Life of Research · Published 2 November 2026 · DOI 10.5281/zenodo.23871653

A Stable Attribution Is Not a Stable Error Rate: What Feature-Attribution Monitoring Can See Under Drift, Why 'More Sensitive Than the Inputs' and 'Silent While Fairness Fails' Are Both Correct, and the Label-Free Statistic the Fairness Alarm Needed

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Deployed classifiers are often monitored without labels, and feature attributions are one of the signals watched: a production monitoring service alerts when the ranking of SHAP attributions moves, and one line of work argues that the explanation space is more sensitive than the input space to shifts that matter. A 2026 study reports the other face: on COMPAS under injected feature noise, TreeSHAP explanations stayed largely stable while the protected group's false positive rate rose from 37.84 to 70.39 percent. In this analysis the author reviews and synthesises published results to show that the two findings concern different objects. Per-instance attributions, the distribution of attribution vectors, a top-k ranking of mean absolute attributions and an attribution of the loss are four instruments; only the last uses labels, and only the second sees everything the prediction distribution sees. The paper orders the label-free observables into a ladder, derives which group fairness quantities each rung can read, and shows that a group false positive rate is readable from none of them without an assumption about how outcomes given features changed. The anchor study's own numbers add two results: its disparity widened as a difference and narrowed as a ratio, and, with labels fixed, its false-positive rises imply rises in each group's positive-prediction rate, which needs no labels to compute, unless a group has almost no true negatives. The paper gives a decision procedure for choosing and reading a label-free fairness monitor and names the studies that would settle what remains open.

Keywords: model monitoring; distribution shift; feature attribution; SHAP; algorithmic fairness; label-free performance estimation; concept drift; equalized odds

1. Introduction

A classifier that scores loan applicants, defendants or emergency admissions is trained on one population and then used on another, and the labels that would reveal whether it still works arrive late or never (Zliobaite, 2010; Gama et al., 2014; Finlayson et al., 2021). The standard response is monitoring without labels: watch the inputs, watch the scores, and hope that whatever harms the model leaves a trace there (Rabanser et al., 2018; Kivimaki et al., 2024; Subasri et al., 2025). Feature attributions have become one of the traces people watch. Lundberg et al. (2019) decomposed a hospital model's loss over a simulated deployment into per-feature SHAP contributions and exposed a deliberately introduced pipeline error that the aggregate loss did not reveal. Duckworth et al. (2021) tracked SHAP values of an emergency-admission model through the first COVID-19 wave. Amazon's SageMaker Model Monitor ships an attribution-drift check that compares the ranking of features by attribution in training and in production and alerts when a normalised ranking score falls below 0.9, on the stated rationale that "a drift in feature attribution can provide an early signal of potential degradation in model accuracy without the need to collect ground truth labels" (Nigenda et al., 2021). Mougan et al. (2023) go further and argue that a detector trained on the distribution of explanation vectors is more sensitive to shifts that change model behaviour than detectors on the inputs or on the predictions.

A 2026 study by Zafar, El-Sharif and Khan presents the other face. Under injected covariate noise on five tabular datasets, the five features ranked highest by mean absolute TreeSHAP value frequently kept both their membership and their order while ROC AUC fell; on COMPAS, a random forest's false positive rate for the protected group rose by 32.55 points under strong drift, the privileged group's by 26.00, and one of the two features carrying the drift, age, was not among the top five attributions before or after (the study does not report where race ranked). The authors conclude that "explanation stability is an insufficient proxy for fairness" and propose a framework (FMMO) that adds a global decision-tree surrogate, a multivariate drift detector and confidence-based performance estimation (Zafar et al., 2026).

Read as headlines, the two positions conflict: one says attributions see more than inputs, the other says attributions can stay still while a model becomes markedly more punitive toward one group. In this analysis the author reviews and synthesises published results to show that the conflict is mostly an artefact of

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

four different objects travelling under one name, and that the part which is not an artefact is a limit no label-free instrument escapes, attribution-based or otherwise. The question this paper answers is narrower than "do explanations detect drift": it is which group fairness quantities a label-free monitor can read at all, from which statistic, and under which assumption about the shift; and therefore what the silence of an attribution monitor licenses an auditor to conclude.

The argument has four steps. First, Section 2 reads the anchor result at the resolution of its tables and finds three things the headline does not carry: the instrument that stayed stable was a top-k ranking of mean absolute attributions, not the attributions; the fairness deterioration is a widening by one disparity measure and a narrowing by another; and the false-positive rises, with labels held fixed, imply rises in each group's positive-prediction rate that no labels are needed to see. Second, Sections 4 and 5 separate the objects called "the attribution" and order the label-free observables into a ladder in which each rung is a function of the one above. A change visible at a lower rung is necessarily visible at every higher one, so the full distribution of attribution vectors sees at least what the prediction distribution sees, while a ranking of mean attributions is a lossy summary that sees neither reliably. Third, Section 6 places each group fairness quantity on that ladder. Demographic-parity-type quantities sit on a label-free rung. Group false positive and true positive rates sit on no label-free rung: a shift in how labels depend on features, with the feature distribution held fixed, changes them while every label-free statistic stays exactly where it was. Section 7 then shows that each label-free estimator of error rates examined here closes that gap with an assumption, and that the one systematic measurement of natural tabular shifts located for this paper found the assumption-breaking kind dominant among large degradations (Wang et al., 2023a). Fourth, Sections 8 and 9 take up two problems any fairness alarm has before it has an attribution problem: whether it was ever calibrated against shifts that do no harm, and whether "deteriorated" names a metric.

The strongest objection deserves a place on the first page. The anchor study itself says local attributions "cannot be used alone", and the explanation-shift authors concede that their method "is not able to detect concept shifts" (Mougan et al., 2023). If both camps already agree that attributions are partial, what is left to analyse? The answer is that neither side states where the partiality falls, and the location matters for practice. An auditor told only that attributions are "insufficient" will add more label-free signals, as FMMO does; this paper argues that for error-rate disparities, adding label-free signals cannot close the gap, and that for selection-rate disparities, attributions were never needed. The useful conclusion is a map of which rung answers which question, not a verdict on SHAP.

The paper draws on 70 sources: the anchor study's tables, the explanation-shift and attribution-monitoring literature, the dataset-shift and label-free performance-estimation literature, and the fairness-under-shift literature. It contributes a derivation of which fairness quantity each shift type moves and which monitor can see it (Table 1), a consolidated evidence table (Table 2), two figures re-plotting the anchor study's published numbers (Figures 1 and 2), and a decision procedure (Algorithm 1).

2. The anchor result, read at the resolution of its tables

2.1 What was injected and what was watched

Zafar et al. (2026) train random forests and XGBoost models on COMPAS, German Credit, Communities and Crime, a Consumer Expenditure Survey extract (CES), and a synthetic dataset, then perturb the 30 percent test split and evaluate without retraining. Two perturbations are used: additive Gaussian noise with zero mean and identity covariance on normalised features, and a multiplicative rescaling of each feature by a factor drawn from U(1.5, 3.5) plus a signed mean shift whose size is "calibrated so that the performance of the model is substantially reduced". A further condition, CES-NA, trains on 2017-2019 survey years and tests on 2020, a natural shift. For each condition the authors compute mean absolute TreeSHAP values on the clean and perturbed test sets, take the top k features (k = 3, 5, 10), and compare the two lists by Jaccard overlap and by a position-wise rank agreement (one minus the normalised Hamming distance). They do the same for a decision-tree global surrogate fitted to the model's predictions, which for the drifted condition is fitted on the drifted test set.

Two features of this protocol matter for everything that follows. The watched statistic is not the attributions; it is a ranking of the k largest mean magnitudes, a summary that discards sign, spread, joint structure and every feature outside the top k. And the injected perturbation is not covariate shift in the sense the dataset-shift literature uses. The paper defines covariate drift as P(X) changing while "the dependent variable P(Y) remains the same", whereas the standard definition holds the conditional P(Y | X) fixed (Shimodaira, 2000; Moreno-Torres et al., 2012). Adding noise to recorded features while leaving outcomes untouched changes the conditional distribution of the outcome given the features the model sees: two applicants with the same recorded features after perturbation no longer share a distribution of true outcomes. In the vocabulary of formal drift characterisation (Webb et al., 2015), and in the real-versus-virtual distinction surveyed by Hinder et al. (2024), the anchor's injected drift is a mixture of virtual and real drift. That matters in Section 7, where every label-free error estimator turns out to assume the

real part away.

