Fairness Certificates via a Lean-Backed Trustworthy Probabilistic Typed Natural Deduction

Fabio Aurelio D'Asaro[1,2,*], Giuseppe Primiero[3]

[1]Department of Human Sciences, University of Verona, Italy
[2]Department of Information Studies, University College London, UK
[3]Department of Philosophy "Piero Martinetti," University of Milan, Italy

Abstract

We present *TPTND-Lean*, a Lean4 implementation of Trustworthy Probabilistic Typed Natural Deduction (TPTND) [1]. The problem addressed is the certification of group-fairness claims, i.e., whether a decision process does or does not treat two demographic groups differently, expressed as the difference between group-conditional proportions tested on observed counts. TPTND-Lean checks TPTND derivation trees whose leaves are empirical observations and declared model benchmarks, and whose internal nodes are rules for probabilistic connectives. Eventually the check produces *fairness certificates*. The kernel is small and auditable; each accepted derivation is a transferable, re-checkable certificate that binds together counts, statistical test, band, and verdict in a single typed tree. We describe a four-stage workflow from group-level summary statistics to a certificate. Two case studies on public data illustrate the approach: (i) the ProPublica COMPAS recidivism data (6172 defendants, reproducing the headline racial-bias finding), and (ii) HMDA mortgage-lending data for Delaware (54505 Black/White applications across 2022–2023, with multi-year and intersectional comparisons). Certificates are *conditional*: they verify the derivation relative to the supplied counts, the chosen statistical rule, and the checker, not the upstream data pipeline. The full source, data pipelines, and all derivations are publicly available at https://github.com/dasaro/tptnd-lean.

Keywords

algorithmic fairness, typed natural deduction, proof checking, Lean4, trust certificates

1. Introduction

Automated decision-making systems increasingly affect high-stakes outcomes (criminal sentencing, credit scoring, mortgage lending), and whether their outputs are *fair* across demographic groups has been a central concern of AI ethics for a decade [9]. Statistical tests for group-fairness properties (equalized odds, calibration, disparate impact) are routinely applied in auditing practice, yet the resulting claims are typically embedded in narrative reports or Jupyter notebooks, with no formal guarantee that the stated assumptions, data, and statistical procedures are consistent. We address the certification of such claims (e.g., equalities or inequalities of group-conditional outcome rates) relative to stated counts.

Trustworthy Probabilistic Typed Natural Deduction (TPTND) [1] offers a proof-theoretic verification protocol for fairness. TPTND is a natural-deduction calculus whose types are probability distributions over finite output alphabets and whose terms carry sample-size and provenance annotations. Introduction rules for *trust* and *untrust* internalize an acceptance-band membership test: if the output probability p lies inside the interval $[\ell, h]$ computed from observed data, the calculus derives a trustworthiness predicate **Trust**; otherwise, the current run of the model is considered untrustworthy **UTrust**. A second pair of certificates added here, **Excess** and **NoExcess**, compares two observed groups directly (Section 2). Elimination rules extract the interval into the typing context, enabling further composition.

In this paper we describe **TPTND-Lean**, the first executable kernel for TPTND, implemented in Lean4 and demonstrated on real-world fairness audits. Our contributions are:

1. A **small, auditable Lean4 kernel** for TPTND,

8th International Workshop on Artificial Intelligence and fOrmal VERification, Logic, Automata, and sYnthesis. OVERLAY 2026 @ FLoC, July 18–19, Lisbon, Portugal
*Corresponding author.
fabioaurelio.dasaro@univr.it (F. A. D'Asaro); giuseppe.primiero@unimi.it (G. Primiero)
0000-0002-2958-3874 (F. A. D'Asaro); 0000-0003-3264-7100 (G. Primiero)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

CEUR-WS.org/Vol-4324/paper3.pdf

2. **Fairness audits as transferable certificates**: a typed object that a regulator can re-verify without rerunning the analysis or trusting the analyst,
3. A **four-stage auditing workflow** that takes summary statistics through claim selection and derivation construction to such a certificate (Section 3.4),
4. **Two case studies** on public data (Section 4).

2. The TPTND calculus

Informally, TPTND turns a statistical audit into type checking: a judgement records what has been observed of a probabilistic process, what is claimed of it, and through a *trust* judgement whether the two are statistically compatible. The test itself is a typing rule, so counts, test, and verdict travel together in one derivation. Rule names follow natural-deduction practice: IT and IUT introduce **Trust** and **UTrust** respectively, ET and EUT eliminate them, and the suffix 2 versions are their two-sample variants; the same scheme names the comparison rules (IEx/INEx introduce **Excess**/**NoExcess**, EEx/ENEx eliminate them). The comparison rules exist because an audit asks a directional question. **Trust** asks whether an observation is compatible with a stated benchmark; **Excess** asks which of two groups fares worse and by how much, so the certificate names a direction and carries a magnitude rather than a yes/no verdict. **NoExcess** is correspondingly weaker than a claim of equality: it records that the data do not establish a gap, not that the groups are alike.

