A statistical measurement instrument for the Fairness & Non-Discrimination dimension in AI system red-teaming

Definition, adaptation of an existing instrument, and conceptual validation

Juan F. Cobo *Author*[1] and Leonel Leenen *Reviewer*[1]

[1]EthiCompass

2026-09-08

Abstract

We present a statistical instrument for converting the output of a per-protected-attribute bias detector — in an AI red-teaming system that evaluates the "Fairness & Non-Discrimination" dimension — into an auditable composite rate, together with its confidence interval. The instrument is derived from one already defined and validated for a sibling dimension, "Safety & Harmful Content," but the adaptation is not a direct translation: the nature of the bias detector (an already-resolved binary decision, with no prior continuous score) removes the need to calibrate a decision threshold, and an explicit methodological decision of this work — not requiring a sensitivity/specificity calibration of the detector against a reference set, assuming instead that its performance is sufficiently reliable — also removes the imperfect-instrument correction that the sibling dimension does require. The result is a simpler version of the original instrument: it reports the observed rate of a composite — the union, measured directly, of bias detections over the set of protected attributes evaluated — together with its Wilson interval, without any sensitivity/specificity correction. We explicitly discuss the consequences of that decision and the conditions under which the full instrument (with calibration) should be reincorporated.

Keywords: AI red-teaming, algorithmic fairness, bias detection, Wilson interval, metric definition.

1 Introduction and motivation

AI red-teaming systems that evaluate fairness need to convert the verdict of a bias detector — does this conversation treat a protected attribute unequally? — into an auditable figure: a violation rate, with its uncertainty, that a client or a regulator can reason about. The problem has a structure very similar to that of other risk dimensions of the same system, and in particular to that of harmful content ("Safety & Harmful Content"), for which a statistical instrument has already been defined, calibrated, and validated in earlier work in this series.

The bias detector, however, has a property that sets it apart: unlike the harmful-content judge, which produces a continuous degree of violation per conversation turn, the bias detector produces an already-resolved **binary decision** — biased or not biased, per protected attribute — with no intermediate score to calibrate. This structural difference, together with an additional methodological decision declared explicitly in this document, determines where the instrument defined here resembles its predecessor and where it departs from it.

This document consolidates that adaptation: what is retained from the original instrument, what is dropped and why, the formal definition of what remains, and a complete illustrative example.

2 Related work

- **Wilson-type interval** (Wilson, 1927): a confidence interval for a binomial proportion with better behavior than the naive normal interval at the extremes (low rates), which is the expected regime for a violation rate. It is the only piece of the classical diagnostic-statistics literature this instrument retains unmodified.
- **Youden's index** (Youden, 1950) and the **Rogan-Gladen correction** (Rogan & Gladen, 1978): both are part of the instrument this work derives from, and both were explicitly considered for this adaptation. They are dropped here for a structural reason, not out of preference: Youden's index calibrates a decision threshold over a continuous score, and the bias detector under evaluation does not produce that score — it already delivers a binary decision — so there is nothing over which to calibrate a threshold. The Rogan-Gladen correction, which depends on knowing the detector's sensitivity and specificity as measured against a reference set, is left out for a distinct, non-structural methodological decision, discussed in detail in Section 4.

3 Formal definition of the instrument

3.1 The detection matrix

Let $i \in \{1, \ldots, N\}$ be the trace index (a complete conversation against an attack vector), $t \in \{1, \ldots, T\}$ the turn index within the trace, and $c \in \{1, \ldots, K\}$ the index of a protected attribute within the set the detector evaluates. The detector produces, for each combination, a binary decision:

$$d_{i,t,c} \in \{0, 1\} \tag{1}$$

interpreted as whether turn t of trace i treats protected attribute c unequally.

Layer independence. The detector is, in itself, a black box to the instrument: it does not matter what detection mechanism produces it, nor what internal statistical method it uses to resolve its own uncertainty. The instrument consumes $d_{i,t,c}$ and, when the detector reports one, an already-aggregated rate per attribute — but it **recomputes its own confidence interval over the raw counts**, without depending on the method the detector used to compute its own. This extends, one level further in, the same layer-independence principle of the original instrument: the aggregation layer does not care what produced the number, nor how the uncertainty accompanying it was computed.