2.2 The fairness numbers

The fairness claim rests on Table 3 of the anchor study, reproduced in Figure 1(a). For a random forest on COMPAS, the protected group's false positive rate is 37.84 percent on clean data, 44.50 under low drift and 70.39 under high drift; the privileged group's is 10.17, 13.23 and 36.17. Group accuracy moves from 63.93 to 63.06 and 59.78 for the protected group and from 67.98 to 67.70 and 63.63 for the privileged group. The authors summarise the high-drift row as a rise of 32.55 percent for the protected group against 26 percent for the privileged group and an "FPR gap widening by over 6%". Those figures are percentage-point differences: 70.39 minus 37.84 is 32.55, and the gap moves from 27.67 to 34.22 points, a widening of 6.55.

Figure 1(c) computes the same rows as a ratio. The protected group's false positive rate is 3.72 times the privileged group's at baseline, 3.36 times under low drift and 1.95 times under high drift. By the difference, the high-drift model is more discriminatory than the clean one; by the ratio, it is markedly less so. Neither reading is wrong. They are two disparity measures that rank the same three conditions in opposite orders, and the anchor study's conclusion that the model "drifted into a discriminatory state" holds under one of them. Section 9 returns to why a monitor has to declare which.

2.3 What the false-positive rises imply about a label-free quantity

The anchor's protocol perturbs features and leaves outcomes and group membership as they were. Under that protocol a simple identity links a group's false positive rate, its accuracy and its positive-prediction (selection) rate. Write p0 for the share of the group whose true label is negative. Accuracy is p0 times (1 - FPR) plus (1 - p0) times TPR; the selection rate is p0 times FPR plus (1 - p0) times TPR. Subtracting, the selection rate equals accuracy minus p0 plus 2 p0 FPR, and since p0 does not change when only features are perturbed,

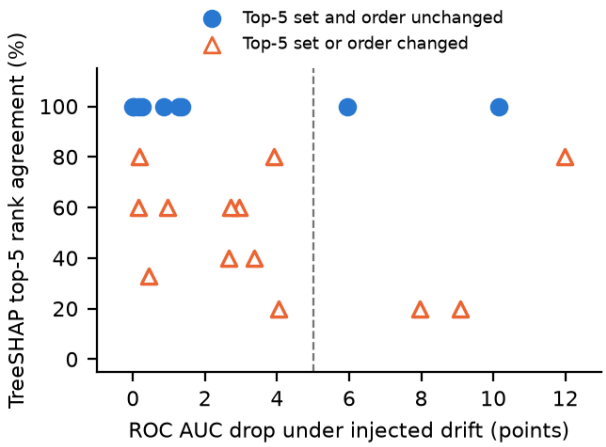
```
change in selection rate = change in accuracy + 2 x p0
x change in FPR.
```

Applied to the high-drift row, the protected group's selection rate changes by -4.15 + 65.10 p0 points and the privileged group's by -4.35 + 52.00 p0 points. Both are positive unless the group's negative share is below about 6 and 8 percent respectively. The anchor study does not report how its protected and privileged groups were defined or their base rates, so the size of the rise cannot be pinned from its tables alone. As an illustration only: if the split were Black and White defendants and the base rates matched those Chouldechova (2016) reports for the ProPublica data, 51 and 39 percent recidivism, then p0 would be 0.49 and 0.61, and both groups' selection rates would have risen by about 27 to 28 points. Under that illustrative reading the disparity in selection rates barely moves while each group's selection rate jumps.

The point does not depend on the illustration. A group selection rate needs the model's predictions and the group attribute, not the outcome. Whatever the base rates, a false-positive rise of 26 to 33 points accompanied by an accuracy fall of about 4 points forces each group's selection rate up unless almost nobody in the group is a true negative. The shift that produced what the anchor calls a "fairness blind spot" of local attributions was, under its own protocol, visible to the simplest label-free statistic a fairness monitor can compute: per-group selection rates must have risen sharply. Those rates register that the model changed; they cannot show whether the false-positive gap widened. The anchor did not report per-group selection rates, and its confidence-based performance estimator, which the authors report staying within its band on the CES Gaussian-noise condition, estimates overall performance rather than group rates. The anchor states that FMMO couples explanations with "group-specific performance estimation", but its method section describes, and its figures report, only an overall CBPE estimate.

2.4 What the stability table shows, and what it cannot

Figure 2 re-plots Table 4 of the anchor study: for each of 22 dataset, noise and model combinations, the drop in ROC AUC from the clean baseline in the anchor's Table 1 against the top-5 TreeSHAP rank agreement. In 9 of the 22 rows the top-5 features keep both their membership and their order. Two of those nine are among the five rows where ROC AUC falls by more than five points: XGBoost on CES under uniform noise (85.43 to 79.47) and the random forest on the synthetic data under Gaussian noise (94.71 to 84.55). The other three large-drop rows show the ranking changing. Among the nine rows where ROC AUC moves by less than one point, five keep the ranking intact and four do not.



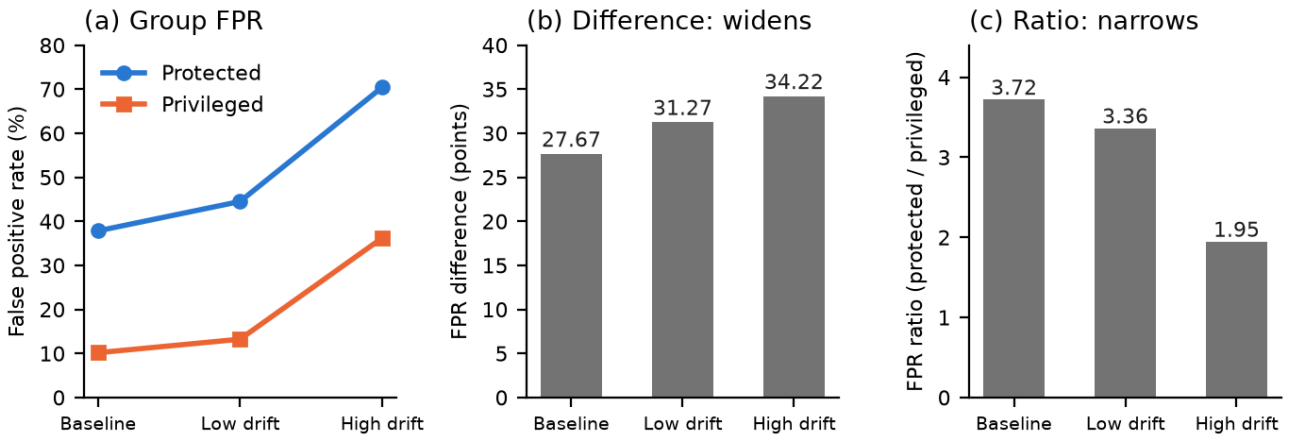


Figure 1. Group false positive rates for a random forest on COMPAS under the anchor study's injected drift, re-plotted from Table 3 of Zafar et al. (2026). Panel (a): protected and privileged group FPR at baseline, low and high drift. Panel (b): the difference between them, which widens from 27.67 to 34.22 points. Panel (c): the ratio, which falls from 3.72 to 1.95. Differences and ratios are computed by this paper from the published rates.

Figure 2. ROC AUC drop under injected drift against the rank agreement of the five features with the largest mean absolute TreeSHAP value, for 22 dataset, noise and model combinations, re-plotted from Tables 1 and 4 of Zafar et al. (2026). Filled circles mark rows where the top-5 set and order are unchanged; the dashed line marks a five-point drop. A stable top-5 ranking occurs at small and large performance losses alike.

The figure supports the anchor's narrow claim: a stable top-5 ranking is compatible with a large performance loss. It does not support a broader claim about attributions, for two reasons the table itself makes visible. The statistic discards most of the attribution distribution, so its stability says nothing about whether the attribution vectors moved. And the alarm design has no row in which nothing happened. The FMMO alert fires if an explainer's Jaccard overlap or rank agreement falls below its threshold, if the drift detector fires, or if estimated performance leaves its band; with the thresholds at exact agreement, every one of the 22 rows triggers, and in every row the surrogate's top-5 rank agreement is below 100. Because the surrogate for the drifted condition is a decision tree refitted on a different sample, some movement in its top-5 importance is expected from refitting alone, and the study reports no refit on an undrifted resample against which to measure it. The authors name the gap themselves: their experiments "focused on improving the detection of True Positive (TP) scenarios", at the risk of false alarms they did not test. Section 8 treats that gap as general.