We treat each atomic output a as an event in a finite sample space, with $P(a)$ we denote its marginal probability. The output algebra mirrors standard event operations: $\overline{\alpha}$ is the complement, $\alpha+\beta$ is disjoint union (the calculus enforces independence of α and β syntactically, so $P(\alpha+\beta) = P(\alpha) + P(\beta)$), $\alpha \times \beta$ is the joint event under an explicit independence assumption ($P(\alpha \times \beta) = P(\alpha)P(\beta)$), and $\alpha \Rightarrow_p \beta$ encodes the conditional $P(\beta \mid \alpha) = p$. Two modalities of judgement accompany distributions: *expected mode* $t_n : \alpha_{\tilde{p}}$ records a model-side prediction, *frequency mode* $t_n : \alpha_f$ the empirical fraction; the trust rules reconcile the two via the acceptance bands of Section 2.2.

Below we summarize the syntax, judgement forms, and key rules. The full calculus appears in [1]; here we present only the rules central to the case studies, whose trust and comparison rules are stated in full in Tables 1 and 2.

2.1. Syntax

Outputs are generated by $\alpha, \beta ::= a \mid \overline{\alpha} \mid \alpha + \beta \mid \alpha \times \beta \mid \alpha \Rightarrow_p \beta$, where a ranges over a finite set of atoms and $p \in [0,1]$ is the conditional probability annotation on arrows. *Probability constraints* are $c ::= p \mid [\ell, h] \mid \neg[\ell, h]$, with $p, \ell, h \in [0,1]$ and $\ell \leq h$. The constraint $\neg[\ell, h]$ records that the relevant probability lies *outside* the interval. *Terms* are $t ::= a \mid \langle t, u\rangle \mid \mathsf{fst}(t) \mid \mathsf{snd}(t) \mid [x]t \mid (t \cdot u)$. A *context* Γ is a finite list of assumptions $x : \alpha_c$, each equipped with a nonempty *support* annotation naming the population the assumption describes. Term sequents carry an explicit provenance index σ: a finite set of identifiers naming the data batches that support the claim. Rules that pool independent data require disjoint provenances; rules that combine different events observed on one batch require equal ones. We write $\Gamma \#_w \Delta$ when the derivation carries an explicit independence assumption w.

2.2. Judgement forms

The calculus uses the sequent forms below:

$\vdash \alpha :: \mathsf{output}$	output declaration
$\vdash \Gamma$	distribution declaration
$\Gamma \vdash x : \alpha_c$	identity (exact model import)
$\Gamma \vdash_\sigma t_n : \alpha_{\tilde{p}}$	expected-mode claim
$\Gamma \vdash_\sigma t_n : \alpha_f$	frequency-mode claim
$\Gamma \vdash \mathbf{Trust}_{\mathcal{P}}/\mathbf{UTrust}_{\mathcal{P}}(t_n : \alpha_f; p, [\ell, h])$	(un)trust certificate
$\Gamma \vdash \mathbf{Excess}_{\mathcal{Q}}/\mathbf{NoExcess}_{\mathcal{Q}}(t_n : \alpha_f \succ / \sim u_m : \alpha_g; d, [\ell, h])$	(no-)excess certificate

Here d is the certified rate difference; two-sample trust certificates carry the subscript $\mathcal{Q}$ in place of $\mathcal{P}$, and both bands are defined below.

Under the null hypothesis that the true rate is the model probability p, the observed frequency f has standard deviation $\sqrt{p(1-p)/n}$; for a binomial mean this is exact rather than an approximation. The trust rules reject p when f lies more than k such deviations away, and solving $|f-p| \le k\sqrt{p(1-p)/n}$ for p rewrites that condition as membership in the *acceptance band* $\mathcal{P}(n,f,p) = f \pm k\sqrt{p(1-p)/n}$, clamped to $[0,1]$: the band is centred on the observation but scaled by the hypothesis under test, which is why p occurs on both sides. With $k = \sqrt{20}$, Chebyshev's inequality bounds the probability of rejecting a correct model, giving significance level at most $1/k^2 = 5\%$ at every sample size and with no normal approximation. IT2/IUT2 apply the same test to a difference: under a shared rate the difference $f-g$ has mean 0 and variance $p(1-p)(1/n+1/m)$, and bounding $p(1-p)$ by $\frac{1}{4}$ gives $\mathcal{Q}(n,m,f,g) = (f-g) \pm k\sqrt{\frac{1}{4}(1/n+1/m)}$, accepted when $0 \in \mathcal{Q}$. Widening a band only makes the test more conservative, so the bound survives while the unknown rate leaves the width entirely. Neither band is a confidence interval: they are acceptance regions, and the models not rejected by an observation are those $p \mid p \in \mathcal{P}(n,f,p)$.

2.3. Key rules

Observation. The data-entry rule, and the only way a frequency enters a derivation is an axiom, with side conditions:

$$\frac{}{\Gamma \vdash_{\sigma} t_n : \alpha_f}\ \textsc{Obs} \qquad \sigma \neq \varnothing, \quad n > 0, \quad nf \in \mathbb{N},$$

where exactly one entry of Γ matches both t's variable and the observed output α, so the benchmark it carries is unambiguous. Requiring $nf \in \mathbb{N}$ keeps f a genuine count out of n, and uniqueness fixes which declared process the claim describes.