It is likewise not this instrument's concern to decide how many, or which, protected attributes the detector evaluates — the metric is agnostic to that list, and operates on whatever set the detector reports in each case.

3.2 Turn → trace collapse

Unlike the original instrument, where the collapse is resolved by taking the maximum of a continuous score, the equivalent operation over a binary decision here is a union over turns: the trace treats

attribute c unequally if **any** turn did.

$$m_{i,c} = \bigvee_{t=1}^{T} d_{i,t,c} \tag{2}$$

3.3 Composite union rule

The trace is flagged as violating the composite if it fires on **at least one** of the protected attributes evaluated:

$$\text{comp}_i = \bigvee_{c=1}^{K} m_{i,c} \tag{3}$$

This decision is measured **directly** on the trace, counting how many traces fire on at least one attribute — it is not derived algebraically by combining the marginal rates of each attribute separately. The reason is that co-occurrence between attributes within a single trace cannot be reconstructed from the marginal rates without knowing it directly, and only whoever measures per trace has that data — not whoever only sees the aggregated rates.

3.4 Observed rate and its interval

$$p_{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} \text{comp}_i \tag{4}$$

Wilson-type confidence interval, with the finite-sample variance correction $s^2 = \frac{n}{n-1}\, p(1-p)$:

$$\text{center} = \frac{p + \frac{z^2}{2n}}{1 + \frac{z^2}{n}} \qquad \text{half-width} = \frac{z}{1 + \frac{z^2}{n}} \sqrt{\frac{s^2}{n} + \frac{z^2}{4n^2}} \tag{5}$$

This interval is computed **once, over the composite** — not for each attribute separately. The input to comp_i is the per-trace, per-attribute detail from the composite union rule above, not the marginal rates; these are a separate product, computed over the same raw data, and not a result the instrument reports with its own uncertainty.

This holds even if the detector were to provide, in addition to the marginal rate per attribute, its own confidence interval over that rate. That marginal interval — even if perfectly reliable — would measure the uncertainty of a single, isolated attribute, not that of the composite: it says nothing about the union without further assuming how the attributes relate to one another, an additional assumption this instrument does not adopt. This is why the reported interval is always computed over the composite, never inherited from an already-computed marginal interval.

4 Why this instrument does not correct for sensitivity/specificity

This is the document's most methodologically consequential decision, and it is declared in explicit contrast with the instrument it derives from: that one establishes that a detector's raw rate must never be taken as evidence without correcting it for the instrument's imperfection, through a prevalence correction that requires knowing the detector's sensitivity and specificity, measured against a reference set with known ground truth.

For the instrument defined here, the opposite is decided: **that calibration is not required**, assuming that the detector's performance is sufficiently reliable to report its observed rate without

correction. This is not a minor simplification — it is, within this series of work, the first time a dimension departs from the principle of never trusting a detector's raw rate. It is stated explicitly, rather than left implicit, so that whoever evaluates this document can judge whether the decision is acceptable for the use the instrument will be put to.

A direct consequence: without measured sensitivity/specificity, there is no "detection floor" criterion — the point below which a detector lacks sufficient discriminative power for its rate to be reliable — to evaluate. The instrument declares none.

Should a reference set with known ground truth become available in the future to measure the detector's sensitivity and specificity, the rest of the original instrument's machinery — prevalence correction, detection floor — could be reincorporated without needing to redesign what this document defines: the only piece that would change is that the observed rate per attribute, instead of being used directly as input to the composite union rule, would first pass through that correction.

5 Illustrative example

(Illustrative construction, not an actual run; the attribute names and numerical values were chosen to complete the example and do not come from any measurement.)

5.1 The scenario

The evaluation checks whether an assistant offers different credit terms depending on the stated age of the person asking. An attack vector — a conversation that first establishes favorable terms and later reveals an advanced age — is replicated across $N = 25$ independent conversations.

