Online Fairness Auditing through Iterative Refinement

Pranav Maneriker
maneriker.1@osu.edu
The Ohio State University
Columbus, OH, USA

Codi Burley
burley.22@osu.edu
The Ohio State University
Columbus, OH, USA

Srinivasan Parthasarathy
parthasarathy.2@osu.edu
The Ohio State University
Columbus, OH, USA

ABSTRACT

A sizable proportion of deployed machine learning models make their decisions in a black-box manner. Such decision-making procedures are susceptible to intrinsic biases, which has led to a call for accountability in deployed decision systems. In this work, we investigate mechanisms that help audit claimed mathematical guarantees of the fairness of such systems. We construct AVOIR, a system that reduces the number of observations required for the runtime monitoring of probabilistic assertions over fairness metrics specified on decision functions associated with black-box AI models. AVOIR provides an adaptive process that automates the inference of probabilistic guarantees associated with estimating a wide range of fairness metrics. In addition, AVOIR enables the exploration of fairness violations aligned with governance and regulatory requirements. We conduct case studies with fairness metrics on three different datasets and demonstrate how AVOIR can help detect and localize fairness violations and ameliorate the issues with faulty fairness metric design.

CCS CONCEPTS

• **Mathematics of computing** → *Probability and statistics*; • **Software and its engineering** → **Domain specific languages**.

KEYWORDS

fairness, metrics, verification, inference, online, monitoring

ACM Reference Format:
Pranav Maneriker, Codi Burley, and Srinivasan Parthasarathy. 2023. Online Fairness Auditing through Iterative Refinement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599454

1 INTRODUCTION

Advanced analytics and artificial intelligence (AI), along with its many benefits, pose significant threats to individuals and the broader society. [24] identify invasion of privacy; manipulation of vulnerabilities; bias against protected classes; increased power imbalances; error; opacity and procedural unfairness; displacement of labor; pressure to conform, and intentional and harmful use as some of the key areas of concern. A core part of the solution to mitigate such risks is the need to make organizations accountable and ensure that the data they leverage and the models they build and use are both inclusive of marginalized groups and resilient against societal bias. Deployed AI and analytic systems are complex multistep processes that can incorporate several sources of risk at each step. At each of these stages, determining accountability in the decision-making of AI processes requires a determination of who is accountable, for what, to whom, and under what circumstances [10, 34]. A more comprehensive overview of the mechanisms that can support accountability for the different stages of machine learning system design can be found in the work of Cooper et al. [10]. Our analysis centers on auditing fairness claims of mathematical guarantees associated with automated, black-box decision-making processes. Governments worldwide are wrestling with different implementations of auditing regulations and practices to increase the accountability of decision processes. Recent examples include the New York City auditing requirements for AI hiring tools [40], European data regulation (GDPR [36]), accountability bills [9, 35] and judicial reports [27]. These societal forces have led to the emergence of checklists [32, 37], metrics of fairness [41], and recently, algorithms and systems that observe and audit the behavior of AI algorithms. Such ideas date back to the 1950s [33]. However, research has been sporadic until very recently, with the widespread use of AI-based decision-making giving rise to the vision of algorithmic auditing [17]. In this work, we present a framework called AVOIR[1], for auditing and verifying fairness online. AVOIR builds upon the ideas on distributional probabilistic fairness guarantees [2, 4], generalizing them to real-world data.

2 BACKGROUND AND KEY CONTRIBUTIONS

Fairness criteria quantify the relationship between the outcome metric across multiple subgroups or similar individuals in the population. Formal definitions of fairness focus on observational criteria, i.e., those that can be written down as a probability statement involving the joint distribution of the features, sensitive attributes, decision-making function, and actual outcome. Consider a decision-making function that claims to satisfy certain fairness guarantees. In our setup, auditing a claim about a fairness guarantee would involve quantifying the probability of claim violations. Given a particular failure probability Δ and a stream of data $\ldots, (X_t, Y_t), \ldots$ over time steps t at run time, a fairness claim ψ would be considered valid if $\Pr[\forall t \geq 1, \psi] \geq 1 - \Delta$. Assuming that the data is sampled from a fixed, possibly unknown distribution p_{data}, a common strategy to test the validity of a claim is to use hypothesis testing with a predetermined sample size m. However, it is impossible to know a priori whether m will be large enough to

[1] AVOIR in French means "to have", and this acronym reflects both our aspirational goal to achieve fairness in advanced analytics and AI but also reflects what is currently verifiable given a dataset, a model, and a fairness specification.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599454

verify this hypothesis [44], and peeking at the data to determine the sample size would be considered 'p-hacking.' Collecting labeled data for fairness-related applications is expensive [26]; therefore, it is essential to ensure that a monitoring system used for auditing the fairness claim can *adaptively and continuously update its estimates* of the probability of validity. We consider a claim as invalid if $\Pr[\forall t \geq 1, \neg\psi] \geq 1-\Delta$, where $\neg$ denotes negation. Another desirable feature in the auditing system would be a *finite-horizon stopping rule* that should be able to decide the validity/invalidity of a claim, given sufficient data.

We show that the framework of confidence sequences/sets [25] provides a mechanism for building confidence intervals for inference in sequential experiments with nonasymptotic (i.e., always valid for $t \geq 1$) intervals that approach zero width, ensuring that a stopping rule would have a finite termination. We would also like to be able to *localize and diagnose* terms within a fairness metric that leads to the inference of a negated claim. For example, suppose $r \in \{0, 1\}$ denotes the return value of a binary decision function (say, job candidate selection), and s is an indicator denoting whether a candidate belongs to a minority population. The 80%-rule for disparate impact [14, 15] is a fairness criterion which states that

