Position: Trustworthy AI Suffers from Invariance Conflicts and Causality is the Solution

Ruta Binkyte [*1] **Ivaxi Sheth** [*1] **Zhijing Jin** [2 3 4] **Mohammad Havaei** [5] **Bernhard Schölkopf** [2] **Mario Fritz** [1]

Abstract

As artificial intelligence (AI), including machine learning (ML) models and foundation models (FMs), are increasingly deployed in high-stakes domains, ensuring their trustworthiness has become a central challenge. However, the core trustworthy AI objectives, such as fairness, robustness, privacy, and explainability, are hard to achieve simultaneously, especially while preserving utility. **This position paper argues that causality is necessary to understand and balance trade-offs in performance and multiple objectives of trustworthy AI.** We ground our arguments in re-interpreting trustworthy AI trade-offs as incompatible invariance requirements under different changes to the data-generating process. We then illustrate that causality provides a unifying framework for understanding how trade-offs in trustworthy AI arise, and how they can be softened or resolved through selective invariance. This perspective applies to both classical ML models and large-scale FMs. Our paper discusses how causal assumptions may be applied explicitly or implicitly in modern large-scale systems. Finally, we outline open challenges and opportunities for using causality to build both trustworthy and high-performing AI.

1. Introduction

Machine learning (ML) models have driven remarkable advances in natural language processing, computer vision, and decision-making, enabling large-scale deployment in domains such as healthcare, finance, education, and social media. More recently, foundation models (FMs), including large language models (LLMs) and vision language models (VLMs), have demonstrated unprecedented generality across tasks (Achiam et al., 2023; Rohan et al., 2023; Radford et al., 2023). Given their influence, ensuring ethical and trustworthy ML and FM systems has become a global priority. Many international regulations and frameworks (European Commission, 2021) seek to establish guidelines for AI that would avoid harmful social impact. In this work, we focus on four trustworthy dimensions - fairness, explainability, robustness, and privacy protection - as well as their relationship with model performance.

Trade-offs in trustworthy AI. The objectives in trustworthy AI[1] are rarely independent. Improving one aspect often comes at the expense of another, or model performance. For example, **privacy** noise added to protect data reduces model accuracy (Xu et al., 2017; Carvalho et al., 2023). Similarly, achieving **fairness** frequently requires sacrificing predictive performance or resolving conflicts between competing fairness notions, such as demographic parity and equalized odds (Friedler et al., 2021; Kim et al., 2020). Performance is also frequently traded off against **explainability**, as complex deep models excel in accuracy but are not comprehensible by humans (Crook et al., 2023). Even fairness and privacy can conflict as the loss in performance disproportionately hurts sensitive groups (Pujol et al., 2020). Similarly, overfitting on spurious signals in the data gives accuracy at the cost of **robustness** (Tsipras et al., 2018). However, much of the machine learning research and development has historically prioritized improving predictive accuracy and performance or treated these trade-offs as empirical side effects.

Causality and trustworthy AI. Unlike correlation-based approaches, causal models represent the mechanisms that produce data, enabling selective invariance: they allow models to be invariant to spurious pathways while remaining sensitive to stable, causally meaningful signals. Causal methods have already shown promise in auditing and mitigating unfairness (Kim et al., 2021;

[1]CISPA Helmholtz Center for Information Security [2]Max Planck Institute for Intelligent, Tübingen [3]ETH Zürich [4]University of Toronto [5]Google Research. Correspondence to: Ruta Binkyte <ruta.binkyte@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]Throughout the paper, we use *AI* as an umbrella term for learning-based models, including both classical machine learning models and foundation models.

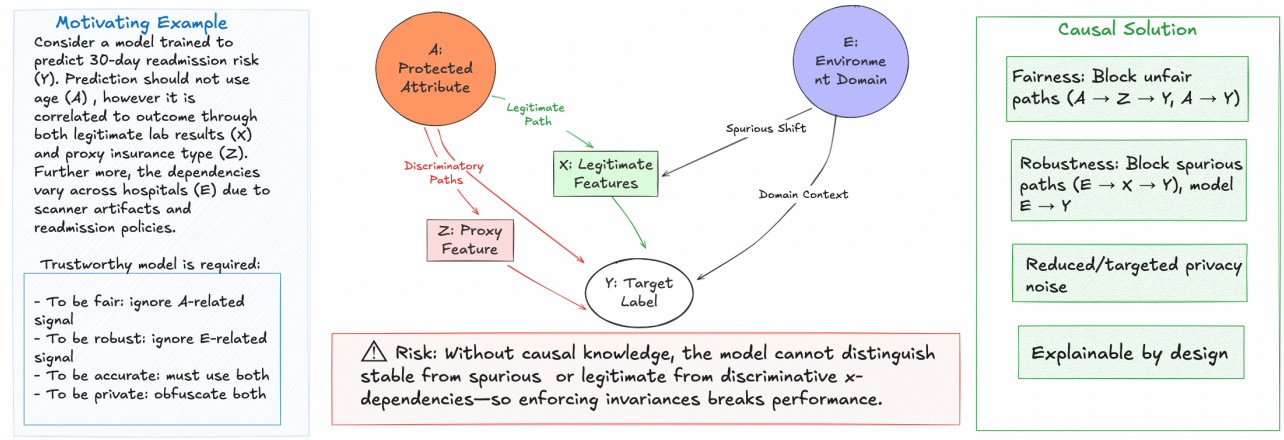


Figure 1. Causal resolution to trade-offs in trustworthy AI. See Appendix A for the details.

Kilbertus et al., 2017; Loftus et al., 2018) and in improving robustness under distribution shift (Schölkopf et al., 2022). In addition, recent theoretical results show that under certain conditions robustness implies causality (Richens & Everitt, 2024). While the connection between causality and privacy is still less explored, early work indicates that causal structure can improve privacy–utility trade-offs (Tschantz et al., 2020; Tople et al., 2020). Finally, explainability is inherently aligned with causality, as explanations are naturally expressed in terms of interventions and counterfactuals. However, despite recognition of causal solutions to individual challenges of trustworthy AI (Liu et al., 2023; Rawal et al., 2025; Ganguly et al., 2023), its role in reconciling the trade-offs between multiple goals and model performance remains underexplored.

