

Bridging Distribution Shift and AI Safety: Conceptual and Methodological Synergies

CHENRUO LIU*, Center for Data Science, New York University, USA

KENAN TANG*, Computer Science Department, University of California, Santa Barbara, USA

YAO QIN, Department of Electrical and Computer Engineering, University of California, Santa Barbara, USA

QI LEI, Courant Institute for Mathematical Sciences & Center for Data Science, New York University, USA

This paper bridges distribution shift and AI safety through a comprehensive analysis of their conceptual and methodological synergies. While prior discussions often focus on narrow cases or informal analogies, we establish two types connections between specific causes of distribution shift and fine-grained AI safety issues: (1) methods addressing a specific shift type can help achieve corresponding safety goals, or (2) certain shifts and safety issues can be formally reduced to each other, enabling mutual adaptation of their methods. Our findings provide a unified perspective that encourages fundamental integration between distribution shift and AI safety research.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Distribution Shift, AI Safety, Selection Bias, Spurious Correlation

1 Introduction

Machine learning (ML) models have achieved remarkable success across complex tasks in various fields, including high-stakes applications such as autonomous driving, healthcare, finance, and legal domains [32, 52, 155, 171]. Traditionally, ML models are trained on historically collected data and subsequently deployed in a testing domain. While these ML models succeed in the case where the training and testing data share the same distribution, performance drops are expected when distribution shifts from training to testing domains [81].

Beyond performance concerns, the failure of ML models to robustly handle distribution shifts in real-world settings can lead to severe safety issues. For instance, such failures have been reported to cause malfunctions in autonomous driving systems [169], inappropriate responses from large language models (LLMs) [183], and incidents of medical misdiagnosis [113]. As artificial intelligence (AI) systems continue to evolve, minimizing their safety risks has emerged as a critical priority [4]. Consequently, distribution shift, which is inherently linked to safety, has become a central challenge in AI safety research.

A deep understanding of the relationship between distribution shift and AI safety is undoubtedly beneficial for the safe deployment of ML models in real-world applications. However, although previous studies have frequently

*Both authors contributed equally to this research.

Authors' Contact Information: Chenruo Liu, cl7758@nyu.edu, Center for Data Science, New York University, New York, New York, USA; Kenan Tang, kenantang@ucsb.edu, Computer Science Department, University of California, Santa Barbara, Santa Barbara, California, USA; Yao Qin, yaoqin@ucsb.edu, Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, California, USA; Qi Lei, ql518@nyu.edu, Courant Institute for Mathematical Sciences & Center for Data Science, New York University, New York, New York, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

mentioned the connection between these two fields [4, 62, 73, 90], they exhibit several limitations. First, they focus narrowly on the relationship between distribution shift and a single aspect of AI safety (e.g., robustness, trustworthiness, or alignment) but lacks a more comprehensive framework. Second, these studies often treat both distribution shift and the targeted AI safety aspect as general concepts, rather than breaking them down into specific categories. Third, existing work discuss the connection between distribution shift and AI safety as an informal analogy without identifying the mathematical relationships among their rigorous definitions. As a result, there remains a lack of detailed exploration into the connections between specific topics in distribution shift and AI safety.

In this paper, we offer the first comprehensive, one-to-one aligned, and mathematically grounded analysis of the relationship between distribution shift and AI safety. Specifically, according to the underlying causes, we examine distribution shifts from three main categories: *selection bias* [151, 164], *spurious correlation* [166, 200], and *label shift* [44, 100, 172] and six more fine-grained categories. Meanwhile, we focus on the following categories of AI safety issues: *AI security* [152], *AI fairness* [221], *AI trustworthiness* [79], and *AI democracy* [49]. For each specific topic discussed within these main categories, we provide its rigorous definition (see Figure 1, Figure 2 and Section 2 for details). Next, we systematically identify connections between specific distribution shift causes and concrete AI safety issues by analyzing the underlying mathematical relationships between their definitions. Based on the identified connections, we then review the relevant literature to establish methodological links between these specific topics.

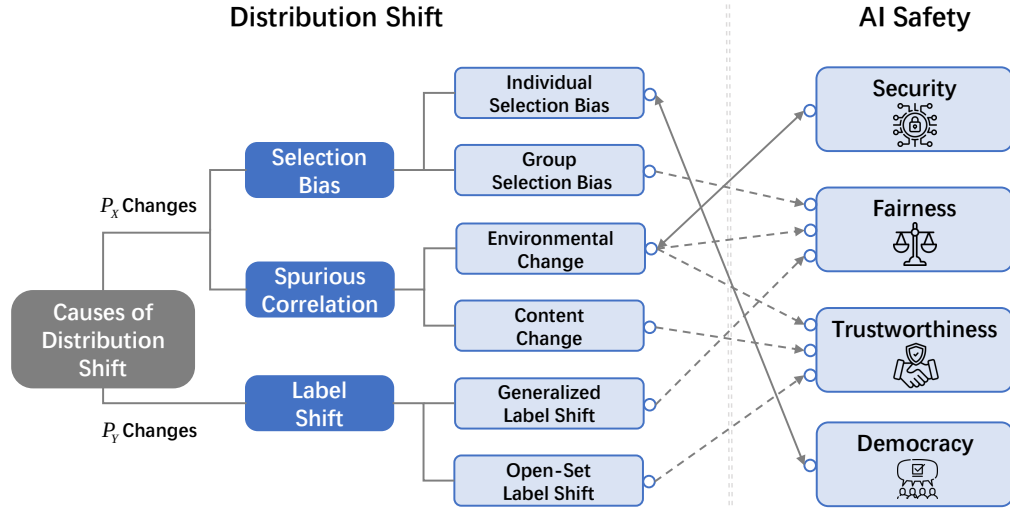


Fig. 1. Connections between distribution shift causes and AI safety issues. Dashed arrows: Methods addressing a specific cause of distribution shift can also achieve a particular safety goal. Bidirectional solid arrows: Due to inherent consistency between a certain distribution shift cause and a corresponding safety issue, the two can be reduced from one another and methods derived for them can be mutually adapted.

Our analysis exhibits three characteristics that distinguish it from previous work:

- **Comprehensive:** We cover a diverse range of aspects of distribution shifts and AI safety, providing a more comprehensive discussion of the relationship between these two fields.
- **One-to-one aligned:** We directly align each specific distribution shift cause with relevant concrete AI safety issues to enable a clear, one-to-one comparison.

- **Mathematically grounded:** The connections we identify are rooted in rigorous mathematical relationships, which guide our subsequent review of the literature.

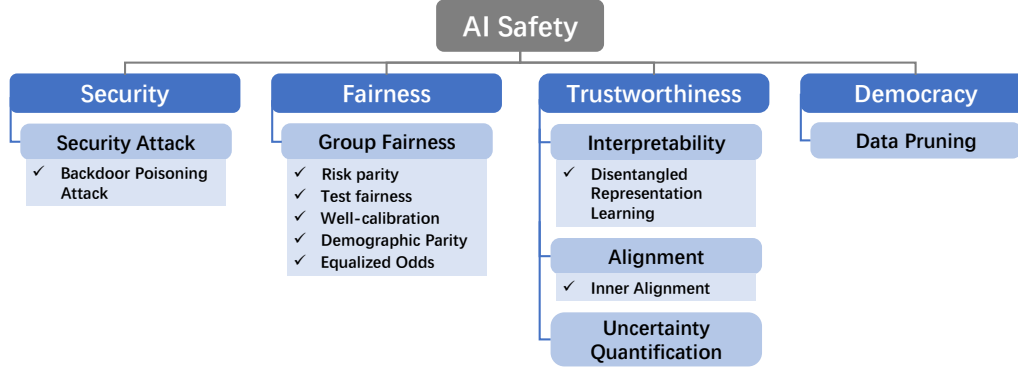


Fig. 2. Summary of safety issues related to distribution shift covered in this paper.

Building on these features, our work reveals two key relationships between distribution shifts and AI safety. First, methods addressing a specific cause of distribution shift can achieve a concrete AI safety goal (indicated by dashed arrows in Figure 1). Second, methods designed for a specific distribution shift cause and a concrete AI safety issue can be mutually adapted because of their definitional consistency (indicated by bidirectional solid arrows in Figure 1). These two types of connections highlight shared research interests and create opportunities for methodological exchange across distribution shift and AI safety. Our findings establish a more unified perspective on distribution shifts and AI safety, fostering closer collaboration between these two research communities.

The remainder of this paper is organized as follows. Section 2 introduces the notations used throughout the paper and outlines the specific topics related to distribution shift and AI safety. In Sections 3, 4, and 5, we examine how the specific causes of distribution shift within main categories—selection bias, spurious correlation, and label shift—are connected to various concrete AI safety issues. Finally, Section 6 concludes the paper with suggestions for future research directions.

2 Background

This section introduces the notations, categorizations, and logical structure that underpin our systematic analysis of the connections between distribution shift and AI safety. In Section 2.1, we present the necessary notations with a brief introduction to distribution shift. In Section 2.2, we offer an overview of the main categories under both distribution shift and AI safety. Then, we summarize the definition numbering for specific distribution shift causes and AI safety issues, along with the section references where they are discussed in detail.

2.1 Notations

Since this paper is organized from the perspective of distribution shift, we begin by introducing common notations for distribution shift. Unless otherwise specified, we consider the task of predicting the label Y from the input variable X , and we denote the ranges of random variables with calligraphic letters. The prediction model is denoted by f . When necessary, we view f as a composition of a feature extractor Φ and a prediction head (also called a classifier) ω , i.e., $f = \omega \circ \Phi$.

Table 1. Common notations for distribution shift.

Notation	Meaning
X	Input variable
Y	Label
E	Domain/Environment variable
G	Group variable
$f(\cdot)$	Prediction model
$\omega(\cdot)$	Classifier/Prediction head
$\Phi(\cdot)$	Feature extractor
$l(\cdot, \cdot)$	Loss function
$\mathcal{R}_d(\cdot)$	Population risk under the distribution specified by d
$\hat{\mathcal{R}}_d(\cdot)$	Empirical risk under the distribution specified by d

Within the context of distribution shift, we introduce a domain (also known as environment¹) variable E indexing any possible domain. It is typically assumed that the training data are drawn from a source domain e_s with joint distribution $P(X, Y | E = e_s)$, while the test data are drawn from a target domain e_t with joint distribution $P(X, Y | E = e_t)$. Under this formulation, the distribution shift between the training and test domains can be expressed as

$$P(X, Y | E = e_s) \neq P(X, Y | E = e_t).$$

Due to this distribution shift, the model learned by minimizing the population risk in the training domain $\mathcal{R}_{e_s}(f) = \mathbb{E} [l(f(X), Y) | E = e_s]$, may suffer significant performance degradation when evaluated on the test data, because it is not optimized for the test domain's risk $\mathcal{R}_{e_t}(f)$. Since the true population distribution is unavailable in practice, we often learn the model f via empirical risk minimization (ERM), i.e., we approximate $\mathcal{R}_{e_s}(f)$ with the empirical risk $\hat{\mathcal{R}}_{e_s}(f) = \frac{1}{n} \sum_{i=1}^n l(f(x_{e_s}^{(i)}), y_{e_s}^{(i)})$, where each $(x_{e_s}^{(i)}, y_{e_s}^{(i)})$ is independently and identically distributed sampled from $P(X, Y | E = e_s)$.

Sometimes, there are multiple training domains with different joint distributions. In this case, distribution shift means that the test domain's distribution differs from that of any training domain. Moreover, if a single environment or domain is considered as a mixture of different groups, we use the variable G to denote the group label. Table 1 lists the common notations used for distribution shift in this paper along with their meanings. More specific definitions will be provided in the corresponding sections.

2.2 Topic Overview

Distribution shift can arise from a wide range of factors. In this paper, according to the underlying causes that lead to distribution shift, we classify it into three main categories: selection bias, spurious correlation, and label shift. In general, selection bias refers to the bias induced by the selection process of input X in the training domain [151, 164]. Spurious correlation refers to the non-causal correlation present in the training data that changes in the test data [166, 200]. Label shift focuses on the shift in the distribution of label Y between the training and test domains [44, 100, 172]. Within each main category, we further consider the specific distribution shift causes. Table 2 summarizes the specific causes of distribution shift with their definition numbering and corresponding section references. Figure 3 illustrates the relationships among these causes.

¹In this paper, domain and environment are used interchangeably. We will use the term that fits the specific context or convention.

Table 2. Causes of distribution shift with definition and section references.

	Cause	Definition	Section
Selection Bias	Individual Selection Bias	Definition 1	Section 3.1
	Group Selection Bias	Definition 3	Section 3.2
Spurious Correlation	Environmental Change	Definition 5	Section 4.1
	Content Change	Definition 14	Section 4.2
Label Shift	Generalized Label Shift	Definition 15	Section 5.1
	Open-set Label Shift	Definition 16	Section 5.2

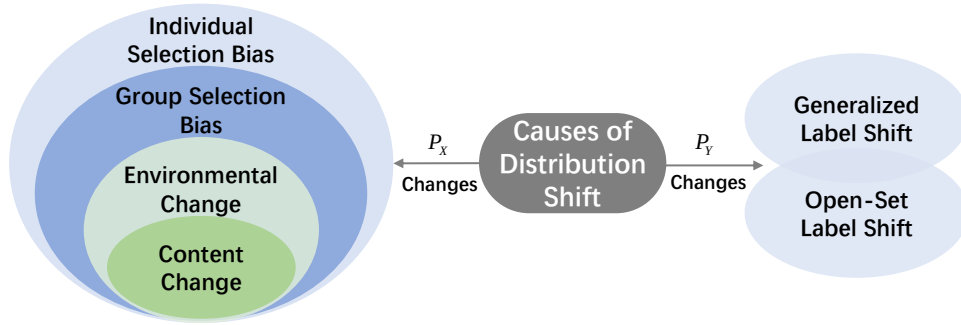


Fig. 3. Illustration of the inclusion relationships among different causes of distribution shift.

In each section corresponding to a specific cause in Table 2, we examine relevant AI safety issues based on the mathematical relationships among the definitions. The concrete AI safety issues under consideration cover four main categories: security, fairness, trustworthiness, and democracy. AI security refers to using AI to protect systems and data from cyber-threats [152]. AI fairness ensures that models do not exhibit prejudice against individuals or groups [221]. AI trustworthiness requires that AI systems meet user expectations in a verifiable manner [79]. AI democracy involves the equitable spread of AI applications to enable more people to access and benefit from the technology [49]. Specifically, Table 3 outlines the concrete safety issues within each of these four main categories, along with their definition numbering and corresponding section references.

3 Selection Bias

Selection bias occurs when the process of selecting training data leads to a dataset that does not accurately represent the target population. This misrepresentation creates a discrepancy between the training and test distributions, often resulting in a drop in performance during deployment [105, 106, 163]. For instance, consider an image classifier designed to recognize everyday objects like chairs and bikes. If the model is trained on real-world photographs but later evaluated on paintings of these objects, it may fail to generalize effectively. While humans can easily recognize objects despite stylistic changes, the classifier struggles with the distribution shift, leading to lower accuracy on the test set.