5.2 Detection per attribute

The detector reports, aggregated over the 25 conversations, two protected attributes evaluated in this vector:

Attribute	k (conversations with unequal treatment)	N
age	9	25
gender	4	25

Of those 25 conversations, 2 fired on both attributes at once (a response that treated the person unequally on both age and gender within the same conversation).

5.3 Building the composite and its interval

Of the 25 conversations, 11 fired on at least one of the two attributes (the 9 for age and the 4 for gender overlap in 2): $k_{\text{comp}} = 11$, $N = 25$, $p_{\text{obs}} = 0.44$.

Wilson interval, with $z = 1.96$:

$$
\begin{aligned}
s^2 &= \tfrac{25}{24}(0.44)(0.56) = 0.2567 \\
\text{denominator} &= 1 + \tfrac{1.96^2}{25} = 1.153664 \\
\text{center} &= \tfrac{0.44+0.076832}{1.153664} = 0.448 \\
\text{half-width} &= \tfrac{1.96}{1.153664}\sqrt{\tfrac{0.2567}{25} + \tfrac{1.96^2}{2500}} = 1.6989 \times 0.1086 = 0.185
\end{aligned}
$$

Result: 0.44 ± 0.185, i.e., $(0.263,\ 0.633)$.

Final reported result: $p_{\text{obs}} = 0.44$, Wilson interval $(0.263,\ 0.633)$, over the composite of the two attributes evaluated in this vector. The individual per-attribute rates (0.36 for age, 0.16 for gender) are not reported with their own interval — they remain as diagnostic input for identifying which attribute is driving the pattern, not as the instrument's result.

6 Discussion

6.1 What this work establishes

The instrument is fully defined: what is measured (binary detection per turn and protected attribute), how it is collapsed (union over turns), how it is composed across attributes (union measured directly, never derived algebraically), and what interval is reported (Wilson over the composite, computed by the instrument itself from raw counts). It is, deliberately, a simpler version than that of the dimension it derives from, and that simplicity has an identified structural cause (absence of a continuous score) and a declared methodological cause (no requirement of sensitivity/specificity calibration).

6.2 What this work does not establish

No figure in this document should be read as an actual measurement: the example in Section 5 is entirely illustrative. Beyond that, two limitations remain explicitly open and unresolved by this work:

- **There is no correction for detector quality.** The instrument assumes, without being able to verify it with the information available, that the detector is sufficiently reliable. If that assumption does not hold in practice, the reported rate will be biased in a direction this document cannot quantify — precisely because quantifying it is what the correction left out (Section 4) would do.
- **The set of protected attributes the detector evaluates falls outside the scope of this document.** The instrument is defined to operate over whatever set it is handed, but which attributes those are, and against what corpus or criterion they were decided, is a question that belongs to whoever builds the detector, not to whoever defines this metric.

6.3 Generalization

The reasoning in this document — adapting an instrument designed for a continuous-score detector to one that already delivers a binary decision — is, in principle, applicable to any other evaluation dimension whose detector shares that same property. This is not an extension evaluated here, and should not be read as one.

7 Conclusion

The instrument defined here for Fairness & Non-Discrimination retains from its predecessor the directly-measured union rule and the Wilson interval, and deliberately leaves out threshold calibration — because the nature of the detector does not require it — and the sensitivity/specificity correction — by explicit methodological decision, not by technical limitation. The result is a simple

instrument, whose principal limitation — not knowing how reliable the detector feeding it is — is declared openly rather than resolved by assumption.

References

[1] Wilson, E.B. (1927). *Probable inference, the law of succession, and statistical inference.* Journal of the American Statistical Association, 22(158), 209–212.

[2] Youden, W.J. (1950). *Index for rating diagnostic tests.* Cancer, 3(1), 32–35.

[3] Rogan, W.J. & Gladen, B. (1978). *Estimating prevalence from the results of a screening test.* American Journal of Epidemiology, 107(1), 71–76.

red-teaming system whose Fairness & Non-Discrimination dimension motivates this work. The instrument defined here is of general application and does not depend on any specific proprietary implementation, but this affiliation is declared explicitly so the reader may take it into account when evaluating the work.