Position. In this paper, *we argue that the trade-offs in trustworthy AI are not incidental, but arise from incompatible invariance requirements imposed by individual trustworthy AI objectives*. Robustness demands stability under distributional shifts; fairness demands stability when protected attributes or group membership change; privacy demands stability when individual data points are added or removed; and explainability demands stable attribution of changes under feasible actions. *At their core, trustworthy AI objectives can be reformulated as demands that a model's behavior remains stable under changes.* When different trust objectives demand stability under different and potentially conflicting sets of changes, trade-offs become unavoidable. For instance, accuracy under the observational distribution may rely on correlations that are unstable under interventions on the sensitive attributes or environmental variables, while fairness and robustness explicitly demand invariance to such interventions. Based on the above, **we position causality as a unifying framework for understanding how trade-offs in trustworthy AI arise, and how they can be softened or resolved.**

When causality is necessary? Not all invariance requirements immediately necessitate causal reasoning. If a single trust objective is considered in isolation and the admissible changes are limited, purely statistical methods may suffice to enforce invariance. For example, a classifier can be trained to satisfy independence between the sensitive attribute and the outcome on a fixed dataset, or to be robust to a specific, observed form of covariate shift, without any explicit causal model. However, this breaks down as soon as additional objectives or accuracy constraints are imposed. When multiple invariances must hold simultaneously, it becomes necessary to distinguish which dependencies are spurious and which reflect stable mechanisms. This distinction cannot, in general, be made from observational data alone. Importantly, this necessity is agnostic to model class. It applies equally to classical machine learning models and to large-scale foundation models. The distinction lies not in whether causal reasoning is needed, but in how causal assumptions are encoded and operationalized at scale.

How causality resolves the trade-offs in trustworthy AI? Causal reasoning is therefore required, not merely to enforce a single constraint, but to reason about which invariances can coexist, which must conflict, and how trade-offs can be selectively navigated. More precisely, causality allows:
(i) *Selective invariance through structural constraints.* Causal graph clarifies pathways that are subject to invariance requirements, allowing constraints to be enforced selectively rather than uniformly. This enables the suppression of normatively unacceptable effects (e.g., direct or proxy discrimination) while preserving causally justified signal, which is impossible under purely associational cri-

teria.
(ii) *Shift focus from observational accuracy to interventional validity.* Causal reasoning distinguishes correlations that merely predict outcomes from causal mechanisms and reduces overfitting to spurious correlations. This resolves the tension between predictive accuracy and trustworthy objectives by shifting the optimization target from performance under the observational distribution to correctness under interventions.

We provide a schematic illustration of causal approach to trustworthy AI in Figure 1.

Paper map and contributions. We reinterpret trustworthy AI literature to illustrate how trustworthy AI objectives translate into competing invariance requirements under interventions. We then show how causal structure enables principled resolution of these conflicts and discuss how causality can be encoded in both classical ML systems and foundation models. We discuss the conceptual and practical limitations of resolving trustworthy AI trade-offs using a causal approach. Finally, we draw future directions and call for actions necessary to overcome these limitations.

Our contributions are the following: 1.) We reinterpret major trade-offs in trustworthy AI through a conflicting invariance requirements lens; 2.) We formulate when causal reasoning is necessary and how it resolves or softens trade-offs in trustworthy AI.

2. Background and Preliminaries

We introduce the core dimensions of trustworthy AI, emphasizing formal characterizations that are relevant for understanding trade-offs and motivating causal reasoning.

2.1. Interventions and Invariance

Let $\mathcal{D}$ denote the observational data distribution over (X, Y), and let $\{\mathcal{D}^I\}_{I\in\mathcal{I}}$ denotes a family of distributions obtained by applying a set of admissible changes I to the data-generating process. These changes may correspond to modifications of inputs, environments, data collection procedures, population characteristics, or other aspects of the data-generating process. A model is said to satisfy an invariance requirement with respect to a set of admissible changes $\mathcal{I}$ if its behavior or performance remains stable across the corresponding distributions $\{\mathcal{D}^I\}_{I\in\mathcal{I}}$.

2.2. Trustworthy ML Objectives

We now instantiate the invariance principle defined above for common trust dimensions. Consider a supervised learning setting with observed features $X \in \mathcal{X}$, a target variable $Y \in \mathcal{Y}$, and a learned predictor $f : \mathcal{X} \to \hat{\mathcal{Y}}$. Each trust objective corresponds to requiring f to satisfy an invariance requirement with respect to a specific class of admissible changes.

Fairness (invariance to protected attributes). Let $A \in \mathcal{A}$ denote a sensitive or protected attribute. Fairness requires that the behavior of f remain *invariant* under admissible changes to A, either in distribution (e.g., Demographic parity) or conditional on Y (e.g., Equalized odds).

Privacy (invariance to data inclusion). Let D denote a training dataset and let D' be a neighboring dataset that differs from D in a single individual. Privacy requires that the behavior of a randomized mechanism M remain approximately *invariant* under the admissible change $D \to D'$. In other words, the output distribution of $M(D)$ should be stable with respect to the inclusion, removal, or modification of any single data point.

Robustness (invariance to distribution shifts). Let $\{\mathcal{D}_e\}_{e\in\mathcal{E}}$ denote a family of distributions indexed by environments e. Robustness requires that the behavior or performance of a model remain *invariant* across these environments. Formally, this is captured by bounded loss across the family $\{\mathcal{D}_e\}_{e\in\mathcal{E}}$, where each e may correspond to a shift in population, measurement process, or data-generating mechanism.

Explainability (selective invariance under changes). Let $\mathcal{I}_{\text{rel}}$ and $\mathcal{I}_{\text{irr}}$ denote sets of admissible relevant and irrelevant changes to the input or context, respectively. Explainability concerns the ability to understand and justify a models predictions through how they respond to these changes. From the invariance perspective, $f(X)$ should be *invariant* under changes in $\mathcal{I}_{\text{irr}}$, and responsive under changes in $\mathcal{I}_{\text{rel}}$ in a predictable and meaningful way. This definition equally applies to modern interpretability methods.

2.3. Accuracy

Accuracy measures predictive performance under the observational distribution $\mathcal{D}$, typically via the expected loss $\mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(f(X), Y)]$. In practice, accuracy often conflicts with invariance-based trust requirements, giving rise to the trade-offs studied in this work.

2.4. Interventional Accuracy

Interventional accuracy measures predictive performance under the interventional distribution $\mathcal{D}^{\mathcal{I}}$, and can be expressed as the expected loss $\mathbb{E}_{(X,Y)\sim\mathcal{D}^{\mathcal{I}}}[\ell(f(X), Y)]$. In contrast to observational accuracy, which rewards correla-

tions that hold in the training environment, interventional accuracy prioritizes predictors whose performance is stable across the family of interventions.

2.5. Causality

Purely associational learning methods operate on the observational distribution $\mathcal{D}$ and are sufficient for answering queries of the form $P(Y \mid X)$. However, many trustworthy ML objectives require reasoning about how predictions would change under hypothetical or interventional modifications to the data-generating process. To formalize such changes, we adopt Pearls structural causal framework (Pearl, 2009).