In this example, using training and test data with different styles illustrates a selection bias that causes a distribution shift. In the following subsections, we will categorize selection bias into two types, namely individual selection bias (Section 3.1) and group selection bias (Section 3.2). For each type, we will first provide the definition. Then, we compare a relevant AI safety issue with the specific type selection bias. By highlighting the similarities between the causes of

Table 3. AI safety topics with definition and section references.

	Topic	Definition	Section
Security	Backdoor Poisoning Attack	Definition 13	Section 4.1.3
	Risk Parity	Definition 4	Section 3.2.1
Fairness	Test Fairness	Definition 7	Section 4.1.1
	Well-Calibration	Definition 8	Section 4.1.1
	Demographic Parity	Definition 9	Section 4.1.1
	Equalized Odds	Definition 10	Section 5.1.1
Trustworthiness	Disentangled Representation Learning	Definition 11	Section 4.1.2
	Inner Alignment	Definition 12	Section 4.2.1
	Uncertainty Quantification	Definition 17	Section 4.1.2
Democracy	Data Pruning	Definition 2	Section 5.2.1

distribution shift and these AI safety issues, we draw connections between the methodologies in the two research areas and discuss their potential applicability in both areas.

3.1 Individual Selection Bias

One type of selection bias is the individual selection bias. We use the word “individual” to emphasize its difference from group selection bias. Individual selection bias often appears for reasons such as variations in dataset recollection [143], evolving clinical conditions [149], or shifts in image styles [46]. Additionally, researchers can simulate individual selection bias through controlled modifications like image coloring or corruption [123], distortion [139], and patch operations [140]. In all these cases that we consider as individual selection bias, the selection bias is applied individually to each data point. In contrast, in group selection bias, the selection bias is applied according to the group.

We follow a similar definition for individual selection bias as in Shimodaira [164]:

DEFINITION 1 (INDIVIDUAL SELECTION BIAS). *Consider the prediction task of inferring Y from X under the source domain e_s and the target domain e_t . Individual selection bias causes a distribution shift where*

$$P(X | E = e_s) \neq P(X | E = e_t),$$

$$P(Y | X, E = e_s) = P(Y | X, E = e_t).$$

Figure 4(a) illustrates individual selection bias. When individual selection bias exists, input distributions of the source (blue curve) and target (green curve) domains differ.

Addressing individual selection bias is essential to mitigate distribution shifts that lead to performance drops during model testing. Methods designed to address this selection bias not only improve model performance but also intersect with broader AI safety issues, particularly the principles of AI democracy. In the following discussion, we explore the connection between methods for addressing individual selection bias and strategies for promoting AI democracy.

3.1.1 Democracy. In this paper, AI democracy denotes the principle of broadening access to AI research [50]. With the rapid growth in the sizes of training datasets and models, researchers with limited computational resources often find themselves unable to develop or analyze state-of-the-art systems. This barrier not only restricts innovation but also limits the comprehensive evaluation of AI safety, as fewer experts are able to investigate the safety implications of these systems. Thus, AI democracy improves AI safety by allowing a diverse range of researchers to scrutinize AI systems.

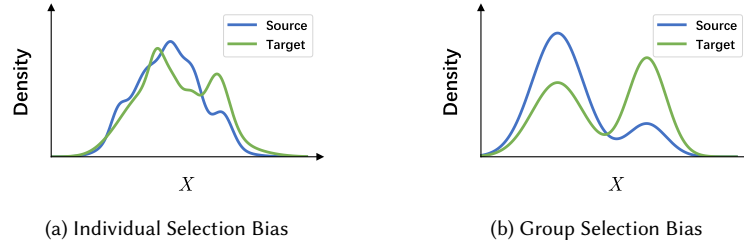


Fig. 4. An illustration of individual selection bias and group selection bias. The blue curves represent the input distribution of the source domain, and the green curves represent the input distribution of the target domain. For individual selection bias, the two distributions are known to differ. However, it is hard to attribute this difference to a certain subpopulation in the data. In contrast, for group selection bias, apparent changes in different subpopulations causes the distribution shift.

Data pruning is a promising method to democratize model training. By carefully selecting a subset of the full dataset, this method maintains a balance between the complexity and representativeness of examples, ensuring that models trained on the pruned data can perform on par with those trained on the entire dataset. Moreover, the reduced dataset size lowers computational demands, which enables researchers with limited resources to train and evaluate advanced models [173].

Data pruning can be seen as a dual problem of individual selection bias. To better clarify this idea, we first formally define the data pruning task as follows [173]:

DEFINITION 2 (DATA PRUNING). *Data pruning describes the following task. We are given a full training set D_{train} sampled from the original data distribution $P_{original}(X, Y)$. Data pruning is to design a (stochastic) selection strategy $S : x \rightarrow \{0, 1\}, x \sim P_{original}(X)$ with the following goal. Now the selected dataset $D_{selected} = \{x \in D_{original} \mid S(x) = 1\}$ can be viewed as sampled from another distribution $P_{selected}(X, Y)$ that depends on the pruning strategy S and the original distribution $P_{original}(X, Y)$. By training a model f on the subset $D_{selected}$, the goal is to enhance the performance of f on $P_{original}(X, Y)$.*

From this definition, there is a natural correspondence between addressing individual selection bias and data pruning, if we interpret $P_{original}(X, Y)$ as the target distribution and $P_{selected}(X, Y)$ as the source distribution. As illustrated in Figure 5, techniques for addressing individual selection bias focus on refining the training process so that a model trained on source data performs well on target data. In contrast, data pruning methods select a representative subset from the full dataset, ensuring that a model trained on this subset achieves performance comparable to one trained on the entire dataset. In both cases, the objective is to bridge the

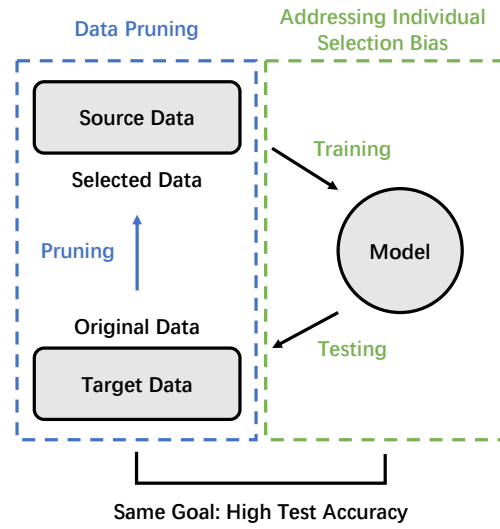


Fig. 5. Data pruning is a dual problem of addressing individual selection bias. To address individual selection bias, methods focus on refining the training process so that a model trained on source data performs well on target data. As a comparison, data pruning methods prune less informative data from the original dataset, so that a model trained on the remaining subset still performs well. Both problems share the same goal of improving test accuracy under distribution differences.

performance gap caused by differing data distributions. Therefore, data pruning can be viewed as the dual problem to addressing individual selection bias.

Below, we compare techniques for addressing individual selection bias with those used in data pruning. Owing to the duality in these problems and the similarities in their underlying approaches, we explore how methods developed for one research direction can be adapted and applied interchangeably to the other.

To improve performance under individual selection bias, methods generally assign different weights to examples in the training set. Importance reweighting [37, 164] is a category of methods that mitigate the effect of individual selection bias. The basic importance reweighting method uses weights derived from the distributions $P(X, Y \mid E = e_s)$ and $P(X, Y \mid E = e_t)$. Instead of minimizing the empirical risk $\hat{\mathcal{R}}_{e_s}(f) = \frac{1}{n} \sum_{i=1}^n l(f(x_{e_s}^{(i)}), y_{e_s}^{(i)})$, with importance reweighting, we minimize the weighted risk

$$\frac{1}{n} \sum_{i=1}^n \frac{P(X = x_{e_s}^{(i)}, Y = y_{e_s}^{(i)} \mid E = e_t)}{P(X = x_{e_s}^{(i)}, Y = y_{e_s}^{(i)} \mid E = e_s)} l(f(x_{e_s}^{(i)}), y_{e_s}^{(i)}),$$

which is an unbiased estimator of the target risk $\mathcal{R}_{e_t}(f)$. Intuitively, when the probability density around a data point is larger in the target domain than in the source domain, the weight of the data point is increased, so that the model is encouraged to minimize the loss regarding this data point.

When dealing with data from complex modalities, deriving an exact analytical form of the distributions is often unfeasible. In such cases, importance reweighting methods instead use weights from other sources. For example, the Learning-from-Failure (LfF) method [123] trains two models simultaneously. The first model, optimized with a generalized cross entropy loss, primarily learns from easier examples, thus performing poorly on more challenging ones. These hard examples are then assigned higher weights during the training of the second model. Similarly, the Just-Train-Twice (JTT) approach [101] divides the training process into two stages. In the first stage, a low-complexity model is trained using empirical risk minimization (ERM), which typically misclassifies challenging inputs. In the second stage, these misclassified examples are upweighted to train the final, more robust classifier.

Other than importance reweighting, many methods to address individual selection bias fall under the framework of distributionally robust optimization (DRO) [11]. With DRO, we minimize the risk

$$\sup_{q \in \mathcal{Q}} \hat{\mathcal{R}}_q(f) = \sup_{q \in \mathcal{Q}} \frac{1}{n} \sum_{i=1}^n l(f(x_q^{(i)}), y_q^{(i)}),$$

where $\mathcal{Q}$ is a set of distributions called the uncertainty set, and $(x_q^{(i)}, y_q^{(i)})$ are drawn i.i.d. from q . Depending on the choice of $\mathcal{Q}$, various DRO methods have been proposed, including Conditional-Value-at-Risk (CVaR) DRO [89], χ^2 -DRO [57], and their variants [205].

In data pruning, each training data point is assigned a score using certain metrics, and those data points falling within a selected score range are chosen to form a small yet representative subset. Researchers have proposed a variety of metrics for scoring, such as representations [115], probability vectors [10], predicted labels [117], confidence scores [25], variance [58], gradients [173], and Hessians [195]. Each of these metrics captures distinct aspects of the data's importance or complexity, enabling a selection that preserves the essential characteristics of the full dataset while greatly reducing the computational load.

Due to the dual nature of data pruning and handling individual selection bias, methods developed for one can often be repurposed for the other. For instance, the metrics used to score data points in data pruning could serve as weights in importance reweighting techniques for addressing individual selection bias. Conversely, weights obtained through

importance reweighting can be used in data pruning strategies. Furthermore, the way DRO-based methods optimize over an uncertainty set can also be regarded as a process of assigning sample weights, and therefore is closely related to data pruning approaches. However, the interchangeability remains under-explored in the existing literature. By establishing the connection between individual selection bias and data pruning, we highlight the potential that the two research directions benefit from each other.

3.2 Group Selection Bias

Group selection bias is a specific form of individual selection bias (see Figure 3). Unlike biases that occur at the individual level, group selection bias arises when the composition of groups differs between the source and target domains. Although both domains may contain the same groups, the relative proportions of these groups vary. This phenomenon has been observed in various contexts, such as racial groups [76] or different time periods [125]. We define group selection bias and the distribution shift it causes as follows:

DEFINITION 3 (GROUP SELECTION BIAS). *Consider the prediction task of inferring Y from X under the source domain e_s and the target domain e_t . Group selection bias causes a distribution shift where*

$$\exists U_g \in \mathcal{G} : P(X \in U_g \mid E = e_s) \neq P(X \in U_g \mid E = e_t),$$

$$P(Y \mid X, E = e_s) \neq P(Y \mid X, E = e_t),$$

where U_g denotes the range of X corresponding to group g , i.e., $U_g \subset X$ and $\bigcup_{g \in \mathcal{G}} U_g = X$. Here, $\mathcal{G}$ represents the range of group labels G .

Figure 4(b) illustrates group selection bias. When group selection bias exists, the most intuitive difference between the source and target domains lies in the varying group sizes across the two domains.

3.2.1 Fairness. Compared to the ideal case of perfect alignment between source and target distributions, group selection bias not only degrades overall model performance, but it also hurts group fairness. Figure 7 shows different types of group fairness. In this section, we primarily consider group fairness defined in terms of the population risk of each group [192], which we refer to as risk parity. More notions of group fairness that are related to other causes of distribution shift will be introduced in Sections 4.1.1 and 5.1.1.

DEFINITION 4 (RISK PARITY). *Let X denote the covariates and let Y denote the target variable. We say a predictive model f satisfies risk parity with respect to the loss function l and attribute R if, for all values r_i, r_j in the range of R , $\mathcal{R}_{r_i}(f) = \mathcal{R}_{r_j}(f)$, i.e.,*

$$\mathbb{E}[l(f(X), Y) \mid R = r_i] = \mathbb{E}[l(f(X), Y) \mid R = r_j].$$

According to Definition 4, risk parity is achieved when a model performs uniformly across all values of the protected attribute R . However, when group selection bias exists between training and testing domains, the model f tends to focus on the majority groups while neglecting minority ones, leading to unequal accuracies of the groups. Considering the group G that a data point belongs to as the protected attribute R , group selection bias naturally results in a model that does not satisfy risk parity.

Since group selection bias is a specific instance of individual selection bias, many mitigation techniques overlap between the two. The key difference lies in the availability of group labels for group selection bias. With these labels, we can explicitly enhance the influence of underrepresented groups during training. In the following sections, we illustrate how representative methods for addressing group selection bias help to achieve risk parity.

Importance reweighting [147] is a such a method, where the examples from each group are assigned a weight inversely proportional to the group size. Specifically, the objective of importance reweighting is to minimize the risk

$$\sum_{g \in \mathcal{G}} \hat{\mathcal{R}}_g(f) = \sum_{g \in \mathcal{G}} \frac{1}{n_g} \sum_{i=1}^{n_g} l(f(x_g^{(i)}), y_g^{(i)}),$$

where the examples $(x_g^{(i)}, y_g^{(i)})$ are drawn i.i.d from distribution $P(X, Y \mid G = g)$, and n_g is the number of examples drawn. By assigning a larger weight to the minority group due to its smaller size, importance reweighting encourages the model to update in a way that treats all groups equally, thus promoting risk parity. Similarly, applying subsampling techniques to the majority group [70] effectively reduces the classification loss of the minority group during training.