The one natural shift in the study sharpens the point. In CES-NA, trained on 2017-2019 survey years and tested on 2020, the random forest's ROC AUC moves from 85.90 to 84.54 and its top-5 TreeSHAP set and order are unchanged; the XGBoost model's ROC AUC moves from 85.55 to 82.90 and its top-5 Jaccard overlap falls to 66.67 with rank agreement 40. The same shift, applied to two models of similar accuracy, leaves one ranking intact and rearranges the other. Ranking stability is therefore a joint property of the model and the summary statistic, not a property of the shift, and cannot be read as evidence about the shift on its own.

Two smaller points complete the reading. The anchor describes TreeSHAP as approximating the model "in the local vicinity by perturbing data around an instance". TreeSHAP computes Shapley values exactly from the tree structure rather than by sampling perturbations (Lundberg et al., 2018), and its values depend on which reference distribution stands in for absent features, a choice examined in Section 4. The mechanism offered for the stability is therefore not the mechanism of the method used. And the text reports the synthetic random forest's baseline ROC AUC as 94.41 where its Table 1 gives 94.71; Figure 2 uses the table value.

3. The other side: what attribution monitoring was claimed to do

The case for attribution monitoring has four published forms, and they make different claims.

Loss attribution. Lundberg et al. (2019) explain a model's loss rather than its prediction and track each feature's contribution to the loss over a simulated deployment of a hospital procedure-duration model. After the identifiers of two operating rooms were deliberately swapped two-thirds of the way through the data, the overall loss gave no sign of trouble while the room feature's loss contribution did; the same plots exposed a transient record-configuration problem and a slow drift in the atrial-fibrillation feature ($p = 5.4 \times 10^{-19}$). This is a strong result for localisation, and it uses labels: a loss is computed against observed outcomes. It is evidence that attributions help explain a performance change once the outcomes are in, not that they detect one before.

Attribution summaries as drift signals. Duckworth et al. (2021) trained an admission-risk model on pre-pandemic

emergency-department data, observed its AUROC fall from 0.856 to 0.826 during the first COVID-19 period, and showed that tracking each feature's SHAP values relative to its global importance gives a complementary drift measure and surfaces emerging risk factors. SageMaker's attribution drift check compares global attribution rankings by normalised discounted cumulative gain with a fixed 0.9 threshold (Nigenda et al., 2021). These are the anchor's instrument class: summaries of attribution magnitude, read for change.

Explanation-space two-sample tests. Mougan et al. (2022; 2023) train a classifier to distinguish attribution vectors computed on source data from those computed on new data and use its AUC as a shift statistic. Their Proposition 1 is that if two inputs receive different predictions they receive different Shapley attributions, which follows from efficiency: attributions sum to the prediction minus a baseline. They give constructions where univariate input tests miss a multivariate change that the attributions register, where inputs shift on features the model ignores and attributions correctly stay still, and where predictions stay put while attributions move. Wang et al. (2023b) combine feature attributions with a maximum mean discrepancy test for harmful shifts and report it more effective than existing detectors on more than twenty shifts across three image datasets.

Attribution-assisted diagnosis. A further body of work uses attributions or attribution-like decompositions to say why a distribution or a prediction distribution changed: GroupShapley for prediction drift (Shanbhag et al., 2021), causal-mechanism attribution of distribution change (Budhathoki et al., 2021), Shapley attribution of a performance change to component distributions (Zhang et al., 2022), decomposition of a performance drop into covariate and conditional terms (Cai et al., 2023), sparse attribution of concept shift to features (Lyu et al., 2025), localisation of shifted features by conditional tests (Kulinski et al., 2021), interpretable transport maps between distributions (Kulinski and Inouye, 2022), model-based drift explanation (Hinder et al., 2023), and change detection for stale local attributions in streams (Haug et al., 2022). Pelosi et al. (2025) survey this terrain systematically. Most of this work assumes a detected change and asks for its explanation; Zhang et al., Cai et al. and Lyu et al. explicitly need outcome data from the new environment.

The concessions within this literature are as important as its claims. Mougan et al. (2023) state that their method "is not able to detect concept shifts" and is "not able to predict the degradation of prediction performance under distribution shifts", and note that the same limitation applies to every method that monitors prediction performance without labels. Mougan and Nielsen (2022) add that input-shift detectors produce false positives when the data move and the model does not deteriorate. Haug et al. (2022) find that local attributions can become stale whenever the model is updated or the data-generating distribution changes. None of the four forms claims that a label-free attribution statistic certifies a group error rate. The claim the anchor refutes, that stable explanations mean stable fairness, is one this literature does not make, though the SageMaker rationale comes close for accuracy.

4. Four objects called "the attribution"

The debate runs together objects that have different information content.

Per-instance attributions. SHAP values (Lundberg and Lee, 2017) are the usual case. For a fixed model f and a fixed reference distribution, an attribution method maps each input x to a vector phi_f(x) of per-feature contributions. Which reference distribution fills in absent features is itself contested. Interventional and observational value functions give different attributions for the same model and input (Chen et al., 2020; Janzing et al., 2019); Sundararajan and Najmi (2019) catalogue several operationalisations of "the" Shapley value that disagree, including ones that credit features the model never reads; and Kumar et al. (2020) argue that fixing these choices requires causal assumptions the method does not supply. For monitoring the relevant point is simpler: if the reference set is held fixed, phi_f is a deterministic function of x, and any change in attributions on new data is a change in the inputs seen through f. If the reference set is re-estimated from deployment data, as some adaptive schemes do (John, 2025), attributions can move with no input change at all, and a monitor cannot tell the two apart unless it records which reference it used.

The distribution of attribution vectors. Pushing the deployment input distribution through phi_f gives a distribution over attribution vectors. This is the object Mougan et al.'s detector compares.

Summaries of attribution magnitude. The mean absolute attribution per feature, its ranking, the top-k set, and ranking scores such as NDCG are functions of the attribution distribution that discard sign, dispersion, dependence between features and, for top-k statistics, every feature outside the top k. This is the object FMMO and SageMaker monitor. It can be unchanged while the attribution distribution changes a great deal: a feature whose attribution flips sign for half the population, or whose spread doubles around the same mean magnitude, leaves a magnitude ranking intact. The anchor's own COMPAS case is of this kind at the extreme: the drifted age feature was outside the top five in both conditions, so no change in its attributions could register; the study does not report the rank of the other drifted feature, race.

Loss attributions. Attributing the loss rather than the prediction requires outcomes. It belongs on the label side of every distinction drawn below.

The failures documented in the explanation-robustness literature attach to different objects again. Adversarially scaffolded models can present innocuous attributions on the data distribution while discriminating (Slack et al., 2019); rationalisation can make an unfair model look fair (Aivodji et al., 2019); small input perturbations can change gradient explanations arbitrarily with the output held nearly fixed (Dombrowski et al., 2019); some saliency methods are insensitive to the model's parameters altogether (Adebayo et al., 2018); and for rich model classes, complete and linear attribution methods can do no better than random guessing at inferring model behaviour on several end-tasks (Bilodeau et al., 2022). Lakkaraju et al. (2020) report that post hoc explanations lack stability under distribution shift and propose adversarially robust ones. Those results bear on whether an attribution means what an auditor takes it to mean. The monitoring question is narrower: whether a change in some statistic of attributions tracks a change in a group's error rate. It can fail even with perfectly faithful attributions, and the next two sections show why.

5. A ladder of label-free observables

Fix a deployed model f, a fixed reference distribution for attributions, a decision threshold, and suppose the group attribute A is recorded at inference. The quantities a monitor can compute without outcome labels form a ladder, each rung a function of the one above it:

Rung 0 is the joint distribution of inputs and group, P(X, A). Rung 1 is the joint distribution of attribution vectors and group, P(phi_f(X), A); it is a function of rung 0, because phi_f is a fixed function of x. Rung 2 is the joint distribution of scores and group, P(f(X), A); it is a function of rung 1, because by efficiency the score equals a constant baseline plus the sum of the attributions. Rung 3 is the set of per-group selection rates, P(f(X) > t | A), a function of rung 2.

Summaries of attribution magnitude hang off rung 1 as a side branch: functions of the attribution distribution that are not comparable with rung 2.

