Protected Heterogeneity: A Variance-Based Framework for Fair Algorithmic Personalization

Noah M. Ahmadi Eva Ascarza Ayelet Israeli

November 16, 2025

Abstract

Effective personalization requires heterogeneity — the differences between users in their characteristics, preferences, behaviors, and contexts — that makes differential treatment valuable. Some of this heterogeneity, however, may stem from legally or ethically protected attributes such as race, gender, or age. We introduce *protected heterogeneity* as the share of variation in an algorithmically produced score explained by protected attributes and their proxies. We derive a corresponding diagnostic measure, R^2_{prot}, which is bounded, interpretable, and does not depend on arbitrary decision thresholds. This model-agnostic diagnostic tool can be used to identify and quantify fairness concerns by measuring the extent to which protected attributes influence algorithmic predictions. It is a variance-based, pre-decision measure of fairness that can be used to audit both existing systems and new algorithms prior to deployment. In addition, by formalizing the estimation of R^2_{prot}, we introduce a straightforward method for removing protected heterogeneity through the construction of residualized scores. These residualized scores exhibit both individual- and group-level fairness guarantees, enabling decision makers to preserve the value of leveraging individual differences while ensuring that decisions are not systematically shaped by protected characteristics. Beyond a technical contribution, the diagnostic supports regulatory compliance, ethical objectives, and reputational risk management in high-stakes domains such as healthcare, finance, and employment.

Keywords: Fairness, algorithmic bias, personalization, heterogeneity, variance decomposition.

1 Introduction

Algorithmic decision-making systems are commonplace across diverse domains, from recidivism prediction (Corbett-Davies et al. 2017) to healthcare triage (Obermeyer et al. 2019) to loan approvals and credit scoring (Kozodoi et al. 2022), and marketing applications such as programmatic advertising delivery and algorithmic pricing (e.g., Aparicio and Misra 2023, Rafieian and Yoganarasimhan 2023). These systems typically rely on individual-level characteristics to produce a score predictive of a desired outcome (e.g., creditworthiness, click probability). This score, tailored to each individual, is then compared to a threshold to generate a *personalized* decision — personalized meaning the decision depends explicitly on the score of each individual.

Crucially, the effectiveness of such personalization hinges on the heterogeneity captured by the score itself: by scoring individuals according to relevant differences, organizations can make scalable, tailored allocation decisions. The score therefore serves as a unified mechanism to quantify those differences, distilling the heterogeneity that matters for the decision at hand into a single, actionable metric.

Not all forms of heterogeneity, however, may be appropriate for decision-making. Certain attributes (e.g., race, gender, age, or socioeconomic status) represent characteristics that may be considered unsuitable to leverage, either due to regulatory constraints or ethical reasons. In applications such as lending, hiring, and criminal-justice, regulations in many countries explicitly prohibit decisions that discriminate based on protected characteristics (Barocas et al. 2023). In applications where regulations do not explicitly prohibit discriminatory decision making, such as marketing, personalization based on sensitive demographics may breach privacy norms, clash with brand values, or trigger reputational fallout (Ascarza and Israeli 2022). When algorithmic systems incorporate these protected attributes — either directly or indirectly through correlated variables — the resulting decisions may raise fairness concerns because they systematically benefit or harm certain groups or disproportionately affect individuals based on their group membership (Mehrabi et al. 2021, Fu et al. 2020).

The scholarly response to algorithmic bias has yielded a "zoo" of fairness metrics (Castelnovo et al. 2022) ranging from group-level criteria (e.g., statistical parity, equalized odds, predictive parity, calibration) to individual-level criteria (e.g., individual fairness, counterfactual fairness)

(Mitchell et al. 2021, Mehrabi et al. 2021). Although each metric captures a legitimate ethical intuition, the metrics are often mutually incompatible when groups differ in base rates (Chouldechova 2017), and their magnitude can change substantially with a small shift in the decision threshold (Corbett-Davies et al. 2023). Consequently, practitioners must navigate multiple indicators, regulators struggle to choose a single yardstick, and stakeholders receive conflicting signals about the fairness of an algorithm. This proliferation of metrics creates a fundamental challenge: the absence of a fairness measure that is both threshold-invariant and easily interpretable undermines coherent auditing and governance.

We address this challenge by introducing *protected heterogeneity*, defined as the share of variation in an algorithmic score that can be attributable to protected attributes and their proxies.[1] We introduce a useful method to quantify protected heterogeneity that yields a metric that is bounded, interpretable, and directly actionable. Our approach, grounded in variance decomposition, provides stakeholders, including decision-makers, regulators, and auditors, with a framework to measure the extent to which algorithmic scores reflect protected characteristics. Unlike existing fairness metrics that depend on specific thresholds, conflate individual- and group-level concerns, or prove difficult to satisfy simultaneously (Chouldechova 2017, Castelnovo et al. 2022, Corbett-Davies et al. 2023), protected heterogeneity offers a pre-decision assessment that is threshold-invariant and connects both individual- and group-fairness perspectives (Dwork et al. 2012). This makes it especially valuable in regulatory and operational settings where decisions must be justified across diverse stakeholders and contexts.

Our framework not only enables the measurement of protected heterogeneity of any algorithmic score; it also provides a straightforward way to act upon it in practice. Specifically, by extending the same measurement framework, one can remove the influence of protected attributes to construct a residualized score — one that allows decision-makers to retain the benefits of personalization while ensuring that differences in algorithmic outcomes cannot be attributed to protected characteristics.

Our contributions are threefold. Conceptually, we introduce protected heterogeneity as a framework for assessing fairness in algorithmic decision-making and connect it to existing definitions of fairness. Methodologically, we propose a variance-based metric bounded between 0 and 1, making

[1] Proxies of protected attributes are variables that are not themselves protected attributes but are related to them, and therefore may effectively reveal or substitute for sensitive information (e.g., the single variable zip code might be a proxy for socioeconomic status, or the combination of zip code and income might together be a proxy for gender).

it interpretable and comparable across domains. Practically, we offer a tool for both diagnosing and mitigating unfair algorithmic outcomes.