Update. Pools two frequency claims with disjoint provenance into a combined weighted frequency:

$$\frac{\Gamma \vdash_{\sigma_1} t_n : \alpha_f \quad \Gamma \vdash_{\sigma_2} t_m : \alpha_g \quad \sigma_1 \cap \sigma_2 = \varnothing}{\Gamma \vdash_{\sigma_1 \uplus \sigma_2} t_{n+m} : \alpha_{\frac{nf+mg}{n+m}}}\ \textsc{Update}$$

Identity. Imports a model probability from the context:

$$\frac{x : \alpha_p \in \Gamma}{\Gamma \vdash x : \alpha_p}\ \textsc{Identity}$$

The remaining rules appear in [1] and are implemented in TPTND-Lean; Table 3 summarizes the composition rules (output formation, weakening, and contraction are administrative). Although it is not presented here in detail, the Bayesian pair I-P/E-P is demonstrated in the COMPAS case study (Section 4.1) below.

2.4. Running example

Consider a simplified lending audit. A context Γ declares an output type Denied and a benchmark denial rate of 20%: $x : \mathsf{Denied}_{0.2}$. An auditor observes $n = 500$ applicants from group A, of whom 175 are denied, recording this via Obs:

$$\Gamma \vdash_{\{src_A\}} t_{500} : \mathsf{Denied}_{0.35}$$

To test whether this observed rate is consistent with the 20% benchmark, we apply IUT. The one-sample band is $\mathcal{P}(500, 0.35, 0.2) = 0.35 \pm \sqrt{20}\sqrt{0.2 \cdot 0.8/500} = [0.27, 0.43]$. Since $0.2 \notin [0.27, 0.43]$, the calculus derives $\Gamma \vdash \mathbf{UTrust}_{\mathcal{P}}(t_{500} : \mathsf{Denied}_{0.35};\ 0.2,\ [0.27, 0.43])$.[1] The certificate is self-contained: anyone with the checker can verify it without re-running the test or accessing the raw data.

[1] We display the exact band; the stored certificate carries the slightly wider rational over-approximation of Section 3.2, which the checker matches verbatim.

$$\frac{\Gamma \vdash x : \alpha_p \quad \Delta \vdash_\sigma u_n : \alpha_f \qquad \mathcal{P}(n, f, p) = [\ell, h] \quad p \in [\ell, h]}{\Gamma, \Delta \vdash \textbf{Trust}_{\mathcal{P}}(u_n : \alpha_f; p, [\ell, h])}\ \textsf{IT}$$

$$\frac{\Gamma \vdash x : \alpha_p \quad \Delta \vdash_\sigma u_n : \alpha_f \qquad \mathcal{P}(n, f, p) = [\ell, h] \quad p \notin [\ell, h]}{\Gamma, \Delta \vdash \textbf{UTrust}_{\mathcal{P}}(u_n : \alpha_f; p, [\ell, h])}\ \textsf{IUT}$$

$$\frac{\Gamma \vdash_\sigma t_n : \alpha_f \quad \Delta \vdash_\tau u_m : \alpha_g \qquad \sigma \cap \tau = \varnothing \qquad \mathcal{Q}(n, m, f, g) = [\ell, h] \quad 0 \in [\ell, h]}{\Gamma, \Delta \vdash \textbf{Trust}_{\mathcal{Q}}(t_n : \alpha_f; g, [\ell, h])}\ \textsf{IT2}$$

$$\frac{\Gamma \vdash_\sigma t_n : \alpha_f \quad \Delta \vdash_\tau u_m : \alpha_g \qquad \sigma \cap \tau = \varnothing \qquad \mathcal{Q}(n, m, f, g) = [\ell, h] \quad 0 \notin [\ell, h]}{\Gamma, \Delta \vdash \textbf{UTrust}_{\mathcal{Q}}(t_n : \alpha_f; g, [\ell, h])}\ \textsf{IUT2}$$

Table 1
Trust-introduction rules. IT/IUT are one-sample (model p is a known constant); IT2/IUT2 are two-sample (both f and g estimated; the test is whether 0 lies in the band for $f-g$). The two-sample rules are oriented ($f \geq g$); the opposite direction swaps the premises.

$$\frac{\Gamma \vdash \textbf{Trust}_{\mathcal{P}}(u_n : \alpha_f; p, [\ell, h])}{\Gamma, x_u : \alpha_{[\ell,h]} \vdash_\sigma u_n : \alpha_f}\ \textsf{ET} \qquad \frac{\Gamma \vdash \textbf{UTrust}_{\mathcal{P}}(u_n : \alpha_f; p, [\ell, h])}{\Gamma, x_u : \alpha_{\neg[\ell,h]} \vdash_\sigma u_n : \alpha_f}\ \textsf{EUT}$$

$$\frac{\Gamma, x_u : \alpha_c \vdash \textbf{Trust}_{\mathcal{P}}(u_n : \alpha_f; p, [\ell, h]) \qquad p \in c}{\Gamma, x_u : \alpha_p \vdash_\sigma u_n : \alpha_{\tilde{p}}}\ \textsf{ET}^{\textsf{ex}}$$