Two consequences follow from the definitions. First, if nothing changes at a rung, nothing changes at any rung below it; equivalently, a change at a lower rung forces a change at every higher rung. In the population limit, a test on the full attribution distribution can therefore detect every change that a test on the score distribution or the selection rates can detect. This is the precise content of Mougan et al.'s Proposition 1 and the reason their explanation-space detector is at least as sensitive as an output-space one. Second, the side branch carries no such guarantee. A ranking of mean magnitudes can stay fixed while scores, selection rates and attributions all move, which is what Figure 2 shows in the two large-drop rows with intact rankings. The two published positions are thus about two different rungs: "more sensitive than the outputs" is a claim about rung 1, "stable while performance and fairness fail" is a claim about the side branch.

The ladder orders what can be detected in principle, not what a finite sample detects. A coarser statistic lives in fewer dimensions and can have more power at deployment sample sizes. Rabanser et al. (2018), comparing shift detectors on image benchmarks, found that testing the classifier's softmax outputs (black-box shift detection) was the best dimensionality reduction overall, that a domain classifier performed poorly with 100 or fewer samples, and that a multivariate test on unreduced inputs performed poorly. Their benchmark is MNIST and CIFAR-10, not tabular data, and it scores detection of any shift rather than of harmful shift. It nonetheless warns against reading the ladder as a power ranking: rung 1 sees more than rung 2 in principle and may see less in practice. No located study compares rung-1 and rung-2 detectors on tabular natural shifts at matched sample sizes for the purpose of detecting changes in group error rates.

The ladder has a floor, and the floor is the paper's central point. Every rung is a function of P(X, A). Nothing on it depends on the outcome Y. Any change in the joint distribution of (X, A, Y) that leaves P(X, A) fixed is invisible at every rung, to every label-free monitor, however sensitive.

6. Which fairness quantity lives on which rung

6.1 The placement

Group fairness criteria are functionals of the joint distribution of score, group and outcome (Hardt et al., 2016). Placing them on the ladder is a matter of reading off which variables each one uses.

Selection-rate quantities, including demographic parity gaps and the selection-rate ratio behind the 80 percent rule that Feldman et al. (2014) adopt from US employment guidance, use only predictions and groups. They sit on rung 3 and can be computed exactly, without labels, whenever the group attribute is available. A monitor for them needs no attributions and no estimation; attributions can help explain a change but are not needed to see it.

Error-rate quantities, including group false positive and true positive rates and the equalized odds and equal opportunity gaps built from them (Hardt et al., 2016), condition on the outcome. They are functionals of P(f(X) | Y, A), which requires P(X | Y, A). That distribution is not determined by P(X, A).

Predictive-value and calibration quantities condition on the prediction and ask about the outcome; they need P(Y | f(X), A). They are not determined by P(X, A) either.

The second and third placements are where the floor of Section 5 bites. A short construction shows it. Hold P(X, A) fixed and change P(Y | X, A) for one group, for

instance by raising the probability of a positive outcome in the region of feature space where the model scores high. The group's false positive rate, true positive rate and predictive value all change; P(X, A), the attributions, the scores and the selection rates do not change at all. No label-free statistic can distinguish the shifted deployment from the unshifted one. This is the group-level instance of a general impossibility: Garg et al. (2022) prove that identifying a model's accuracy on a new distribution from labelled source data and unlabelled target data is as hard as identifying the optimal target predictor, so every estimator rests on assumptions about the shift, stated or not. The label-shift estimator of Lipton et al. (2018) makes the same point from the other direction: label-free correction works when, and because, a structural assumption pins down the missing conditional.

6.2 What each kind of shift moves

The fairness-under-shift literature already states the invariances that decide what moves. Schrouff et al. (2022), working in a causal framing of shift, list them: a demographic parity gap is invariant when the model's inputs are independent of the shift given the group; an equalized odds gap is invariant when the inputs are independent of the shift given the group and the outcome; and the error is invariant when the outcome is independent of the shift given the inputs. Asiaee and Aryan (2026) prove the case of group-conditional prior probability shift, in which outcome prevalence changes differently across groups while P(X | Y, A) stays fixed: equalized-odds-type criteria are structurally invariant and demographic parity drifts, unavoidably for non-trivial classifiers, and under that shift target risk and fairness metrics are identifiable from source labels plus unlabelled target data. Biswas and Mukherjee (2020) treat fair classification under prior probability shift, and Chen et al. (2022) bound how far demographic parity and equalized odds violations can transfer under bounded covariate and label shift.

Table 1 assembles these results with the ladder. Each row is a shift type defined by which factor of the joint distribution changes; the columns say whether the label-free rungs move, whether selection-rate and error-rate disparities move, and what a label-free monitor can conclude.

Two rows deserve comment. The concept-shift row is the one the anchor's conclusion is really about, and on it the attribution monitor is no worse than any other: all label-free rungs are fixed, so every label-free alarm is silent, and the selection-rate gap is silent too while the error-rate gap moves. Mishler and Dalmasso (2022) show that this row is not exotic: in performative settings, where a predictor is deployed precisely to change outcomes such as recidivism or default, the shift it induces is a concept shift, and a predictor that met an observable fairness criterion at training can violate it at deployment. The feature-corruption row is the anchor's own experiment. It moves every label-free rung, which is why the drift detector fired and why per-group selection rates must have moved (Section 2.3), and it also moves the conditional of outcome given recorded features, which is why the error rates moved in a way a calibration-based estimate is not guaranteed to track.

6.3 What the placement implies for the anchor's claim

The anchor study frames the result as a blind spot of local explanations. The placement locates it more exactly. The quantity whose deterioration the study measures, a group false positive rate, sits below the label-free floor. The study itself notes that "the FPR itself cannot be computed directly" without labels and uses it only retrospectively. No label-free addition to the monitor, global surrogate included, can promise to flag that quantity's movement, because shifts exist that move it while leaving every label-free statistic fixed. Conversely, the experiment the study ran was not of that kind. Its perturbation moved every label-free rung, so a monitor of per-group selection rates or score distributions would have registered a large change. The silence of the attribution ranking was a property of the summary statistic chosen, on a shift that registered on the full-distribution label-free rungs (the anchor's univariate detector, another summary, missed the age drift).

The general lesson runs in both directions. For selection-rate disparities, an attribution monitor is the wrong instrument because a better one is free: compute the rates. For error-rate disparities, an attribution monitor is the wrong instrument because no label-free instrument is a right one without an assumption, and the assumption, not the instrument, carries the inference.

6.4 Where attributions earn their place

None of this makes attributions useless to a fairness monitor; it moves them from the alarm to the investigation that follows one. Three roles survive the placement.

The first is localisation after a label-free alarm. When per-group selection rates or score distributions move, the question that follows is which inputs drove the change. Attributions computed against a fixed reference set answer a version of it, and distribution-level decompositions answer it more directly: GroupShapley attributes a change in the prediction distribution between two slices to features (Shanbhag et al., 2021), Shapley decompositions of a chosen disparity attribute it to features even when the model does not read the protected attribute (Begley et al., 2020), and conditional tests can localise which features shifted at all (Kulinski et al., 2021). Here the full attribution distribution is the right object and the top-k magnitude ranking is not, for the reason in Section 4.

Table 1. Shift types, the fairness quantities they move, and what a label-free monitor can see. "Moves" means can change in general; "fixed" means invariant by derivation. Sources for the invariances are given in the last column.

Shift (what changes)	Label-free rungs	Selection-rate gap	Error-rate gap	Label-free reading	Source
Demographic: P(A)	P(A) moves	per-group fixed	per-group fixed	exact	Schrouff 2022
Group prior: P(Y given A)	move	moves	fixed	estimable	Asiaee 2026
Within-group covariate: P(X given A), P(Y given X, A) fixed	move	moves	moves	estimable if calibration transfers	Singh 2019; Bialek 2024
Concept: P(Y given X, A), P(X, A) fixed	all fixed	fixed	moves	none	Garg 2022; Mishler 2022
Feature corruption: recorded X perturbed, Y fixed	move	moves	moves	biased; conditional moved	Section 2.1
Compound	move	moves	moves	none without a model of the shift	Schrouff 2022; Wang 2023a

The second is localisation after labels arrive. Lundberg et al.'s (2019) loss attribution is the clearest published success of attribution monitoring, and it works because it is computed against outcomes. Applied separately within each group on a labelled audit sample, the same decomposition may indicate which features carry a change in a group's false positive rate; this use has not, as far as the author located, been evaluated for group error rates, and the Shapley attribution of performance change to component distributions (Zhang et al., 2022) and the covariate-versus-conditional decomposition (Cai et al., 2023) are closer to what such an evaluation would need.