Another popular approach for addressing group selection bias is Group DRO [147]. In Group DRO, the model is trained towards minimizing the highest empirical risk across all groups, that is

$$\max_{g \in \mathcal{G}} \hat{\mathcal{R}}_g(f) = \max_{g \in \mathcal{G}} \frac{1}{n_g} \sum_{i=1}^{n_g} l(f(x_g^{(i)}), y_g^{(i)}).$$

Under this worst-case risk framework, the model is optimized to minimize the risk of the group with the highest risk, ultimately aiming for uniform risk across all groups. To further ensure that the model optimizes to reduce loss on the groups with larger generalization gaps, the adjusted group DRO approach [147] adds an extra loss term that is inversely proportional to the square root of the group’s size. In a similar vein, Krueger et al. [83] introduce two risk extrapolation (REx) methods, namely Minimax REx and Variance REx. These two methods can enhance model robustness under larger distribution shifts compared to traditional DRO. Minimax REx achieves this by adding an extra penalty to the group with the highest risk. Variance REx integrates the variance of group risks into the loss function, in order to penalize groups that deviate more from the average risk. Consequently, both REx methods effectively work toward equalizing risk across groups.

The underlying connection between methods for addressing group selection bias and the concept of risk parity is that risk parity fundamentally guides the design of these methods. Even though various approaches aim to balance risks across groups, empirical evidence shows that worst-group accuracy frequently falls below the average accuracy [147], suggesting that true risk parity is not fully realized at convergence. Therefore, there is ample room for future research in developing optimization procedures that can reliably achieve risk parity in practice.

4 Spurious Correlation

Different from selection bias which focuses on distribution shifts caused by the sample selection process of input X , some types of distribution shift are closely related to the emergence of spurious correlations among different features in the training data.

Spurious correlation is the statistical relationship where two or more variables are associated but not causally related [166]. In parallel with the rapid development of ML techniques in recent years, the sensitivity of ML models to spurious correlations has been frequently observed in various fields, such as computer vision [72, 136], natural language processing [179, 186], and reinforcement learning (RL) [30, 112].

Consider the classification of waterbirds and landbirds in images as an example. If almost all waterbirds appear with water backgrounds and landbirds with land backgrounds during training, the ML classifier might learn to predict bird types based on the image’s background due to inductive bias [198]. However, in the test environment, waterbirds may be found in land backgrounds, which could significantly reduce the model’s classification accuracy. Therefore, the

spurious correlation between the background Z and the label Y , i.e., their non-causal relationship, leads to inconsistency in the relationship between Z and Y across the training and testing environments. This inconsistency results in poor generalization performance of the ML model.

From the example above, we observe that the spurious correlation can generally be regarded as arising from group selection bias. This is because, from the training to test environments, there is a change in the distribution of groups formed by different kinds of birds and their corresponding backgrounds. Specifically, under group selection bias, there are two causes directly associated with the occurrence of spurious correlations, namely environmental change and content change (see Figure 3 for an illustration of the relationship among group selection bias, environmental change, and content change). In Sections 4.1 and 4.2, environmental change and content change will be introduced, respectively. We will discuss how environmental change and content change are closely connected to spurious correlations, and further demonstrate they are inherently and naturally connected to different aspects of AI safety, such as fairness, trustworthiness, and security.

4.1 Environmental Change

Distribution shifts occur when there are changes in the environment between training and testing. In tasks where the input variable X is used to predict the label Y , when there is an environmental change, some features of X experience a change in their relationship with Y while others maintain their relationship with Y . Specifically, environmental change is defined as follows:

DEFINITION 5 (ENVIRONMENTAL CHANGE). *Consider the prediction task of inferring Y from X under the environment variable E with possible values $\mathcal{E}$. Denote the Bayes error [39] for predicting Y from X in environment $e \in \mathcal{E}$ as $\epsilon_e(X)$, and define the set of feature extractors that (approximately) preserve the Bayes error in environment e as $\mathcal{F}_e = \{\Phi : \epsilon_e(\Phi(X)) \approx \epsilon_e(X)\}$. Then, environmental change with respect to environment variable E causes a type of distribution shift where there exist a content (invariant) feature $C(X) \in \bigcap_{e \in \mathcal{E}} \mathcal{F}_e$ and an environmental (spurious) feature $Z(X)$, such that*

$$\exists e_{i_0}, e_{j_0} \in \mathcal{E} : P(Y | Z(X), E = e_{i_0}) \neq P(Y | Z(X), E = e_{j_0}),$$

$$\forall e_i, e_j \in \mathcal{E} : P(Y | C(X), E = e_i) = P(Y | C(X), E = e_j).$$

In concise conditional independence notation, $Y \not\perp\!\!\!\perp E | Z(X)$ and $Y \perp\!\!\!\perp E | C(X)$.

From Definition 5, the inconsistent relationship between $Z(X)$ and Y indicates the presence of environment-sensitive spurious correlation between them. Many models that predict label Y from X employ an function Φ to extract features $\Phi(X)$, which are then used for prediction. If the feature extractor Φ captures information from environmental features $Z(X)$, the model will inevitably learn these spurious correlations and suffer performance degradation under environmental changes. Figure 6 specifically illustrates how environmental changes occur in the previously example of waterbird versus landbird classification, and how the spurious correlation under environmental change affects the model's generalization performance.

Various methods have been employed to address the challenge of environmental change [1, 5, 15, 80, 82, 185]. Notably, many of these methods are closely related to several key aspects in AI safety, such as fairness, trustworthiness, and security. We will systematically analyze how the inherent similarities between environmental change and AI safety principles help bridge their methodologies.

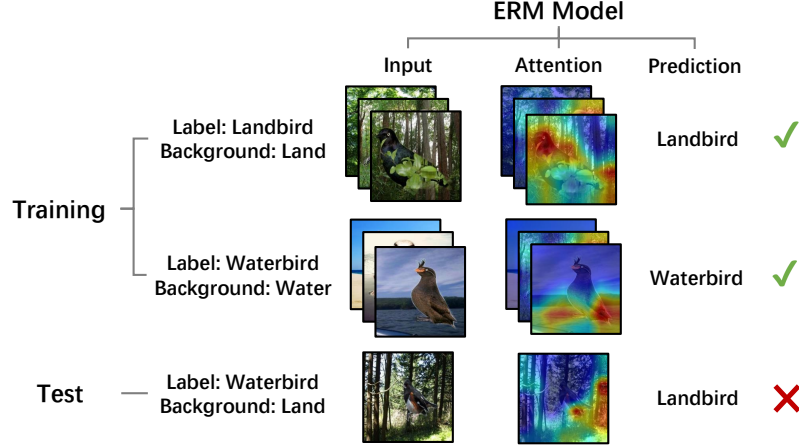


Fig. 6. Illustration of the negative impact of environmental change on a model trained via empirical risk minimization (ERM), using the example of waterbird and landbird classification. The first column shows sample images from the training and test environments. In the training environments, most waterbirds have water backgrounds, and most landbirds have land backgrounds. However, in the test environment, the background is not highly correlated with the bird category; for instance, a waterbird can have a land background. Thus, $Z(X)$ represents the background feature, while $C(X)$ represents the bird feature. As shown in the second column with the Grad-CAM visualization [153] of the ERM model, the feature extractor Φ may focus on backgrounds $Z(X)$ to predict Y . Consequently, in the third column, the model's predicted labels indicate that the test model may misclassify a waterbird with a land background as a landbird based on $Z(X)$.

4.1.1 Fairness. From the perspective of the independence among random variables, group fairness mainly includes sufficiency, independence, and separation [142]. Specifically, for a model that uses X to predict Y , if $\hat{Y}$ is the model's prediction and R is the protected attribute, then the definitions of sufficiency, independence, and separation are as follows:

DEFINITION 6 (SUFFICIENCY, INDEPENDENCE, AND SEPARATION). Let $\hat{Y}$ be a prediction derived from covariates X , let Y denote the true outcome, and let R denote the protected attribute. Then we have:

$$\text{Sufficiency: } Y \perp\!\!\!\perp R \mid \hat{Y}, \quad \text{Independence: } \hat{Y} \perp\!\!\!\perp R, \quad \text{Separation: } \hat{Y} \perp\!\!\!\perp R \mid Y.$$

In Sections 4.1.1 and 5.1.1, we will provide a detailed analysis of the relationships between different types of distribution shifts and these three notions of group fairness. For clarity, we first present a unified list of the specific fairness definitions involved. Among these definitions, test fairness (Definition 7; see also [24]) and well-calibration (Definition 8; see also [180]) belong to the group sufficiency notion, demographic parity (Definition 9; see also [118]) falls under group independence, and equalized odds (Definition 10; see also [56]) corresponds to group separation. Unlike the approaches surveyed by Shao et al. [156], which treat fairness with respect to a certain protected attribute as an additional constraint while attempting to maintain accuracy under distribution shift, we demonstrate that many methods for addressing distribution shift are inherently related to these fairness notions.

DEFINITION 7 (TEST FAIRNESS). Let $S = S(x)$ be the predictive score computed from the covariates $X = x$, and let $Y \in \{0, 1\}$ denote the true outcome indicator. Then, S is test-fair with respect to the attribute $R \in \{r_1, r_2\}$ if for all values s of S , $P(Y = 1 \mid S = s, R = r_1) = P(Y = 1 \mid S = s, R = r_2)$.

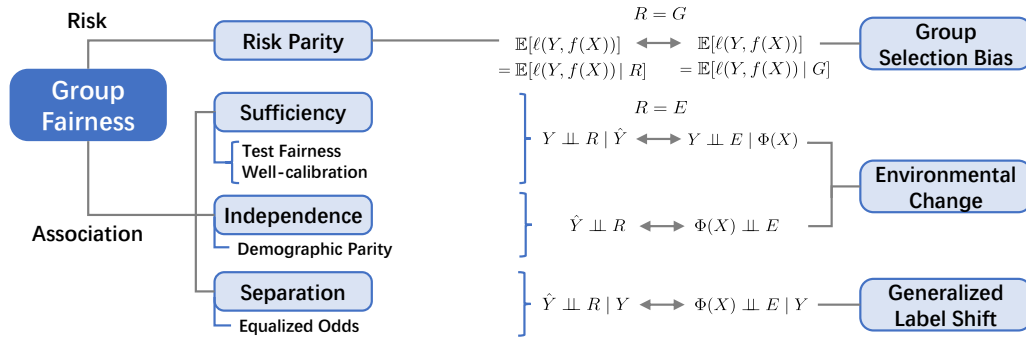


Fig. 7. Summary of group fairness and their relationships with different causes of distribution shift. For risk-based fairness notions, when considering the protected attribute R as groups G in distribution shift, risk parity (Definition 4) and group selection bias (Definition 3) are closely related. From the perspective of the association between random variables, if R is considered as the environmental variable E , then sufficiency and independence (Definition 6) are closely related to environmental change (Definition 5), while separation (Definition 6) and generalized label shift (Definition 15) are closely related. Detailed relationships between different fairness definitions and the causes of distribution shift will be analyzed in the corresponding sections (Sections 3.2.1, 4.1.1, and 5.1.1).

DEFINITION 8 (WELL-CALIBRATION). Let $S = S(x)$ be the predictive score computed from the covariates $X = x$, and let $Y \in \{0, 1\}$ denote the true outcome indicator. Then, S is well-calibrated with respect to the attribute $R \in \{r_1, r_2\}$ if for all values s of S , $P(Y = 1 \mid S = s, R = r_1) = P(Y = 1 \mid S = s, R = r_2) = s$.

DEFINITION 9 (DEMOGRAPHIC PARITY). Let $\hat{Y} \in \{0, 1\}$ be a prediction computed from the covariates X . Then, $\hat{Y}$ satisfies demographic parity with respect to the attribute $R \in \{r_1, r_2\}$ if $P(\hat{Y} = 1 \mid R = r_1) = P(\hat{Y} = 1 \mid R = r_2)$.

DEFINITION 10 (EQUALIZED ODDS). Let $\hat{Y} \in \{0, 1\}$ be a prediction computed from the covariates X , and let $Y \in \{0, 1\}$ denote the true outcome indicator. Then, $\hat{Y}$ satisfies equalized odds with respect to the attribute $R \in \{r_1, r_2\}$ if for every $c \in \{0, 1\}$, $P(\hat{Y} = 1 \mid Y = c, R = r_1) = P(\hat{Y} = 1 \mid Y = c, R = r_2)$.

As shown in Figure 7, in this section, our primary focus is on the connection between environmental change and the sufficiency and independence conditions. Specifically, when environmental changes occur, it is crucial for models to capture appropriate features $\Phi(X)$ of X for predicting Y . The Definition 5 suggests that models should avoid extracting environmental features $Z(X)$, and instead focus on capturing content features whose relationships with Y remain invariant. This underpins the core idea of invariant learning.

Invariant learning aims to discover features $\Phi(X)$ that satisfy

$$\forall e_i, e_j \in \mathcal{E} : P(Y \mid \Phi(X), E = e_i) = P(Y \mid \Phi(X), E = e_j),$$

where a classifier ω subsequently makes predictions based on $\Phi(X)$. The learned Φ avoids capturing spurious correlations that vary under environmental changes.

Furthermore, this definition of features $\Phi(X)$ is closely related to the notion of test fairness (Definition 7), which is a specific case of group sufficiency [107]. In this setting, if we consider the environment variable E as representing the protected attribute R , then the representation $\Phi(X)$ in invariant learning is formally consistent with the predictive score $S(X)$ in test fairness, and can thus be regarded as an extension of sufficiency at the feature level. Moreover, the actual predictive score $f(x) := (\omega \circ \Phi)(x)$ of the model satisfies the definition of test fairness when $\Phi(X)$ meets the

minimal sufficiency condition. This means that $\Phi(X)$ is a sufficient statistic for Y that is minimal in the sense that it contains no redundant or extraneous information.

Therefore, due to the mathematical connection between environmental change and test fairness, invariant learning techniques developed to address environmental change can be applied to achieve test fairness. Previous works implement or approximate invariant learning in different ways. While contributing to achieving test fairness, some of these methods are also related to other fairness notions.² We will introduce the methods related to invariant learning principles and provide detailed discussions on how they are connected with test fairness and other related fairness notions.

4.1.1.a Conditional Distribution Invariance. To pursue invariant learning, we define conditional distribution invariance as the objective of methods that directly aim to keep the conditional probability $P(Y | \Phi(X), E = e)$ unchanged across different e .

Invariant Optimal Classifier Learning. When predicting Y through $\Phi(X)$ in various environments, invariant optimal classifier learning requires the model to learn using the same optimal classifier ω in different environments. Notably, the invariance of ω has a close connection with learning $\Phi(X)$ such that $P(Y | \Phi(X), E = e)$ remains invariant across different environments e . Peters et al. [135] point out that when $\Phi(X)$ is a subset of X , invariant learning is equivalent to the existence of an invariant classifier ω across environments and an error distribution ϵ_e , such that

$$Y = \omega(\Phi(X), \epsilon_e), \quad \epsilon_e \perp\!\!\!\perp \Phi(X),$$

for X and Y in all environments $e \in \mathcal{E}$. Based on this equivalence, Invariant causal prediction (ICP) [135] estimates valid $\Phi(X)$ using causal inference under the assumption that ω is linear. Heinze-Deml et al. [61] further extend the method of ICP to nonlinear settings.