$$\frac{\Gamma \vdash_\sigma t_n : \alpha_f \quad \Delta \vdash_\tau u_m : \alpha_g \qquad \sigma \cap \tau = \varnothing \quad \mathcal{Q}(n, m, f, g) = [\ell, h] \quad 0 \notin [\ell, h]}{\Gamma, \Delta \vdash \textbf{Excess}_{\mathcal{Q}}(t_n : \alpha_f \succ u_m : \alpha_g;\ f-g,\ [\ell, h])}\ \textsf{IEx}$$

$$\frac{\Gamma \vdash_\sigma t_n : \alpha_f \quad \Delta \vdash_\tau u_m : \alpha_g \qquad \sigma \cap \tau = \varnothing \quad \mathcal{Q}(n, m, f, g) = [\ell, h] \quad 0 \in [\ell, h]}{\Gamma, \Delta \vdash \textbf{NoExcess}_{\mathcal{Q}}(t_n : \alpha_f \sim u_m : \alpha_g;\ f-g,\ [\ell, h])}\ \textsf{INEx}$$

Table 2
Trust elimination and comparison rules. ET, EUT, and ET$^{\text{ex}}$ (which re-enters the expected layer at the trusted value) are one-sample eliminations: they are not admissible for the two-sample certificates, whose interval concerns a difference $f-g$ rather than the marginal probability of α. EEx and ENEx (not shown) re-enter the frequency layer under a comparison certificate, shifting the left observation by a model-derived interval. IEx/INEx share the orientation $f \geq g$.

3. TPTND-Lean: implementation in Lean 4

3.1. Kernel

TPTND-Lean consists of a kernel surfaced through a single entry point (`checkDerivation`), which traverses a derivation tree and either returns successfully, in which case the tree *is* the certificate, or yields an error message naming the failing premise. Three design choices reduce the trusted surface. A probability is wrapped in a `Prob` structure, which ensures an ill-formed probability cannot be represented. The checking monad forces every rule to return a checked result or an explicit error, ruling out silent failures. The dispatcher matches rule names against a closed enumeration and rejects any derivation whose rule name falls outside it. A machine-checked soundness theorem (`checker_sound`) proves that every derivation the checker accepts is derivable in an inductive specification of the rules stated independently of the checker; the trusted base is the Lean kernel, the specification, and the

Rules	Reading
I+/E+$_{L,R}$	From $t_n : \alpha_f$, $t_n : \beta_g$ on one batch (equal σ; α, β disjoint) derive $t_n : (\alpha{+}\beta)_{f+g}$; E+ recovers a summand by subtraction. Aggregates mutually exclusive events.
I×/E×$_{L,R}$	From expected-mode $t_n : \alpha_{\tilde{p}}$, $u_n : \beta_{\tilde{q}}$ under an independence assumption derive $\langle t, u\rangle_n : (\alpha{\times}\beta)_{\tilde{p}\tilde{q}}$; E× recovers a factor by division. Null baselines for joint frequencies.
I→/E→	I→ discharges an assumption $x : \alpha_a$, forming $[x]t_n : (\alpha \Rightarrow_a \beta)$; E→ applies the abstraction to $u_n : \alpha_{\tilde{q}}$, multiplying the values. The natural encoding of conditional criteria, which the case studies instead reach by subsetting (Section 3.5).
I-P/E-P	I-P packages hypotheses $a_1, \ldots, a_k$ with prior weights b_i ($\sum_i b_i = 1$) into a prior family; E-P conditions it on an observation $t_n : \alpha_f$, yielding the exact posterior of the context-selected hypothesis (Section 4.1).

Table 3
Composition rules implemented in TPTND-Lean beyond the trust family.

external pipeline producing the leaf counts.

The kernel implements the full rule set of the calculus, including rules the case studies do not exercise; their audit roles are summarized in Table 3. The comparison eliminations EEx/ENEx re-enter the frequency layer carrying a model-shifted interval (Tree C in Section 4.2). Apart from Tree C and the COMPAS Bayesian update, the case studies stay inside the trust and comparison sub-fragment: every audit they encode reduces to a difference of marginal proportions on a single atomic event.

3.2. Rational arithmetic and provenance

All arithmetic is rational ($\mathbb{Q}$); the $[0, 1]$ bounds are enforced as type-level proof obligations on every probability value. Square roots for the bands' standard-error term are approximated by a rational Newton iteration (15 steps), followed by rounding to six decimal places to keep denominators bounded. The multiplier $k = \sqrt{20}$ is stored as the rational $559017/125000$, which over-approximates $\sqrt{20}$. The Newton iterate over-approximates the square root and the rounding is a ceiling, so the stored band always contains the exact band: rounding can widen an acceptance but never manufacture a rejection. None of the case-study derivations sits within rounding distance of a boundary.

Provenance and probability conservation. The provenance index σ has two regimes: I+, E+$_{L,R}$, I×, E×$_{L,R}$, E→ require all three claims to share σ (disjoint events on one batch); Update, IT2/IUT2, IEx/INEx require the two premise provenances to be *disjoint* with the conclusion's σ their union (independent batches). Without this discipline the textbook attack of [1],

$$\frac{\Gamma \vdash_{\sigma_A} t_4 : \alpha_{3/4} \quad \Gamma \vdash_{\sigma_B} t_4 : \beta_{3/4}}{\Gamma \vdash_{?} t_4 : (\alpha + \beta)_{3/2}}\ \text{I+}$$

would derive a probability of $3/2$. The rule layer rejects the $\sigma_A \neq \sigma_B$ premises before any arithmetic runs; as a backstop, the value-level primitives (`probAdd`, `probSub`, `probDiv`, `weightedFreq`) return `Option Prob` and fail outside $[0, 1]$. A same-σ attack with $p + q > 1$ is blocked by the value layer; a mismatched-σ attack with $p + q \leq 1$ is caught only by the rule layer.