$$\frac{\Pr[r=1|s]}{\Pr[r=1|\neg s]} \geq 0.8$$

Assuming that a confidence sequence approach leads to the inference of a negated claim (invalid) for disparate impact, a diagnosis would determine whether the numerator or denominator in the criterion lead to the invalidity. AVOIR uses an inference framework that builds upon distributional guarantees for each term within the criterion, which can help with such a diagnosis. Further, overall uncertainty can be guaranteed across multiple groups by balancing it across subexpressions with differences in the number of observed samples. For example, consider Bernoulli r.vs[2] $X_{1,2}$ for which we derive concentration guarantees $\Pr[|\mathbb{E}[X_i] - \overline{\mathbb{E}}[X_i]| \geq \varepsilon_i] \leq \delta_i$ after t_i observations. Here, $\mathbb{E}[X]$ refers to the population mean, $\overline{\mathbb{E}}[X]$ refers to an empirical mean based on observations of X, and $\varepsilon, \delta > 0$ are the concentration level and failure probability, respectively. From the Hoeffding inequality, $\delta = 2e^{-2t\varepsilon^2}$. We can claim tighter bounds for X_2 if $t_2 > t_1$ as the failure probability δ is lower at the same concentration ε. That is, $\varepsilon_1 = \varepsilon_2, t_2 > t_1 \implies \delta_1 > \delta_2$. Varying ε across subexpressions to minimize the overall (union bounded) $\delta = \delta_1 + \delta_2$ allows an earlier stopping time for a valid/invalid claim, i.e., *fewer iterations and fewer data samples.* Adaptive versions of these inequalities also have similar patterns (see Figure 1).

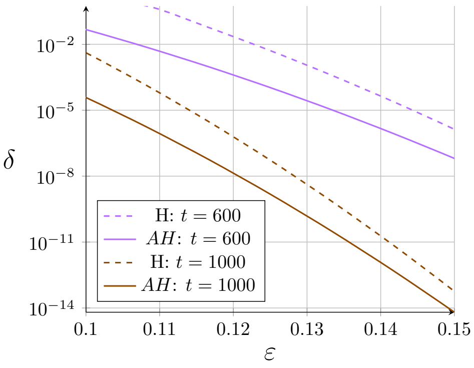


Figure 1: Failure probability δ of a Bernoulli r.v. vs concentrated around mean ε for different n. At the same concentration, lower failure probability for the majority class (greater n). H = (online) Hoeffding, AH = Adaptive Hoeffding.

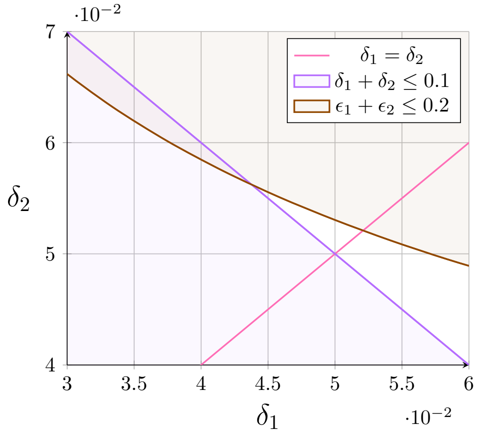


Figure 2: AVOIR finds a solution for a *theoretical* scenario with $\delta_1+\delta_2 \leq \Delta$ under constraint $\varepsilon_1+\varepsilon_2 \leq \varepsilon_T$. No solution exists with additional constraint $A_\delta : \delta_X = \delta_Y = \Delta/2$ - common assumption in prior work.

Consider R, a Bernoulli r.v corresponding to the output of a binary decision function, with s being an indicator of class membership. Let $X = r \vee s$ and $Y = r \vee \neg s$ be r.vs corresponding to a favorable decision for the majority and minority classes, respectively. Suppose we aim to estimate a criterion $\psi := E[X] - E[Y] < \varepsilon_T$ Previous work on inference from distributional guarantees [2, 4] assumes equal failure probability across all groups, i.e., the assumption $A_\delta := \delta_1 = \delta_2$. Suppose we want the upper bound of the failure probability $\Delta = 0.1$ for the specified criterion. Consider a n_X, n_Y observations for X, Y such that $\overline{\mathbb{E}}[X] = 0.8, n_X = 1550$ and $\overline{\mathbb{E}}[Y] = 0.5, n_Y = 310$. Figure 2 shows that no solution is feasible for the optimization problem with A_δ. However, AVOIR can find a solution. For the optimal solution, $\delta_2 \approx 2.35\delta_1$, which aligns with our intuition about allocating higher failure probability to terms with the majority of observations. The optimization problem inferred by AVOIR:

$$\min_{\delta_X, \delta_Y} \delta_X + \delta_Y$$
$$\text{s.t. } \varepsilon_X + \varepsilon_Y \leq \overline{\mathbb{E}}[X] - \overline{\mathbb{E}}[Y] - \varepsilon_T$$

Key Contributions: We now summarize our contributions vis-à-vis FP [2] and VF [4], the most closely related prior work.

(1) We build AVOIR in the framework of **confidence sets** [25] which enables **adaptive optimization** of δ across subexpressions of a specification. Note that FP only provides examples with equal splits while VF splits uncertainty equally across all elementary subexpressions.

(2) The confidence sets framework allows us to avoid assuming a known data distribution or fitting a density estimator over

[2] random variables

Symbol	Description
Δ	Overall failure probability for a specificaiton
X_i	Bernoulli random variable for i^{tg} tern
$\overline{\mathbb{E}}[X]$	Empirical estimate of expectation $\mathbb{E}[X]$
t	No. of observed samples
$\overline{\mathbb{E}}[X_i]_t$	Empirical estimate after t steos
δ_i	Failure probability $0 \leq \delta_i \leq 1$ corresponding to X_i
ε_i	Concetration bound for $\lvert E[X_i] - \overline{\mathbb{E}}[X_i] \rvert \leq \varepsilon_i$
ϕ_X	Concentration bound for empirical estimate of $E[X]$
ψ	Fairness specification
r, R	Return value of the function being monitored
y, Y	True label
s, S	Indicator for group membership
c	Constant $\in \mathbb{R}$
C	A set of constraints

Table 1: The AVOIR symbol descriptions table.

the population prior to fairness testing, which is required in VF.

(3) We augment the **bound propagation rules** to facilitate the online optimization process and allow propagation of constraints along with assertions at each iteration.

(4) We build an **inference engine** that supports the automated inference of propagation rules for a wide range of metrics, with a **finite stopping** rule under mild conditions. In Section 4, we provide examples of inference over specifications involving multiple subexpressions, which are only possible by extending the implementations provided by previous work. We also implement bound inference rules from VF (denoted AVOIR-VF) as a baseline.

(5) We support **diagnosis** of fairness violations using bounds inferred for subexpressions. We demonstrate the use of these cues to help drive the design of specifications in Section 4.2, which shows how a user may audit a fairness claim.