Another common approach for invariant optimal classifier learning is to approximate the probability distribution in invariant learning using expectations. That is, for all the possible value α ,

$$\forall e_i, e_j \in \mathcal{E} : \quad \mathbb{E}[Y | \Phi(X) = \alpha, E = e_i] = \mathbb{E}[Y | \Phi(X) = \alpha, E = e_j].$$

The representation $\Phi(X)$ is said to satisfy the environment invariance constraint (EIC) [26] if it meets the above condition. EIC is a weaker constraint for invariant learning and can itself be interpreted as a fairness notion under group sufficiency [26]. Moreover, when Y is binary, EIC is equivalent to invariant learning. Thus, it is also closely related to test fairness.

EIC is intimately connected to invariant optimal classifier learning because, for a given environment e and feature $\Phi(x) = \alpha$, the Bayes classifier $\omega := \omega_e^*$ that minimizes the population risk $\mathcal{R}_e(\omega \circ \Phi) = \mathbb{E}[\ell((\omega \circ \Phi)(X), Y) | E = e]$ can be expressed as

$$(\omega_e^* \circ \Phi)(x) = \mathbb{E}[Y | \Phi(X) = \alpha, E = e]$$

when the loss function ℓ is the mean squared error [5, 144]. This shows that achieving EIC is equivalent to identifying a invariant optimal classifier ω across all environments.

Among various algorithms that implement EIC, Invariant Risk Minimization (IRM) [5] is one of the most representative methods. Its core idea precisely lies in finding an invariant optimal classifier ω when using $\Phi(X)$ to predict Y across different environments. In the subsequent work of IRM, Ensemble Invariant Risk Minimization Games (EIRM) [3] attempt to find the Nash equilibrium of an ensemble game among different environments and prove that EIRM satisfies EIC under appropriate conditions. Teney et al. [175] minimize the risk across different environments while using the

²Throughout this paper, when a method is associated multiple fairness-related causes of distribution shift, we assign it to the section corresponding to the last cause of distribution shift.

variance of the classifiers ω under different environments as a regularizer, so that the model is encouraged to reach a stationary point that satisfies the IRM principle. In Regret Minimization [74], the IRM constraint is substituted with a predictive regret formulation, which can be minimized if there exists a single optimal classifier under all environments.

Furthermore, some works have extended IRM by incorporating additional constraints. Ahuja et al. [2] argue that EIC alone is not sufficient to guarantee good generalization performance, and they introduce the information bottleneck constraint to address the issue that $\Phi(X)$ might be fully informative about Y . Wald et al. [181] analyze the relationship between EIC and calibration of predictive functions. Building upon IRM, they require the model $f = \omega \circ \Phi$ be calibrated across all environments. That is, for every α in the range of f :

$$\forall e_i, e_j \in \mathcal{E} : \mathbb{E}[Y | f(X) = \alpha, E = e_i] = \mathbb{E}[Y | f(X) = \alpha, E = e_j] = \alpha.$$

This requirement is closely related to the fairness notion named well-calibration (Definition 8), which is an extension of test fairness. When spurious feature $Z(X)$ can be accessed,³ the model $f = \omega \circ \Phi$ in Spurious-Correlation-strata Invariant Learning with Label-balance [21] achieves EIC while additionally enforcing the spurious-free constraint (SFC):

$$P(f(X) | Z(X) = \alpha, Y = c) = P(f(X) | Z(X) = \alpha', Y = c)$$

for all class label c and α, α' in the support of $Z(X)$. SFC is equivalent to the definition of equalized odds (Definition 10) in fair ML if spurious feature is considered as a protected attribute.

Conditional Probability Matching. Different from methods that learn an invariant optimal classifier and focus on the EIC, some models directly estimate $\Phi(X)$ by matching the conditional distributions $P(Y | \Phi(X), E = e)$ across different environments e .

The Maximal Invariant Predictor (MIP) [82] aligns the conditional distributions $P(Y | \Phi(X), E = e)$ and $P(Y | \Phi(X))$ for each e by minimizing their Kullback-Leibler (KL) divergence, while simultaneously maximizing the Shannon mutual information between $\Phi(X)$ and Y . Heterogeneous Risk Minimization (HRM) [103] builds on MIP to address situations where environmental labels are unknown. Müller et al. [122] leverage conditional normalizing flows to identify an equivalent form of the conditional independence $Y \perp\!\!\!\perp E | \Phi(X)$. The Latent Causal Invariance Model (LaCIM) [170] directly employs a Variational Auto-Encoder to obtain invariant $P(Y | \Phi(X), E = e)$ across environments e , under the additional assumption of invariance in $P(X | \Phi(X), Z(X), E = e)$.

Furthermore, a series of methods use mutual information techniques to achieve conditional probability matching. For instance, Invariant Rationalization [17] reformulates the conditional independence $Y \perp\!\!\!\perp E | \Phi(X)$ via Shannon entropy: $H(Y | \Phi(X), E) = H(Y | \Phi(X))$, i.e., the mutual information $I(Y; E | \Phi(X))$ equals zero. At the same time, it maximizes the mutual information between $\Phi(X)$ and Y . Similarly, Li et al. [91] propose an invariant information bottleneck principle based on minimizing $I(Y; E | \Phi(X))$.

When explicit environmental labels are unavailable, but knowledge of spurious features $Z(X)$ under environmental changes is given, some models instead learn $\Phi(X)$ such that $P(Y | \Phi(X), Z(X) = \alpha)$ remains consistent across α .⁴ Puli et al. [138] propose finding a distribution where $Y \perp\!\!\!\perp Z(X)$. Then it learns $\Phi(X)$ by minimizing the mutual information between Y and $Z(X)$ given $\Phi(X)$, such that $Y \perp\!\!\!\perp Z(X) | \Phi(X)$, while requiring that $\Phi(X)$ has maximal

³Throughout this paper, “access to $Z(X)$ ” refers to either direct knowledge of $Z(X)$ [116, 126, 137, 138] or reference models capturing $Z(X)$ -related information [21, 210, 217]. We treat them uniformly since we focus on how the model uses $Z(X)$ (regardless of acquisition method) to analyze connections with ML fairness. Similarly, “access to environment E ” is treated in the same way.

⁴Unless otherwise specified, we define all notions of invariance with respect to the environment label E . Whenever E is unavailable but the spurious feature $Z(X)$ is known, we naturally refer to the same notion of invariance across different values of $Z(X)$.

mutual information with Y . Since annotations for $Z(X)$ are seldom available, Nuisances via Negativa [137] constructs semantic corruptions $T(X)$ of X and demonstrates that it serves well as a proxy for $Z(X)$.

4.1.1.b Marginal Distribution Invariance. Marginal distribution invariance refers to the objective of learning $\Phi(X)$ such that the marginal distribution $P(\Phi(X) | E = e)$ remains invariant across different environments e . Under the additional condition that the invariance of $P(Y | \Phi(X), E = e)$ also holds, the joint distribution

$$P(\Phi(X), Y | E = e) = P(Y | \Phi(X), E = e) P(\Phi(X) | E = e)$$

is also invariant across different environments e , which implies that minimizing the population risk in different environments yields the same solution ω . With the same prediction function ω , the invariance of $P(\Phi(X) | E = e)$ also implies that the model $f = \omega \circ \Phi$ satisfies the invariance of $P(f(X) | E = e)$. This is equivalent to the group independence fairness definition known as demographic parity (Definition 9).

To achieve invariant learning based on the invariant marginal distribution, various approaches directly assume or enforce the invariance of $P(Y | \Phi(X), E = e)$, and then further learn feature $\Phi(X)$ across different environments such that the discrepancy between them is minimized.

Specifically, Maximum Mean Discrepancy [14] (MMD) is widely used as a tool for minimizing the distance between different marginal distributions. Early methods such as Maximum Mean Discrepancy Embedding [128], Transfer Component Analysis [129], Domain Invariant Projection [8], Transfer Joint Matching [111], and Deep Adaptation Networks [110] use MMD to ensure that $P(\Phi(X), E = e)$ remains consistent across various environments. Zellinger et al. [202, 203] propose the Central Moment Discrepancy as a new distance metric between $P(\Phi(X) | E = e)$ across different environments. It outperforms and is more computationally efficient than MMD. Domain-Adversarial Neural Networks [41] aim to minimize the $\mathcal{H}$ -divergence between $\Phi(X)$ across different environments while ensuring good predictive performance in environments with known labels. Shen et al. [159] use the Wasserstein distance for minimizing the distance between the distributions of $\Phi(X)$ from different environments. Since the pairwise distances between $P(\Phi(X), Y | E = e)$ for different e can be difficult to optimize, Chevalley et al. [22] introduce a softer constraint that offers a better trade-off between invariance and predictability.

In addition to requiring the invariance of marginal distributions and assuming the invariance of $P(Y | \Phi(X), E = e)$, some methods directly maintain the invariance of both $P(Y | \Phi(X), E = e)$ and $P(\Phi(X) | E = e)$ across different environments e . Domain Invariant Component Analysis [121] minimizes the distance between the distributions of $\Phi(X)$ across environments and ensures an invariant conditional relationship of Y given $\Phi(X)$. Zhao et al. [215] achieve invariant marginal distributions through adversarial training, and incorporates entropy regularization to minimize the divergence between conditional distributions across different e . Nguyen et al. [124] simultaneously align these two distributions by leveraging Domain Density Transformation Functions. While enforcing the invariance of the marginal distribution, Katz et al. [77] introduce the cross-domain conditionals agreement error to quantify the discrepancy of $P(Y | \Phi(X), E = e)$ across different environments.

Similar to the case when achieving conditional distribution invariance, there may also be a lack of environmental labels E , but spurious features $Z(X)$ under environmental changes can be obtained. In this case, Makar et al. [116] first find a distribution under which $Y \perp\!\!\!\perp Z(X)$. Then, based on this distribution, Makar et al. [116] instead learn Φ satisfying $\Phi(X) \perp\!\!\!\perp Z(X)$ by minimizing the MMD of $P(\Phi(X) | Z(X))$ under different $Z(X)$. Zheng and Makar [217] extend the work of Makar et al. [116] to more general settings regarding the dimensionality and type of spurious features and labels.

4.1.2 Trustworthiness. To ensure the safety of AI systems, many guidelines have been proposed for building trustworthy AI [79, 90, 102]. According to International Organization for Standardization and International Electrotechnical Commission [71], a trustworthy AI should be able to meet user expectations in a verifiable way. In other words, AI trustworthiness requires that the model effectively encodes the user’s goals to satisfy their expectations, and that its decision-making process can be understood and verified. Consequently, both alignment and interpretability are essential for achieving AI trustworthiness. In this section, we examine the relationship between distribution shift and AI trustworthiness. Specifically, we discuss how interpretability and alignment are related to environmental change from a definitional perspective.

4.1.2.a Interpretability. With the rapid advancement of deep learning models, their increasingly complex architectures and the inherent black-box nature of neural networks make them difficult to understand. This has positioned interpretability as a crucial research topic within the AI safety communities [93]. Interpretability can be broadly defined as the ability to explain or present information in terms that are understandable to a human [34]. Interpretability paradigms primarily include behavioral interpretability, attributional interpretability, concept-based interpretability, and mechanistic interpretability [12]. The primary focus of this section is the interpretability of latent representations, which falls under mechanistic interpretability.

To extract latent features that are more interpretable to humans, various disentangled representation learning methods have been proposed. According to Wang et al. [187], disentangled representation can be defined as:

DEFINITION 11 (DISENTANGLED REPRESENTATION). *Disentangled representation should separate the distinct, independent and informative generative factors of variation in the data. Single latent variables are sensitive to changes in single underlying generative factors, while being relatively invariant to changes in other factors.*

Specifically, in tasks aimed at mitigating spurious correlation under environmental change, although there might not be strict constraints on the conditional probabilities Y with respect to different features of X , disentangled representation can aid in identifying and separating the environmental feature $Z(X)$ and content feature $C(X)$.

Many works have achieved this separation by disentangling the features of X based on their relationship with target labels Y and the environment E . Note that different works may refer to the environment E as domain and the label Y as class, identity, or category, depending on the specific context. For example, the Disentangled Semantic Representation model [15] employs gradient reversal layers to construct domain and label classifiers, which are used to partition the latent variables into semantic and domain components. Domain2Vec [133] achieves the disentanglement of domain-specific and category-specific features through Gram matrix. Peng et al. [132] decompose features into domain-invariant and domain-specific components while removing information from class-irrelevant features. MDTA-ITA [48] uses two distinct encoders to capture domain-specific and shared information across different domains e . Zhang et al. [207] separate semantic and variation factors of X through solving a constrained optimization problem. Chroma-VAE [196] mitigates spurious correlations by constructing shortcut-invariant and shortcut-encoded latent spaces. Wang et al. [185] separate domain-related and domain-unrelated semantic features, and then refine the semantic features with causality.

While all aforementioned methods incorporate autoencoder architectures and their variants, Disentangled Feature Augmentation [87] leverages feature-level data augmentation to embed X into intrinsic and bias attributes. Generalizable Person Re-Identification [213] extracts identity-specific and domain-specific features, and then employs backdoor adjustment to mitigate the influence of domain-specific features on identity-specific features.

In summary, these feature disentanglement methods that are investigated in this section typically partition the features of X into two components to represent environmental features $Z(X)$ and content features $C(X)$ separately. After isolating these components, by predicting Y based on $C(X)$, the models can effectively mitigate the impact of spurious correlations under environmental changes. In this manner, disentangled representation learning enables a more interpretable extraction and analysis of features. Moreover, we argue that, from another view of the cause of spurious correlation, the features of X can be disentangled in a more structured and granular manner. Further details will be provided in Section 4.2.1.

4.1.2.b Alignment. As society increasingly delegates decision-making to AI systems, a fundamental challenge surfaces: How can we ensure these technologies authentically reflect human ethics and societal norms? This question lies at the heart of AI alignment research, which seeks to ensure AI systems to behave in ways that are in accordance with human intentions and values [73], particularly in complex systems such as LLMs [160] and RL applications [78]. It is widely recognized that inner and outer alignment constitute two critical components of AI alignment research [158]. In general, outer alignment refers to the alignment of human values with the AI system’s training objectives, and inner alignment refers to the AI system’s actual achievement to these training objectives during operation [68]. In other words, outer alignment ensures that the training goals of AI systems are those we value as beneficial, while inner alignment ensures that the systems faithfully pursue these goals in practice.