Central to Pearl's framework, an SCM consists of a set of endogenous variables V, exogenous variables U, and structural equations F that determine how each variable in V is generated from its parents and noise. This structure is summarized by a causal graph $\mathcal{G} = (V, \mathcal{E})$, which is a directed acyclic graph (DAG) where $\mathcal{E}$ is a set of edges connecting the nodes V.

Given a causal graph $\mathcal{G}$, a *causal path* from a variable X to a variable Y is any directed sequence of edges in $\mathcal{G}$ that begins at X and ends at Y. Causal paths characterize how the effect of one variable on another is transmitted through intermediate variables, and provide a decomposition of the total causal effect into distinct mechanistic components.

3. Trade-offs as Invariance Conflicts

In this section, we build on concrete examples and reinterpret well-known results from the literature to show that many trade-offs in trustworthy AI arise from fundamentally incompatible invariance requirements imposed by different objectives. We then highlight how causal reasoning provides a principled way to diagnose and soften or resolve these conflicts.

3.1. Fairness–Accuracy Trade-off

Statistical origin of the trade-off. Predictive accuracy under the observational distribution $\mathcal{D}$ often relies on correlations between features and sensitive attributes A. When these correlations are predictive, suppressing them to satisfy fairness constraints can reduce accuracy. This phenomenon is well documented across fairness-aware ML learning methods (Zliobaite, 2015; Zhao & Gordon, 2022). Similarly, in generative models, careless fairness interventions can backfire; for instance, efforts to enhance diversity have led to historically inaccurate outputs, as seen in critiques of Google's Gemini (Vincent, 2024).

Invariance conflict. Many fairness notions can be expressed as invariance requirements under interventions on A. For instance, Demographic parity can be interpreted as requiring invariance of the decision distribution $f(X)$ for all values of A. In contrast, maximizing accuracy under $\mathcal{D}$ encourages sensitivity to any predictive signal, including those downstream of A.

This creates a conflict between fairness and accuracy goals. Without distinguishing normatively unacceptable pathways from causally justified ones, these requirements are incompatible.

Causal resolution. Causal models distinguish between admissible and inadmissible causal pathways from A to Y (Chiappa, 2019). By explicitly modeling the data-generating process, causal reasoning enables selective invariance: suppressing only those paths deemed unfair (e.g., direct or proxy discrimination), while preserving legitimate causal effects (e.g., through explaining variables, such as gender dependent disease risks).

This logic applies across model classes. In classical ML, causal constraints operate on input features and representations. In foundation models, the same causal pathways are encoded implicitly in embeddings, attention patterns, or generation dynamics. Causal interventions, such as counterfactual data augmentation, causal disentanglement, or SCM-based regularization, can mitigate biased generative behavior while preserving task-relevant predictive structure (Zhou et al., 2023; Madhavan et al., 2023). Entity substitutions can selectively block unfair or spurious causal paths while preserving task-relevant information, illustrating how causal intervention softens the fairness-accuracy trade-off (Wang et al., 2023).

3.2. Privacy–Utility (and Attribution)

Statistical origin of the trade-off. Privacy-preserving mechanisms such as differential privacy reduce the influence of individual data points or sensitive attributes on model outputs by introducing privacy noise. While this limits information leakage, it also degrades utility by attenuating informative signals. In large-scale generative models, privacy risks additionally arise through memorization and indirect reproduction of personally identifiable information, further intensifying the tension between privacy and model usefulness (Xu et al., 2017).

Invariance conflict. Privacy requires invariance to interventions on private or individual-level variables, such that small changes to sensitive information do not substantially affect outputs. Utility, in contrast, requires sensitivity to variables and pathways that encode task-relevant information. In foundation models, this conflict extends

to attribution: models must be insensitive to whether a specific individuals data is present in the training set, while remaining sensitive to generalizable patterns.

Causal resolution. Causal models mitigate the privacy–utility trade-off by reducing reliance on spurious correlations that lead to overfitting and memorization, which are primary drivers of re-identification and membership inference attacks. By aligning learning with stable causal mechanisms, such models are less susceptible to privacy attacks that exploit dataset-specific artifacts. In particular, Tople et al. (Tople et al., 2020) show that causal models provide stronger defenses against membership inference attacks while requiring a smaller privacy budget ϵ under differential privacy, thereby achieving comparable privacy guarantees with less degradation in utility.

In addition, explicitly modeling how private attributes propagate through the system, causal structure makes it possible to suppress only privacy-sensitive effects while preserving non- sensitive causal relationships. We highlight this form of targeted causal obfuscation as a promising design direction that can lower the impact of data obfuscation on performance. Similarly, in foundation models, causal auditing can enable principled attribution by distinguishing data memorization from incidental stylistic similarity (Sharkey et al., 2024).

3.3. Robustness–Accuracy

Statistical origin of the trade-off. High predictive accuracy under a fixed training distribution often relies on shortcut correlations that are specific to that environment. When deployment conditions change, these correlations may break, leading to degraded performance. Methods that improve robustness by discouraging such shortcuts often reduce in-distribution accuracy.

Invariance conflict. Robustness requires invariance across a family of environments, which can be formalized as distributions related by interventions on latent or observed environmental variables. Accuracy under the observational distribution, however, rewards sensitivity to all predictive correlations, including those that are unstable under such interventions.

Causal resolution. Causal reasoning resolves this conflict by distinguishing invariant causal mechanisms from spurious associations. Features that are causal parents of the target remain predictive under interventions, while non-causal shortcuts do not. In this way, causal reasoning shifts focus to interventional accuracy and provides a principled way to prioritize stable features.

In foundation models, shortcut correlations are often encoded implicitly in learned representations. Causal invariance and regularization techniques can reduce reliance on unstable patterns while preserving task-relevant structure (Zhou et al., 2023). Moreover, recent theoretical results show that agents robust across environments must implicitly learn causal world models (Richens & Everitt, 2024), further highlighting the link between robustness and causality.

3.4. Explainability–Performance

Statistical origin of the trade-off. Highly expressive models often achieve strong performance by exploiting complex, distributed correlations that are difficult for humans to interpret. As a result, improvements in predictive capability frequently come at the cost of interpretability and explainability (London, 2019; van der Veer et al., 2021).

Invariance conflict. Explainability requires stability of model behavior under interventions on semantically meaningful variables, enabling counterfactual reasoning about why a prediction was made. Performance-oriented models, in contrast, may rely on opaque correlations that change unpredictably under such interventions.

Causal resolution. Causal models address this tension by explicitly representing how inputs influence outputs through causal mechanisms, enabling counterfactual explanations and actionable recourse (Wachter et al., 2017).