3.3. Two-sample trust

The one-sample rules IT/IUT compare an observed frequency against a known model probability; the two-sample variants IT2/IUT2 (Table 1) apply when the benchmark is itself estimated from data, taking two frequency observations as premises. Their disjoint-provenance requirement $\sigma \cap \tau = \varnothing$ prevents a group from being compared with (a subset of) itself.

3.4. User workflow: producing a fairness certificate

Below we describe how an auditor could produce a TPTND-Lean certificate in four stages.

1. Group-level summary statistics. The auditor extracts sample sizes and event counts for the demographic strata of interest. No TPTND-Lean-specific tooling is required at this stage: SQL, pandas, or R suffice. For COMPAS (Section 4.1) the inputs are sample sizes and high-risk counts for each (race, sex, recidivism-status) cell; for the Home Mortgage Disclosure Act (HMDA) (Section 4.2), data are denial counts indexed by year, race, and sex. The product of this stage is, for each stratum, a triple (n, f, σ): the sample size, the event frequency, and a fresh provenance label naming the data source. This is the only point at which the auditor needs raw data.

2. Claim selection. The fairness question fixes which TPTND-Lean rule closes the derivation: a rate against a known benchmark $\rightarrow$ IT/IUT; two populations statistically indistinguishable $\rightarrow$ IT2/IUT2; a significant change between two periods $\rightarrow$ IUT2 on year-stratified Obs leaves; persistence under pooling $\rightarrow$ Update chained with one of these. The natural-language statement of the audit fixes the rule signature, with the trust rules playing the role of classical hypothesis tests in a statistical-software workflow.

3. Derivation construction. The derivation tree mirrors the structure of the auditor's natural language description. Each leaf is an Obs node carrying one (n, f, σ) triple from stage 1; internal nodes apply the rules from stage 2. Rule arity, output-type matching, and provenance disjointness are fixed by the rule, leaving only the data values as free parameters.

4. Verification and certification. A single call to `checkDerivation` traverses the tree, applies the per-rule checker at each node, and returns either Ok (the derivation *is* the certificate) or an error message that identifies the failing premise. The certificate is a self-contained inductive value, independent of the data and tools used to produce it.

The rule signatures constrain stages 2–3 and the kernel mechanises stage 4; stage 1 remains entirely the auditor's responsibility.

3.5. Fairness measures expressible in TPTND

TPTND directly encodes any group-fairness criterion that reduces to a difference-of-proportions test on group-conditional event probabilities. We follow [9], writing $\hat{Y}$ for the model's decision, Y for the true outcome, and A for the protected attribute: **Statistical parity** ($P(\hat{Y}{=}1 \mid A{=}a)$ equal across groups) is encoded as IT2/IUT2 on raw acceptance rates; Tree A in Section 4.2 is a one-sample variant against a fixed reference. **Equalized odds** [10] ($P(\hat{Y}{=}1 \mid A{=}a, Y{=}y)$ equal across a for each y) is encoded on the Y-conditioned subsets; the ProPublica headline derivation (Section 4.1) is the false-positive case ($y{=}0$), and the *equal-opportunity* relaxation uses only $y{=}1$. **Predictive parity/calibration** [11] ($P(Y{=}1 \mid \hat{Y}{=}1, A{=}a)$ equal across a) is encoded as IT2 on the $\hat{Y}$-conditioned subsets; under unequal base rates, calibration and equalized odds cannot both hold [11], so an audit must make explicit which property is being certified.

Two important families fall outside this fragment. *Disparate-impact ratio* (an 80% rule on a ratio of acceptance rates) is multiplicative and would require an Excess-style family with division. *Individual* and *counterfactual* fairness require causal structure, addressed by the causal-label extensions of TPTND [4, 5] and will be included in future versions of TPTND-Lean.

4. Case studies

The certificates produced below are *conditional*: the checker verifies that the supplied counts and the chosen rule form a valid TPTND derivation, not that the upstream data extraction (filtering, demographic encoding, outcome construction, hypothesis selection) was done correctly. Those steps remain part of the external audit trail, and we inherit the standard caveats of the underlying tests (binomial sampling, independence, no multiple-testing correction). The case studies assess expressiveness (whether the

fairness questions arising in two real audits can be phrased as TPTND rules), validation (whether the certified verdicts agree with published analyses of the same or related data), and checking cost.