To provide a more precise definition of inner and outer alignment, Hubinger et al. [68] introduce four relevant concepts. The base objective is defined as the initial set of goals or tasks that the AI system is programmed to optimize, and the base optimizer is the algorithmic framework or mechanism through which these goals are pursued. The mesa-objective is the objective discovered by the base optimizer that shows effective performance in the training environment. Consequently, at this point, the base optimizer turns into the mesa-optimizer to achieve the mesa-objective. With these concepts, Hubinger et al. [68] define inner alignment and outer alignment as follows (see Figure 8):

DEFINITION 12 (INNER AND OUTER ALIGNMENT). *Inner Alignment: The alignment between the base objective and the mesa-objective. Outer Alignment: The alignment between the base objective and the human values.*

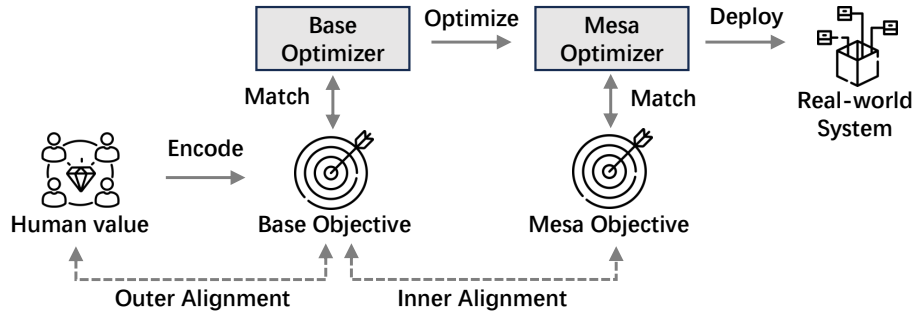


Fig. 8. The relationship between inner alignment and outer alignment.

Failure to achieve proper alignment in AI systems can lead to various detrimental outcomes, including behaviors that deviate from the principles of helpfulness [36], honesty [130], and harmlessness [55], which together constitute the HHH alignment framework [6]. Goal misgeneralization, which is one of the numerous potential forms of misalignment,

is considered to be caused by distribution shift [29, 154]. Specifically, we will focus on spurious correlations under environmental changes and reveal their close relationship with inner alignment.

Spurious correlation under environmental change can be viewed as one of the triggers for the failure of inner alignment. In Definition 5, consider the task of predicting Y from input X as the base objective. Then, the presence of a spurious correlation between the environmental feature $Z(X)$ and Y may lead the model to learn to predict Y using $Z(X)$. Predicting Y using the environmental feature $Z(X)$ is precisely the mesa-objective that the model learns. When an environmental change occurs, the inconsistency in the relationship between $Z(X)$ and Y leads to the misalignment between the base objective and the mesa-objective. Therefore, the approaches mentioned in Section 4, which are designed for environmental change and directly connected to specific AI safety topics, as well as any efforts aimed at mitigating spurious correlations caused by environmental changes [28, 80, 141, 191, 194, 199], can contribute to improving the model’s inner alignment.

Beyond the task of predicting Y , in the realms of RL and LLMs, spurious correlation also serves as a major cause of inner alignment failure. For example, in the context of RL, if the agent designed to assist with driving is trained in an environment where traffic jams only occur during the daytime, it might learn a spurious correlation between light intensity and traffic congestion [30]. When deployed in a new environment where traffic jams occur at night, due to the misalignment between the mesa-objective and the base objective, the agent may fail to effectively predict or respond to traffic conditions. In the case of LLMs, models primarily trained on multiple-choice safety questions may associate safe answers with a particular answer style. This spurious correlation can result in the models failing to generate safe outputs when facing open-ended queries [188]. Furthermore, with the emergence of advanced AI systems, deceptive alignment has been observed [67]. In deceptive alignment, AI models’ mesa-objectives are aware of the existence of the base objective and engage in deceptive behavior, pretending to align with the base objective to avoid being modified during training. However, once deployed in a new environment, the model will no longer follow the base objective and may exhibit undesirable behaviors.

Several works have started to adapt concepts from spurious correlation under environmental changes, for example invariant learning [146, 167, 219], to RL and LLMs. Moreover, research on deceptive alignment is still largely confined to high-level discussions and empirical identification [16, 67, 68, 160]. Although these topics do not align perfectly with the definition of environmental change in prediction tasks (see Definition 5), they likewise involve misalignment between the base and mesa-objectives caused by changes in the environment. Therefore, these inherent similarities suggest significant opportunities to apply methods addressing spurious correlation under environmental changes to RL and LLMs. Such methods hold promise for resolving deceptive alignment as well as other inner alignment challenges in these domains.

4.1.3 Security. While ML models have achieved remarkable success across various fields, they also face significant security challenges. ML attacks refer to malicious attempts to subvert ML systems through adversarial inputs [65, 131]. These attacks enable adversaries to compromise model integrity, availability, and privacy [13]. Specifically, backdoor attacks [53] constitute a critical focus within ML attack research, posing substantial risks to computer vision [216], natural language processing [197], and RL applications [184].

Backdoor attacks aim to manipulate a model’s behavior by introducing carefully crafted malicious patterns during training. These malicious patterns cause the model to exhibit unexpected behaviors when they are present during inference. While backdoors can be injected at various attack surfaces of the ML pipeline, including data, model, and code [42], this section primarily focuses on data poisoning-based backdoor attacks during the model training process.

Specifically, based on whether the attacker modifies the ground-truth labels of the poisoned samples, backdoor poisoning attacks can be categorized into dirty-label attack [20, 208] and clean-label attack [178, 204]. In dirty-label attacks, adversaries typically select samples from multiple classes, embed a backdoor trigger in them, and then relabel these samples with a target class designated by the attacker. In contrast, clean-label attacks maintain the original labels of the poisoned samples while still embedding the trigger (see Figure 9 for a simplified and intuitive illustration of dirty-label and clean-label attacks). In the context of prediction, we formalize backdoor poisoning attack as:

DEFINITION 13 (BACKDOOR POISONING ATTACK). ⁵ Let X and Y be random variables representing clean data with joint distribution $P(X, Y)$ and range $\mathcal{X}$ and $\mathcal{Y}$. Define a set of possible triggers $\mathcal{Z}$ (where $0 \in \mathcal{Z}$ denotes no trigger), and let Z be a random variable over $\mathcal{Z}$. Define a data transformation function $T : \mathcal{X} \times \mathcal{Z} \rightarrow \mathcal{X}$, where $T(x, 0) = x$ (no trigger inserted) and $T(x, z)$ modifies x by inserting the trigger $z \in \mathcal{Z} \setminus \{0\}$.

In a poisoned dataset, any sample can be written as $(T(x, z), y)$. With probability $p \in (0, 1)$, the sample comes from the clean distribution, i.e., $z = 0$ and (x, y) is sampled from $P(X, Y)$. With probability $1 - p$, the sample is poisoned, meaning $z \neq 0$ is sampled from a trigger distribution $P(Z | Z \neq 0)$ and (x, y) is sampled from a specified distribution: For clean-label attack, $P(X | Y = y_t) \times \mathbf{1}_{\{Y = y_t\}}$; For dirty-label attack, $P(X) \times \mathbf{1}_{\{Y = y_t\}}$, where y_t is the target class for backdoor poisoning attack. The goal of backdoor poisoning attack is to choose proper T , p , and $P(Z)$, such that after training a prediction model $f : \mathcal{X} \rightarrow \mathcal{Y}$ by minimizing a loss function on the training poisoned dataset, both $P((f \circ T)(X, Z) = Y | Z = 0)$ (Clean data accuracy) and $P((f \circ T)(X, Z) = y_t | Y \neq y_t, Z \neq 0)$ (Attack success rate) can be maximized during test time.

According to Definition 13, the objective of backdoor poisoning attacks is to achieve both accurate prediction on clean data and targeted misclassification on triggered data.

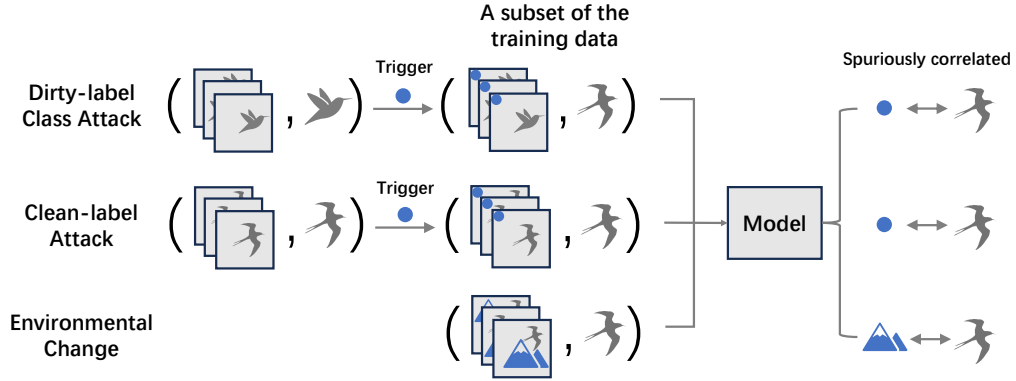


Fig. 9. The connection between backdoor poisoning attacks and environmental change. By injecting a trigger, both dirty-label attacks and clean-label attacks introduce a spurious correlation between the trigger and the target label. In environmental change, the spurious correlation between the environmental feature $Z(X)$ and Y can lead to poor model performance in the test environment. Similarly, the spurious correlation between the trigger and the label in a backdoor attack can cause the model to incorrectly predict Y during testing. Note that not all dirty-label and clean-label attacks strictly follow the illustrated case; this is a simplified general case for intuitive explanation.

It is worth noting that the attack succeeds by creating a spurious correlation between the trigger pattern Z and the target class y_t during training. Considering the poisoned dataset as drawn from the poisoned environment e_p , the

⁵Due to the diversity of actual implementations of backdoor poisoning attacks, a few attacks [148, 168] may not fully satisfy this definition; the one provided here represents the general case.

poisoning process defined in Definition 13 precisely establishes a strong relationship: $P(Y = y_t \mid Z \neq 0, E = e_p) = 1$. However, in the real environment e_r , regardless of whether a trigger Z appears or what the trigger is, it should not affect the label Y . Therefore, compared with the poisoned environment,

$$P(Y \mid Z, E = e_p) \neq P(Y \mid Z, E = e_r).$$

Interpreting the trigger Z as a feature of $T(X, Z)$, then according to Definition 5, Z is precisely a spurious feature of Y under environmental change. Figure 9 provides a more intuitive illustration of the connection between environmental change and backdoor poisoning attacks.

Despite the substantial similarities between backdoor poisoning attacks and spurious correlation under environmental change, these two areas largely remain separate domains of research, with only a few works exploring their connections. He et al. [60] explicitly posit a strong connection between backdoor poisoning attacks and spurious correlations. Causality-inspired Backdoor Defense [214] first trains a biased model to capture the spurious feature, then learns a clean model to learn features that are independent of the spurious feature. Front-door Adjustment for Backdoor Elimination [109] introduces front-door variables as intermediate variables to eliminate all spurious correlations. These methods view backdoor attacks as a form of spurious correlation, suggesting that their defense strategies can also be applied in settings involving environmental changes.

More generally, while each area has its unique characteristics, due to the intrinsic relationship between spurious correlations and backdoor poisoning attacks, we suggest that existing methodologies developed for addressing spurious correlations could be adapted to mitigate backdoor poisoning attacks, and vice versa.

To elaborate, there are two main types of existing defense strategies against backdoor poisoning attacks based on their goals: triggers in the dataset and backdoor in the model [99]. At the dataset level, the impact of triggers can be mitigated through various techniques in either training or inference stage, including input purification [33, 162] (i.e., neutralizing potential trigger patterns in the input data), input filtering [19, 176, 182] (i.e., filtering out suspected poisoned data), and input perturbation [98, 206] (i.e., applying transformations such as adding noise or geometric augmentations to the data). At the model level, depending on the different defense timings, approaches include training a clean model from a poisoned data set [66, 95, 177], and eliminating existing backdoors within the model architecture of a trained model [96, 104, 222].

Since few studies have approached backdoor poisoning attacks from the perspective of spurious correlations, the methodologies that are relevant to addressing spurious correlation under environmental change, such as learning invariant features $\Phi(X)$ and feature disentanglement as discussed in Sections 4.1.1 and 4.1.2, could be adapted and applied to backdoor poisoning attacks. Particularly, as triggers become more diverse and stealthy, spurious correlation-based methods may hold significant potential for more challenging applications.

Conversely, defense strategies against backdoor poisoning attacks could also be repurposed to address spurious correlations under environmental change. For instance, many backdoor detection methods operate on the premise that triggers exhibit distinctive patterns in model behavior and metrics, such as spectral signature [176], neuron activation [19], modification cost [182], training loss [95], etc. Consequently, it is common to identify poisoned samples based on the trigger’s characteristics [43, 189] or to construct the trigger in reverse [108, 157] for backdoor mitigation. Although triggers are maliciously introduced which makes them different from general spurious features, these methods suggest that addressing spurious correlations could benefit from focusing on the patterns and differences in models or certain features across different environments e . Moreover, since the permissions of defenders differ, defenses for backdoor poisoning attack may be deployed at various stages of the model lifecycle [7]. The corresponding diverse

strategies could also help mitigate spurious correlations across multiple stages, including data processing, model training, hyperparameter tuning, and inference.

4.2 Content Change

In Section 4.1, we posit the existence of content feature $C(X)$ whose relationship with the target label Y remains invariant across different environments E . Consequently, the goal of the approach described in Section 4.1 is to extract features that maintain a consistent relationship with Y for prediction. However, are these extracted features truly sufficient? In this section, we further examine the causes of spurious correlation from another perspective beyond environmental change.

DEFINITION 14 (CONTENT CHANGE). Consider the prediction task of inferring Y from X under the environment variable E with possible values $\mathcal{E}$. Content Change under environmental variable E causes a type of distribution shift in which there exist an environmental feature $Z(X)$ and a content feature $C(X)$ that satisfy the conditions in Definition 5, and if both $Z(X)$ and $C(X)$ can be further decomposed, then the pairwise correlations among their components exhibit variations across different environments. More formally, assume that $Z(X) = [Z_1(X), Z_2(X), \dots, Z_m(X)]$ and $C(X) = [C_1(X), C_2(X), \dots, C_k(X)]$. Then, for any two distinct components $\Phi_i(X)$ and $\Phi_j(X)$ with $\Phi(X) = [\Phi_1(X), \Phi_2(X), \dots, \Phi_{m+k}(X)] = [Z_1(X), \dots, Z_m(X), C_1(X), \dots, C_k(X)]$,

$$\exists e_{i_0}, e_{j_0} \in \mathcal{E} : \rho_{e_{i_0}}(\Phi_i(X), \Phi_j(X)) \neq \rho_{e_{j_0}}(\Phi_i(X), \Phi_j(X)).$$

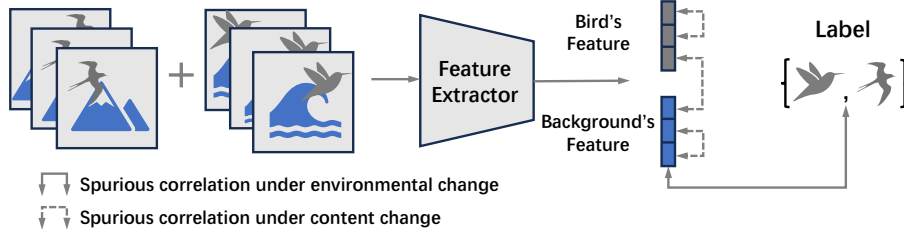


Fig. 10. Difference between environmental change (Definition 5) and content change (Definition 14) in the Waterbirds example. Environmental change focuses on the spurious correlation between the background feature $Z(X)$ (background) and the label Y (bird class), whereas content change emphasizes the cross-feature correlations among different features in the image X . Note that content change often naturally induces environmental change.