In foundation models, causal structure can be probed within internal components such as embeddings, attention heads, and logits. Methods for causal path analysis and intervention-based probing enable step-by-step explanations of generative behavior (Bagheri et al., 2024; Conmy et al., 2023). In addition, the causal approach in mechanistic interpretability provides a principled way to faithfully translate complex generative processes to human-understandable abstractions (Geiger et al., 2024).

3.5. Towards Multi-Objective Trade-Off Resolution

Origins of multi-objective trade-offs. The preceding sections show that many trade-offs between fairness, privacy, robustness, and accuracy or model performance arise when invariance constraints are enforced uniformly over all statistical dependencies. In addition, privacy mechanisms often worsen fairness for minority groups. When obfuscation is applied uniformly, the signal-to-noise ratio of underrepresented populations collapses first, because their causal pathways are already weakly supported by data. This creates a fairness-privacy trade-off. Similarly, data obfuscation obscures the relationships between predictive variables and the outcome, thus creating a privacy-explainability

trade-off.

Towards causal resolution. We argue that the causal approach leads to the improvement of multi-objective trade-offs in trustworthy AI, or, in other words, allows to improve multiple dimensions simultaneously. First, by decomposing total effects into causal pathways, causal models allow invariance requirements to be targeted rather than global. For example, privacy can be enforced only on paths that transmit individual identifying information. Second, causal models are inherently explainable, robust, and require less privacy noise for preventing leakage of sensitive information (Tople et al., 2020). As a result, they should better preserve global accuracy as well as accuracy for sensitive and underrepresented groups. Finally, in algorithmic recourse, achieving fairness and robustness under counterfactual changes requires a causal structure to specify which interventions are feasible and admissible (Karimi et al., 2021). This makes causality a favorable design principle for future trustworthy AI systems and motivates systematic study of how these dimensions interact and can be reconciled through causal modeling.

4. Integrating Causality into ML and Foundation Models

In this section, we discuss how causality can be encoded in learning systems, spanning both classical ML models and modern foundation models (FMs). We focus on the high-level principles relevant to the feasibility of the causal approach. For a detailed overview of the methods see Appendix B.

4.1. Defining Explicit And Implicit Causal Integration

Causal assumptions can be incorporated into learning systems either explicitly, through structural representations, or implicitly, through inductive biases and training regimes that encourage causal behavior (Rawal et al., 2025). Based on this distinction, we define *Explicit* and *Implicit* approaches to causality in ML and FMs.

Explicit causal integration. Explicit approaches represent causal structure directly, typically via structural causal models (SCMs), causal graphs, or path-specific constraints, and learning objectives are defined to satisfy causal properties under specified interventions.

Formally, given a causal graph G, a model f_θ is trained to satisfy invariance or sensitivity constraints derived from G.

The primary strength of explicit causal integration lies in its transparency and auditability: causal constraints are interpretable, and violations can be traced to specific variables or pathways. This makes explicit approaches particularly well-suited to high-stakes settings where trade-offs between fairness, privacy, robustness, and explainability must be justified and inspected. By distinguishing spurious, as well as admissible or inadmissible causal effects, explicit models allow selective invariance-preserving task-relevant mechanisms while suppressing normatively unacceptable ones.

Implicit causal integration. Implicit approaches do not represent explicit causal structure, but instead encourage models to behave causally through training objectives, data diversity, sparsity, or environmental interaction.

Formally, models are trained to satisfy causal properties without explicitly encoding the causal graph G, but rather a set of constraints C, e.g., sparsity.

Implicit causal integration, by contrast, offers flexibility. It allows models to approximate causal behavior through inductive biases, multi-environment data, or counterfactual data augmentation, even when causal structure is incomplete or unknown. While they offer weaker formal guarantees, they can still soften trade-offs in practice by discouraging reliance on unstable, unethical, or spurious correlations.

Resolving trade-offs. From the perspective of trustworthy AI trade-offs, explicit and implicit causal integration tend to be effective in different regimes. Explicit approaches are particularly well-suited to objectives that require normative judgments about which causal pathways are admissible, such as explainable vs. proxy discrimination paths in fairness. Implicit approaches, by contrast, are often more effective for trade-offs driven by spurious correlations or distributional instability, such as robustness or privacy-utility trade-offs, especially in high-dimensional or large-scale settings where full causal structure is unavailable. In practice, many systems combine both modes, using explicit causal constraints to target specific effects while relying on implicit mechanisms to scale causal behavior more broadly.

Integration into AI. The necessity of causality does not depend on model scale. The key distinction lies in *how* these assumptions are integrated. In smaller or more structured ML models, causal integration is often *explicit*: causal assumptions can be embedded directly into the model class or training objective, for example, by enforcing consistency with a specified causal graph during learning (Berrevoets et al., 2024). Such explicit integration enables strong, transparent causal guarantees, but relies on the availability of reliable causal structure.

Foundation models, by contrast, rely predominantly on *implicit* causal integration. Their scale, opacity, and fixed

pre-training pipelines make global explicit causal specification impractical. Instead, causal assumptions are encoded indirectly through data curation, representation learning, inductive biases, and post-training objectives that encourage invariant behavior without explicitly representing causal structure. This reliance on implicit mechanisms is not unique to foundation modelslarge neural models in classical ML settings also employ sparsity constraints or counterfactual data augmentation to induce causal behavior (Burgess et al., 2018; Kim et al., 2021; Kaddour et al., 2022)but it becomes the dominant mode at scale.

Nevertheless, recent work demonstrates that explicit causal structure can still be leveraged in foundation models in a targeted or partial manner, for example, through SCM-guided entity interventions or causal regularization applied to specific components or behaviors (Wang et al., 2023; Madhavan et al., 2023). This suggests that effective causal integration in foundation models often takes a hybrid form, combining implicit causal learning at scale with localized explicit constraints where interpretability or normative guarantees are required.

4.2. Application to Foundational Models

For FMs, causal integration is best understood as a lifecycle process, with different leverage points at pre-training, post-training, or auditing. Importantly, these stages differ not only in when causal assumptions are introduced, but also in what causal object is being manipulated.

During *pre-training*, causal priors are introduced by intervening on the data-generating process or the representation space before task-specific objectives are defined. This includes counterfactual data generation and causal representation learning methods that aim to disentangle underlying generative factors (Rajendran et al., 2024; Jiang et al., 2024; Chen et al., 2022). At this stage, causality operates primarily at the distributional level: interventions reshape the joint data distribution so that invariant mechanisms become statistically identifiable, while spurious correlations are attenuated. Pre-training, therefore, offers a natural opportunity to encode broad structural assumptions that generalize across downstream tasks.