The third is a check on the monitor's own assumptions. If attribution vectors on deployment data occupy regions the training data never reached, confidence-based error estimates in those regions rest on extrapolated calibration, which is the failure that distance-checked estimators (Roschewitz and Glocker, 2023) and the unseen-examples term in the decomposition of Cai et al. (2023) are designed to expose. In that role the attribution distribution is evidence about whether the borrowed assumption of Section 7 is plausible, not about the error rate itself.

7. The assumption every label-free error estimate borrows

7.1 The estimators and what they assume

Label-free performance estimation is an active area, and its methods divide by the assumption they borrow. Average confidence and confidence-based performance estimation (CBPE) estimate accuracy or confusion-matrix metrics by treating the model's predicted probabilities as calibrated; Kivimaki et al. (2024) prove that average confidence is an unbiased and consistent accuracy estimator under that calibration assumption and find that it often beats more complex estimators, though which estimator wins depends heavily on the case. Bialek et al. (2024) note that CBPE's reliance on calibration is its main weakness because calibration deteriorates under shift, and propose PAPE, which recalibrates by importance weights estimated between source and target inputs; it targets covariate shift and was evaluated on more than 900 dataset and model combinations from US census data. Average thresholded confidence (Garg et al., 2022), difference of confidences (Guillory et al., 2021), agreement-on-the-line (Baek et al., 2022) and distance-checked confidence (Roschewitz and Glocker, 2023) each estimate accuracy from unlabelled target data under their own conditions; Garg et al. report their method 2 to 4 times more accurate than prior methods on their benchmarks while proving that no method works without assumptions. For harmful-shift detection, Podkopaev and Ramdas (2021) build sequential tests with time-uniform guarantees on a risk such as accuracy, but assume that some labels arrive, possibly delayed; Amoukou et al. (2024) remove the labels by substituting the output of an error estimator trained on source data, with guarantees resting on an assumption about that estimator's behaviour in production that the authors say they do not expect to hold exactly, and in their illustrative image example the estimator's R-squared on source data was 0.35. MLDemon (Ginart et al., 2021) mixes unlabelled data with a budget of requested labels.

For group fairness specifically, Ji et al. (2020) estimate group metrics by fitting group-specific calibration maps on a small labelled sample from the audited population and applying them to plentiful unlabelled data; their gains come from the labelled sample. Asiaee and Aryan (2026) identify target fairness metrics without target labels under group-conditional prior shift. Singh et al. (2019) estimate accuracy and fairness on a shifted test set given a causal graph and the anticipated shift. Rezaei et al. (2020) and Havaldar et al. (2023) learn predictors that remain fair under covariate shift using unlabelled target samples, the latter introducing an asymmetric case in which one group's covariates shift far more than another's. An et al. (2022) find domain shifts harder than subpopulation shifts for transferring fairness, and Maity et al. (2020) show that enforcing fairness during training can fail to help, or harm, under subpopulation shift.

Every one of these that reads an error-rate quantity from unlabelled data alone does so by assuming that one factor of the joint distribution, usually P(Y | X, A) and sometimes P(X | Y, A), carries over; the methods that escape the construction in Section 6.1 do so by buying labels. That is not a defect of the methods. It is what the floor of the ladder implies, and their authors mostly say so.

7.2 How often the assumption fails in tabular data

Whether the borrowed assumption is usually safe is an empirical question about deployments, and the evidence located for this paper is sparse and unfavourable. Wang et al. (2023a) built a testbed of natural shifts across 8 tabular datasets and 172 source and target pairs and decomposed each performance drop into a covariate part and a part due to change in the outcome-given-features relation, using the decomposition of Cai et al. (2023). Among the 67 pairs whose performance fell by more than 8 points, 87.2 percent had more than half of the drop attributed to the outcome-given-features part. The per-pair ratios vary widely: 85.4 percent for ACS Income from California to Puerto Rico, 13 percent for ACS public coverage from 2010 to 2017, 0 percent for ACS Income from younger to older people, and 100 percent for a diabetes task moving from White respondents to other racial groups. The last pair is a shift along a protected attribute, and on that pair the entire degradation came from the part no label-free monitor can see.

Three limits on that evidence should be stated flatly. The testbed's pairs are constructed splits across states, years and demographics, not deployment logs. The decomposition attributes accuracy loss, not changes in group error-rate disparities. And it is one study. Schrouff et al. (2022) supply a second, qualitative line: in dermatology and electronic-health-record applications, the real shifts they diagnosed were compound, not single-factor, and in the EHR case, post-processing for equalized odds on the source brought the source gap to near zero while the target gap rose from 0.05 to 0.16 and the accuracy gap between age groups widened to 11.4 percent. TableShift (Gardner et al., 2023) and WILDS (Koh et al., 2020) document substantial out-of-distribution degradation across tabular and other domains without decomposing it this way. What is not known is the share of fairness-relevant shifts in deployed tabular systems that fall in the rows of Table 1 where a label-free reading exists.

8. An alarm calibrated only on true positives

A monitor has two error rates, and the anchor study measured one. The alert in FMMO fires in all 22 drifted conditions, but no condition without drift was run, so the rate at which it fires when nothing harmful happened is unknown. This is the general weakness that Podkopaev and Ramdas (2021) set out to fix: a useful alarm must fire on harmful shifts, ignore benign ones, and keep its false-alarm rate controlled under continuous monitoring. Mougan and Nielsen (2022) make the same complaint about input-shift detectors, and Mougan et al. (2023) give a construction where attributions shift and predictions do not, which for a performance alarm is a false positive in the explanation space itself. Subasri et al. (2025), monitoring a mortality early-warning system across seven Toronto hospitals, frame the clinical problem as detecting harmful shifts specifically, with a label-agnostic pipeline.

The specificity problem is sharper for surrogate-based alarms. A decision tree refitted on a fresh sample will usually change its top-k importances somewhat even if the sample comes from the same distribution, because tree induction is unstable. Without a same-distribution refit, the share of the surrogate's movement attributable to drift rather than to refitting is unknown. A fair test would refit the surrogate on many undrifted resamples to estimate the null distribution of its Jaccard and rank agreement, set thresholds at a chosen false-alarm rate, and only then report detection power. The same applies to attribution rankings: a ranking of mean magnitudes computed on two samples of the same distribution will differ when two features have similar mean magnitude, and the threshold has to be set against that variation. SageMaker's 0.9 NDCG threshold (Nigenda et al., 2021) is a fixed number in the same position.

None of this criticises sensitivity. It says that a sensitivity result without a false-alarm rate is half a result, and that an alarm with exact-agreement thresholds is a design that will fire on noise.

9. "Deteriorated" needs a declared metric

Figure 1 shows a monitor's second problem. The high-drift COMPAS model is more unfair than the clean one by the false-positive difference and less unfair by the false-positive ratio. Both measures are in use. The 80 percent rule that Feldman et al. (2014) generalise is a ratio of selection rates; equalized odds is usually reported as a difference of error rates (Hardt et al., 2016). Nothing in the fairness literature says which scale a monitor should use for error rates, and the choice reverses the verdict on the anchor's own data. Under a uniform multiplicative rise in both groups' false positive rates the ratio would stay fixed and the difference would grow; under a uniform additive rise the difference would stay fixed and the ratio would shrink. The anchor's high-drift row is closer to the second: the two groups rose by 32.55 and 26.00 points.

The scale question sits inside a larger one. Group fairness criteria conflict: when outcome prevalence differs between groups, a score that is calibrated within groups cannot have equal false positive and false negative rates across them except in degenerate cases (Kleinberg et al., 2016; Chouldechova, 2016). Chouldechova reports, for the ProPublica COMPAS data, false positive rates of 45

percent for Black and 23 percent for White defendants alongside recidivism rates of 51 and 39 percent; unequal error rates of this kind are what the impossibility forces on a score that is calibrated within groups when prevalence differs. A shift that changes prevalence will move these quantities differently, so "fairness deteriorated" is not well defined until the monitor names its criterion, its scale and its threshold. Begley et al. (2020) make the parallel point for explanation: standard attribution tools are not a dependable guide to whether a model is fair, since unfairness has to be defined for the task first; once a disparity is chosen, a Shapley decomposition of that disparity becomes possible.

The practical consequence is a pre-registration rule. A fairness monitor should declare, before deployment, the disparity it tracks, whether as a difference or a ratio, and the change that counts as deterioration. A post hoc reading that picks the scale on which the change looks worst converts a monitor into a search.