4.1. COMPAS: recidivism risk scoring

The first case study uses the ProPublica COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset [7], filtered to 6172 defendants by the published criteria. We construct fourteen Lean-checked derivations, including:

ProPublica's headline (UTrust). ProPublica's article reports false-positive rates (FPR) of 44.9% for Black and 23.5% for White non-recidivists [7]. On the published filter (yielding the same 6172 cases) the corresponding observed rates are $641/1514 \approx 42.3\%$ for Black and $282/1281 = 94/427 \approx 22.0\%$ for White, preserving the $\sim 1.9\times$ disparity. The White FPR is itself a sample statistic, so the statistically more appropriate comparison is two-sample; the suite cross-checks it as IEx and IUT2. We present the IUT version here for continuity with ProPublica's fixed-benchmark framing. Let Γ_{BN} carry the non-exact assumption $\mathsf{HighRisk}_{[0,1]}$ for the Black cohort and $\Gamma^{\text{mdl}}_{WN} \vdash x_{WN} : \mathsf{HighRisk}_{94/427}$ import the White benchmark; write $\Theta_{FP} = \Gamma^{\text{mdl}}_{WN}, \Gamma_{BN}$ and $\sigma_B = \sigma_m \uplus \sigma_f$ for the male $(510/1168)$ and female $(131/346)$ sub-cohorts. We display deep derivations as numbered lines, each justified on the right by its rule and the lines it consumes. The TPTND-Lean-verified derivation is:

1	$\Gamma_{BN} \vdash_{\sigma_m} u_{1168} : \mathsf{HighRisk}_{255/584}$	Obs
2	$\Gamma_{BN} \vdash_{\sigma_f} u_{346} : \mathsf{HighRisk}_{131/346}$	Obs
3	$\Gamma_{BN} \vdash_{\sigma_B} u_{1514} : \mathsf{HighRisk}_{641/1514}$	Update $(1, 2)$
4	$\Gamma^{\text{mdl}}_{WN} \vdash x_{WN} : \mathsf{HighRisk}_{94/427}$	Identity
5	$\Theta_{FP} \vdash \mathbf{UTrust}_{\mathcal{P}}(u_{1514} : \mathsf{HighRisk}_{641/1514};\ 94/427,\ [0.3757, 0.4710])$	IUT $(4, 3)$
6	$\Theta_{FP},\ x_{BN} : \mathsf{HighRisk}_{\neg[0.3757,0.4710]} \vdash_{\sigma_B} u_{1514} : \mathsf{HighRisk}_{641/1514}$	EUT (5)

Line 5 fires because $94/427 \notin \mathcal{P}(1514, 641/1514, 94/427) = [0.3757, 0.4710]$; line 6 re-enters the complement interval as an assumption on the observed variable. The resulting term claim (line 6) is what a regulator would receive and re-check.

Bayesian belief update. A pilot auditor entertaining three hypotheses about the FPR (a_1=1/6, a_2=1/3, a_3=1/2) with prior weights $b = (1/2, 1/3, 1/6)$ (summing to 1) observes s=2 flagged among n=5 pilot defendants (an illustrative micro-example of I-P/E-P rather than a COMPAS-derived audit). I-P introduces the prior family from three exact model-import premises; E-P produces the posterior under hypothesis H_j given the s hits among n:

$$P(H_j \mid s, n) = \frac{a_j^s (1-a_j)^{n-s}\, b_j}{\sum_i a_i^s (1-a_i)^{n-s}\, b_i}.$$

For H_2 this gives $256/565 \approx 45.3\%$ (up from 33.3%) while H_1 drops from 50% to $75/226 \approx 33.2\%$.

A genuine non-finding (Trust). Among Black recidivists, the female lower-risk rate is $62/203$ and the male benchmark is $411/1458 = 137/486$. Since $\mathcal{P}(203, 62/203, 137/486) = [0.1641, 0.4466]$ and $137/486 \approx 0.2819 \in [0.1641, 0.4466]$, the calculus derives **Trust** and ET$^{\text{EX}}$ re-enters the expected layer with the trusted exact value.

4.2. HMDA: mortgage lending fairness

The second case study uses HMDA data for Delaware, years 2022 and 2023, downloaded from the CFPB Data Browser.[2] HMDA is a federally mandated dataset under Regulation C (12 CFR 1003), so the "transferable certificate" framing is grounded in actual regulatory practice.

After restricting to originated (`action_taken` $= 1$) and denied (`action_taken` $= 3$) applications, we obtain the summary in Table 4. We build seven derivation trees, organised in five cases that each highlight a different aspect of the certificate framework:

[2] https://ffiec.cfpb.gov/data-browser/?category=states&items=DE

Year	Group	Denied	Total	Rate
2022	White	4914	24614	19.96%
	Black	2409	7142	33.73%
2023	White	3908	17355	22.52%
	Black	2035	5394	37.73%
2023	Black Male	741	1825	40.60%
	White Male	1466	5450	26.90%
	Black Female	909	2341	38.83%
	White Female	1234	4634	26.63%

Table 4
HMDA Delaware denial rates, 2022–2023.

Tree A: reference-rate trust. An IDENTITY leaf pins a 20% reference denial rate; an OBS leaf records the White DE 2022 frequency ($4914/24614$); IT derives **Trust**. Every assumption (reference value, data source, sample size, test) is explicit in the certificate and machine-checkable.