In Section 4, we introduced the Waterbirds example, where in classifying waterbirds and landbirds, most waterbirds have water backgrounds and most landbirds have land backgrounds. While environmental change leads to a spurious correlation between the background feature $Z(X)$ and the label Y , as shown in the Figures 3 and 10, the definition of content change further suggests focusing on the statistical association between different features of X .

Intuitively, Definition 14 indicates that when features are improperly extracted from X , different features may become mixed together for predicting the label Y , thereby leading to spurious correlations among different parts of the feature representation. Furthermore, content change implies that instead of only identifying features $C(X)$ that maintain invariant relationships with Y , the model can use more semantically meaningful and distinct representations when predicting Y . All these aspects are closely related to fine-grained disentangled representation learning as defined in Definition 11, and are therefore intimately connected to interpretability, which, as mentioned in Section 4.1.2, is a crucial component of AI trustworthiness.

4.2.1 Trustworthiness. In contrast to the interpretability-related discussion in Section 4.1.2 where disentangled representation learning techniques mainly separate features of X into two parts and rely on the content features $C(X)$ for prediction, we aim for a more fine-grained separation of meaningful features in the content change setting.

Montero et al. [119] and Dittadi et al. [31] empirically observe that proper feature disentanglement of X can improve the generalization performance of models under distribution shift. Furthermore, several feature decorrelation approaches achieve feature disentanglement by removing statistical dependencies between latent features. These methods implies that for the different components of $\Phi(X)$ in Definition 14, $\rho_e(\Phi_i(X), \Phi_j(X)) = 0$ in every training environment e . In doing so, the model can leverage diverse features for prediction while avoiding spurious correlations under content change. Specifically, while Shen et al. [161] and Kuang et al. [84] effectively address the linear cases, StableNet [211] and Region Aware Proposal reweighTing [212] extends feature decorrelation to deep learning frameworks using random fourier features. Different from pursuing statistical orthogonality, Project and Probe [18] enforces geometric orthogonality through a two-step strategy: first, it projects the pre-trained features of X onto orthogonal and diverse components; then, it uses these components for classifying Y .

Another line of work enables feature disentanglement through diverse model training in the same environment. They capture different features in $\Phi(X)$ by training distinct models f , and enforce dissimilarity among these models to encourage decorrelation between different features. Teney et al. [174] measure gradient alignment to encourage different classifiers ω to focus on distinct features while maintaining accuracy. Diversity-By-disAgreement Training [127] learns ensembles of diverse models f by minimizing their agreement in the target environment. Diversify and Disambiguate [88] first minimizes the mutual information between each pair of predictions, then selects the optimal model under limited supervision.

Sparsity is also recognized as beneficial for extracting features that capture the same semantic meaning across different environments e , thereby facilitating feature disentanglement. With sparse classifiers, models are more likely to learn independent features that are shared across environments, which in turn helps models avoid capturing spurious correlations under content change. As shown by Lachapelle et al. [85], the use of sparse classifiers contributes to effective feature disentanglement. Fumero et al. [40] further demonstrate that disentangled representations naturally emerge when features are sparsely activated across environments while being shared among them.

Furthermore, through appropriate disentanglement and a more structured utilization of the diverse features in $\Phi(X)$, models can achieve better generalization under content change. For instance, Zheng et al. [218], Rosenfeld et al. [145], Eastwood et al. [35] and Zhong et al. [220] demonstrate that beyond solely relying on features that exhibit a stable association with Y across different environments e for prediction, it is feasible to incorporate additional features from X to enhance predictive accuracy in test environments. This observation indicates that the distinction between $C(X)$ and $Z(X)$ is not binary; the appropriate components within $Z(X)$ can also assist in predicting Y more effectively. More specifically, it highlights an inherent trade-off between estimation and approximation error: a more granular disentanglement and inclusion of features may increase estimation error, yet simultaneously reduce approximation error. Therefore, reasonable utilization of additional features could lead to improved generalization in novel environments.

5 Label Shift

Label shift happens when the distribution of labels Y changes between the source and target domains. As illustrated in Figure 11, this phenomenon arises from variations in the label distribution rather than changes in the input features X . For example, consider a binary classifier trained to predict whether an individual is negative (y_1) or positive (y_2) for pneumonia. If the classifier is trained on data from a time period when pneumonia is rare (source domain) but then

tested later when the disease becomes more prevalent (target domain), the resulting shift in the label distribution will lead to decreased accuracy [100].

In the following sections, we focus on two specific types of label shift, namely generalized label shift and open-set label shift. These two types of label shift are closely connected to the AI safety topics of fairness and trustworthiness.

5.1 Generalized Label Shift

When label shift occurs, it is commonly assumed that $P(X | Y, E = e)$ remains invariant across different environments e . However, this assumption does not always hold in real-world scenarios. Therefore, generalized label shift (GLS) [172] relaxes the conditions of label shift:

DEFINITION 15 (GENERALIZED LABEL SHIFT). *Consider the prediction task of inferring Y from X under the environment variable E with possible values $\mathcal{E}$. Generalized label shift indicates a type of distribution shift where there exist a feature $\Phi(X)$ such that*

$$\forall e_i, e_j \in \mathcal{E} : P(\Phi(X) | Y, E = e_i) = P(\Phi(X) | Y, E = e_j).$$

In other words, GLS instructs the model to extract feature $\Phi(X)$ such that, for each given class label, the distribution of $\Phi(X)$ remains invariant across environments. Noted that we adopt Definition 15 to guide the identification of methods addressing GLS. As shown in Figure 3, GLS and the causes of distribution shift discussed in Sections 3 and 4 focus on two different directions, namely, Y and X , respectively. Consequently, works related to GLS may also appear in settings such as selection bias or spurious correlation.

5.1.1 Fairness. When using the same classifier ω , GLS also ensures that $P(f(X) | Y, E = e)$ remains invariant across different environments e , where $f = \omega \circ \Phi$. As illustrated in the Figure 7, this is equivalent to the predicted output $\hat{Y}$ satisfying equalized odds (Definition 10) with respect to the environment E , which is a fairness notion of group separation (Definition 6).

Conditional Invariant Components [51] achieves matching the class-conditional distributions $P(\Phi(X) | Y, E = e)$ across different environments e under mild conditions. Tachet des Combes et al. [172] transform the conditions of GLS to enable direct application to existing domain adaptation algorithms. Shui et al. [165] introduce the Wasserstein Aggregation Domain Network to minimize the differences between class-conditional distributions across environments using Wasserstein distance. Mahajan et al. [114] point out the insufficiency of the invariance of class-conditional distribution, and instead learns $\Phi(X)$ such that the distance between the representations of the same object across different environments is minimized. Wu et al. [193] combine the invariance of $P(\Phi(X) | Y, E = e)$ with importance reweighting for label shift correction. Unlike these methods, Predictive Group Invariance [1] directly matches the class-conditional distribution of the model output $f(X) = (\omega \circ \Phi)(X)$ by minimizing the KL divergence of $\mathbb{E}[P(f(X) | \Phi(X), Y = c, E = e)]$ across different environments e .

Recent advances in contrastive learning techniques have also been used to help enhance the similarity of $P(\Phi(X) | Y, E = e)$ across different environments e . The central idea behind contrastive learning is to embed features of similar samples closer to each other while pushing features of dissimilar samples further apart [86]. Specifically, Motiian et al. [120] both require the feature distributions of the same labels to be similar across different environments e , and encourages the feature distributions of different labels across different environments to be as separated as possible. Similarly, Contrastive Adaptation Network [75] proposes Contrastive Domain Discrepancy which helps reduce the divergence between $P(\Phi(X) | Y, E = e)$ across source and target environments. To enhance robustness of foundation

models (FMs) for zero-shot classification, Zhang and Ré [209] infer different environments based on the distance between X after FM embedding, and constructs a contrastive adapter that pushes together $\Phi(X)$ with the same label Y but from different e , while separating $\Phi(X)$ with the same environment e but different labels Y . In GLS settings, we may sometimes lack environmental labels and only have access to spurious features $Z(X)$ as mentioned in Section 4. In this case, the method Correct-n-Contrast [210], encourages closeness between $\Phi(X)$ with the same label Y but different spurious features $Z(X)$, and pulls apart $\Phi(X)$ for samples with the same spurious features $Z(X)$ but differs in label Y .

Furthermore, some algorithms not only learn features $\Phi(X)$ that satisfy equalized odds under GLS, but also impose additional fairness-related constraints that have been previously mentioned. Regarding EIC (see conditional distribution invariance in Section 4.1.1), Guo et al. [54] combine IRM with the invariance of $P(\Phi(X) \mid Y, E = e)$ to address the issue of IRM yielding undesired solutions under strong triangle spuriousness. Moreover, a number of works require that both $P(\Phi(X) \mid Y, E = e)$ and $P(\Phi(X) \mid E = e)$ remain invariant across different environments. This additional condition aligns with the Marginal Distribution Invariance discussed in Section 4.1.1. Specifically, in Scatter Component Analysis [47] and Domain Invariant Class Discriminative learning [92], the authors propose methods to use MMD in different ways to align both the class-conditional distribution $P(\Phi(X) \mid Y, E = e)$ and the marginal distribution $P(\Phi(X), E = e)$ across different e . The Conditional Invariant Adversarial Networks [97] and Conditional Invariant Domain Generalization [94] leverage adversarial networks and kernel mean embedding, respectively, to ensure that both the class-conditional distribution and the normalized marginal distributions remain unchanged across environments. Alleviating Semantic-level Shift [190] similarly encourages the model to achieve alignment of both $P(\Phi(X) \mid Y, E = e)$ and $P(\Phi(X), E = e)$ through adversarial learning and the construction of semantic-level feature vectors.

5.2 Open-Set Label Shift

Open-set label shift (OSLS) [44] describes a case where certain classes present in the target domain are entirely absent from the source domain e_s . For example, consider an image classifier trained only on images of dogs and cats. If this classifier encounters images of monkeys in the test set, it will be forced to incorrectly classify them as either a dog or a cat because the monkey class was absent during training. Such misclassification is problematic and should be avoided. Formally, OSLS is defined as follows [44]:

DEFINITION 16 (OPEN-SET LABEL SHIFT). *Consider the prediction task of inferring Y from X under the source domain e_s and the target domain e_t . OSLS indicates a type of distribution shift in which there exists $y \in \mathcal{Y}$ such that*

$$P(Y = y \mid E = e_s) = 0 \quad \text{and} \quad P(Y = y \mid E = e_t) > 0.$$

Furthermore, for any $y \in \mathcal{Y}$ with $P(Y = y \mid E = e_s) \neq 0$, it is required that

$$P(X = x \mid Y = y, E = e_s) = P(X = x \mid Y = y, E = e_t).$$

Figure 11 illustrates OSLS. Even though the model is trained only on seen classes from the source domain, we aim to use its outputs to detect and correctly handle unseen classes. Approaches to tackling OSLS are closely tied to uncertainty quantification, as measuring prediction uncertainty can help identify inputs that likely belong to classes absent from the training set.

5.2.1 Trustworthiness. Beyond interpretability and alignment (see Sections 4.1.2 and 4.2.1), a model's trustworthiness can also be enhanced through well-quantified uncertainty. In the ML pipeline, uncertainty can be categorized into different types, including aleatoric and epistemic uncertainty [69]. Here, we focus on the uncertainty in the model's

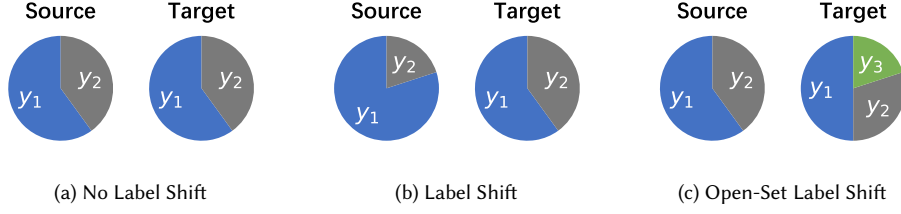


Fig. 11. An illustration of label shift and open-set label shift (OSLS). Label shift is the general setting where the proportions of labels differ between the source and target domains. OSLS is the setting where a new class (shown in green) appears in the target domain.

predictions, which can arise from various factors. For instance, consider an autonomous driving system that uses an image classification model to identify objects around the vehicle [23]. The system relies on these predictions to make critical decisions, such as yielding to pedestrians. However, in cases where an ambiguous object-like a realistic depiction of a human on an advertising board—appears, the model should express uncertainty rather than overconfidence. Such calibrated predictive uncertainty improves the reliability and safety of the system.

It is crucial that this uncertainty is quantified, as the model’s prediction is further used to guide the system’s decision. Various definitions of uncertainty quantification have been proposed [59]. In our paper, we define uncertainty quantification at a high level as follows [63]:

DEFINITION 17 (UNCERTAINTY QUANTIFICATION). *In a classification problem, for each input x , uncertainty quantification derives a score $u \in \mathbb{R}$ from the outputs of the classification model f . A higher score indicates that the model is less certain in the prediction.*

With this definition, we establish a connection between OSLS and uncertainty quantification. We continue to use the autonomous driving example to demonstrate this connection. The vision model in the system predicts a label for an observed object, yet if the true label is not included in the training set’s label set $\mathcal{Y}$, the prediction will inevitably be wrong. If the system acts on this prediction, an accident could occur. To address OSLS, existing methods typically follow a two-step process [38, 45, 150, 201]. First, an uncertainty score is computed for each input. This score reflects the model’s confidence regarding whether the input belongs to a seen or an unseen class. The uncertainty score can be calculated based on quantities derived from model outputs, such as prediction entropy [150, 201] or softmax confidence [38, 45]. Next, a threshold is selected to decide when to reject the model’s decision. If the uncertainty score exceeds the threshold, the input is classified as belonging to an unseen class rather than any of the seen classes. This threshold should be carefully chosen, as it balances the risks [9] of type I errors (misclassifying an input from a seen class as unseen) and of type II errors (assigning an unseen input to one of the seen classes).

Therefore, since the unseen label is the primary cause of the model’s high uncertainty, OSLS naturally includes the process of determining whether new data under distribution shift belongs to seen classes, and their methods contribute to quantifying uncertainty. Notably, while this section primarily discusses how methods for addressing OSLS are related to uncertainty quantification, more techniques in uncertainty quantification [59] can be applied to identify data points with unseen labels.