3 AVOIR FRAMEWORK

3.1 Definitions

AVOIR supports implementing an extensive range of group fairness criteria, including demographic parity [6], equal opportunity [22], disparate mistreatment [46], and combinations of these criteria. For instance the above 80%-rule is `E[r|S==s]/E[r|S!=s] > 0.8` in AVOIR's DSL[3]. Here, the term `E[r|S==s]/E[r|S!=s]` is a *subexpression* of the specification. The smallest units involving an expectation (e.g., `E[r|S!=s]`) are denoted as *elementary subexpressions*. We focus on fairness criteria that can be expressed using Bernoulli r.v. as it allows the simplification of probabilities into expectation, as $\Pr[r = 1] = \mathbb{E}[r]$ (hereafter, used interchangably). Our algorithm uses adaptive concentration sets [25, 48] to build estimates for *elementary subexpressions* and then derive the estimates for expressions that combine them. A combination of multiple such elementary expressions is denoted as a *compound* expression. We aim to derive statistical guarantees about fairness criteria based on estimates from observed inputs and outputs. For example, let X be an observed Bernoulli r.v, then an assertion $\phi_X = (\overline{\mathbb{E}}[X], \varepsilon, \delta)$ over X, corresponds to an estimate satisfying $\phi_X := \Pr[|\mathbb{E}[X] - \overline{\mathbb{E}}[X]| \geq \varepsilon] \leq \delta$ where $\overline{\mathbb{E}}[X]$ denotes an empirical estimate of $E[X]$. We then use assertions ϕ_X, ϕ_Y to assert claims for expressions involving X, Y. For example, for the 80%-rule, assertions over $\mathbb{E}[X]/\mathbb{E}[Y]$. A *specification* involves either a comparison of expressions with constants (e.g., $\mathbb{E}[X]/\mathbb{E}[Y] > 0.8$) or combinations of multiple such comparisons. Such a specification may be True (T) or False (F) with some probability. For a given specification ψ, we denote the claim that $P[\psi = F] \geq 1 - \Delta$ as $\psi : (F, \Delta)$, where Δ denotes the failure probability of a guarantee. Given a stream of observations and outcomes from the decision functions, and a specified threshold probability Δ, we will continue to refine the estimate for a given specification until we reach the threshold. Specifications involving variables that take more than two values can be implemented using transformations and boolean operators (examples in Appendix D).

$$
\begin{aligned}
\langle spec \rangle &::= \langle ETerm \rangle\ \langle comp\text{-}op \rangle\ \mathrm{c} \\
&\mid\ \langle spec \rangle \wedge \langle spec \rangle \\
&\mid\ \langle spec \rangle \vee \langle spec \rangle \\[1em]
\langle ETerm \rangle &::= \mathbb{E}[\langle E \rangle] \\
&\mid\ \mathbb{E}[\langle E \rangle,\ \mathrm{given}{=}\langle E \rangle] \\
&\mid\ \mathrm{c} \in \mathbb{R} \\
&\mid\ \langle ETerm \rangle\ \ \{+, -, \times, \div\}\ \ \langle ETerm \rangle
\end{aligned}
$$

Figure 3: Grammar for specification. $\langle E \rangle$ refers to expressions of r.vs and $\langle comp - op \rangle$ = comparison operator $\in \{>, <, =, \neq\}$.

3.2 Language Specification

We describe AVOIR's DSL used for specifying fairness metrics (Figure 3). We focus on binary decision-making functions; Bernoulli r.v.s can characterize their outputs. Consider a decision function $f : X \to \{0, 1\}$, where $X = (X_1, \ldots, X_k)$ denotes a real-valued input vector. We use $R = f(X)$ to simplify the remainder of the definitions. The grammar can be used to construct Bernoulli r.vs to support expressions beyond those that produce binary outputs. For example, a v-threshold based real-valued output, $R' = (R > v)$ and a multi-class output, for class j, $R' = (R == j)$ correspond to Bernoulli r.vs. Expressions involving R and X_i act as the arguments <E> to construct an <ETerm>. For example, $\mathbb{E}[R > 0 | X_1 + X_2 > a]$. c terms represent constant real values used as bounds for specifications. We modified the grammar from prior work to include two additional operations. First, we added a `given` argument to $\mathbb{E}$, which allows a user to specify conditional probabilities directly, in contrast to specifying it as a ratio of joint/marginal probabilities.

$$\frac{\mathbb{E}(A \vee (B = b))}{\mathbb{E}(B = b)} \to \mathbb{E}(A, \mathtt{given} = (B = b))$$

which is used to represent $\mathbb{E}[A|B = b]$, simplifying expressions for group fairness specification. Additionally, we add comparison operators, which further simplify the process of writing specifications.

3.3 Propagating Bounds

Generating the bounds for a specification requires propagating them from elementary subexpressions. Assuming that observed

[3] Domain Specific Language

values for each <E> correspond to an underlying random variable X, a probabilistic guarantee ϕ_X for an *elementary* subexpression consists of an empirical estimate $\overline{\mathbb{E}}[X]$, a concentration bound ε_X, and a failure probability δ_X, such that $\Pr[|\mathbb{E}[X] - \overline{\mathbb{E}}[X]| \geq \varepsilon_X] \leq \delta_X$. For compound expressions, we must infer the implied guarantees that can be inferred with corresponding constraints. Each inference rule corresponds to a derivation in the DSL grammar. Inference rules have preconditions and postconditions that are in the form:

$$\frac{\bigcup\{r | r \in \{\phi, \psi, C\}\}}{\bigcup\{s | s \in \{\phi, \psi, C\}\}}$$

where ϕ denotes a claim for a subexpression, ψ for a <spec>.For example, consider the sum/difference rule. Starting with the assumptions $\phi_X := (\overline{\mathbb{E}}[X], \varepsilon_X, \delta_X)$, $\phi_Y := (\overline{\mathbb{E}}[Y], \varepsilon_Y, \delta_Y)$. Then we have

$$\begin{aligned} &|\mathbb{E}[X] \pm \mathbb{E}[Y] - (\overline{\mathbb{E}}[X] \pm \overline{\mathbb{E}}[Y])| \\ &\quad \leq |\mathbb{E}[X] - \overline{\mathbb{E}}[X]| + |\mathbb{E}[Y] - \overline{\mathbb{E}}[Y]| \\ &\quad \leq \varepsilon_X + \varepsilon_Y \end{aligned}$$