2 Related Work

The proliferation of algorithmic decision-making has sparked a surge of research on fairness, yielding a diverse set of metrics — yet no unified, interpretable standard has emerged. Formal definitions of algorithmic fairness generally fall under two categories: *group* (statistical) fairness and *individual* fairness.

Group fairness metrics assess whether outcomes are balanced across protected strata. The most widely cited is statistical parity (also known as demographic parity), which requires equal positive outcome rates across groups regardless of ground truth (Feldman et al. 2015). Variants such as equalized odds and its relaxation, equality of opportunity, equalize specific error rates (e.g., true-positive or false-positive rates) across groups (Hardt et al. 2016). Predictive parity requires that the positive predictive value (precision) be equal across groups, while calibration demands that a given score correspond to the same empirical outcome probability within each group.

Although directly auditable through observed outcomes (Beutel et al. 2019), these metrics have two key limitations. First, when base rates differ, they are mathematically incompatible: no classifier can satisfy all criteria simultaneously (Chouldechova 2017). Second, they are threshold-dependent, complicating assessment and producing inconsistent signals across operating points (Berk et al. 2021, Corbett-Davies et al. 2023).

Individual fairness metrics emphasize person-to-person consistency. Fairness through awareness states that "similar individuals should be treated similarly" and formalizes this through a Lipschitz constraint on a task-specific distance function (Dwork et al. 2012). Counterfactual fairness extends this principle by requiring that predictions remain unchanged in a hypothetical world where only the protected attribute is changed relative to the real world (Kusner et al. 2017). While conceptually appealing, these definitions depend on similarity metrics or causal models that are rarely available, and neither ensures group-level parity.

A key challenge in the fairness literature is the incompatibility of existing metrics — no single decision rule can satisfy all criteria simultaneously (e.g., Chouldechova 2017). Because group

and individual fairness metrics pursue distinct goals, researchers have proposed hybrid objectives, multi-metric dashboards, and constrained optimization schemes. Yet empirical studies show that optimizing one metric can degrade another or reduce utility (Corbett-Davies et al. 2023, Mitchell et al. 2021), leaving practitioners with partial solutions that do not resolve incompatibilities or threshold sensitivities.

Compounding this is threshold dependence: most fairness metrics vary with decision boundaries, making it difficult to evaluate fairness in the underlying score independently of cutoff choices (Corbett-Davies et al. 2023). Together, metric proliferation, mutual incompatibility, and threshold sensitivity hinder efforts to manage, regulate, and audit algorithmic fairness. Practitioners are therefore left to navigate competing definitions, often without clear guidance on which metric aligns with institutional goals or policy constraints.

These challenges underscore the need for a holistic framework; one that unifies divergent fairness definitions and provides interpretable, actionable guidance for assessing fairness without being constrained by conflicting or threshold-sensitive metrics. Such a measure should ideally be (i) threshold-invariant, (ii) bounded and interpretable, and (iii) capable of linking group- and individual-level perspectives. Our objective in this research is to meet these needs by introducing and formalizing a novel, variance-based metric that quantifies a core fairness concern: the extent to which algorithmic scores reflect protected attributes.

To achieve this, we focus on *protected heterogeneity* — a framework that captures this concern while supporting both measurement and mitigation. By adopting a variance-based perspective, we directly quantify the proportion of score variation attributable to protected attributes and their proxies, independent of any decision threshold, thereby satisfying the desired properties of a holistic fairness measure.

Variance-based approaches have recently emerged as promising tools for fairness diagnostics. For example, Mukherjee et al. (2020) use variance decomposition to isolate gender bias in deep learning embeddings, demonstrating its power for attributing model behavior to gender features. More broadly, variance decomposition is a well-established technique for partitioning explained variation across covariates (Gelman 2005). Extending this line of work, Bénesse et al. (2024) frame fairness as a problem of global sensitivity analysis and propose a family of indices to quantify the influence of protected features. While these approaches highlight the promise of variance-based

fairness measurement, they rely on multiple indices or context-specific choices rather than a single, interpretable diagnostic.

A complementary line of work develops general measures of conditional dependence, which underpin many fairness definitions. Azadkia and Chatterjee (2021) propose a coefficient of conditional dependence that converges to zero if and only if two variables are conditionally independent given covariates. Although useful as a diagnostic of independence, their measure does not partition variation into fairness-relevant versus fairness-neutral components.

Building on these foundations, we introduce *protected heterogeneity*. Our approach adapts variance decomposition into a single, bounded, and threshold-invariant metric tailored to fairness. By explicitly incorporating both protected attributes and their proxies, our measure provides an interpretable diagnostic and a direct path to mitigation through residualization. Protected heterogeneity therefore offers a useful measure of fairness in algorithmically produced scores. It is simple to compute and apply in practice, requiring only a post-processing step. To our knowledge, no other existing metric is simultaneously threshold-invariant, bounded, and interpretable. Furthermore, we provide guidance on acting upon protected heterogeneity by residualizing these scores, enabling personalization while also providing both individual and group fairness guarantees. In doing so, our framework provides a concrete tool for auditors, regulators, and decision-makers to use when implementing and evaluating fairness standards.

3 Protected Heterogeneity

This section defines and operationalizes protected heterogeneity and provides key properties of this proposed metric.

3.1 Notation and formal definition

Let each individual i be described by covariates $\mathbf{C}_i = \{\mathbf{X}_i, \mathbf{Z}_i\}$, where $\mathbf{Z}_i \in \mathbb{R}^{d_z}$ are the *protected* attributes (e.g., race, gender, age) and $\mathbf{X}_i$ are the remaining *non-protected* attributes. An algorithm outputs a deterministic score $\phi_i = f(\mathbf{X}_i, \mathbf{Z}_i)$, which a downstream policy may compare to a threshold to determine an outcome. This score is central to many algorithmic systems, where individuals are ranked by ϕ and allocation is often determined by their position relative to a cutoff.