6 Conclusions and Future Directions

Distribution shifts pose severe safety challenges to AI systems. To gain an in-depth insight into the relationships between distribution shift and AI safety, we have conducted a systematic analysis of the connections between specific causes of distribution shift and concrete AI safety issues. The main contributions of our work are summarized as follows.

- **A comprehensive, apples-to-apples, and mathematically grounded analysis.** We have thoroughly examined various aspects of distribution shift and AI safety. By breaking them down into specific topics, we provide a detailed, one-to-one correspondence in our discussion. Moreover, the connections between specific distribution shift causes and concrete AI safety issues are established based on the mathematical relationships in their rigorous definitions.
- **Identification of two types of relationship with a corresponding methodological review.** Through our analysis, we have identified two types of relationships between distribution shift and AI safety: (1) methods that address a specific cause of distribution shift can achieve a concrete AI safety goal; and (2) methods designed for a specific distribution shift cause and the corresponding AI safety issue can be mutually adapted due to the inherent consistency in their definitions. Based on these connections, we provide a detailed review of related methodologies to establish a solid link between the specific topics within distribution shift and AI safety.

Based on the connections between distribution shift and AI safety revealed in our work, we suggest two promising directions for future research in these two fields.

- **In-depth integration of methods in distribution shift and AI safety.** Our work demonstrates how methods for handling distribution shifts can effectively achieve specific AI safety goals (i.e., the first type of relationship). However, within AI safety research, techniques from the field of distribution shift have not yet been widely recognized and applied [64, 90, 118]. Furthermore, while we have revealed consistency between the definitions of specific topics in distribution shift and AI safety (i.e., the second type of relationship) and such consistency is already reflected in the similarity of their methods, few works have truly adapted methods from one field to the other. Therefore, there is significant potential for applying methods in the field of distribution shift to AI safety, and vice versa.
- **Broader interconnections between distribution shift and AI safety.** In our work, we focus on three well-studied categories of distribution shift: selection bias, spurious correlation, and label shift. However, as ML models are increasingly deployed across diverse fields, a wider variety of causes of distribution shift emerge. For example, under performative distribution shift [27, 134], the predictions made by deployed ML models influence the data generation process, so that the data distribution is changed. These new forms of distribution shift introduce further safety risks, yet their connections to AI safety remain underexplored. Moreover, while our discussion primarily focuses on distribution shift in prediction tasks, we further suggest that it is necessary to systematically extend our framework to encompass additional settings, such as reinforcement learning and autoregressive learning.

References

- [1] Faruk Ahmed, Yoshua Bengio, Harm Van Seijen, and Aaron Courville. 2020. Systematic generalisation with group invariant predictions. In *International Conference on Learning Representations*.
- [2] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. 2021. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems* 34 (2021), 3438–3450.

- [3] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. 2020. Invariant risk minimization games. In *International Conference on Machine Learning*. PMLR, 145–155.
- [4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).
- [5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).
- [6] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021).
- [7] Yang Bai, Gaojie Xing, Hongyan Wu, Zhihong Rao, Chuan Ma, Shiping Wang, Xiaolei Liu, Yimin Zhou, Jiajia Tang, Kaijun Huang, et al. 2024. Backdoor Attack and Defense on Deep Learning: A Survey. *IEEE Transactions on Computational Social Systems* (2024).
- [8] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. 2013. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE international conference on computer vision*. 769–776.
- [9] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)* 68, 6 (2021), 1–34.
- [10] Emanuel Ben-Baruch, Adam Botach, Igor Kviatkovsky, Manoj Aggarwal, and Gérard Medioni. 2024. Distilling the Knowledge in Data Pruning. *arXiv preprint arXiv:2403.07854* (2024).
- [11] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59, 2 (2013), 341–357.
- [12] Leonard Bereska and Efstratios Gavves. 2024. Mechanistic Interpretability for AI Safety—A Review. *arXiv preprint arXiv:2404.14082* (2024).
- [13] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2154–2156.
- [14] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, 14 (2006), e49–e57.
- [15] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. 2019. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, Vol. 2019. NIH Public Access, 2060.
- [16] Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnub Tandon, and Sanmi Koyejo. 2023. Deceptive alignment monitoring. *arXiv preprint arXiv:2307.10569* (2023).
- [17] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. 2020. Invariant rationalization. In *International Conference on Machine Learning*. PMLR, 1448–1458.
- [18] Annie S Chen, Yoonho Lee, Amrith Setlur, Sergey Levine, and Chelsea Finn. 2023. Project and probe: Sample-efficient domain adaptation by interpolating orthogonal features. *arXiv preprint arXiv:2302.05441* (2023).
- [19] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728* (2018).
- [20] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* (2017).
- [21] Yimeng Chen, Ruibin Xiong, Zhi-Ming Ma, and Yanyan Lan. 2022. When does group invariant learning survive spurious correlations? *Advances in Neural Information Processing Systems* 35 (2022), 7038–7051.
- [22] Mathieu Chevalley, Charlotte Bunne, Andreas Krause, and Stefan Bauer. 2022. Invariant causal mechanisms through distribution matching. *arXiv preprint arXiv:2206.11646* (2022).
- [23] Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. 2024. Open-set domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23943–23953.
- [24] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
- [25] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. 2020. Selection via Proxy: Efficient Data Selection for Deep Learning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=HJg2b0VYDr>
- [26] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. 2021. Environment inference for invariant learning. In *International Conference on Machine Learning*. PMLR, 2189–2200.
- [27] Edwige Cyffers, Muni Sreenivas Pydi, Jamal Atif, and Olivier Cappé. 2024. Optimal classification under performative distribution shift. *arXiv preprint arXiv:2411.02023* (2024).
- [28] Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. 2023. Robust learning with progressive data expansion against spurious correlation. *Advances in neural information processing systems* 36 (2023), 1390–1402.
- [29] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. 2022. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*. PMLR, 12004–12019.
- [30] Wenhao Ding, Laixi Shi, Yuejie Chi, and Ding Zhao. 2024. Seeing is not believing: Robust reinforcement learning against spurious correlation. *Advances in Neural Information Processing Systems* 36 (2024).

- [31] Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. 2020. On the transfer of disentangled representations in realistic settings. *arXiv preprint arXiv:2010.14407* (2020).
- [32] Matthew F Dixon, Igor Halperin, and Paul Bilokon. 2020. *Machine learning in finance*. Vol. 1170. Springer.
- [33] Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. 2020. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Proceedings of the 36th Annual Computer Security Applications Conference*. 897–912.
- [34] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [35] Cian Eastwood, Shashank Singh, Andrei L Nicolicioiu, Marin Vlastelica Poganić, Julius von Kügelgen, and Bernhard Schölkopf. 2024. Spuriousity didn't kill the classifier: Using invariant predictions to harness spurious features. *Advances in Neural Information Processing Systems* 36 (2024).
- [36] Tom Everitt and Marcus Hutter. 2018. The alignment problem for Bayesian history-based reinforcement learners. *Under submission* (2018).
- [37] Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. 2020. Rethinking importance weighting for deep learning under distribution shift. *Advances in neural information processing systems* 33 (2020), 11996–12007.
- [38] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2020. Learning to detect open classes for universal domain adaptation. In *Computer vision—ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XV* 16. Springer, 567–583.
- [39] Keinosuke Fukunaga. 2013. *Introduction to statistical pattern recognition*. Elsevier.
- [40] Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. 2023. Leveraging sparse and shared feature activations for disentangled representation learning. *Advances in Neural Information Processing Systems* 36 (2023), 27682–27698.
- [41] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research* 17, 59 (2016), 1–35.
- [42] Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. 2020. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760* (2020).
- [43] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*. 113–125.
- [44] Saurabh Garg, Sivaraman Balakrishnan, and Zachary Lipton. 2022. Domain adaptation under open set label shift. *Advances in Neural Information Processing Systems* 35 (2022), 22531–22546.
- [45] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. 2017. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418* (2017).
- [46] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bygh9j09KK>
- [47] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. 2016. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence* 39, 7 (2016), 1414–1430.
- [48] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. 2020. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing* 29 (2020), 3993–4002.
- [49] Alice Gomstyn and Alexandra Jonker. 2024. Democratizing AI: What does it mean and how does it work? <https://www.ibm.com/think/insights/democratizing-ai> Online; accessed 2025-02-22.
- [50] Alice Gomstyn and Alexandra Jonker. 2024. Democratizing AI: What does it mean and how does it work? <https://www.ibm.com/think/insights/democratizing-ai> Online; accessed 2025-03-12.
- [51] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. 2016. Domain adaptation with conditional transferable components. In *International conference on machine learning*. PMLR, 2839–2848.
- [52] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of field robotics* 37, 3 (2020), 362–386.
- [53] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733* (2017).
- [54] Ruocheng Guo, Pengchuan Zhang, Hao Liu, and Emre Kiciman. 2021. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732* (2021).
- [55] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.
- [56] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).
- [57] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.
- [58] Muyang He, Shuo Yang, Tiejun Huang, and Bo Zhao. 2024. Large-scale dataset pruning with dynamic uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7713–7722.
- [59] Wenchong He and Zhe Jiang. 2023. A survey on uncertainty quantification methods for deep neural networks: An uncertainty source perspective. *arXiv preprint arXiv:2302.13425* (2023).

- [60] Xuanli He, Qionghai Xu, Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2023. Mitigating backdoor poisoning attacks through the lens of spurious correlation. *arXiv preprint arXiv:2305.11596* (2023).
- [61] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. 2018. Invariant causal prediction for nonlinear models. *Journal of Causal Inference* 6, 2 (2018), 20170016.
- [62] Jose Hernandez-Orallo, Fernando Martinez-Plumed, Santiago Escobar, and Pablo A. M. Casares. 2022. Distributional shift — The TAILOR Handbook of Trustworthy AI. https://prafra.github.io/jupyter-book-TAILOR-D3.2/T3.2/distributional_shift.html Online; accessed 2025-02-22.
- [63] Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459* (2023).
- [64] Yupeng Hu, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. 2021. Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–36.
- [65] Andong Hua, Jindong Gu, Zhiyu Xue, Nicholas Carlini, Eric Wong, and Yao Qin. 2024. Initialization matters for adversarial transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24831–24840.
- [66] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. 2022. Backdoor defense via decoupling the training process. *arXiv preprint arXiv:2202.03423* (2022).
- [67] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566* (2024).
- [68] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820* (2019).
- [69] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning* 110, 3 (2021), 457–506.
- [70] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. 2022. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*. PMLR, 336–351.
- [71] International Organization for Standardization and International Electrotechnical Commission. 2020. *Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*. Standard ISO/IEC TR 24028:2020. International Organization for Standardization and International Electrotechnical Commission. <https://www.iso.org/standard/77608.html>
- [72] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. 2022. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems* 35 (2022), 38516–38532.
- [73] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852* (2023).
- [74] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. 2020. Domain extrapolation via regret minimization. *arXiv preprint arXiv:2006.03908* 3 (2020).
- [75] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4893–4902.
- [76] Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1548–1558.
- [77] Ori Katz, Ronen Talmon, and Uri Shaham. 2024. Supervised Domain Adaptation Based on Marginal and Conditional Distributions Alignment. *Transactions on Machine Learning Research* (2024).
- [78] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925* (2023).
- [79] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)* 55, 2 (2022), 1–38.
- [80] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2022. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937* (2022).
- [81] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*. PMLR, 5637–5664.
- [82] Masanori Koyama and Shoichiro Yamaguchi. 2020. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883* (2020).
- [83] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*. PMLR, 5815–5826.
- [84] Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. 2020. Stable prediction with model misspecification and agnostic distribution shift. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (Apr. 2020), 4485–4492.
- [85] Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. 2023. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*. PMLR, 18171–18206.
- [86] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *Ieee Access* 8 (2020), 193907–193934.

- [87] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. 2021. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems* 34 (2021), 25123–25133.
- [88] Yoonho Lee, Huaxiu Yao, and Chelsea Finn. 2023. Diversify and disambiguate: Out-of-distribution robustness via disagreement. In *The Eleventh International Conference on Learning Representations*.
- [89] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. 2020. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems* 33 (2020), 8847–8860.
- [90] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From principles to practices. *Comput. Surveys* 55, 9 (2023), 1–46.
- [91] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Colorado Reed, Dongsheng Li, Kurt Keutzer, and Han Zhao. 2022. Invariant Information Bottleneck for Domain Generalization. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 7 (2022), 7399–7407.
- [92] Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. 2018. Domain invariant and class discriminative feature learning for visual domain adaptation. *IEEE transactions on image processing* 27, 9 (2018), 4260–4273.
- [93] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. 2022. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems* 64, 12 (2022), 3197–3234.
- [94] Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. 2018. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [95] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. 2021. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems* 34 (2021), 14900–14912.
- [96] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. 2021. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930* (2021).
- [97] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*. 624–639.
- [98] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. 2020. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692* (2020).
- [99] Yudong Li, Shigeng Zhang, Weiping Wang, and Hong Song. 2023. Backdoor attacks to deep learning models and countermeasures: A survey. *IEEE Open Journal of the Computer Society* 4 (2023), 134–146.
- [100] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*. PMLR, 3122–3130.
- [101] Evan Z Liu, Behzad Haghighi, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*. PMLR, 6781–6792.
- [102] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2022. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2022), 1–59.
- [103] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyuan Shen. 2021. Heterogeneous risk minimization. In *International Conference on Machine Learning*. PMLR, 6804–6814.
- [104] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*. Springer, 273–294.
- [105] Litian Liu and Yao Qin. 2024. Fast Decision Boundary based Out-of-Distribution Detector. *ICML* (2024).
- [106] Litian Liu and Yao Qin. 2025. Detecting Out-of-Distribution through the Lens of Neural Collapse. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2025).
- [107] Lydia T Liu, Max Simchowitz, and Moritz Hardt. 2019. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*. PMLR, 4051–4060.
- [108] Yingqi Liu, Wen-Chuan Lee, Guan hong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. 2019. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 1265–1282.
- [109] Yiran Liu, Xiaoang Xu, Zhiyi Hou, and Yang Yu. 2024. Causality Based Front-door Defense Against Backdoor Attack on Language Models. In *Forty-first International Conference on Machine Learning*.
- [110] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.
- [111] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jianguang Sun, and Philip S Yu. 2014. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1410–1417.
- [112] Clare Lyle, Amy Zhang, Minqi Jiang, Joelle Pineau, and Yarin Gal. 2021. Resolving causal confusion in reinforcement learning via robust exploration. In *Self-Supervision for Reinforcement Learning Workshop-ICLR*, Vol. 2021.
- [113] Wenao Ma, Cheng Chen, Shuang Zheng, Jing Qin, Huimao Zhang, and Qi Dou. 2022. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 313–323.
- [114] Divyat Mahajan, Shruti Tople, and Amit Sharma. 2021. Domain generalization using causal matching. In *International conference on machine learning*. PMLR, 7313–7324.