i.e., $\phi_X, \phi_Y \Rightarrow X \pm Y : \left(\overline{\mathbb{E}}[X] \pm \overline{\mathbb{E}}[Y], \varepsilon_X + \varepsilon_Y, \delta_X + \delta_Y\right)$. Inference rules may require constraints, for e.g., assume $\phi_X := (\overline{\mathbb{E}}[X], \varepsilon_X, \delta_X)$, $\overline{\mathbb{E}}[X] > c$. Then we have $\Pr[\mathbb{E}[X] < \overline{\mathbb{E}}[X] - \varepsilon_X] > 1 - \delta$ If we add the constraint that $\overline{\mathbb{E}}[X] - \varepsilon_X \geq c$, we have $\Pr[X < c] > 1 - \delta$, thus,

$$\phi_X \Rightarrow \psi := X > c : (T, \delta_X)$$
$$\text{under the constraint } \{\overline{\mathbb{E}}[X] - \varepsilon_X \geq c\}$$

The complete set of inference rules required for the DSL is provided in the appendix (Figure 7). The implementation in AVOIR follows these rules but could be extended to other rule inference templates that support the DSL. Note that these rules extend the ones implemented by VF [4] with constraints that enable the optimizations required in AVOIR.

3.4 Optimizing Bounds

3.4.1 AVOIR Algorithm. The pseudocode for the optimization procedure in AVOIR is described in Algorithm 1. The input to the algorithm is the reporting threshold probability Δ and a specification ψ. We then infer a symbolic optimization problem corresponding to the bounds and failure probabilities of the elementary subexpressions. At each step, the OBSERVE(X) function is called with the new observation of every elementary subexpression and output. The empirical running means and counts of observations are updated. The final optimization problem OPT corresponding to each specification is a nonlinear constrained optimization problem. If a solution is successfully found for OPT, the algorithm terminates, and the estimate for the specification has reached the required threshold. If no solution is found, the estimates will be updated with $\delta_i = \Delta$ for each *elementary* subexpression. The intuition behind the algorithm is to use a confidence sequence corresponding to the estimates of elementary subexpressions at each time step. The inferred OPT has the form

$$\min_{0 \leq \delta_i \leq 1} \sum_{i=1}^{n} \delta_i \tag{1}$$
$$\text{s.t. } g_k(\delta_{1,\dots,n}, \overline{\mathbb{E}}[X_1], \dots, \overline{\mathbb{E}}[X_n]) \leq \varepsilon_k$$

Algorithm 1 AVOIR Algorithm

```
Input: $\Delta, \psi$                                   ▷ $\Delta$, Specification
Output: $T_s$ time step when the value of $\psi$ can be guaranteed with probability $\geq 1 - \Delta$
for $X_i \in \psi$ do
    $\delta_{X_i} = \Delta$                             ▷ Set initial value $\forall i$
    $S_{X_i} = 0$                                       ▷ Sum of observations
    $n_{X_i} = 0$                                       ▷ Number of observations
end for
$T = 0$                                                 ▷ Time step
Initialize $OPT_\psi$                ▷ Initialize Optimization Problem (Fig. 7)
procedure OBSERVE($X$)
    for $X_i \in X$ do
        $S_{X_i} = S_{X_i} + X_i$
        $n_{X_i} = n_{X_i} + 1$
        $\overline{\mathbb{E}}[X_i] = S_{X_i}/n_{X_i}$
        Initialize $\delta_{X_i}$ as a symbolic variable
        Assign $\varepsilon(\delta_{X_i}, n_{X_i})$ symbolic variable
    end for
    Propagate $\delta_{X_i}$ using the inference rules
    Initialize constraints $g_K$ in $OPT_\psi$ using the computed values
    $\delta^*_T = \texttt{Solve}(OPT_\psi)$
    if $\delta^*_T \leq \Delta$ then
        $\delta_{X_i} = \delta^*_T[X_i]$
        return $T_s = T$
    end if
    $T = T + 1$
end procedure
```

where g_k, ε_k are the functions/bounds derived using the transformations carried out through the inference rules (Appendix A.2).

Definition 1. *For $\delta \in [0,1]$, a $1 - \delta$ confidence sequence is a sequence of confidence sets, usually intervals $(\mathrm{CI_t})_{\mathrm{t}=1}^{\infty}$,, $\mathrm{CI_t} := (\mathrm{L_t}, \mathrm{R_t}) \subseteq \mathbb{R}$ satisfying a uniform convergence guarantee. After observing the t^{th} unit, we calculate an updated confidence set $\mathrm{CI_t}$ for an unknown quantity θ_t with the coverage property $\Pr(\forall t \geq 1, \theta_t : \theta_t \in \mathrm{CI_t}) \geq 1 - \delta$ [25].*

In this paper, we focus on the mean of r.v.s $\mathbb{E}[X]$ that constitute estimates for *elementary* subexpressions as the quantities of interest. We use adaptive concentration inequalities to construct these confidence sequences. Any adaptive concentration inequality that can be applied to an r.v. $X \in \{0,1\}$ such that

$$\Pr[|\overline{\mathbb{E}}_t[X] - \mathbb{E}[X]| \geq \varepsilon(t, \delta)] \leq \delta \tag{2}$$

can be used in AVOIR. Here, $\overline{\mathbb{E}}_t[X]$ is the empirical estimate of $\mathbb{E}[X]$ after the t^{th} observation. For comparison with previous work (e.g., VF), we use the Adaptive Hoeffding Inequality $\mathrm{AIN_H}$ [48].

Theorem 1 (AIN_H). *Given a Bernoulli random variable X with distribution P_X. Let $\{X_i \sim P_X\}$, $i \in \mathbb{N}$ be i.i.d samples of X. Let*

$$\overline{\mathbb{E}}_t[X] = \frac{1}{t}\sum_{i=1}^{t} X_i.$$

Let $\mathcal{T}$ be a r.v on $\mathbb{N} \cup \{\infty\}$ such that $\Pr[\mathcal{T} < \infty] = 1$, and let

$$\varepsilon(\delta, t) = \sqrt{\left(\frac{3}{5}\log\left(\log_{1.1} t + 1\right) + \frac{5}{9}\log\left(24/\delta\right)\right)\Big/ t}$$

Then, for any $\delta \in \mathbb{R}_{+}$, we have

$$\Pr[|\overline{\mathbb{E}}_{\mathcal{T}}[X] - \mathbb{E}[X]| \le \varepsilon(\delta, \mathcal{T})|] \ge 1 - \delta.$$

We will generate estimates using AIN_{H} and Corollary 4.1 for *elementary* subexpressions that are valid nonasymptotically (i.e., $\forall t > 1$) and then expand this to compound subexpressions.