At the *post-training and alignment* stage, causal assumptions are enforced through fine-tuning, regularization, or alignment objectives that operate directly on model behavior (Xia et al., 2024). Rather than modifying the training distribution, constraints are expressed as requirements on how model outputs should respond to counterfactual perturbations. This makes post-training the primary locus for enforcing trustworthy AI objectives in foundation models. Fairness, or robustness constraints, can be implemented as invariance or sensitivity conditions on outputs, even when the causal structure was not explicitly modeled during pre-training.

Finally, during the *auditing and interpretability* phase, causal reasoning can be applied retrospectively, even when no causal constraints were imposed during training. Intervention-based probing, activation patching, and mechanistic interpretability methods enable estimation of the causal influence of internal components on model behavior, supporting auditing and explanation (Kissane et al., 2024; Conmy et al., 2023; Syed et al., 2023). These approaches allow practitioners to assess whether learned representations and internal circuits satisfy desired invariance properties, and to diagnose failure modes related to fairness, robustness, or spurious correlations without modifying the training pipeline.

5. Challenges and Opportunities

Despite the advantages of a causal approach, practical applications face fundamental limitations that arise from the same design choices that enable causal integration. In particular, trade-offs between explicit and implicit representations, between early and late intervention in the model lifecycle limits what guarantees can be achieved in practice. These limitations are amplified in foundation models due to their scale, opacity, and decoupled training pipelines. We outline key conceptual and practical obstacles and corresponding opportunities for navigating this design space.

5.1. Conceptual Challenges

Potentially unresolvable tensions. Not all tensions in trustworthy AI can always be fully resolved. For instance, stronger privacy protections often reduce model utility (Dwork et al., 2014; Bassily et al., 2014). We also acknowledge that some fairness conflicts stem from deeper normative or value-based disagreements; for example, a causal relationship may exist, but relying on it in decision-making may still be viewed as unfair from an ethical or legal standpoint. In these cases, causality does not eliminate trade-offs; however, it makes their structural origin explicit, shifting the debate from engineering heuristics to transparent normative choices.

Concept superposition. Foundation models suffer from concept superposition, namely, multiple meanings are entangled within a single representation, complicating causal reasoning (Elhage et al., 2022). This limits the granularity at which causal interventions can be applied, shifting causal control toward behavior-level constraints.

5.2. Implementation Challenges

Assumptions and identifiability. The main limitation of explicit causal integration is its reliance on accurate or partially specified causal knowledge, particularly in the form

of DAGs. Misspecified graphs or incorrect structural assumptions can lead to incorrect invariances or unintended behavior. Expert-constructed DAGs may suffer from subjectivity and scalability issues, while ML-based causal discovery is constrained by identifiability assumptions and noise sensitivity. However, recent hybrid approaches combining classical causal discovery with LLM-based reasoning offer promising solutions (Afonja et al., 2024). In addition, approximate causal interventions are possible with a partial causal graph (Zuo et al., 2022).

Implicit approaches make weaker structural assumptions, but they are not assumption-free. They rely on properties such as environmental diversity, intervention coverage, or stability of causal mechanisms across domains. When these conditions fail, e.g., when all training environments share the same confounding structure, implicit methods may converge to spurious but stable correlations.

Scaling. Explicit causal approaches scale poorly with the complexity of the causal structure, rather than with model size alone. As the number of variables, dependencies, or latent confounders grows, specifying and enforcing global causal constraints becomes difficult. In practice, explicit methods are therefore often applied locally or partially, for example, by constraining specific pathways relevant to fairness (Wang et al., 2023). Hybrid strategies that combine partial causal structure with implicit learning objectives offer a practical compromise between interpretability and scalability. Implicit approaches, by contrast, scale naturally with model and data size, as they rely on data augmentation, representation learning, and objective design rather than explicit symbolic causal structure.

Evaluation and benchmarking. Evaluating causal integration remains challenging due to the lack of standardized benchmarks that reflect interventional objectives. Most existing evaluations rely on observational performance, which may obscure failures under interventions. Developing benchmarks and evaluation protocols aligned with causal invariance requirements is, therefore, a critical opportunity for advancing trustworthy AI.

Lack of high-quality causal data. Applying causal integration at the pre-training phase requires high-quality interventional data, which are scarce and expensive to produce. Scalable methods for generating synthetic causal datasets show a promising direction (Webster et al., 2020; Chen et al., 2022). Alternatively, focusing on post-training methods allows causal interventions in a more data-efficient way.

Computational complexity. Integrating causal reasoning into large models introduces additional computational overhead, for example, through counterfactual evaluation or causal regularization during training. Parameter-efficient adaptation methods such as LoRA reduce this burden by restricting updates to low-rank subspaces, enabling efficient causal fine-tuning without modifying the full parameter set (Hu et al., 2021).

6. Alternative Views

Symbolic AI leverages prior knowledge through logic rules, ontologies, and formal representations, supporting structured reasoning, explainability, and generalization from small data (Díaz-Rodríguez et al., 2022). It can also help navigate trade-offs, for example, between accuracy and interoperability, by making the reasoning process more transparent and controllable. While both symbolic AI and causal reasoning rely on prior assumptions, causality uniquely enables reasoning about interventions and counterfactuals. Unlike symbolic systems that model static relationships, causal models capture how changes in one variable affect others - an essential feature in domains like fairness and policy evaluation. A dominant alternative position advocates for improving model performance by scaling, and deems causal approaches as impractical. Our goal therefore, is to illustrate that implicit, explicit, or hybrid causal guidance can provide feasible, data-efficient methods for both performant and trustworthy AI models.

7. Conclusion and Call for Action

We demonstrate that causal models offer a principled approach to trustworthy AI by disentangling conflicting invariance requirements and shifting attention from observational to interventional accuracy. Hence, the causal approach creates models that are not only ethically sound but also practically effective. We further discuss practical ways to apply causality to AI by distinguishing explicit and implicit design. To further advance the application of causality for trustworthy AI, we call for the following actions

✓ *Redefine trustworthy AI as a multi-objective optimization, rather than as a collection of competing constraints.* This requires establishing evaluation frameworks and benchmarks that jointly measure objectives of trustworthy AI, explicitly quantifying the trade-offs between them via multi-objective evaluation metrics.

✓ *Leverage Causality to Resolve or Soften Trade-offs*: Where possible, integrate causal reasoning to disentangle competing objectives and mitigate conflicts.

✓ *Develop Scalable Methods for Causal Data Integration*: Encourage the development of algorithms and pipelines to integrate causal knowledge into foundation models at scale.

✓ *Create and Share High-Quality Causal Datasets*: Foster initiatives to curate, annotate, and share datasets with explicit causal annotations or interventional information.

Impact statement

This paper advances a causal perspective on trustworthy foundation models by framing fairness, privacy, robustness, and explainability as competing invariance requirements under interventions. Rather than treating trade-offs between these objectives as incidental, our analysis clarifies when they are structurally unavoidable and when they can be softened through selective causal invariance. By making the assumptions underlying these trade-offs explicit, the proposed framework supports more transparent, accountable, and principled design of learning-based AI systems, particularly in high-stakes domains such as healthcare, law, and finance.