Protected attributes may influence ϕ both directly via $\mathbf{Z}$ and indirectly via proxies contained in $\mathbf{X}$. We denote

$$h(\mathbf{X}_i) = \big(h_1(\mathbf{X}_i),\ \dots\ , h_{d_z}(\mathbf{X}_i)\big),$$

a d_z-dimensional vector of functions where $h_j(\mathbf{X}) = \mathbb{E}[Z_j \mid \mathbf{X}]$ captures the component of $\mathbf{X}$ acting as a proxy for the protected attribute Z_j.[2] We then define

$$\zeta := \big(\mathbf{Z}, h(\mathbf{X})\big),$$

which captures both the direct and proxy channels through which the protected attributes may influence ϕ. We omit the subscript i for ease of notation and decompose

$$\phi \ = \ \underbrace{\mathbb{E}[\phi \mid \zeta]}_{g(\zeta)} \ + \ \underbrace{\big[\phi - \mathbb{E}[\phi \mid \zeta]\big]}_{\phi^{\text{res}}}, \tag{1}$$

such that $g(\zeta) = \mathbb{E}[\phi \mid \zeta]$ is the portion of ϕ explained by $\mathbf{Z}$ and its proxy $h(\mathbf{X})$, and $\phi^{\text{res}} = \phi - g(\zeta)$ is the residual component of ϕ, uncorrelated with ζ.[3] This decomposition implies the following variance partitioning in ϕ:

$$\mathbb{V}(\phi) = \mathbb{V}\big(g(\zeta)\big) + \mathbb{V}(\phi^{\text{res}}), \tag{2}$$

where $\mathbb{V}\big(g(\zeta)\big)$ captures the variance explained both directly and indirectly by the protected attributes $\mathbf{Z}$. We define this portion of total variance as *protected heterogeneity*. The term $\mathbb{V}(\phi^{\text{res}})$ is the *residual* portion of total heterogeneity. This component is orthogonal to protected individual differences in the score, making it desirable for fair personalization. Because residual heterogeneity is unrelated to protected attributes, it represents the portion of the score that can be leveraged for personalization without amplifying protected-attribute differences.

Finally, we define the share

$$R^2_{\text{prot}} \ := \ \frac{\mathbb{V}\big(g(\zeta)\big)}{\mathbb{V}(\phi)} \ \in \ [0, 1] \tag{3}$$

as the fraction of total variance attributable to protected heterogeneity in the score ϕ. When $R^2_{\text{prot}} = 1$, *all* variation is protected and decisions based on the score will only reflect differences

[2] For example, customer characteristics or browsing behavior may be predictive of gender. The function $h(\cdot)$ captures the extent to which those variables encode information about gender.

[3] This linear framework ensures that the two components are orthogonal, since $\text{Cov}(g(\zeta), \phi - g(\zeta)) = 0$.

in protected attributes. Conversely, when $R^2_{\text{prot}} = 0$ the score is free of all predictable information related to protected attributes, implying that decisions made based on the score will not capture or reflect any meaningful differences in the attributes aimed to be protected.

3.2 Estimating protected heterogeneity in practice

While $\mathbb{V}(\phi)$ can be computed directly from the observed scores, both the proxy component $h(\mathbf{X})$ of ζ and the true conditional mean $g(\zeta) = \mathbb{E}[\phi \mid \zeta]$ are unknown and must be estimated from data. We describe how to estimate these unknowns and compute R^2_{prot} in practice.

Step 1: Estimating proxies $\hat{h}(\mathbf{X})$ and forming $\hat{\zeta}$. For each protected attribute in $\{Z_j \colon j = 1, \ldots, d_z\}$, we estimate a conditional expectation (proxy) function $h_j(\mathbf{X}) = \mathbb{E}[Z_j \mid \mathbf{X}]$ using a flexible predictor $\hat{h}_j(\cdot)$ (e.g., probability forests for binary Z_j, regression forests for continuous Z_j; other well-regularized learners are admissible). Stacking these estimates yields

$$\hat{h}(\mathbf{X}) = \big(\hat{h}_1(\mathbf{X}), \ldots, \hat{h}_{d_z}(\mathbf{X})\big) \quad \text{where} \quad \hat{\zeta} = \big(\mathbf{Z},\, \hat{h}(\mathbf{X})\big).$$

Step 2: Learning the protected component $\hat{g}(\hat{\zeta})$. To further approximate the protected component $g(\zeta) = \mathbb{E}[\phi \mid \zeta]$, we model the score ϕ as a function of the estimated protected information vector $\hat{\zeta}$ (constructed in Step 1) using a flexible model such as a regression forest, neural net, or linear regression. Predictions of such a model would yield the estimated protected component $\hat{g}(\hat{\zeta})$.

While it may seem natural to substitute $\hat{g}(\hat{\zeta})$ directly into Equation (3), for $\mathbb{V}\big(\hat{g}(\hat{\zeta})\big)$ to approximate $\mathbb{V}\big(g(\zeta)\big)$ accurately, two conditions must be met. First, $\hat{g}(\cdot)$ must correctly capture the portion of variation in ϕ attributable to protected attributes — both directly through $\mathbf{Z}$ and indirectly through the proxies components $\hat{h}(\mathbf{X})$. Underfitting will push protected variation into the residual, leading to an underestimate of protected heterogeneity, whereas overfitting will cause $\hat{g}(\hat{\zeta})$ to absorb unrelated variation, producing a spuriously inflated estimate. To mitigate these risks, we employ honest estimation strategies (e.g., cross-fitting or out-of-bag estimation) and use model-fit diagnostics to ensure that the estimated $\hat{g}(\cdot)$ reflects genuine protected heterogeneity rather than artifacts of model misspecification.