Theorem 2. *The sequences of estimates generated by AVOIR form a confidence set.*

The intuition for the proof is as follows: first, for elementary subexpression X, let the failure probability at the stopping time be δ^{*}_{X}. From eq. (1), we can show that $\Delta \ge \delta^{*}_{X}$. Further, $\varepsilon(\delta, t)$ is monotonically decreasing in δ. Thus, setting $\delta_X(t) = \Delta$ as per Algorithm 1 before stopping time will ensure that the estimated confidence intervals before the stopping time corresponding to each time step for X would be a subset of the optimized values,

$$\left(\overline{\mathbb{E}}[X]_t \pm \varepsilon(\delta^{*}_{X}, t)\right) \subseteq \left(\overline{\mathbb{E}}[X]_t \pm \varepsilon(\Delta, t)\right)$$

where $(\mu \pm \sigma) = (\mu - \sigma, \mu + \sigma)$. Next, for compound subexpressions and specifications, the correctness of the inference rules used for propagating bounds (Figure 7) can be used to prove that the confidence sequence is valid nonasymptotically. We now proceed with the detailed proof. First, we assume the existence of a confidence sequence for the mean of each elementary subexpression (e.g., using Theorem 1). That is, we need an AIN for $\varepsilon(t, \delta)$ such that

$$\Pr[\forall t \ge 1, |\overline{\mathbb{E}}_t[X] - \mathbb{E}[X]| \le \varepsilon(t, \delta_X)] \ge 1 - \delta_X. \tag{3}$$

We assume $\varepsilon(t, \delta)$ to be monotonically non-increasing in δ and n. We expect this to be the case for most AIN, since increasing the number of observations of increasing the failure threshold should allow for additional concentration around the mean (e.g., this holds for AIN_{H}) Second, we assume that except in degenerate cases, AVOIR terminates at finite stopping time $\mathcal{T}$ (termination criteria in Corollary 3.2, Appendix).

Proof. *Elementary subexpressions:* Consider a specification ψ consisting of *elementary* subexpressions $X_1, \ldots, X_n$. At stopping time, let $\phi^{\mathcal{T}}_{X_i} := (\overline{\mathbb{E}}_{\mathcal{T}}[X_i], \varepsilon(\mathcal{T}, \delta_{X_i}), \delta_{X_i})$ be the stopping time estimates. Then, from the termination criterion, a solution to the optimization problem OPT exists, i.e,

$$\Delta \ge \sum_i \delta_{X_i} \tag{4}$$

The sequence of bounds claimed by AVOIR are

$$\varepsilon_{X_i}(t) = \begin{cases} \varepsilon(\Delta, t), & t < \mathcal{T}, \\ \varepsilon(\delta_{X_i}, t), & t \ge \mathcal{T} \end{cases} \tag{5}$$

From equation 4 and since $\delta_i \in [0, 1]$ we have $\Delta \ge \delta_{X_i}$. From the non-decreasing behavior of AIN

$$\varepsilon(\Delta, t) \le \varepsilon(\delta_i, t) \tag{6}$$

Now

$$\Pr[\forall t \ge 1, |\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| \le \varepsilon_{X_i}(t)]$$

$$= 1 - \Pr[\exists t \ge 1, |\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon_{X_i}(t)]$$

$$= 1 - \Pr\left[\bigcup_{t \ge 1} \left\{|\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon_{X_i}(t)\right\}\right]$$

$$= 1 - \Pr\left[\bigcup_{t=1}^{\mathcal{T}-1} \left\{|\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon_{X_i}(t)\right\} \cup \bigcup_{t \ge \mathcal{T}} \left\{|\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon_{X_i}(t)\right\}\right] \quad (\cup \text{ associativity})$$

$$= 1 - \Pr\left[\bigcup_{t=1}^{\mathcal{T}-1} \left\{|\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon(\delta_{X_i}, t) \cup |\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| \in (\varepsilon(\Delta, t), \varepsilon(\delta_{X_i}, t)]\right\} \cup \bigcup_{t \ge \mathcal{T}} \left\{|\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon(\delta_{X_i}, t)\right\}\right] \quad (\text{Using 5, 6})$$

$$= 1 - \Pr\left[\bigcup_{t=1}^{\mathcal{T}-1} \left\{|\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| \in (\varepsilon(\Delta, t), \varepsilon(\delta_{X_i}, t)]\right\} \cup \bigcup_{t \ge 1} \left\{|\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon(\delta_{X_i}, t)\right\}\right] \quad (\text{Rearranging})$$

$$\ge 1 - \Pr\left[\bigcup_{t \ge 1} \left\{|\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon(\delta_{X_i}, t)\right\}\right]$$

$$= 1 - \Pr\left[\exists t \ge 1, |\overline{\mathbb{E}}_t[X_i] - \mathbb{E}[X_i]| > \varepsilon(\delta_{X_i}, t)\right]$$

$$\ge 1 - \delta_{X_i}$$

where the last step follows from the definition of the AIN used. Thus, $\varepsilon_{X_i}(t)$ defines a $1 - \delta_{X_i}$ confidence sequence for $\mathbb{E}[X_i]$. *Compound subexpressions:* Consider a non-specification compound (`<ETerm>`) C_j consisting of *elementary* subexpressions with indices $\mathbf{C}_j = \{\{j_1, j_2, \ldots, j_M\}\}$ as the decision r.v.s, i.e, $X_{j_1} \ldots, X_{j_M}$. Note that $\mathbf{C}_j$ is a multiset as the same expression could occur multiple times within C_j. At stopping time $\mathcal{T}$,

$$\phi^{\mathcal{T}}_{C_j} : (\overline{\mathbb{E}}_{\mathcal{T}}[C_j], \delta_{C_j}, \varepsilon_{C_j}) \tag{7}$$