References

Abdulaal, A., Montana-Brown, N., He, T., Ijishakin, A., Drobnjak, I., Castro, D. C., Alexander, D. C., et al. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2023.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv*, 2023.

Afonja, T., Sheth, I., Binkyte, R., Hanif, W., Ulas, T., Becker, M., and Fritz, M. LLM4GRN: Discovering causal gene regulatory networks with LLMs–evaluation through synthetic data generation. *arXiv preprint arXiv:2410.15828*, 2024.

AI4Science, M. R. and Quantum, M. A. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv*, 2023.

Bagheri, A., Alinejad, M., Bello, K., and Akhondi-Asl, A. C2p: Featuring large language models with causal reasoning. *arXiv preprint arXiv:2407.18069*, 2024.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

Berrevoets, J., Kacprzyk, K., Qian, Z., van der Schaar, M., et al. Causal deep learning: encouraging impact on real-world problems through causality. *Foundations and Trends® in Signal Processing*, 18(3):200–309, 2024.

Bi, X., Wu, D., Xie, D., Ye, H., and Zhao, J. Large-scale chemical process causal discovery from big data with transformer-based deep learning. *Process Safety and Environmental Protection*, 173:163–177, 2023.

Brinkmann, J., Sheshadri, A., Levoso, V., Swoboda, P., and Bartelt, C. A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. *arXiv preprint arXiv:2402.11917*, 2024.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in vae. *arXiv preprint arXiv:1804.03599*, 2018.

Carvalho, T., Moniz, N., Faria, P., and Antunes, L. Towards a data privacy-predictive performance trade-off. *Expert Systems with Applications*, pp. 119785, 2023.

Chen, Z., Gao, Q., Bosselut, A., Sabharwal, A., and Richardson, K. Disco: Distilling counterfactuals with large language models. *arXiv preprint arXiv:2212.10534*, 2022.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

Chiappa, S. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Crook, B., Schlüter, M., and Speith, T. Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pp. 316–324. IEEE, 2023.

Díaz-Rodríguez, N., Lamas, A., Sanchez, J., Franchi, G., Donadello, I., Tabik, S., Filliat, D., Cruz, P., Montes, R., and Herrera, F. Explainable neural-symbolic learning (x-nesyl) methodology to fuse deep learning representations with expert knowledge graphs: the monumai cultural heritage use case. *Information Fusion*, 79:58–83, 2022.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Elhage, N., Olsson, C., Henighan, T., Hernandez, D., Joseph, N., Mann, B., Askell, A., DasSarma, N., Tran-Johnson, E., Amodei, D., Brown, T., Clark, J., McCandlish, S., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/.

European Commission. European Union AI act, 2021. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. Proposal for a regulation laying down harmonized rules on artificial intelligence.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021.

Ganguly, N., Fazlija, D., Badar, M., Fisichella, M., Sikdar, S., Schrader, J., Wallat, J., Rudra, K., Koubarakis, M., Patro, G. K., et al. A review of the role of causality in developing trustworthy ai systems. *arXiv preprint arXiv:2302.06975*, 2023.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Preprint*, pp. 9, 2024.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. Causal generative neural networks. *arXiv preprint arXiv:1711.08936*, 2017.

Guo, Y., Yang, Y., and Abbasi, A. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1012–1023, 2022.

Hauser, A. and Bühlmann, P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012. URL https://jmlr.org/papers/v13/hauser12a.html.

He, J., Xia, M., Fellbaum, C., and Chen, D. Mabel: Attenuating gender bias using textual entailment data. *arXiv preprint arXiv:2210.14975*, 2022.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Jiang, Y., Rajendran, G., Ravikumar, P., Aragam, B., and Veitch, V. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*, 2024.

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., and Silva, R. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Kaneko, M. and Bollegala, D. Debiasing pre-trained contextualised embeddings. *arXiv preprint arXiv:2101.09523*, 2021.

Karimi, A.-H., Schölkopf, B., and Valera, I. Algorithmic recourse: from counterfactual explanations to interventions. In *4th Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, pp. 353–362, 2021. doi: 10.1145/3442188.3445899.

Kasetty, T., Mahajan, D., Dziugaite, G. K., Drouin, A., and Sridhar, D. Evaluating interventional reasoning capabilities of large language models. *arXiv preprint arXiv:2404.05545*, 2024.

Khatibi, E., Abbasian, M., Yang, Z., Azimi, I., and Rahmani, A. M. Alcm: Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*, 2024.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research (2023)*, 2023.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pp. 656–666, 2017.

Kim, H., Shin, S., Jang, J., Song, K., Joo, W., Kang, W., and Moon, I.-C. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8128–8136, 2021.

Kim, J. S., Chen, J., and Talwalkar, A. Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pp. 5264–5274. PMLR, 2020.

Kissane, C., Krzyzanowski, R., Bloom, J. I., Conmy, A., and Nanda, N. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*, 2024.

Le, T. D., Hoang, T., Li, J., Liu, L., Liu, H., and Hu, S. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(5):1483–1495, 2016.

Liu, H., Chaudhary, M., and Wang, H. Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives. *arXiv preprint arXiv:2307.16851*, 2023.

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

London, A. J. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, 49(1):15–21, 2019.

Madhavan, R., Garg, R., Wadhawan, K., and Mehta, S. Cfl: Causally fair language models through token-level attribute controlled generation, 2023.

Nauta, M., Bucur, D., and Seifert, C. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

Pearl, J. *Causality*. Cambridge University Press, Cambridge, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161. URL https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. Identifiability of causal graphs using functional models. In Cozman, F. G. and Pfeffer, A. (eds.), *27th Conference on Uncertainty in Artificial Intelligence*, pp. 589–598, Corvallis, OR, 2011. AUAI Press.

Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 189–199, 2020.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.

Rajendran, G., Buchholz, S., Aragam, B., Schölkopf, B., and Ravikumar, P. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint arXiv:2402.09236*, 2024.

Rawal, A., Raglin, A., Rawat, D. B., Sadler, B. M., and McCoy, J. Causality for trustworthy artificial intelligence: status, challenges and perspectives. *ACM Computing Surveys*, 57(6):1–30, 2025.

Richens, J. and Everitt, T. Robust agents learn causal world models. In *2024 The Twelfth International Conference on Learning Representations (ICLR)*. ICLR, 2024.

Rohan, Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv*, 2023.

Schölkopf, B., Hogg, D., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., and Peters, J. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Science*, 113(27):7391–7398, 2016.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kaltenpoth, M., Goyal, A., and Bengio, Y. Towards causal representation learning. *Proceedings of the IEEE, 109(5), 612-634.*, 2022.