Second, the estimated protected component must be orthogonal to the residuals such that its variance cleanly represents the protected share. Flexible learners do not guarantee this property, even under cross-validation. We therefore introduce an OLS projection as a final estimation step, which enforces orthogonality by construction, and ensures that the resulting R^2 corresponds precisely to the share of variation explained by protected attributes.

Step 3: Orthogonalization via OLS and computing R^2_{prot}. In this final step, we project ϕ onto $\hat{g}(\hat{\zeta})$ via OLS such that:

$$\phi \;=\; \beta_0 \;+\; \beta_1\, \hat{g}(\hat{\zeta}) \;+\; \hat{\phi}^{\text{res}}, \tag{4}$$

and take the R^2 from Equation (4) as our operational estimate of R^2_{prot}.[4] This OLS projection serves as a calibration step: it guarantees orthogonality between protected and residual components with $\text{Cov}\big(\hat{g}(\hat{\zeta}), \hat{\phi}^{\text{res}}\big) = 0$ regardless of how $\hat{g}(\cdot)$ was originally obtained. As a result, the R^2 in Equation (4) provides a clean and model-agnostic estimate of the share of variance in ϕ aligned exclusively with the protected component, consistent with the definition in Equation (3).

3.3 Key Properties

Having defined and operationalized R^2_{prot}, we now highlight several properties that make it especially useful for auditing, monitoring, and governance.

3.3.1 Bounded and Interpretable

By construction, $R^2_{\text{prot}} \in [0, 1]$. When $R^2_{\text{prot}} = 0$, the score is completely orthogonal to protected attributes. When $R^2_{\text{prot}} = 1$, all variation is explained by protected attributes and their proxies. This boundedness provides natural interpretive anchors: values near 0 suggest minimal fairness concerns, while values near 1 indicate greater discrimination based on protected features, with a value of 1 indicating discrimination exclusively based on protected features.

A related additional benefit is the ability to measure and compare protected heterogeneity across different protected attributes. For example, a model may have $R^2_{\text{prot}} = 0.10$ when protecting

[4] As $\hat{g}(\cdot) \to g(\cdot)$, we have $\beta_0 \to 0$, $\beta_1 \to 1$, and $\hat{\phi}^{\text{res}} \to \phi - g(\cdot)$.

Z_1 and $R^2_{\text{prot}} = 0.20$ when protecting Z_2, indicating a larger share of total heterogeneity is linked to Z_2. [5]

3.3.2 Threshold-Invariant

Most existing fairness metrics are evaluated at the decision level — that is, *after* a threshold is applied to the score, such as a loan approval cutoff or a hiring shortlist. As a result, fairness assessments are sensitive to the chosen threshold: a model may appear fair under one cutoff and unfair under another. This dependence complicates both diagnosis and governance, since different stakeholders may select different thresholds and obtain conflicting conclusions.

In contrast, R^2_{prot} is computed directly on the score itself, before any thresholding. Its value depends only on the extent to which protected attributes (or their proxies) explain score variation, not on how the score is later translated into binary decisions. Whether the cutoff is strict, lenient, or varies across decisions, the measure remains the same.

This property makes R^2_{prot} especially useful for auditing and regulation: it provides a single, consistent diagnostic that does not hinge on arbitrary implementation details. In practice, this means that once R^2_{prot} is computed, stakeholders can evaluate fairness without needing to agree on a decision threshold.

3.3.3 Flexible Measurement

A distinctive advantage of R^2_{prot} is that it accommodates protected attributes of any dimensionality, including discrete, categorical, and continuous variables, as well as their interactions. In contrast, many existing fairness metrics, such as statistical parity or equalized odds, require the discretization of attributes before measurement, such as defining arbitrary age groups or income brackets. Such discretization not only discards information but also introduces opportunities for manipulation, since parity results may depend heavily on how groups are chosen. This flexibility allows policymakers, regulators, and decision-makers to evaluate fairness comprehensively across diverse contexts and definitions of protected classes.

[5] Note, however, that when protecting both Z_1 and Z_2 simultaneously, covariances across these attributes are accounted for and the overall proportion of protected heterogeneity is not necessarily equal to the sum of the individual contributions. When individual measurement is of interest, protected heterogeneity can be modeled separately for each attribute in $\mathbf{Z}$.

3.3.4 Actionable

Beyond serving as a diagnostic, R^2_{prot} also provides a direct path to mitigation. Removing the portion of the score ϕ explained by ζ generates the residualized score: $\phi^{\text{res}} = \phi - g(\zeta)$, which, by construction, retains only variation orthogonal to protected attributes. This residualized score preserves legitimate, non-protected (residual) heterogeneity that firms or policymakers may wish to act upon, while discarding the protected component that drives unfair disparities. As such, it supports compliance with fairness goals or regulatory constraints without sacrificing all the benefits of personalized scoring. In this sense, the metric R^2_{prot} is not merely diagnostic but also *actionable*: it provides both a measure of fairness concerns and a principled procedure for their removal.

To build intuition, consider the extreme cases. When $R^2_{\text{prot}} = 0$, $g(\zeta) = \mathbb{E}[\zeta]$ and residualization has no effect. In this case, $\phi^{\text{res}} = \phi - \mathbb{E}[\zeta]$ and the distribution of ϕ is preserved since no protected variation is present. When $R^2_{\text{prot}} = 1$, all heterogeneity is protected and $\phi^{\text{res}} = 0$. In this case, residualized scores collapse to zero and score-based personalization is impossible; once protected heterogeneity is removed, there is no heterogeneity left to be leveraged. In other words, *residual heterogeneity* $\mathbb{V}(\phi^{\text{res}})$ from Equation (2) captured by ϕ^{res} represents the share of variation that remains available for equitable resource allocation.