Tree B: temporal composition. Two OBS leaves per racial group (one for 2022, one for 2023) are combined via UPDATE into pooled frequencies. In numbered-line form:

$$
\begin{array}{lll}
1 & \Gamma \vdash_{\sigma_{22}} t_{7142} : \mathsf{Denied}_{2409/7142} & \textsc{Obs} \\
2 & \Gamma \vdash_{\sigma_{23}} t_{5394} : \mathsf{Denied}_{2035/5394} & \textsc{Obs} \\
3 & \Gamma \vdash_{\sigma_B} t_{12536} : \mathsf{Denied}_{4444/12536} & \textsc{Update}\ (1,2) \\
4 & \Gamma \vdash_{\tau_{22}} u_{24614} : \mathsf{Denied}_{4914/24614} & \textsc{Obs} \\
5 & \Gamma \vdash_{\tau_{23}} u_{17355} : \mathsf{Denied}_{3908/17355} & \textsc{Obs} \\
6 & \Gamma \vdash_{\tau_W} u_{41969} : \mathsf{Denied}_{8822/41969} & \textsc{Update}\ (4,5) \\
7 & \Gamma \vdash \mathbf{UTrust}_{\mathcal{Q}}(t_{12536} : \mathsf{Denied}_{4444/12536};\ 8822/41969,\ [0.1215, 0.1670]) & \text{IUT2}\ (3,6)
\end{array}
$$

Since $0 \notin [0.1215, 0.1670]$, line 7 yields **UTrust**; a regulator with the kernel re-checks the whole chain without access to the raw data.

Tree C: comparison elimination. A depth-4 chain OBS→UPDATE→IEX→EEX certifies the pooled racial gap as **Excess** and re-enters the frequency layer: the extracted claim carries the benchmark-shifted bound $\mathsf{Denied}_{[0.3317,\,0.3772]}$ in its typing context.

Trees D.1/D.2: intersectional certificates. Two auditors independently certify Black-vs-White denial gaps for males ($741/1825$ vs. $1466/5450$, $\mathcal{Q} = [0.0765, 0.1975]$) and females ($909/2341$ vs. $1234/4634$, $\mathcal{Q} = [0.0653, 0.1787]$); both derive **UTrust**. Each certificate is verifiable independently of the other auditor.

Trees E.1/E.2: year-over-year monitoring. White denial rose 2022→2023 from 20.0% to 22.5% ($\mathcal{Q} = [0.0033, 0.0477]$, **UTrust**). Black denial rose from 33.7% to 37.7%, but on the smaller cohort the band $\mathcal{Q} = [0.0000, 0.0803]$ contains 0, so the calculus declines to certify the increase (**Trust**). The certified direction agrees with the national-level disparities documented by the Urban Institute [8].

Table 5 collects the principal fairness questions exercised, the verdicts, and the deciding bands (derivation steps such as pooling and interval re-entry are not listed separately). All twenty-one derivations (fourteen COMPAS, seven HMDA) re-check in well under a second each.

5. Related work

TPTND was introduced in [1], with a companion possible worlds-semantics in [2], λ-calculus formulation in [3] and causal-label extensions in [4, 5]. The BRIO tool [6] evaluates bias via distance metrics but does not produce derivation trees. The certifying-algorithms pattern of separating producer from verifier is classical [12, 13]. VeriFair [14] and FairSquare [15] verify fairness properties *a priori* (of the model); TPTND-Lean instead produces *a posteriori* certificates from observed data, checkable without model access. LEAN4 [16] has been used for large-scale mathematical formalization but not,

Dataset	Comparison / setting	Verdict	Acceptance band or value
COMPAS	Caucasian female FNR vs. male benchmark	**Trust**	$p \in \mathcal{P}$
	AA female FNR vs. male benchmark (one-sample)	**Trust**	$p \in \mathcal{P}$
	AA female FPR vs. AA male FPR	**Trust**	$p \in \mathcal{P}$
	AA non-recid. FPR vs. Caucasian baseline	**UTrust**	$p \notin \mathcal{P}$
	Bayesian update of FPR (pilot illustration)	posterior	$\Pr(H \mid s, n)$
	Black vs. White FPR (non-recid.)	**Excess**	$0 \notin \mathcal{Q}$
	AA female vs. male FNR	**NoExcess**	$0 \in \mathcal{Q}$
	4 two-sample cross-checks of the above	varied	—
HMDA	White DE 2022 denial vs. 20% benchmark	**Trust**	$p \in \mathcal{P}$
	Black vs. White denial, pooled 2022–23	**UTrust**	$\mathcal{Q} = [0.1215, 0.1670]$
	Pooled gap, EEx extraction (Tree C)	**Excess**	$0 \notin \mathcal{Q}$
	Black vs. White denial, male subgroup	**UTrust**	$\mathcal{Q} = [0.0765, 0.1975]$
	Black vs. White denial, female subgroup	**UTrust**	$\mathcal{Q} = [0.0653, 0.1787]$
	Black denial 2022→2023 increase	**Trust**	$\mathcal{Q} = [0.0000, 0.0803]$
	White denial 2022→2023 increase	**UTrust**	$\mathcal{Q} = [0.0033, 0.0477]$

Table 5
Fairness questions exercised in the case studies, with the verdict each Lean-checked derivation produces and the band that decides it. $\mathcal{P}$ is the one-sample band on the marginal probability; $\mathcal{Q}$ the two-sample band on the rate difference. FNR/FPR abbreviate false-negative/false-positive rate; AA, African-American.

to our knowledge, for algorithmic-fairness auditing. Toolkits such as Fairlearn [17], AIF360 [18], and What-If [19] compute metrics but do not produce verified certificates.