Sharkey, L., Ghuidhir, C. N., Braun, D., Scheurer, J., Balesni, M., Bushnaq, L., Stix, C., and Hobbhahn, M. A causal framework for ai regulation and auditing. 2024.

Sheth, I., Abdelnabi, S., and Fritz, M. Hypothesizing missing causal variables with llms. *arXiv preprint arXiv:2409.02604*, 2024.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT press, 2000.

Syed, A., Rager, C., and Conmy, A. Attribution patching outperforms automated circuit discovery. *arXiv preprint arXiv:2310.10348*, 2023.

Tople, S., Sharma, A., and Nori, A. Alleviating privacy attacks via causal learning. In *International Conference on Machine Learning*, pp. 9537–9547. PMLR, 2020.

Tschantz, M. C., Sen, S., and Datta, A. Sok: Differential privacy as a causal property. In *2020 IEEE Symposium on Security and Privacy (SP)*, pp. 354–371. IEEE, 2020.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

van der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., Oswald, M., et al. Trading off accuracy and explainability in ai decision-making: findings from 2 citizens juries. *Journal of the American Medical Informatics Association*, 28(10):2128–2138, 2021.

Vashishtha, A., Reddy, A. G., Kumar, A., Bachu, S., Balasubramanian, V. N., and Sharma, A. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

Vincent, J. Google pauses gemini ai image generation after historical inaccuracies spark backlash. *The Verge*, 2024. URL https://www.theverge.com.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law and Technology*, 31(2):841–887, 2017.

Wang, F., Mo, W., Wang, Y., Zhou, W., and Chen, M. A causal view of entity bias in (large) language models. *arXiv preprint arXiv:2305.14695*, 2023.

Wang, Z. and Culotta, A. Identifying spurious correlations for robust text classification. *arXiv preprint arXiv:2010.02458*, 2020.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., and Petrov, S. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.

Xia, Y., Yu, T., He, Z., Zhao, H., McAuley, J., and Li, S. Aligning as debiasing: Causality-aware alignment via reinforcement learning with interventional feedback. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4684–4695, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.262. URL https://aclanthology.org/2024.naacl-long.262/.

Xu, L., Jiang, C., Qian, Y., Li, J., Zhao, Y., and Ren, Y. Privacy-accuracy trade-off in differentially-private distributed classification: A game theoretical approach. *IEEE Transactions on Big Data*, 7(4):770–783, 2017.

Zhao, H. and Gordon, G. J. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552, 2022.

Zhou, F., Mao, Y., Yu, L., Yang, Y., and Zhong, T. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4227–4241, 2023.

Zinati, Y., Takiddeen, A., and Emad, A. Groundgan: Grn-guided simulation of single-cell rna-seq data using causal generative adversarial networks. *Nature Communications*, 15(1):4055, 2024.

Zliobaite, I. On the relation between accuracy and fairness in binary classification. *The 2nd workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15*, 2015.

Zuo, A., Wei, S., Liu, T., Han, B., Zhang, K., and Gong, M. Counterfactual fairness with partially known causal graph. *Advances in Neural Information Processing Systems*, 35:1238–1252, 2022.

A. Motivating Example

Consider a model trained to predict whether a patient will be readmitted within 30 days after discharge. The model uses patient information such as age and laboratory measurements, and is deployed across multiple hospitals. Although age is statistically correlated with readmission risk, it should not influence decisions directly, or indirectly through proxy variables, such as insurance type, which introduce bias and discrimination. Age may affect the prediction only through legitimate clinical signals, such as lab results. In addition, hospitals differ in medical devices, data logging systems, and discharge policies, which induces environment-dependent variation in the observed data. Some of this variation is purely spurious, reflecting measurement artifacts, such as image frames, while other differences reflect real domain-specific mechanisms, such as internal policies. Moreover, patient attributes are sensitive and must be protected, motivating privacy-preserving learning. However, privacy mechanisms often degrade performance by injecting noise. This example illustrates the fundamental challenge of building models that are simultaneously fair, robust to domain shifts, privacy-preserving, and explainable.

Causal model. We consider a structural causal model (SCM) with observed variables A (protected attribute - age), E (environment or domain - hospital), X (legitimate observed features), Z (proxy feature - insurance type), and Y (target label - hospital readmission). The causal graph encodes the following relations: $A \to X \to Y$ (legitimate path based on lab tests and vitals), $A \to Z \to Y$ (proxy-based discriminatory path based on insurance type), $A \to Y$ (direct discrimination based on age) $E \to X$ (environment-dependent spurious measurement artifacts, such as image frames), and $E \to Y$ (domain-dependent outcome mechanism, such as hospital policies).

Fairness. We define fairness as the removal of all unfair causal effects of the protected attribute A on the target Y. This is achieved by blocking the causal paths $A \to Z \to Y$ and $A \to Y$, thereby eliminating both direct and proxy-mediated discriminatory influence. The remaining path $A \to X \to Y$ represents legitimate influence through task-relevant features.

Robustness. We distinguish between spurious and contextual environment effects. Spurious environment effects correspond to measurement artifacts, modeled by the path $E \to X \to Y$. We block this path to prevent the predictor from exploiting non-semantic, environment-induced variation. At the same time, we explicitly model the contextual effect $E \to Y$, which captures genuine domain-dependent mechanisms (e.g., policy differences). Thus, robustness is achieved by enforcing invariance to spurious environment-induced shifts while conditioning on domain context.

Privacy. Focusing on causal features and discarding spurious correlations, the resulting models are less prone to overfitting and exhibit improved robustness to adversarial perturbations. This allows strong privacy guarantees with reduced noise and impact on predictive performance. Moreover, targeted and path specific obfuscation is possible.

Explainability. The proposed framework is explainable by design, since predictions correspond to explicit causal pathways in the underlying causal graph.

In conclusion, this unified causal perspective enables resolving or softening the trade-offs and achieving accurate, robust, fair, private and explainable models.

B. Methods

B.1. Integrating Causality into ML

Integrating causality into ML allows models to move beyond surface-level patterns and capture the underlying mechanisms that generate data. This section introduces methods for incorporating causal knowledge into ML models, as well as techniques for using ML to discover causal relationships from data.

Causally Constrained ML (CCML). CCML refers to approaches that *explicitly* incorporate causal relationships into model training or inference as constraints or guiding principles. Given a causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ represents variables and $\mathbf{E}$ denotes directed edges encoding causal relationships, the goal is to ensure that the learned model $f : \mathbf{X} \to Y$ adheres to the causal structure encoded in $\mathcal{G}$ (Berrevoets et al., 2024; Zinati et al., 2024; Afonja et al., 2024; Schölkopf et al., 2016).

