23]
- **Colorectal Cancer Single-cell Data (GSE81861)** [137]: The authors provide two datasets. The first dataset contains 1,591 single cells RNA-seq data from 11 colorectal cancer patients. The second dataset contains 630 single cells from seven cell lines and can be used to benchmark cell-type identification algorithms. The data are publicly available on the NCBI Gene Expression Omnibus (GEO) server under the accession number GSE81861. → **Used by:** [23]
- **Pulmonary Fibrosis Single-cell Data** [217]: This genomic dataset contains approximately 76,000 single-cell RNA-seq data from healthy lungs and lungs from patients with pulmonary fibrosis. The data are available online and comes with a cluster visualization based on marker gene expressions. → **Used by:** [23]
- **SARS-CoV-2 Host-Pathogen Interaction Map** [83]: A PPI network that maps 27 SARS-CoV-2 proteins to human proteins through 332 high-confidence protein-protein interactions. The online data contain data from the initial study and the CORUM database [265]. → **Used by:** [23]
- **Lung Image Database Consortium image collection (LIDC-IDRI)** [14]: An image dataset comprising annotated thoracic CT Scans of more than 1,000 cases. The data stem from seven academic centers and eight medical imaging companies. Four trained thoracic radiologists provided the image annotations. → **Used by:** [271]
- **MEDLINE** [269]: An online bibliographic database of more than 29 million article references from the field of life science (primarily in biomedicine). MEDLINE is a primary component of PubMed and is hosted and managed by the NLM National Center for Biotechnology Information (NCBI). → **Used by:** [314]
- **Cochrane Central Register of Controlled Trials (CENTRAL)** [53]: A database of reports for randomized and quasi-randomized controlled trials collected from different online databases. Although it does not contain full-text articles, the CENTRAL includes bibliographic details and often an abstract of the report. → **Used by:** [314]

F.2 Interesting Causal Tools

- **Causal Inference 360** [245]: A Python package developed by IBM to infer causal effects from given data. Causal Inference 360 includes multiple estimation methods, a medical dataset, and multiple simulation sets. The provided methods can be used for any complex ML model through a scikit-learn-inspired API.
- **gCastle** [300]: An end-to-end causal structure learning toolbox that is equipped with 19 techniques for Causal Discovery. It also assists users in data generation and evaluating learned structures. Having a firm understanding of the causal structure is crucial for healthcare-related research.

²⁴See https://clue.io/connectopedia/perturbagen_types_and_controls for the definition of this term

- **Benchpress** [220]: A benchmark for causal structure learning allowing users to compare their causal discovery methods with over 40 variations of state-of-the-art algorithms. The plethora of available techniques in this single tool could encourage more causality-based solutions for the healthcare domain.
- **CausalML** [46]: The Python package enables users to analyze the Conditional Average Treatment Effect (CATE) or Individual Treatment Effect (ITE) observable in experimental data. The package includes tree-based algorithms, meta-learner algorithms, instrumental variable algorithms, and neural-network-based algorithms.
- **WhyNot** [167]: A Python package that provides researchers with many simulation environments for analyzing causal inference and decision-making in a dynamic setting. It allows benchmarking of multiple decision-making systems on 13 different simulators. This set of simulators includes environments that simulate HIV treatment effects and system dynamics models of both the Zika epidemic and the US opioid epidemic.

F.3 Prominent Non-Causal Tools

- **Medical Open Network for AI (MONAI)** [38]: A PyTorch-based framework that offers researchers pre-processing methods for medical imaging data, domain-specific implementations of machine learning architectures, and ready-to-use workflows for healthcare imaging. The actively maintained framework also provides APIs for integration into existing workflows.
- **DeepChem** [209]: A Life Science toolbox that provides researchers with deep learning solutions for different fields of Life Science, such as Quantum Chemistry, Biology, or Drug Discovery (with the latter being particularly interesting for comparisons to causality-based solutions). Deepchem supports TensorFlow, PyTorch, and JAX and has an extensive collection of running examples.
- **Curated Lists on Github** [19, 181]: Niemelä et al. [181] host an [up-to-date GitHub repository](#) of relevant open-source healthcare tools and resources. Beam et al. [19] provide an [extensive overview of valuable medical datasets](#) that could be used to assess Causal ML healthcare solutions. Although this list has not been updated since 2020, we still believe it to be a helpful initial overview of relevant datasets.