Acting on residual heterogeneity through the use of ϕ^{res} exhibits desirable properties connected to existing fairness definitions. We discuss these, and additional characteristics, in Section 4.

4 Residual Heterogeneity

In this section, we discuss the benefits and limitations of using residual heterogeneity for fair decision-making. We first establish the relationship between decisions based on the residualized score ϕ^{res} and common definitions of individual and group fairness. Specifically, we show that, when a score ϕ contains protected heterogeneity, deploying the residualized score ϕ^{res} ensures that decisions are both invariant to protected attributes and are weakly more balanced across protected groups, thereby providing operational guarantees of both individual and group fairness. We then discuss the limitations of residualization and highlight additional governance-related benefits.

4.1 Connections to Classical Fairness Definitions

Individual Fairness

Individual fairness requires that individuals who are similar in their non-protected attributes $\mathbf{X}$ receive similar treatment. A natural operationalization of this idea is $\mathbf{Z}$-invariance: holding $\mathbf{X}$ constant, variation in the protected attributes $\mathbf{Z}$ should not affect the score. Under residual heterogeneity, where $\zeta = (\mathbf{Z}, h(\mathbf{X}))$ and $h(\mathbf{X}) = \mathbb{E}[\mathbf{Z} \mid \mathbf{X}]$, the residualized score satisfies

$$\mathbb{E}[\phi^{\text{res}} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}] \;=\; \mathbb{E}[\phi^{\text{res}} \mid \mathbf{X} = \mathbf{x}] \qquad \forall\, \mathbf{x}, \mathbf{z},$$

which ensures that protected attributes do not explain systematic variation in ϕ^{res}, conditional on $\mathbf{X}$. Consequently, the residualized score for an individual with protected attributes $\mathbf{Z}$ would remain unchanged under any counterfactual protected status $\mathbf{Z}'$, such that

$$\phi^{\text{res}}(\mathbf{X}, \mathbf{Z}) = \phi^{\text{res}}(\mathbf{X}, \mathbf{Z}').$$

Note that this property does not imply that $\phi^{\text{res}}(\mathbf{X}, \mathbf{Z})$ equals $\mathbb{E}[\phi \mid \mathbf{X}]$. Rather, ϕ^{res} is the component of ϕ that is orthogonal to all information contained in ζ and individually fair by construction.

This property implies that individuals who are identical in their non-protected attributes are treated identically when decisions are based on ϕ^{res}. For instance, if gender or race are designated as protected attributes in a creditworthiness assessment, two individuals with the same income and credit history but differing in gender or race will have identical residualized scores. Consequently, decisions derived from ϕ^{res} satisfy the criterion of individual fairness.

Statistical Parity

Statistical parity is typically defined for a binary decision $D \in \{0, 1\}$ and a categorical protected attribute Z, requiring that the proportion of individuals receiving a positive outcome be equal across groups defined by Z. In our case, however, fairness is assessed at the pre-decision stage, before any threshold is applied to the score. Because we do not observe the realized decision, we adopt a stricter formulation of statistical parity (e.g., Chzhen et al. 2020, Jiang et al. 2020). This formulation requires that allocation rates be equal across protected groups at *every possible*

decision threshold q:

$$\mathbb{P}(\phi > q \mid Z = z) = \mathbb{P}(\phi > q) \qquad \forall\, z, q.$$

This threshold-independent statistical parity notion ensures that fairness evaluation does not depend on an arbitrary choice of cutoff, an important property since operational thresholds can vary across applications and over time.

Using this definition, we show that replacing the original score ϕ with the residualized score ϕ^{res} guarantees that the upper bound on the maximum group disparity engendered by any threshold-based decision rule is no larger than before. In other words, residualization weakly improves statistical parity across protected groups, regardless of the threshold applied.

The complete derivation of this result appears in the Appendix; the main intuition is as follows. Because fairness is evaluated across all possible thresholds, the relevant comparison is between the *entire score distributions* of scores for different groups, not their outcome at a single cutoff. The maximum possible disparity across thresholds corresponds to the *total variation* (TV) distance between the group-conditional score distributions. Information-theoretic arguments then relate this TV distance to the amount of information the score carries about the protected attribute, a dependence that can be quantified by the *mutual information* $I(Z;\phi)$.

Residualization reduces this dependence: by construction, $I(Z;\phi^{\text{res}}) \leq I(Z;\phi)$. It therefore follows that the upper-bound on worst-case disparity is weakly smaller when decisions are based on ϕ^{res}. Consequently, residualization thus provides an operational guarantee of group fairness: it limits the extent to which any downstream decision can differentially allocate outcomes across protected groups.

4.2 Limitations and Additional Characteristics

What residualization does *not* guarantee about fairness. Residualization removes protected heterogeneity, guarantees individual fairness, and weakly improves group fairness, but it does not by itself ensure other fairness notions. For group fairness, residualization does *not* guarantee that different groups will have identical score distributions. Existing differences in the distribution of $\mathbf{X}$ across protected groups — for example, educational attainment or income — can still lead to differences in the distribution of ϕ^{res} across Z. Threshold-invariant statistical parity can only be

achieved when the distributions of non-protected attributes $\mathbf{X}$ are identical across groups, in which case ϕ^{res} is independent of Z.

Additionally, residualization does not guarantee other fairness notions such as calibration (equal outcome rates conditional on predicted scores), predictive parity (equal precision across groups), or utility optimality. These issues are conceptually distinct from our focus and can instead be addressed through complementary tools such as calibration procedures, post-hoc threshold selection, cost-sensitive objective functions, or multi-metric evaluation.

Efficiency Costs of Residualization. A natural implication of residualization is that, by construction, it removes some of the heterogeneity that drives predictive or allocative performance. Because ϕ^{res} excludes variation explained by the protected attributes and their proxies, it necessarily leverages a smaller share of the total heterogeneity available in ϕ. As a result, decisions based on ϕ^{res} will typically be less efficient than those based on the original score, yielding lower predictive accuracy or allocative efficiency. The magnitude of this efficiency loss would depend on the proportion of residual heterogeneity that remains after residualization. Quantifying this trade-off between fairness and efficiency is ultimately an empirical question that depends on the data environment and domain-specific context, and is therefore beyond the scope of this paper.