- [115] Adyasha Maharana, Prateek Yadav, and Mohit Bansal. 2024. $\mathbb{D}^2$ Pruning: Message Passing for Balancing Diversity & Difficulty in Data Pruning. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=thbtoAkCe9>
- [116] Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D’Amour. 2022. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 739–766.
- [117] Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. 2022. Trivial or Impossible – dichotomous data difficulty masks model differences (on ImageNet and beyond). In *International Conference on Learning Representations*. https://openreview.net/forum?id=C_vsGwEljAr
- [118] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.
- [119] Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. 2020. The role of disentanglement in generalisation. In *International Conference on Learning Representations*.
- [120] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. 2017. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*. 5715–5725.
- [121] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *International conference on machine learning*. PMLR, 10–18.
- [122] Jens Müller, Robert Schmier, Lynton Ardizzone, Carsten Rother, and Ullrich Köthe. 2021. Learning robust models using the principle of independent causal mechanisms. In *DAGM German Conference on Pattern Recognition*. Springer, 79–110.
- [123] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. 2020. Learning from Failure: Training Debiased Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*.
- [124] A Tuan Nguyen, Toan Tran, Yarin Gal, and Atilim Gunes Baydin. 2021. Domain invariant representation learning with domain density transformations. *Advances in Neural Information Processing Systems* 34 (2021), 5264–5275.
- [125] Jianmo Ni, Jiacheng Li, and Julian 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 188–197. doi:10.18653/v1/D19-1018
- [126] Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. 2022. Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1295–1305.
- [127] Matteo Pagliardini, Martin Jaggi, François Fleuret, and Sai Praneeth Karimireddy. 2022. Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202.04414* (2022).
- [128] Sinno Jialin Pan, James T Kwok, Qiang Yang, et al. 2008. Transfer learning via dimensionality reduction.. In *AAAI*, Vol. 8. 677–682.
- [129] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2010. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks* 22, 2 (2010), 199–210.
- [130] Peter S Park, Simon Goldstein, Aidan O’Gara, Michael Chen, and Dan Hendrycks. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns* 5, 5 (2024).
- [131] Cato Pauling, Michael Gimson, Muhammed Qaid, Ahmad Kida, and Basel Halak. 2022. A tutorial on adversarial learning attacks and countermeasures. *arXiv preprint arXiv:2202.10377* (2022).
- [132] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. 2019. Domain agnostic learning with disentangled representations. In *International conference on machine learning*. PMLR, 5102–5112.
- [133] Xingchao Peng, Yichen Li, and Kate Saenko. 2020. Domain2vec: Domain embedding for unsupervised domain adaptation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI* 16. Springer, 756–774.
- [134] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.
- [135] Jonas Peters, Peter Bühlmann, and Nicolai 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.
- [136] Hoang Phan, Andrew Gordon Wilson, and Qi Lei. 2024. Controllable Prompt Tuning For Balancing Group Distributional Robustness. *Proceedings of Machine Learning Research* 235 (2024), 40667–40687.
- [137] Aahlad Puli, Nitish Joshi, Yoav Wald, He He, and Rajesh Ranganath. 2022. Nuisances via negativa: Adjusting for spurious correlations via data augmentation. *arXiv preprint arXiv:2210.01302* (2022).
- [138] Aahlad Puli, Lily H Zhang, Eric K Oermann, and Rajesh Ranganath. 2021. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. *arXiv preprint arXiv:2107.00520* (2021).
- [139] Yao Qin, Xuezhi Wang, Balaji Lakshminarayanan, Ed H Chi, and Alex Beutel. 2023. What are effective labels for augmented data? improving calibration and robustness with autolabel. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 365–376.
- [140] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. 2022. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *Advances in Neural Information Processing Systems* 35 (2022), 16276–16289.
- [141] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. 2023. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*. PMLR, 28448–28467.

- [142] Tim R  z. 2021. Group fairness: Independence revisited. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 129–137.
- [143] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet?. In *International conference on machine learning*. PMLR, 5389–5400.
- [144] Mateo Rojas-Carulla, Bernhard Sch  lkopf, Richard Turner, and Jonas Peters. 2018. Invariant models for causal transfer learning. *Journal of Machine Learning Research* 19, 36 (2018), 1–34.
- [145] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. 2022. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856* (2022).
- [146] Sorawit Saengkyongam, Nikolaj Thams, Jonas Peters, and Niklas Pfister. 2023. Invariant policy learning: A causal perspective. *IEEE transactions on pattern analysis and machine intelligence* 45, 7 (2023), 8606–8620.
- [147] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=ryxGuJrFvS>
- [148] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. 2020. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11957–11965.
- [149] Berkman Sahiner, Weijie Chen, Ravi K Samala, and Nicholas Petrick. 2023. Data drift in medical machine learning: implications and potential remedies. *The British Journal of Radiology* 96, 1150 (2023), 20220878.
- [150] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. 2020. Universal domain adaptation through self supervision. *Advances in neural information processing systems* 33 (2020), 16282–16292.
- [151] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. 2020. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859* (2020).
- [152] Iqbal H Sarker, Md Hasan Furhad, and Raza Nowrozy. 2021. Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. *SN Computer Science* 2, 3 (2021), 173.
- [153] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [154] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal misgeneralization: Why correct specifications aren’t enough for correct goals. *arXiv preprint arXiv:2210.01790* (2022).
- [155] K Shailaja, Banoth Seetharamulu, and MA Jabbar. 2018. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 910–914.
- [156] Minglai Shao, Dong Li, Chen Zhao, Xintao Wu, Yujie Lin, and Qin Tian. 2024. Supervised algorithmic fairness in distribution shifts: A survey. *arXiv preprint arXiv:2402.01327* (2024).
- [157] Guangyu Shen, Yingqi Liu, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, Shiqing Ma, and Xiangyu Zhang. 2021. Backdoor scanning for deep neural networks through k-arm optimization. In *International Conference on Machine Learning*. PMLR, 9525–9536.
- [158] Hua Shen, Tiffany Kneare, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. *arXiv preprint arXiv:2406.09264* (2024).
- [159] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [160] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025* (2023).
- [161] Zheyang Shen, Peng Cui, Tong Zhang, and Kun Kunag. 2020. Stable learning via sample reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5692–5699.
- [162] Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. 2023. Black-box backdoor defense via zero-shot image purification. *Advances in Neural Information Processing Systems* 36 (2023), 57336–57366.
- [163] Zhouxing Shi, Nicholas Carlini, Ananth Balashankar, Ludwig Schmidt, Cho-Jui Hsieh, Alex Beutel, and Yao Qin. 2023. Effective robustness against natural distribution shifts for models with different training data. *Advances in Neural Information Processing Systems* 36 (2023), 73543–73558.
- [164] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.
- [165] Changjian Shui, Zijian Li, Jiaqi Li, Christian Gagn  , Charles X Ling, and Boyu Wang. 2021. Aggregating from multiple target-shifted sources. In *International Conference on Machine Learning*. PMLR, 9638–9648.
- [166] Herbert A Simon. 1954. Spurious correlation: A causal interpretation. *Journal of the American statistical Association* 49, 267 (1954), 467–479.
- [167] Anoopkumar Sonar, Vincent Pacelli, and Anirudha Majumdar. 2021. Invariant policy optimization: Towards stronger generalization in reinforcement learning. In *Learning for Dynamics and Control*. PMLR, 21–33.
- [168] Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. 2022. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. *Advances in Neural Information Processing Systems* 35 (2022), 19165–19178.
- [169] Benjamin Stoler, Ingrid Navarro, Meghdeep Jana, Soonmin Hwang, Jonathan Francis, and Jean Oh. 2024. SafeShift: Safety-informed distribution shifts for robust trajectory prediction in autonomous driving. In *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1179–1186.

- [170] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. 2021. Recovering latent causal factor for generalization to distributional shifts. *Advances in Neural Information Processing Systems* 34 (2021), 16846–16859.
- [171] Harry Surden. 2021. Machine learning and law: An overview. *Research handbook on big data law* (2021), 171–184.
- [172] Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. 2020. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems* 33 (2020), 19276–19289.
- [173] Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. 2024. Data pruning via moving-one-sample-out. *Advances in Neural Information Processing Systems* 36 (2024).
- [174] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van den Hengel. 2022. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16761–16772.
- [175] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. 2021. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1417–1427.
- [176] Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral signatures in backdoor attacks. *Advances in neural information processing systems* 31 (2018).
- [177] Loc Truong, Chace Jones, Brian Hutchinson, Andrew August, Brenda Praggastis, Robert Jasper, Nicole Nichols, and Aaron Tuor. 2020. Systematic evaluation of backdoor data poisoning attacks on image classifiers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 788–789.
- [178] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. 2019. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771* (2019).
- [179] Victor Veitch, Alexander D’Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems* 34 (2021), 16196–16208.
- [180] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.
- [181] Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. 2021. On calibration and out-of-domain generalization. *Advances in neural information processing systems* 34 (2021), 2215–2227.
- [182] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 707–723.
- [183] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095* (2023).
- [184] Lun Wang, Zaynah Javed, Xian Wu, Wenbo Guo, Xinyu Xing, and Dawn Song. 2021. Backdoorl: Backdoor attack against competitive reinforcement learning. *arXiv preprint arXiv:2105.00579* (2021).
- [185] Shanshan Wang, Yiyang Chen, Zhenwei He, Xun Yang, Mengzhu Wang, Quanzeng You, and Xingyi Zhang. 2023. Disentangled representation learning with causality for unsupervised domain adaptation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 2918–2926.
- [186] Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2021. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736* (2021).
- [187] Xin Wang, Hong Chen, Zihao Wu, Wenwu Zhu, et al. 2024. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [188] Yixu Wang, Yan Teng, Kexin Huang, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, Yu-Gang Jiang, Yu Qiao, and Yingchun Wang. 2023. Fake Alignment: Are LLMs Really Aligned Well? *arXiv preprint arXiv:2311.05915* (2023).
- [189] Zhenting Wang, Hailun Ding, Juan Zhai, and Shiqing Ma. 2022. Training with more confidence: Mitigating injected and natural backdoors during training. *Advances in Neural Information Processing Systems* 35 (2022), 36396–36410.
- [190] Zhonghao Wang, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. 2020. Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 936–937.
- [191] Tao Wen, Zihan Wang, Quan Zhang, and Qi Lei. 2025. Elastic Representation: Mitigating Spurious Correlations for Group Robustness. *arXiv preprint arXiv:2502.09850* (2025).
- [192] Robert Williamson and Aditya Menon. 2019. Fairness risk measures. In *International conference on machine learning*. PMLR, 6786–6797.
- [193] Keru Wu, Yuansi Chen, Woosuk Ha, and Bin Yu. 2023. Prominent Roles of Conditionally Invariant Components in Domain Adaptation: Theory and Algorithms. *arXiv preprint arXiv:2309.10301* (2023).
- [194] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. 2020. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 6502–6509.
- [195] Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. 2023. Dataset Pruning: Reducing Training Data by Examining Generalization Influence. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=4wZiAXD29TQ>
- [196] Wanqian Yang, Polina Kirichenko, Micah Goldblum, and Andrew G Wilson. 2022. Chroma-vae: Mitigating shortcut learning with generative classifiers. *Advances in Neural Information Processing Systems* 35 (2022), 20351–20365.
- [197] Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021. Rethinking stealthiness of backdoor attack against nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5543–5557.

- [198] Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. 2024. Identifying spurious biases early in training through the lens of simplicity bias. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2953–2961.
- [199] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. 2022. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*. PMLR, 25407–25437.
- [200] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715* (2024).
- [201] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2019. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2720–2729.
- [202] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschlager, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811* (2017).
- [203] Werner Zellinger, Bernhard A Moser, Thomas Grubinger, Edwin Lughofer, Thomas Natschlager, and Susanne Saminger-Platz. 2019. Robust unsupervised domain adaptation for neural networks via moment alignment. *Information Sciences* 483 (2019), 174–191.
- [204] Yi Zeng, Minzhou Pan, Hoang Anh Just, Lingjuan Lyu, Meikang Qiu, and Ruoxi Jia. 2023. Narcissus: A practical clean-label backdoor attack with limited information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 771–785.
- [205] Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. 2021. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*. PMLR, 12345–12355.
- [206] Shengfang Zhai, Qingni Shen, Xiaoyi Chen, Weilong Wang, Cong Li, Yuejian Fang, and Zhonghai Wu. 2023. Ncl: Textual backdoor defense using noise-augmented contrastive learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
- [207] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Scholkopf, and Eric P Xing. 2022. Towards principled disentanglement for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8024–8034.
- [208] Jie Zhang, Chen Dongdong, Qidong Huang, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. 2022. Poison ink: Robust and invisible backdoor attack. *IEEE Transactions on Image Processing* 31 (2022), 5691–5705.
- [209] Michael Zhang and Christopher Re. 2022. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems* 35 (2022), 21682–21697.
- [210] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. 2022. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517* (2022).
- [211] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyang Shen. 2021. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5372–5382.
- [212] Xingxuan Zhang, Zekai Xu, Renzhe Xu, Jiashuo Liu, Peng Cui, Weitao Wan, Chong Sun, and Chen Li. 2022. Towards domain generalization in object detection. *arXiv preprint arXiv:2203.14387* (2022).
- [213] Yi-Fan Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. 2022. Learning domain invariant representations for generalizable person re-identification. *IEEE Transactions on Image Processing* 32 (2022), 509–523.
- [214] Zaixi Zhang, Qi Liu, Zhicai Wang, Zepu Lu, and Qingyong Hu. 2023. Backdoor defense via deconfounded representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12228–12238.
- [215] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. 2020. Domain generalization via entropy regularization. *Advances in neural information processing systems* 33 (2020), 16096–16107.
- [216] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. 2020. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14443–14452.
- [217] Jiayun Zheng and Maggie Makar. 2022. Causally motivated multi-shortcut identification and removal. *Advances in Neural Information Processing Systems* 35 (2022), 12800–12812.
- [218] Kecheng Zheng, Jiawei Liu, Wei Wu, Liang Li, and Zheng-jun Zha. 2021. Calibrated feature decomposition for generalizable person re-identification. *arXiv preprint arXiv:2111.13945* (2021).
- [219] Rui Zheng, Wei Shen, Yuan Hua, Wenbin Lai, Shihan Dou, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Haoran Huang, Tao Gui, et al. 2023. Improving generalization of alignment with human preferences through group invariant learning. *arXiv preprint arXiv:2310.11971* (2023).
- [220] Ziliang Samuel Zhong, Xiang Pan, and Qi Lei. 2025. Bridging Domains with Approximately Shared Features. In *The 28th International Conference on Artificial Intelligence and Statistics*.
- [221] Jianlong Zhou, Fang Chen, and Andreas Holzinger. 2020. Towards explainability for AI fairness. In *International workshop on extending explainable AI beyond deep models and classifiers*. Springer, 375–386.
- [222] Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, and Baoyuan Wu. 2023. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4466–4477.