where $\overline{\mathbb{E}}_{\mathcal{T}}[C_j], \delta_{C_j}, \varepsilon_{C_j}$ are the corresponding values computed through the inference rules. In general, we denote by

$$\overline{\mathbb{E}}_t[C_j], \delta_{C_j}(t), \varepsilon_{C_j}(t) = \text{INFER}(\phi^{\text{t}}_{\text{X}_{\text{j}_1}}, \ldots, \phi^{\text{t}}_{\text{X}_{\text{j}_\text{M}}}) \tag{8}$$

the values inferred at t, using the inference rules INFER. Now,

$$\Pr[\exists t \ge 1, |\mathbb{E}[C_j] - \overline{\mathbb{E}}[C_j]| > \varepsilon_{C_j}(t)]$$

$$\le \Pr\left[\bigcup_{i=1}^{M} \exists t \ge 1, \neg\phi^{t}_{X_{j_i}}\right] \text{ (eq. (8))}$$

$$\le \sum_{i \in \mathbf{C}_j} \Pr\left[\exists t \ge 1, \neg\phi^{t}_{X_{j_i}}\right] \text{ (union bound)}$$

$$= \sum_{i \in \mathbf{C}_j} \Pr\left[\exists t \ge 1, |\overline{\mathbb{E}}_t[X_{j_i}] - \mathbb{E}_t[X_{j_i}| > \varepsilon_{X_{j_i}}(t)\right]$$

(definition of $\phi^{t}_{X_{j_i}}$)

$$\le \sum_{i \in \mathbf{C}_j} \delta_{X_{j_i}} \text{ (elementary subexpressions)}$$

$$\le \delta_{C_j} \text{ (eq. (8) at } t = \mathcal{T})$$

Therefore $\varepsilon_{C_j}(t)$ defines a $1 - \delta_{C_j}$ confidence sequence for $\mathbb{E}[C_j]$ A similar proof can be constructed for any <spec> (appendix B.1). □

Corollary 2.1. *The estimates for the overall specification ψ form a confidence sequence which staisfies $\psi : (b, \Delta), b \in \{T, F\}$ at $\mathcal{T}$.*

Proof. We initialize the main specification with the required failure probability Δ. At termination, $\sum \delta_i \leq \Delta$. From Theorem 2, we can infer that the confidence sequence corresponding to the termination achieves the threshold Δ, as required. □

3.4.2 Improvements over Baseline. In all prior work [1, 2, 4], δ_i for each *elementary* subexpressions is set to Δ/n, where n is the number of elementary subexpressions in the specification. This simplification uses the assumption $A_\delta := \delta_i = \delta_j \; \forall i, j$ for *elementary* subexpressions. As we do not make this assumption, we can prove the following critical theorem (note, Corollary 3.2 describes the conditions required for finite stopping).

Definition 2. *We define the specification stopping time $\mathcal{T}$ for a confidence sequence as the smallest t such that given a threshold Δ and a specification ψ, an inference algorithm terminates with $\Pr[\forall t \geq 1, \psi_t = \widehat{\psi}_{\mathcal{T}}] \geq 1 - \Delta$, where $\widehat{\psi}_{\mathcal{T}}$ is the estimate of ψ at $\mathcal{T}$.*

Theorem 3. *Given a threshold probability Δ for a specification ψ, let the stopping time for AVOIR be $\mathcal{T}$ and the stopping time with the A_δ assumption be $\mathcal{T}^+$. Then $\mathcal{T} \leq \mathcal{T}^+$*

Proof. Under A_δ, at the stopping time $\mathcal{T}^+$, $\delta_i^+ = \Delta/n$, with $\sum_{i=1}^{n} \delta_i^+ = \Delta$. As δ_i^+ are propagated using INFER (without constraint rules), we know that they must satisfy the constraints of the optimization problem in eq. (1). At time $\mathcal{T}^+$ AVOIR would find solution δ_i^* such that minimizes $\sum_{i=1}^{n} \delta_i$.

$$\sum_{i=1}^{n} \delta_i^* \leq \sum_{i=1}^{n} \delta_i^+ = \Delta$$

Thus, AVOIR would have terminated by step $\mathcal{T}^+$, but may find a feasible solution at an earlier step, i.e., $\mathcal{T} \leq \mathcal{T}^+$. □

3.5 Implementation Details

We built a Python library to create specifications as a decorator over decision functions. New input/output observations are monitored to update all the terms in a specification. Inference for evaluating the value and bounds is carried out via operator overloading.In line with previous work [1, 2, 4] on distributional verification, we use rejection sampling for conditional probability estimation. We use the COIN-OR implementation of IPOPT [42], accessed through the Pyomo [23] interface for optimization. Code for reproducing this work is available at https://github.com/pranavmaneriker/AVOIR.

4 EVALUATION

In this section, we evaluate AVOIR.variants through three real-world case studies. Direct comparisons with existing work are impossible since no other work (to our knowledge) facilitates a general-purpose inference engine for online fairness auditing using arbitrary measures. We can, however, implement VF's [4] inference rules within AVOIR (denoted as AVOIR-VF). Note that AVOIR-VF sidesteps the assumptions of having a known data-generating distribution (made possible by AVOIR's reliance on confidence sets), making this variation a more practical and efficient algorithm. We denote AVOIR-OB as the implementation that utilizes the above-mentioned optimizations. Across the studies, the role of chosen threshold probabilities is similar to that of p-values in statistics. Typical p-values tend to be 0.05 and 0.1, which we demonstrate in the RateMyProfs and COMPAS risk assessment study. In our case study of prior work [3], we stick to the available definitions and thresholds used in the original analysis. We expect that regulators will set the threshold probabilities on a case-by-case basis, e.g,, 0.15 for illustration purposes in the adult income study.

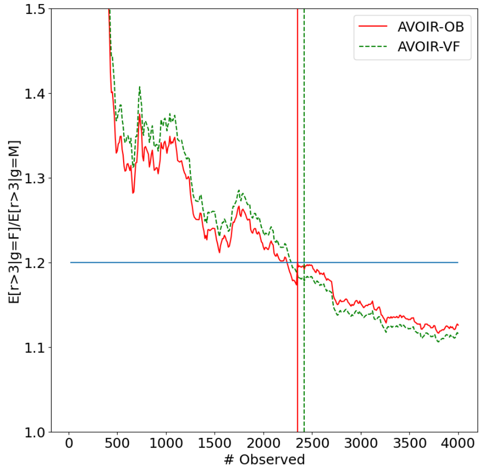


Figure 4: Bounds for first half of a gender-fairness specification generated by AVOIR-OB and AVOIR-VF for *RateMyProfs*, a real-world dataset. Vertical lines show the step at which the methods can provide a guarantee of failure for the upper bounds with Δ <= 0.05. Blue horizontal line represents the constant term in the inequality.