Dependence on model and data. The practical performance of our diagnostic and mitigation steps depends on the fidelity of the score model ϕ and the nuisance estimators for $h(\cdot)$ and $g(\cdot)$. Estimation error or drift can lead to imperfect orthogonalization in finite samples; we therefore recommend routine audits (e.g., testing predictability of $\mathbf{Z}$ from ϕ^{res} and from any post-processed score) and, if needed, additional alignment or independence-penalized training.

Compliance with privacy. An additional benefit of leveraging residual heterogeneity is that it yields a *privacy-preserving* score. Because ϕ^{res} is orthogonal to $\mathbf{Z}$ (and to proxy channels $h(\mathbf{X})$), it weakly reduces information about protected attributes. Decisions based on ϕ^{res} therefore carry less information about protected characteristics, aligning the approach with privacy regulations that restrict the use or disclosure of sensitive information. Strong privacy requires that $\mathbf{Z}$ is not predictable from the released score; in practice, we advise adversarial or holdout audits of $\mathbf{Z}$-predictability and,

if needed, distributional alignment or explicit independence penalties to drive leakage toward zero. This is not a fairness guarantee per se, but it represents an important governance advantage.

Taken together, these properties underscore the complementary role of residual heterogeneity in our framework. It can be leveraged as a constructive solution that can satisfy compliance with classical fairness definitions and is actionable through the residualization of ϕ. Beyond fairness, leveraging residual heterogeneity also offers ancillary compliance benefits, making it a practical tool for managers and regulators concerned with both equity and governance.

5 Implications and Discussion

This paper introduces *protected heterogeneity* as a variance-based framework for understanding and evaluating algorithmic fairness. We formalize a corresponding metric, R^2_{prot}, which quantifies the share of variation in algorithmic scores that is explained by protected attributes and their proxies. The measure is bounded, interpretable, and threshold-invariant, and it enables stakeholders to audit, compare, and mitigate fairness concerns in algorithmic systems without relying on post-hoc thresholded decisions. We show how leveraging residual heterogeneity through ϕ^{res} can be used to attenuate protected variation while preserving the benefits of personalization in decision-making. These residualized scores are both individually fair and weakly more balanced across protected groups relative to the original algorithmic scores.

The simplicity and generality of R^2_{prot} make it a practical diagnostic for regulators, corporate stakeholders, and institutional auditors. In regulatory settings, R^2_{prot} provides a model-agnostic and threshold-independent summary of group- and individual-level disparities. It can be applied to discrete, categorical, and continuous protected attributes of any dimension, and measured at any point in the algorithmic lifecycle. In audit contexts, our metric enables transparent documentation of fairness properties, offering a single, interpretable quantity that summarizes the extent to which outcomes may reflect protected status.

Because R^2_{prot} is expressed as a percentage of explained variation, it is readily comparable across domains (e.g., credit scoring vs. hiring) and across scoring systems. This comparability supports the development of domain-specific benchmarks and legal standards. In this way, regulators can

treat R^2_{prot} as a quantitative yardstick for compliance, firms can integrate residualization into model deployment to reduce liability and reputational risk, and auditors can incorporate the measure into standardized fairness dashboards.

Related work has explored the idea of using a threshold ε to define acceptable levels of imbalance (Nabi and Shpitser 2018). Our metric lends itself naturally to such extensions: values of R^2_{prot} falling below a policy-defined threshold could be deemed sufficiently fair in legal, ethical, or managerial terms. This opens the door for R^2_{prot} to serve as a policy-relevant fairness constraint in algorithm certification or risk assessments.

Like any statistical diagnostic, R^2_{prot} has limitations. The metric assumes that protected attributes $\mathbf{Z}$ are observed and accurately measured. In many real-world settings, protected characteristics may be noisy, missing, or only partially observed, limiting the reliability of the estimate. In addition, correct modeling approaches must be used such that the estimated protected heterogeneity is not inflated or misleading.

It is important to emphasize that R^2_{prot} is a *diagnostic tool* for quantifying fairness-related variation, not as a mechanism for determining what constitutes an acceptable level of bias. While R^2_{prot} provides a clear signal of protected variation, it does not distinguish between different sources of heterogeneity. For instance, a protected attribute may be highly predictive of outcomes due to systemic inequality in access or opportunity. Whether such associations should be preserved or removed is a normative and context-dependent question. Accordingly, this paper does not prescribe what level of protected heterogeneity is acceptable, nor does it take a position on when residualization should be applied. These decisions depend on domain-specific legal, ethical, and operational considerations and should remain with policymakers, regulators, and institutional decision-makers. Our protected heterogeneity framework provides both the measurement and remediation infrastructure; decisions about thresholds for action rest with policymakers, regulators, and institutions.

Our framework opens several promising avenues for future research. A first direction concerns methodological extensions. A clear methodological direction would be to explicitly study the trade-offs between equity and performance, linking R^2_{prot} to measures of predictive accuracy, utility, or welfare (cf. Fu et al. (2022)). Such analysis could provide systematic guidance on how much protected variation can be removed before predictive power is substantially compromised. Connecting fairness and utility also sets the stage for elicitation of individual preferences on fairness,

where applicable. Similar to consequentialist frameworks that formalize fairness in terms of downstream welfare or utility (Chohlas-Wood et al. 2024), our approach invites analysis of how removing protected heterogeneity affects both predictive accuracy and societal outcomes.