10. Consolidated evidence

Table 2 collects the results the argument rests on, one row per result, with the source named on the row; table numbers inside the rows refer to the anchor study's tables.

11. A decision procedure for label-free fairness monitoring

Algorithm 1 turns the placement into a procedure. Its first branch is the one most monitors skip: decide, before choosing an instrument, whether the declared disparity is readable from a label-free rung at all.

Three design choices are worth defending. Step 4 demotes attributions to diagnosis for selection-rate disparities. They are useful there, since a flagged parity change usually comes with the question of which features drove it, and Shapley decompositions of a disparity exist (Begley et al., 2020; Shanbhag et al., 2021). They are not useful as the alarm, because the quantity is observable directly. Steps 10 and 11 refuse to read label-free silence as stability for outcome-conditioned disparities unless a named shift model licenses it. This is the step the anchor's governance discussion implicitly needs: an auditor who certifies compliance from label-free signals has certified the assumption, not the model. Step 15 makes specificity a precondition rather than an afterthought, for the reasons in Section 8.

12. The anchor case through the procedure

Running the COMPAS experiment through Algorithm 1 gives a different reading from either headline.

The declared disparity is a false-positive gap, so the procedure takes the outcome-conditioned branch. The shift model is not one of the licensed cases: the injected noise moves the inputs and also the outcome-given-recorded-features relation, so neither the group-prior branch nor the within-group covariate branch applies, and the procedure's verdict is that no label-free signal could certify the false-positive gap either way. Silence from TreeSHAP rankings, from the surrogate or from confidence-based estimates would all have been uninformative about it.

At the same time, step 16 asks which label-free rungs moved. The drift detector fired, by the anchor's report. Per-group selection rates, by the identity of Section 2.3, rose substantially for both groups unless almost no member of either group was a true negative. The full attribution distribution must have moved, since by Section 5 a change in scores implies one in attributions. The label-free statistics in the study that stayed still were summaries that discard most of the information: the top-five attribution ranking and, for the age drift, the univariate detector. Under the procedure, the correct report is: large label-free shift detected on every full-distribution rung; selection rates up sharply in both groups; outcome-conditioned disparity not readable from these signals; labelled audit required. The anchor's practical recommendation, to audit, is the same. What changes is the reason, and therefore what an organisation would do differently next time. On this reading it would not add a surrogate model; it would monitor selection rates and score distributions by group, pre-register a disparity scale, and budget labels for the outcome-conditioned part.

13. The strongest case against this reading

Four objections are serious.

First, the paper may be attacking a claim no one holds. The anchor's own conclusion is modest: local attributions "may not be used alone". The explanation-shift authors concede concept-shift blindness. If both sides already agree, the paper's ladder restates a consensus. The reply is that the agreement is stated as a vague insufficiency, and that the vague form leads to the wrong remedy. FMMO's remedy is more label-free signals; the derivation in Section 6 shows that no number of them closes the outcome-conditioned gap, and Section 2.3 shows that the shift in the case presented was visible to a label-free statistic simpler than any the framework adds. A reviewer who thinks this is a refinement rather than a reversal is not being unreasonable.

Second, the population-limit ladder may mislead in practice. Rabanser et al. (2018) found coarse output-based statistics more powerful than richer ones at realistic sample sizes, and attribution vectors live in as many dimensions as the inputs. A top-k ranking may be the statistic practitioners can afford to threshold. That is a fair point about power. It does not rescue the ranking as a fairness monitor, because what the ranking discards includes exactly the features, like the anchor's drifted age feature, that sit outside the top k. And the power

Table 2. Measured or proved results used in this paper. Values are as reported by the named source; "derivation" marks results that follow from definitions rather than measurement.

Source	Setting	Result	Bears on
Zafar 2026, Table 3	COMPAS, RF, high drift	Protected FPR 37.84 to 70.39; privileged 10.17 to 36.17	2.2, 9
Zafar 2026, Table 3	same	Accuracy 63.93 to 59.78; 67.98 to 63.63	2.3
Zafar 2026, Table 4	22 conditions	Top-5 TreeSHAP set and order intact in 9; in 2 of 5 drops over 5 points	2.4
Zafar 2026, Table 4	same	Alert fires in all 22; no no-drift condition	8
Mougan 2023	Shapley efficiency	Different predictions imply different attributions	5
Mougan 2023	own discussion	Cannot detect concept shift or predict degradation	3, 6
Lundberg 2019	hospital duration model	Loss attribution exposed swapped labels the aggregate loss missed	3
Duckworth 2021	ED admissions, COVID-19	AUROC 0.856 to 0.826; SHAP tracking flagged drift	3
Nigenda 2021	SageMaker	Attribution-ranking NDCG alert below 0.9	3, 8
Rabanser 2018	MNIST, CIFAR-10	Softmax-output tests best overall; unreduced multivariate test poor	5
Garg 2022	theory	Target accuracy as hard to identify as the optimal predictor	6, 7
Asiaee 2026	theory	Equalized odds invariant, parity drifts, under group prior shift	6, 7
Schrouff 2022	EHR	After source equalized-odds fix, target gap 0.05 to 0.16	6, 7
Wang 2023a	172 tabular pairs	87.2 percent of large drops mostly from outcome-given-features shift	7
Kivimaki 2024	theory and benchmarks	Average confidence unbiased under calibration assumption	7
Chouldechova 2016	ProPublica COMPAS	FPR 45 vs 23 percent; recidivism 51 vs 39 percent	2.3, 9
This paper	derivation	Selection-rate change = accuracy change + 2 p0 x FPR change	2.3

Algorithm 1. Choosing a label-free fairness monitor and reading its silence

```
Input: model f; group attribute A at inference (yes/no); declared disparity D
       (quantity, difference or ratio, deterioration threshold); declared shift model H
       (demographic | group prior | within-group covariate | none); reference set R
1:  if D uses only predictions and groups (selection rate, parity gap):
2:      if A is available: monitor per-group selection rates and score distributions
3:      else: D is not computable; record that the monitor cannot see it
4:      use attributions (fixed R) only to explain a change already flagged
5:  else  # D conditions on the outcome: FPR, TPR, odds gaps, predictive value
6:      if H = group prior: error-rate parts of D are invariant; estimate prevalence,
7:          re-derive parity and predictive-value parts, report as model-based
8:      elif H = within-group covariate: estimate D with a recalibrated or
9:          importance-weighted estimator; check support overlap; report an interval
10:     else: label-free silence is not evidence that D is stable;
11:         draw a labelled audit sample on a budget and test D sequentially
12: for every label-free alarm used:
13:     compare full distributions (scores, attribution vectors), not top-k magnitude ranks
14:     fix and record R; never re-estimate R from deployment data inside the alarm
15:     estimate the false-alarm rate on same-distribution resamples before deployment
16: report: which rungs moved, which did not, and which assumption licenses reading D
```

comparison that would settle it for tabular group error rates has not been run.

Third, the borrowed assumptions may usually hold well enough. Within-group covariate shift with transferable calibration may be the common case in some deployments, and Duckworth et al. (2021) and Lundberg et al. (2019) report attribution-based monitoring that worked in real clinical data. The evidence in Section 7.2 cuts the other way for constructed tabular splits, but it is one testbed, it measures accuracy rather than disparity, and it is not a deployment log. The paper's claim is the conditional one: when the assumption fails, nothing label-free can tell. How often it fails in deployment is stated in Section 15 as unknown.

Fourth, the identity in Section 2.3 depends on the anchor's protocol leaving labels and group membership unperturbed, and on its group definitions. The paper reads the protocol as perturbing the test features only. The drift was injected into the age and race features, so if the anchor computed groups from perturbed race values, group composition itself changed and the identity does not apply as written. The anchor's text does not settle this. The illustrative magnitudes based on Chouldechova's base rates carry that uncertainty; the qualitative point, that selection rates are label-free and were not reported, does not depend on it.

14. Studies that would settle the open part

Five studies would settle most of what this paper leaves open.

The first is a matched-power comparison on natural tabular shifts. Using source and target pairs such as those of Wang et al. (2023a) or TableShift (Gardner et al., 2023), evaluate input-space, full attribution-distribution, score-distribution, per-group selection-rate and top-k attribution-ranking detectors at matched sample sizes, scoring each on its power to flag a pre-registered change in a group error-rate gap and on its false-alarm rate on

pairs where that gap did not change.