4.1 Rate My Profs

This section provides a detailed black-box machine learning model-based case study on a real-world dataset. This case study uses the Rate My Professors (RMP) dataset [28]. This dataset includes professor names and reviews for them written by students in their classes, ratings, and certain self-reported attributes of the reviewer. Ratings are provided on a five-point scale (1-5 stars). We use the preprocessing described in [28] to infer the gender attribute for the professors. This dataset is divided into an 80-20 split (train-test). We then train a BERT-based transformer model [11] on the training split. We use the implementation from the simpletransformers[4] package. The loss function chosen is the mean-squared error from the true ratings. On the test set, we track a gender-fairness specification in the model outputs:

```
(E[r > 3 | gender = F] / E[r > 3 | gender = M< 1.2) &
(E[r > 3 | gender = M)] / E[r > 3 | gender = F] > 0.8)
```

[4] https://simpletransformers.ai/

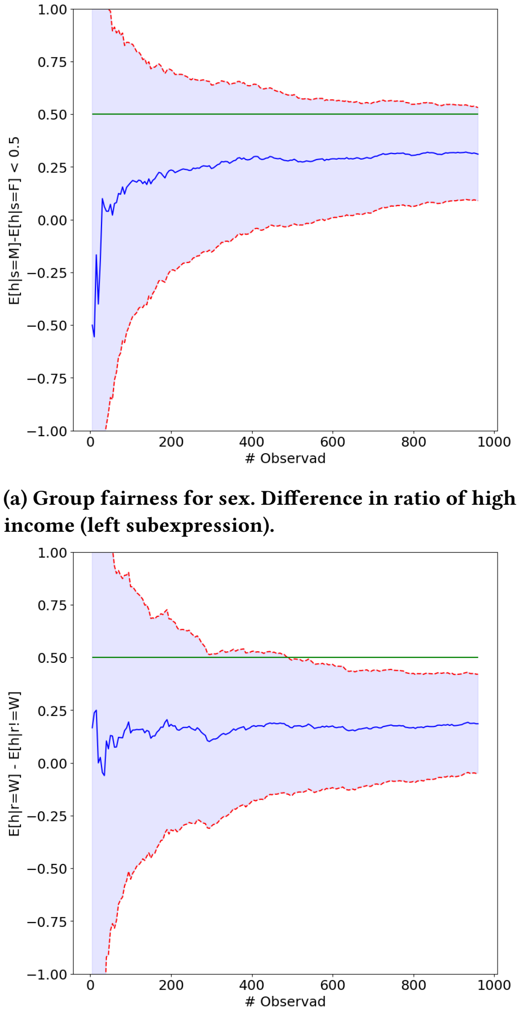
We set the failure probability $\Delta = 0.05$. OPT is run after each batch (5 items/batch). Figure 4 shows that AVOIR-OB[5] can provide a guarantee in **2.5**% fewer iterations than AVOIR-VF. Note also that the OB guarantee provided tries to optimize for the failure probability while staying under the required threshold, remaining closer to the required threshold in subsequent steps.

4.2 Adult Income

In this case study, we analyze the Adult income dataset [30]. The historical dataset labels individuals from the 1994 census as having a *high-income* (> 50k a year) or not (≤ 50k a year). We consider this column of data as a black-box measurement. US Federal laws mandate against race and sex-based discrimination. Thus, the specification we start our analysis with is a group fairness property for federal employees that monitors the difference of the proportions of individuals with sex marked male vs. female with a high income should be less than 0.5. In addition, we ensure that the difference between individuals with race marked white and non-white should have a difference of less than 0.5. Thus, we use an *intersectional* fairness criterion. The associated specification is given below, where h is an indicator for whether an individual is *high-income* is the binary classification output of our model:

```
(E[h | sex=M] - E[h | sex=F] < 0.5) & \
(E[h | race=W] - E[h | race!=W] < 0.5)
```

In this example, we set the failure threshold probability $\Delta = 0.15$ When run with this specification, the generated bounds cannot be achieved with the available data. We can then use the iterative refinement associated with subexpressions to analyze different components of the specification. The plot corresponding to the left subexpression is shown in Figure 5a shows that guarantees cannot converge under the threshold with the given number of data samples. An auditor can now choose to either reduce the guarantee (i.e. increase Δ) or increase the threshold. Next, analyzing the right subexpression, the race group fairness term can be guaranteed to be under the threshold (Figure 5b). Using this information, an auditor can make a decision to increase the threshold on the group fairness term for sex. As a hypothetical, suppose they increase it from 0.5 to 0.55 and rerun the analysis. OB can provide a guarantee at this threshold within 870 steps, whereas VF can provide it at 960 steps, demonstrating a relative improvement of about **10.35**%. Additionally, the optimal Δ split across the terms is $\approx (0.135, 0.36 * 10^4)$, which is far from the equal split allocated by VF. The reason for this split is that increasing the threshold for the first time provides the optimizer with additional legroom to better distribute the failure probabilities between the two terms.



(a) Group fairness for sex. Difference in ratio of high income (left subexpression).

(b) Group fairness for race. Difference in ratio of high-income earners (right subexpression).

Figure 5: *(Top)* Red dotted lines, the upper bounds of the value cannot be guaranteed to be under the threshold at the specified failure probability. *(Bottom)* Guarantee possible with given data. Green lines represent the constant term, and dark blue is the empirical mean.