6. Conclusion

We have presented TPTND-Lean, the first Lean4 implementation of TPTND. Its small, auditable kernel turns every accepted derivation tree into a transferable certificate binding counts, test, band, and verdict, re-checkable by any third party without rerunning the analysis; a four-stage workflow takes group-level statistics to such certificates, exhibited on two public datasets from criminal justice and mortgage lending. The main limitations are the oriented form of the two-sample comparisons and the absence of two-sample eliminations. Signed-difference types, such eliminations, the causal-label extensions of [4], and a front-end for producing certificates without writing Lean are left to future work.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: Grammar and spelling check, Improve writing style, Paraphrase and reword, Peer review simulation. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgments

This research was funded by Ministero dell'Università e della Ricerca (MUR) through PRIN 2022 Project SMARTEST – Simulation of Probabilistic Systems for the Age of the Digital Twin (20223E8Y4X), and Project "Departments of Excellence 2023-2027" awarded to the Department of Philosophy "Piero Martinetti" of the University of Milan.

References

[1] F. A. D'Asaro, F. A. Genco, G. Primiero, "Checking trustworthiness of probabilistic computations in a typed natural deduction system," *Journal of Logic and Computation*, vol. 35, no. 6, article exaf003, 2025. https://doi.org/10.1093/logcom/exaf003.

[2] E. Kubyshkina, G. Primiero. "A possible worlds semantics for trustworthy non-deterministic computations," *International Journal of Approximate Reasoning*, 172:109212, 2024.

[3] F. A. Genco, G. Primiero, "A typed lambda-calculus for establishing trust in probabilistic programs," arXiv:2302.00958, 2023.

[4] L. Ceragioli, G. Primiero, "A Proof System with Causal Labels (Part I): checking Individual Fairness and Intersectionality," in *Short Paper Proceedings of the 7th International Workshop on Artificial Intelligence and Formal Verification, Logic, Automata, and Synthesis (OVERLAY 2025)*, A. Montanari, A. Orlandini, N. Saccomanno, S. Tonetta (Eds.), CEUR-WS vol. 4142, 2025, pp. 173–179. https://ceur-ws.org/Vol-4142/paper19.pdf.

[5] L. Ceragioli, G. Primiero, "A Proof System with Causal Labels (Part II): checking Counterfactual Fairness," in *Short Paper Proceedings of the 7th International Workshop on Artificial Intelligence and Formal Verification, Logic, Automata, and Synthesis (OVERLAY 2025)*, A. Montanari, A. Orlandini, N. Saccomanno, S. Tonetta (Eds.), CEUR-WS vol. 4142, 2025, pp. 181–187. https://ceur-ws.org/Vol-4142/paper20.pdf.

[6] G. Coraglia, F. A. D'Asaro, F. A. Genco, D. Giannuzzi, D. Posillipo, G. Primiero, C. Quaggio, "BRIOx-Alkemy: a Bias Detecting Tool," in *Proceedings of the 2nd Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE 2023) co-located with AI*IA 2023*, CEUR-WS vol. 3615, 2023, pp. 44–60. https://ceur-ws.org/Vol-3615/paper4.pdf.

[7] J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine bias," *ProPublica*, May 23, 2016.

[8] J. H. Choi, P. J. Mattingly, "What Different Denial Rates Can Tell Us About Racial Disparities in the Mortgage Market," Urban Institute, Urban Wire, January 13, 2022. https://www.urban.org/urban-wire/what-different-denial-rates-can-tell-us-about-racial-disparities-mortgage-market.

[9] S. Barocas, M. Hardt, A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*, fairmlbook.org, 2019.

[10] M. Hardt, E. Price, N. Srebro, "Equality of opportunity in supervised learning," in *NIPS*, 2016, pp. 3315–3323.

[11] A. Chouldechova, "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[12] G. C. Necula, "Proof-carrying code," in *POPL '97*, ACM, 1997, pp. 106–119.

[13] R. M. McConnell, K. Mehlhorn, S. Näher, P. Schweitzer, "Certifying algorithms," *Computer Science Review*, vol. 5, no. 2, pp. 119–161, 2011.

[14] O. Bastani, X. Zhang, A. Solar-Lezama, "Probabilistic Verification of Fairness Properties via Concentration," *Proceedings of the ACM on Programming Languages*, vol. 3, OOPSLA, article 118, October 2019, 27 pp. https://doi.org/10.1145/3360544.

[15] A. Albarghouthi, L. D'Antoni, S. Drews, A. V. Nori, "FairSquare: probabilistic verification of program fairness," *Proceedings of the ACM on Programming Languages*, vol. 1, OOPSLA, article 80, October 2017. https://doi.org/10.1145/3133904.

[16] L. de Moura, S. Ullrich, "The Lean4 theorem prover and programming language," in *CADE-28*, LNCS vol. 12699, 2021, pp. 625–635.

[17] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft Research, Tech. Rep. MSR-TR-2020-32, May 2020.

[18] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.

[19] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, "The What-If Tool:

Interactive Probing of Machine Learning Models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2020. https://doi.org/10.1109/TVCG.2019.2934619.