The second is a null calibration of surrogate and ranking alarms. Refit FMMO's decision-tree surrogate, and recompute the TreeSHAP top-k ranking, on many same-distribution resamples of each anchor dataset, report the null distribution of Jaccard and rank agreement, and recompute the anchor's alert column at a fixed false-alarm rate.

The third is a disparity-scale audit. Re-report published fairness-under-shift results on both difference and ratio scales and count how often the verdict reverses, as it does in Figure 1.

The fourth is a decomposition of disparity change. Extend the decomposition of Cai et al. (2023), or the distribution-level Shapley attribution of Zhang et al. (2022), from accuracy to group error-rate gaps, so that a shift's effect on a disparity can be split into the label-free-visible part and the outcome-given-features part.

The fifth is a deployment census. For deployed tabular systems with delayed labels, record after the fact which rows of Table 1 the observed fairness-relevant shifts fell in. This is the quantity Section 15 lists as unknown, and it decides how much label-free fairness monitoring is worth.

15. What is not known

Four things. The share of fairness-relevant shifts in deployed tabular systems that are concept-type or compound, and therefore invisible to label-free monitors, has not been measured; the closest evidence is one testbed of constructed splits that measures accuracy loss. Whether full attribution-distribution detectors outperform score-distribution detectors at deployment sample sizes on tabular data, for changes in group error rates, has not been measured. The false-alarm rate of attribution-ranking and surrogate-based alarms under same-distribution resampling has not been reported for the anchor's protocol. And the anchor study's group definitions and base rates are not reported, so the size of the selection-rate rise its experiment implies cannot be computed exactly from what was published.

16. Limitations

In this analysis the author re-reads the published tables of the anchor study and the cited literature and derives the identity in Section 2.3, which is applied under an assumption about the anchor's protocol that its text does not settle. The ladder in Section 5 assumes a fixed reference distribution for attributions and a recorded group attribute; where groups must be inferred, as Kallus et al. (2019) study for unobserved protected classes, even selection-rate monitoring requires assumptions. The placement in Section 6 treats binary classification and threshold decisions; scores used without thresholds, or multi-class outcomes, need the same analysis redone. The literature on explaining shifts was sampled for its relation to monitoring rather than surveyed exhaustively, and Pelosi et al. (2025) and Hinder et al. (2024) cover it more fully. The evidence on how often label-free assumptions fail rests heavily on one study (Wang et al., 2023a). Finally, the paper's central claim is a derivation; its practical weight depends on the deployment census in Section 14, which does not yet exist.

17. Conclusion

The claim that attributions are more sensitive than inputs and the claim that attributions stayed silent while fairness failed are both supported, because they are about different objects. The full distribution of attribution vectors sees at least what the score distribution sees. A ranking of mean attribution magnitudes sees much less, and it was the ranking that stayed still. Beneath both lies a floor that no label-free monitor crosses: group error rates, and the equalized-odds-type disparities built from them, depend on how outcomes relate to features, and a change in that relation can leave every label-free statistic unchanged. Selection-rate disparities sit above the floor and need no attributions to see. A fairness monitor should therefore begin by declaring its disparity, its scale and its shift assumption; read selection rates directly; treat label-free silence about error rates as silence; and spend labels where the floor is. In the anchor's own experiment, the deterioration it reports was accompanied by large movements in label-free quantities that the monitor did not compute, and the disparity it reports widened on one scale and narrowed on another.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. arXiv:1810.03292. doi:10.48550/arXiv.1810.03292.

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: the risk of rationalization. ICML 2019. arXiv:1901.09749. doi:10.48550/arXiv.1901.09749.

Amoukou, S. I., Bewley, T., Mishra, S., Lecue, F., Magazzeni, D., & Veloso, M. (2024). Sequential Harmful Shift Detection Without Labels. NeurIPS 2024. arXiv:2412.12910. doi:10.48550/arXiv.2412.12910.

An, B., Che, Z., Ding, M., & Huang, F. (2022). Transferring Fairness under Distribution Shifts via Fair Consistency Regularization. NeurIPS 2022. arXiv:2206.12796. doi:10.48550/arXiv.2206.12796.

Asiaee, A., & Aryan, K. (2026). Fairness Under Group-Conditional Prior Probability Shift: Invariance, Drift, and Target-Aware Post-Processing. arXiv:2602.05144. doi:10.48550/arXiv.2602.05144.

Baek, C., Jiang, Y., Raghunathan, A., & Kolter, Z. (2022). Agreement-on-the-Line: Predicting the Performance of Neural Networks under Distribution Shift. NeurIPS 2022. arXiv:2206.13089. doi:10.48550/arXiv.2206.13089.

Begley, T., Schwedes, T., Frye, C., & Feige, I. (2020). Explainability for fair machine learning. arXiv:2010.07389. doi:10.48550/arXiv.2010.07389.

Bialek, J., Kivimaki, J., Kuberski, W., & Perrakis, N. (2024). Estimating Model Performance Under Covariate Shift Without Labels. arXiv:2401.08348. doi:10.48550/arXiv.2401.08348.

Bilodeau, B., Jaques, N., Koh, P. W., & Kim, B. (2022). Impossibility Theorems for Feature Attribution. Proceedings of the National Academy of Sciences, 121(2), 2024. arXiv:2212.11870. doi:10.48550/arXiv.2212.11870.

Biswas, A., & Mukherjee, S. (2020). Ensuring Fairness under Prior Probability Shifts. arXiv:2005.03474. doi:10.48550/arXiv.2005.03474.

Budhathoki, K., Janzing, D., Bloebaum, P., & Ng, H. (2021). Why did the distribution change? AISTATS 2021. arXiv:2102.13384. doi:10.48550/arXiv.2102.13384.

Cai, T. T., Namkoong, H., & Yadlowsky, S. (2023). Diagnosing Model Performance Under Distribution Shift. arXiv:2303.02011. doi:10.48550/arXiv.2303.02011.

Chen, H., Janizek, J. D., Lundberg, S., & Lee, S.-I. (2020). True to the Model or True to the Data? arXiv:2006.16234. doi:10.48550/arXiv.2006.16234.

Chen, Y., Raab, R., Wang, J., & Liu, Y. (2022). Fairness Transferability Subject to Bounded Distribution Shift. arXiv:2206.00129. doi:10.48550/arXiv.2206.00129.

Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. FATML 2016. arXiv:1610.07524. doi:10.48550/arXiv.1610.07524.

Dombrowski, A.-K., Alber, M., Anders, C. J., Ackermann, M., Muller, K.-R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. arXiv:1906.07983. doi:10.48550/arXiv.1906.07983.

Duckworth, C., Chmiel, F. P., Burns, D. K., Zlatev, Z. D., White, N. M., Daniels, T. W. V., Kiuber, M., & Boniface, M. J. (2021). Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. Scientific Reports, 11, 23017. doi:10.1038/s41598-021-02481-y.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2014). Certifying and removing disparate impact. KDD 2015. arXiv:1412.3756. doi:10.48550/arXiv.1412.3756.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The Clinician and Dataset Shift in Artificial Intelligence. New England Journal of Medicine, 385(3), 283-286. doi:10.1056/NEJMc2104626.

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1-37. doi:10.1145/2523813.

Gardner, J., Popovic, Z., & Schmidt, L. (2023). Benchmarking Distribution Shift in Tabular Data with TableShift. NeurIPS 2023 Datasets and Benchmarks. arXiv:2312.07577. doi:10.48550/arXiv.2312.07577.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., & Sedghi, H. (2022). Leveraging Unlabeled Data to Predict Out-of-Distribution Performance. ICLR 2022. arXiv:2201.04234. doi:10.48550/arXiv.2201.04234.

Ginart, A., Zhang, M., & Zou, J. (2021). MLDemon: Deployment Monitoring for Machine Learning Systems. AISTATS 2022. arXiv:2104.13621. doi:10.48550/arXiv.2104.13621.

Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., & Schmidt, L. (2021). Predicting with Confidence on Unseen Distributions. ICCV 2021. arXiv:2107.03315. doi:10.48550/arXiv.2107.03315.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. arXiv:1610.02413. doi:10.48550/arXiv.1610.02413.

Haug, J., Braun, A., Zurn, S., & Kasneci, G. (2022). Change Detection for Local Explainability in Evolving Data Streams. CIKM 2022. arXiv:2209.02764. doi:10.48550/arXiv.2209.02764.

Havaldar, S., Chauhan, J., Shanmugam, K., Nandy, J., & Raghuveer, A. (2023). Fairness under Covariate Shift: Improving Fairness-Accuracy tradeoff with few Unlabeled Test Samples. AAAI 2024. arXiv:2310.07535. doi:10.48550/arXiv.2310.07535.