A second direction involves empirical investigation. Benchmarking typical ranges of R^2_{prot} across industries and domains would parallel the role of predictive accuracy benchmarks in applied machine learning, providing reference points for fairness-related variance that would aid both practitioners and regulators. Another frontier concerns settings with partial or noisy observability of $\mathbf{Z}$, where proxies must be inferred under uncertainty. Studying how protected heterogeneity behaves in such environments would enhance the applicability of our framework to real-world data limitations. Moreover, future empirical work could examine the trade-off between fairness and efficiency introduced by residualization, quantifying how predictive or allocative performance changes with the share of residual heterogeneity across different domains and data environments.

Finally, our framework has direct implications for policy and practice. Because R^2_{prot} is simple to compute and directly interpretable, it can be incorporated into standardized fairness dashboards, risk assessments, and certification regimes. Policy-driven extensions could set acceptable thresholds for R^2_{prot} analogous to statistical tolerances in audit practice, providing regulators with actionable benchmarks and offering firms clear targets for compliance. Embedding R^2_{prot} into governance processes would strengthen accountability while preserving the benefits of personalization.

By reframing fairness in terms of protected heterogeneity, our approach provides a single, interpretable diagnostic and a constructive remedy. In doing so, we aim to offer a common language for managers, regulators, and researchers to evaluate, compare, and govern algorithmic personalization, while charting a foundation for future methodological, empirical, and policy advances. We hope this variance-based perspective can help move the field toward greater clarity, accountability, and actionable fairness in practice.

References

Aparicio D, Misra K (2023) Artificial intelligence and pricing. Sudhir K, Toubia O, eds., *Artificial intelligence in marketing*, 103–124 (Emerald Publishing Limited).

Ascarza E, Israeli A (2022) Eliminating unintended bias in personalized policies using bias-eliminating adapted trees (beat). *Proceedings of the National Academy of Sciences* 119(11):e2115293119.

Azadkia M, Chatterjee S (2021) A simple measure of conditional dependence. *The Annals of Statistics* 49(6):3070–3102.

Barocas S, Hardt M, Narayanan A (2023) *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press).

Bénesse C, Gamboa F, Loubes JM, Boissin T (2024) Fairness seen as global sensitivity analysis. *Machine Learning* 113(5):3205–3232.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021) Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50(1):3–44.

Beutel A, Chen J, Doshi T, Qian H, Woodruff A, Luu C, Kreitmann P, Bischof J, Chi EH (2019) Putting fairness principles into practice: Challenges, metrics, and improvements. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 453–459.

Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC (2022) A clarification of the nuances in the fairness metrics landscape. *Scientific reports* 12(1):4209.

Chohlas-Wood A, Coots M, Zhu H, Brunskill E, Goel S (2024) Learning to be fair: A consequentialist approach to equitable decision making. *Management Science* .

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2):153–163.

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020) Fair regression with wasserstein barycenters. Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H, eds., *Advances in Neural Information Processing Systems*, volume 33, 7321–7331 (Curran Associates, Inc.), URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf`.

Corander J, Remes U, Koski T (2021) On the jensen-shannon divergence and the variation distance for categorical probability distributions. *Kybernetika* 57(6):879–907.

Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S (2023) The measure and mismeasure of fairness. *Journal of Machine Learning Research* 24(312):1–117.

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Fu R, Aseri M, Singh PV, Srinivasan K (2022) "un" fair machine learning algorithms. *Management Science* 68(6):4173–4195.

Fu R, Huang Y, Singh PV (2020) Ai and algorithmic bias: Source, detection, mitigation and implications. *Detection, Mitigation and Implications (July 26, 2020)* .

Gelman A (2005) Analysis of variance: Why it is more important than ever. *The Annals of statistics* 33(1):1–31.

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29.

Jiang R, Pacchiano A, Stepleton T, Jiang H, Chiappa S (2020) Wasserstein fair classification. Adams RP, Gogate V, eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, 862–872 (PMLR), URL `https://proceedings.mlr.press/v115/jiang20a.html`.

Kozodoi N, Jacob J, Lessmann S (2022) Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research* 297(3):1083–1094.

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. *Advances in neural information processing systems* 30.

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54(6):1–35.

Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application* 8(1):141–163.

Mukherjee D, Yurochkin M, Banerjee M, Sun Y (2020) Two simple ways to learn individual fairness metrics from data. *International conference on machine learning*, 7097–7107 (PMLR).

Nabi R, Shpitser I (2018) Fair inference on outcomes. *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453.

Rafieian O, Yoganarasimhan H (2023) Ai and personalization. Sudhir K, Toubia O, eds., *Artificial intelligence in marketing*, 103–124 (Emerald Publishing Limited).

Appendix: Residualization and Threshold-Independent Statistical Parity

This appendix derives the theoretical foundation for the claim that residualization reduces the dependence of the score on protected attributes and, consequently, bounds the maximum disparity that can arise from any threshold-based decision rule.

We proceed in four steps. First, we express the maximum threshold disparity as the Kolmogorov-Smirnov distance between the distributions of two subgroups of a protected attribute. Second, we show that this KS distance is bounded by the total variation (TV) distance. Third, we relate TV to the mutual information between the score and the protected attribute, which quantifies their statistical dependence. Finally, we demonstrate that residualization, by construction, weakens this dependence, thereby tightening the upper bound on the maximum possible disparity.

Setup and Notation

Let $Z \in \{0, 1\}$ denote a binary protected attribute (the argument extends naturally to multi-category Z). Let ϕ denote a real-valued score. We denote by $P_z = \mathcal{L}(\phi \mid Z = z)$ the conditional probability law of the score given group z, and by P the overall mixed distribution of ϕ, defined as

$$P = (1 - \pi)P_0 + \pi P_1, \qquad \text{with } \pi = \mathbb{P}(Z = 1). \tag{5}$$

Then, for any threshold $q \in \mathbb{R}$, let $F_z(q) = \mathbb{P}(\phi \leq q \mid Z = z)$ denote the group-conditional cumulative distribution functions (CDFs).