Invariant feature learning (IFL). IFL relies on discovered *implicit* or latent causal features and structures. The task of Invariant Feature Learning (IFL) is to identify features of the data $\mathbf{X}$ that are predictive of the target Y across a range of

environments $\mathcal{E}$. From a causal perspective, the causal parents $\text{Pa}(Y)$ are always predictive of Y under any interventional distribution (Kaddour et al., 2022). IFL can be achieved by regularizing the model or providing causal training data that is free of confounding.

Disentangled VAEs. VAEs aim to decompose the data $\mathbf{X}$ into disentangled latent factors $\mathcal{Z}$ that correspond to distinct underlying generative causes, especially when trained with causal constraints or interventional data (Burgess et al., 2018; Kim et al., 2021). This aligns with causal reasoning, as each latent dimension can be interpreted as an independent causal factor influencing the observed data.

Double Machine Learning (DML). DML provides another causal approach by leveraging modern ML techniques for estimating high-dimensional nuisance parameters while preserving statistical guarantees in causal inference (Chernozhukov et al., 2018). DML decomposes the estimation problem into two stages: (1) predicting confounders using ML models and (2) estimating the causal effect using residualized outcomes.

Causal Discovery. Finally, ML can be leveraged for causal inference or to discover causal knowledge from observational data. For instance, methods for causal discovery use conditional independence, score, or noise asymmetry-based methods, to infer causal relationships, with notable examples including (Spirtes et al., 2000; Shimizu et al., 2006; Janzing & Schölkopf, 2010; Peters et al., 2011; Hauser & Bühlmann, 2012; Le et al., 2016). More recent methods leverage Deep Neural Networks (DNNs) in combination for causal discovery (Nauta et al., 2019; Bi et al., 2023; Goudet et al., 2017).

Summary and Use Cases.

The methods presented above represent a spectrum of causal integration strategies in ML. CCML requires explicit or semi-explicit causal structures, offering high interpretability but often demanding strong prior knowledge. However, explicit causal guardrails are beneficial for high-stakes applications requiring clear accountability. In contrast, IFL, DML, and Disentangled VAEs operate with weaker structural assumptions and are well-suited for settings with limited causal annotations as well as exploratory tasks or scalable representation learning in unstructured data. Causal Discovery lies at the other end of the spectrum, aiming to uncover causal relationships from data alone. Combining methodsfor example, using causal discovery to inform CCML also help balance rigor and practicality.

B.2. Integrating Causality in Foundation Models

This section delves into practical applications of causality in FMs across three key stages: pre-training, post-training, and auditing. We conclude with a discussion of the practical advantages and limitations of the proposed approaches.

Causal Representation Learning. By disentangling causal factors from non-causal ones, models can learn representations that separate meaningful causal features from irrelevant associations. Techniques such as causal embedding methods (Rajendran et al., 2024; Jiang et al., 2024), use training data annotated with causal labels. This has been shown to reduce reliance on spurious correlations, such as gender-biased occupational associations (Zhou et al., 2023). Fine-tuning models on embeddings pre-trained with debiased token representations has shown promise for causal learning (Kaneko & Bollegala, 2021; Guo et al., 2022; He et al., 2022; Wang et al., 2023).

Entity interventions. SCMs can be used to determine interventions on specific entities (e.g., replacing “Joe Biden" with “ENTITY-A") during pre-training (Wang et al., 2023), thus reducing entity-based spurious associations while preserving causal relationships in the data.

Counterfactual Data Augmentation. Synthetic datasets with explicit causal structures or counterfactual examples, that introduce scenarios where causal relationships differ from spurious correlations, help models learn true causal dependencies instead of misleading patterns (Webster et al., 2020; Chen et al., 2022). Synthetic causal data can also be used for fine-tuning the models, similar to pre-training, but with better sample size efficiency. Frameworks like DISCO (Chen et al., 2022) generate diverse counterfactuals during fine-tuning to enhance OOD generalization for downstream tasks.

SCM-based Methods. Causally Fair Language Models (CFL) (Madhavan et al., 2023) use SCM-based regularization to detoxify outputs or enforce demographically neutral generation during post-training. Wang & Culotta (2020) use causal reasoning to separate genuine from spurious correlations by computing controlled direct effects, ensuring robust performance.

Causal RLHF Alignement. RLHF can be adapted to include causal interventions, allowing feedback to act as instrumental variables that correct biased model behavior. Causality-Aware Alignment (Xia et al., 2024) incorporates causal

interventions to reduce demographic stereotypes during fine-tuning with alignment objectives. Extending RLHF with causal alignment to support dynamic, context-sensitive interventions could help address biases that evolve. Integrating causal reasoning into the reward model's decision-making process, by critiquing the output of LLM using a reward model or a mixture of reward models that control for specific confounders or spurious correlations, can potentially improve the downstream reasoning abilities, potentially mitigating hallucinations.

Latent Representation Discovery. Recent work in mechanistic interpretability has employed sparse autoencoders to extract interpretable latent representations aligned with internal circuit features (Kissane et al., 2024; Conmy et al., 2023). These methods enable the identification of functional substructures within models by isolating meaningful directions in activation space.

Targeted Interrogation Methods. Techniques based on interventions, such as ablation, activation patching, or causal scrubbing, allow us to move beyond passive observation by directly testing the functional roles of model components (Brinkmann et al., 2024; Syed et al., 2023). These methods provide a powerful framework for identifying how specific neurons or pathways contribute to behavior, enabling rigorous, theory-driven interpretability.

Causal Discovery. The recent advancements in large language models (LLMs) have inspired their use in causal discovery (Kıcıman et al., 2023; Kasetty et al., 2024; Vashishtha et al., 2023; AI4Science & Quantum, 2023; Abdulaal et al., 2023; Khatibi et al., 2024). Most of the above methods involve the refinement of the statistically inferred causal graph by LLM. However, emerging research shows that LLMs can be useful in constructing full causal graphs based on parametric knowledge of scientific literature in diverse domains (Sheth et al., 2024; Afonja et al., 2024).

Summary and Use Cases.

These methods span both causal inference and integration. Causal data augmentation, representation learning, and entity interventions are methods used in pretraining, where broad structural priors can be encoded. In contrast, causal fine-tuning, alignment, and auditing are post-training methods requiring less causally annotated data, but more explicit goals and causal structures. Latent representation discovery and targeted interrogation methods stand out as a retrospective techniques that can help to assess fairness, privacy, and robustness without modifying the training pipeline. Causal discovery can support both pre- and post-training by informing model design or evaluating learned representations. Selecting among these approaches depends on the availability of causal knowledge, the interpretability needs, and the constraints of the training pipeline.