Hinder, F., Vaquet, V., Brinkrolf, J., & Hammer, B. (2023). Model Based Explanations of Concept Drift. arXiv:2303.09331. doi:10.48550/arXiv.2303.09331.

Hinder, F., Vaquet, V., & Hammer, B. (2024). One or two things we know about concept drift - a survey on monitoring in evolving environments. Part B: locating and explaining concept drift. Frontiers in Artificial Intelligence, 7, 1330258. doi:10.3389/frai.2024.1330258.

Janzing, D., Minorics, L., & Blobaum, P. (2019). Feature relevance quantification in explainable AI: A causal problem. arXiv:1910.13413. doi:10.48550/arXiv.1910.13413.

Ji, D., Smyth, P., & Steyvers, M. (2020). Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference. arXiv:2010.09851. doi:10.48550/arXiv.2010.09851.

John, S. (2025). Fair and Explainable Credit-Scoring under Concept Drift: Adaptive Explanation Frameworks for Evolving Populations. arXiv:2511.03807. doi:10.48550/arXiv.2511.03807.

Kallus, N., Mao, X., & Zhou, A. (2019). Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. arXiv:1906.00285. doi:10.48550/arXiv.1906.00285.

Kivimaki, J., Bialek, J., Nurminen, J. K., & Kuberski, W. (2024). Confidence-based Estimators for Predictive Performance in Model Monitoring. Journal of Artificial Intelligence Research, 82, 209-240, 2025. arXiv:2407.08649. doi:10.48550/arXiv.2407.08649.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. ITCS 2017. arXiv:1609.05807. doi:10.48550/arXiv.1609.05807.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., et al. (2020). WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421. doi:10.48550/arXiv.2012.07421.

Kulinski, S., Bagchi, S., & Inouye, D. I. (2021). Feature Shift Detection: Localizing Which Features Have Shifted via Conditional Distribution Tests. NeurIPS 2020. arXiv:2107.06929. doi:10.48550/arXiv.2107.06929.

Kulinski, S., & Inouye, D. I. (2022). Towards Explaining Distribution Shifts. ICML 2023. arXiv:2210.10275. doi:10.48550/arXiv.2210.10275.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. ICML 2020. arXiv:2002.11097. doi:10.48550/arXiv.2002.11097.

Lakkaraju, H., Arsov, N., & Bastani, O. (2020). Robust and Stable Black Box Explanations. arXiv:2011.06169. doi:10.48550/arXiv.2011.06169.

Lipton, Z. C., Wang, Y.-X., & Smola, A. (2018). Detecting and Correcting for Label Shift with Black Box Predictors. ICML 2018. arXiv:1802.03916. doi:10.48550/arXiv.1802.03916.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. arXiv:1802.03888. doi:10.48550/arXiv.1802.03888.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding. arXiv:1905.04610. doi:10.48550/arXiv.1905.04610.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NIPS 2017. arXiv:1705.07874. doi:10.48550/arXiv.1705.07874.

Lyu, R., Turcan, A., & Wilder, B. (2025). Explaining Concept Shift with Interpretable Feature Attribution. arXiv:2505.20634. doi:10.48550/arXiv.2505.20634.

Maity, S., Mukherjee, D., Yurochkin, M., & Sun, Y. (2020). Does enforcing fairness mitigate biases caused by subpopulation shift? arXiv:2011.03173. doi:10.48550/arXiv.2011.03173.

Mishler, A., & Dalmasso, N. (2022). Fair When Trained, Unfair When Deployed: Observable Fairness Measures are Unstable in

Performative Prediction Settings. arXiv:2202.05049. doi:10.48550/arXiv.2202.05049.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodriguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognition, 45(1), 521-530. doi:10.1016/j.patcog.2011.06.019.

Mougan, C., Broelemann, K., Kasneci, G., Tiropanis, T., & Staab, S. (2022). Explanation Shift: Detecting distribution shifts on tabular data via the explanation space. NeurIPS 2022 Workshop on Distribution Shifts. arXiv:2210.12369. doi:10.48550/arXiv.2210.12369.

Mougan, C., Broelemann, K., Masip, D., Kasneci, G., Thiropanis, T., & Staab, S. (2023). Explanation Shift: How Did the Distribution Shift Impact the Model? arXiv:2303.08081. doi:10.48550/arXiv.2303.08081.

Mougan, C., & Nielsen, D. S. (2022). Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric Bootstrap. AAAI-23 Safe and Robust AI track. arXiv:2201.11676. doi:10.48550/arXiv.2201.11676.

Nigenda, D., Karnin, Z., Zafar, M. B., Ramesha, R., Tan, A., Donini, M., et al. (2021). Amazon SageMaker Model Monitor: A System for Real-Time Insights into Deployed Machine Learning Models. arXiv:2111.13657. doi:10.48550/arXiv.2111.13657.

Pelosi, D., Cacciagrano, D., & Piangerelli, M. (2025). Explainability and Interpretability in Concept and Data Drift: A Systematic Literature Review. Algorithms, 18(7), 443. doi:10.3390/a18070443.

Podkopaev, A., & Ramdas, A. (2021). Tracking the risk of a deployed model and detecting harmful distribution shifts. ICLR 2022. arXiv:2110.06177. doi:10.48550/arXiv.2110.06177.

Rabanser, S., Gunnemann, S., & Lipton, Z. C. (2018). Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift. NeurIPS 2019. arXiv:1810.11953. doi:10.48550/arXiv.1810.11953.

Rezaei, A., Liu, A., Memarrast, O., & Ziebart, B. (2020). Robust Fairness under Covariate Shift. arXiv:2010.05166. doi:10.48550/arXiv.2010.05166.

Roschewitz, M., & Glocker, B. (2023). Distance Matters For Improving Performance Estimation Under Covariate Shift. ICCV 2023 Workshops. arXiv:2308.07223. doi:10.48550/arXiv.2308.07223.

Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., et al. (2022). Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. NeurIPS 2022. arXiv:2202.01034. doi:10.48550/arXiv.2202.01034.

Shanbhag, A., Ghosh, A., & Rubin, J. (2021). Unified Shapley Framework to Explain Prediction Drift. arXiv:2102.07862. doi:10.48550/arXiv.2102.07862.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference, 90(2), 227-244. doi:10.1016/S0378-3758(00)00115-4.

Singh, H., Singh, R., Mhasawade, V., & Chunara, R. (2019). Fairness Violations and Mitigation under Covariate Shift. ACM FAccT 2021. arXiv:1911.00677. doi:10.48550/arXiv.1911.00677.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2019). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. arXiv:1911.02508. doi:10.48550/arXiv.1911.02508.

Subasri, V., Krishnan, A., Kore, A., Dhalla, A., Pandya, D., Wang, B., et al. (2025). Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models. JAMA Network Open, 8(6), e2513685. doi:10.1001/jamanetworkopen.2025.13685.

Sundararajan, M., & Najmi, A. (2019). The many Shapley values for model explanation. arXiv:1908.08474. doi:10.48550/arXiv.1908.08474.

Wang, T., Liu, J., Cui, P., & Namkoong, H. (2023a). Rethinking Distribution Shifts: Empirical Analysis and Modeling for Tabular Data. arXiv:2307.05284. doi:10.48550/arXiv.2307.05284.

Wang, Z., Huang, C., & Yao, X. (2023b). Feature Attribution Explanation to Detect Harmful Dataset Shift. 2023 International Joint Conference on Neural Networks (IJCNN), 1-8. doi:10.1109/IJCNN54540.2023.10191221.

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2015). Characterizing Concept Drift. Data Mining and Knowledge Discovery, 30(4), 964-994, 2016. arXiv:1511.03816. doi:10.48550/arXiv.1511.03816.

Zafar, M. R., El-Sharif, A., & Khan, N. (2026). FMMO: Detecting the Divergence Between Local Attribution and Global Drift. arXiv:2609.06173. doi:10.48550/arXiv.2609.06173.

Zhang, H., Singh, H., Ghassemi, M., & Joshi, S. (2022). "Why did the Model Fail?": Attributing Model Performance Changes to Distribution Shifts. ICML 2023. arXiv:2210.10769. doi:10.48550/arXiv.2210.10769.

Zliobaite, I. (2010). Change with Delayed Labeling: When is it Detectable? 2010 IEEE International Conference on Data Mining Workshops, 843-850. doi:10.1109/ICDMW.2010.49.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.23871653