From Threshold Parity to Kolmogorov-Smirnov Distance

At a fixed threshold q, the *statistical parity gap* is defined as

$$\Delta(q) = |\mathbb{P}(\phi > q \mid Z = 1) - \mathbb{P}(\phi > q \mid Z = 0)|.$$

Since $\mathbb{P}(\phi > q \mid Z = z) = 1 - F_z(q)$, this parity gap can equivalently be expressed in terms of the cumulative distribution functions (CDFs) as $\Delta(q) = |F_1(q) - F_0(q)|$. Maximizing the parity gap over all possible thresholds yields the Kolmogorov-Smirnov (KS) distance between the two

group-conditional score distributions:

$$\sup_q \Delta(q) = \sup_q |F_1(q) - F_0(q)| = D_{\mathrm{KS}}(P_0, P_1).$$

Therefore, the maximum threshold-based disparity — representing the largest possible difference in positive prediction rates between protected groups — is precisely the KS distance between their conditional score distributions. In the next section, we show that this quantity admits a theoretical upper bound and, importantly, that this bound weakly decreases when the residualized score ϕ^{res} is used instead of the original score ϕ.

KS Distance Bounded by Total Variation

While the Kolmogorov–Smirnov (KS) distance measures the maximal difference between the two cumulative distribution functions F_0 and F_1, the *total variation* (TV) distance measures the maximal difference across all measurable subsets of the support. Formally,

$$||P_1 - P_0||_{\mathrm{TV}} = \sup_{A \in \mathcal{B}(\mathbb{R})} |\mathbb{P}_1(A) - \mathbb{P}_0(A)|$$

where $\mathcal{B}(\mathbb{R})$ is the Borel σ-field on $\mathbb{R}$.

Because every threshold-based set $B_q = \{\phi > q\}$ is a measurable subset of $\mathbb{R}$, the KS distance is necessarily bounded above by the total variation distance:

$$D_{\mathrm{KS}}(P_0, P_1) = \sup_q |\mathbb{P}_1(B_q) - \mathbb{P}_0(B_q)| \leq \sup_{A \in \mathcal{B}(\mathbb{R})} |\mathbb{P}_1(A) - \mathbb{P}_0(A)| = ||P_1 - P_0||_{\mathrm{TV}}.$$

From Total Variation to Mutual Information

The total variation distance can, in turn, be related to the *mutual information* between the protected attribute Z and the score ϕ, denoted $I(Z; \phi)$. Mutual information quantifies the overall statistical dependence between these two variables—that is, the extent to which the score ϕ reveals information about the protected attribute Z.[6]

[6] $I(Z; \phi) = 0$ if and only if ϕ and Z are independent.

Given the mixed distribution of ϕ defined in Equation (5), the mutual information can be expressed as

$$I(Z;\phi) = (1-\pi)D_{\mathrm{KL}}(P_0||P) + \pi D_{\mathrm{KL}}(P_1||P),$$

where $D_{\mathrm{KL}}(\cdot||\cdot)$ is the Kullback-Leibler divergence.

A classical inequality (e.g., Corander et al. 2021) links total variation distance to mutual information: $I(Z;\phi) \geq \frac{\pi(1-\pi)}{2}||P_1 - P_0||^2_{\mathrm{TV}}$. Rearranging gives

$$||P_1 - P_0||_{\mathrm{TV}} \leq \sqrt{\frac{2I(Z;\phi)}{\pi(1-\pi)}}.$$

This bound shows that the potential for group-level disparity across thresholds is fundamentally limited by the amount of information that the score carries about the protected attribute Z. When $I(Z;\phi)$ is small, the conditional distributions P_1 and P_0 must be close, and thus the largest possible statistical parity difference across thresholds must also be small.

Finally, combining this bound with the previous inequality yields:

$$\sup_q \Delta(q) = D_{\mathrm{KS}}(P_0, P_1) \leq ||P_1 - P_0||_{\mathrm{TV}} \leq \sqrt{\frac{2I(Z;\phi)}{\pi(1-\pi)}},$$

establishing a direct link between the worst-case threshold disparity, overall distributional separation, and the underlying informational dependence between ϕ and Z.

Effect of Residualizing ϕ on Mutual Information

Let ϕ^{res} denote the residualized score, obtained by removing variation in ϕ that is explained by the protected attributes and their proxies. Formally, ϕ^{res} is constructed through a function $g : \mathbb{R}^{2d_z} \to \mathbb{R}$ such that $\phi^{\mathrm{res}} = g(\phi, \zeta)$, where ζ represents information derived from Z and its correlated features.

Although the strict data-processing inequality does not apply here (since Z explicitly enters the construction of ϕ^{res} through ζ), a well-specified residualization procedure that removes the variation explained by protected attributes Z and their proxies $h(\mathbf{X})$ implies that $I(Z;\phi^{\mathrm{res}}) \leq I(Z;\phi)$. Intuitively, by purging from ϕ all variation attributable to Z and its correlated features, the residualized score ϕ^{res} necessarily contains less information about Z than the original ϕ.

Applying the chain of inequalities derived earlier to both the original and the residualized scores gives

$$\sup_q \Delta^{\text{res}}(q) = D_{\text{KS}}(P_0^{\text{res}}, P_1^{\text{res}}) \leq ||P_1^{\text{res}} - P_0^{\text{res}}||_{\text{TV}} \leq \sqrt{\frac{2I(Z;\phi^{\text{res}})}{\pi(1-\pi)}} \leq \sqrt{\frac{2I(Z;\phi)}{\pi(1-\pi)}}.$$

Therefore, the worst-case threshold disparity under the residualized score cannot exceed that of the original score.

By construction, residualization reduces the dependence between the score and protected attributes. Therefore, it provides a theoretical guarantee that fairness, in the sense of maximum disparity under threshold-independent statistical parity, cannot deteriorate as a result of applying the residualization procedure.