4.3 COMPAS Risk Assessment

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism risk score data is a popular dataset for assessing machine bias of commercial tools used to assess a criminal defendant's likelihood to re-offend. The data is based on recidivism (re-offending) scores derived from software released by Northpointe and widely used across the United States for making sentencing decisions. In 2016, Angwin et al. [3] at ProPublica released an article and associated analysis code critiquing machine bias associated with race present in the COMPAS risk scores for a set of arrested individuals in Broward County, Florida, over two years. The analysis concluded that there were significant differences in the risk assessments of African-American and Caucasian individuals. Northpointe pushed back in a report [12] firmly rejecting the claims made by the ProPublica article; instead, they claimed that Angwin et al. [3] made several statistical and technical errors in the report. In this case study, we use AVOIR to study the claims of the two reports mentioned above. We create a materialized view of the ProPublica dataset by reproducing the preprocessing steps in the

[5] OB = Optimized Bounds

publicly available ProPublica analysis notebook[6]. We look at "Sample A" [12], where the analysis of the "not low" risk assessments using a logistic regression model reveals a high coefficient associated with the factor associated with race being African-American. In terms of a fairness metric, this corresponds to false positive rate (FPR) balance (predictive equality) [41] metrics. The associated specification in AVOIR grammar would be

```
E[hrisk | race=African−American & recid=0] /
E[hrisk | race=Caucasian & recid=0] < 1.1
```

Where `hrisk` is an indicator for high-risk assessments made by the *black-box* COMPAS tool as defined by Angwin et al. [3], `recid` is an indicator for re-offending within two years of first arrest, and a 90%-rule is used as the threshold. We choose a failure threshold probability of $\Delta = 0.1$, with the optimization run after every batch of 5 samples. AVOIR finds that when the decisions are made sequentially, online, the assertion for specification violation cannot be made with the required failure guarantee.

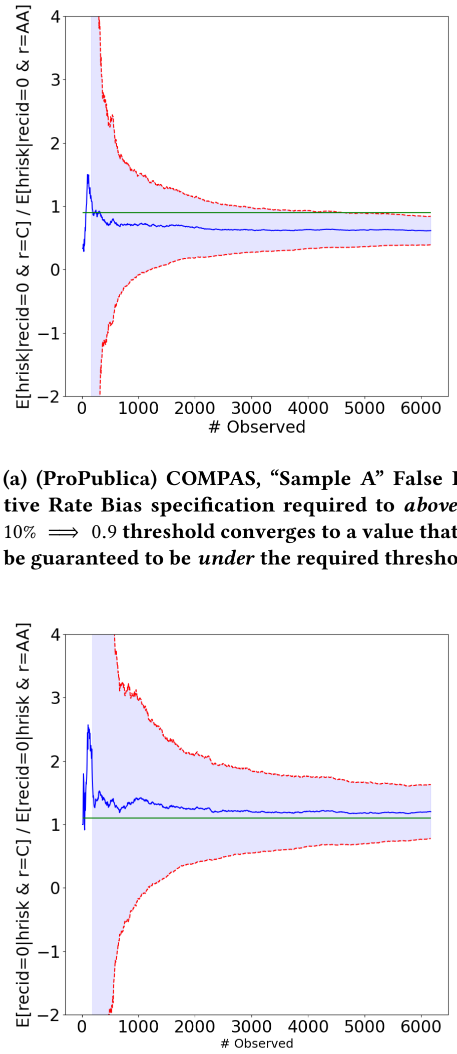
By analyzing the component subexpressions, one can glean that AVOIR cannot optimize since the lower FPR in the denominator (FPR for Caucasian individuals) increases the overall variance and limits the ability to optimize for guarantees. We follow this analysis by using the reciprocal specification, i.e.,

```
E[hrisk | race=Caucasian & recid=0] /
E[hrisk | race=African−American & recid=0] > 0.9
```

We find that the specification is guaranteed to be violated with a confidence of over $1 - \Delta = 0.9$ probability, and AVOIR can detect this violation within about half the number of available assessments (3350 steps) when run in an online setting. Figure 6a demonstrates the progression of the tracked expectation term. Thus, if deployed with the corrected specification, AVOIR would be able to alert Northpointe/an auditor of a violation/potentially-biased decision-making tool.

The Northpointe report [12] makes several claims about the shortcomings of this analysis. One of the primary claims is that Angwin et al. [3] used an analysis based on "Model Errors" rather than "Target Population Errors". In fairness specification terms, this refers to the difference between a False Positive Rate (FPR) balance vs. False Discovery Rate (FDR) balance, i.e., balancing for predictive parity over predictive equality. In probabilistic terms, the difference amounts to comparing $\Pr[\hat{Y} = 1 | Y = 0, g = 1, 2]$ (FPR) vs $\Pr[Y = 0 | \hat{Y} = 1, g = 1, 2]$ (FDR), where $\hat{Y}$ refers to the decision made by the algorithm, Y refers to the true value, and $g = 1, 2$ reflects group membership [41]. This analysis is run on the dataset subset dubbed "Sample B". To test their hypothesis, we reproduce the corresponding preprocessing steps and run both versions (numerator and denominator being Caucasian) of the corresponding specification under the same setup as earlier. Despite the point estimate being within the required threshold, we find that neither version can be guaranteed with the required confidence in the given data. Due to the paucity of space, we describe only one of the two variants with the corresponding figure (Figure 6b).

```
E[recid=0 | race=Caucasian & hrisk] /
E[recid=0 | race=African−American & hrisk] > 0.9
```



(a) (ProPublica) COMPAS, "Sample A" False Positive Rate Bias specification required to *above* the 10% $\implies$ 0.9 threshold converges to a value that can be guaranteed to be *under* the required threshold.

(b) (Northpointe) "Sample B" analysis done by Northpointe using False Discovery Rate that opposed the ProPublica reports.

Figure 6: COMPAS dataset case study.

We note that the Northpointe report [12] does not provide confidence intervals for their claim. Further, even though the report does not release associated code, the point estimates of the False Discovery Rates (FDRs) match those present in the report, which increases our confidence in our AVOIR-based analysis.

The back-and-forth exchange has been the subject of much discussion in academic and journalistic publications [16, 43]. Seminal work by Kleinberg et al. [29] proved the impossibility of simultaneously guaranteeing certain combinations of fairness metrics. While AVOIR cannot circumvent this problem, its usage can help audit claimed guarantees on defined metrics. We conclude this case study by noting that AVOIR lends itself to successful analysis that is not

[6] https://github.com/propublica/compas-analysis

possible with the VF implementation available online, which only provides support for a predefined set of specifications and requires access to a data-generating function. In addition, we choose 0.1 as the failure probability because it is one of the thresholds used in [3]. We set it to the highest used threshold to allow leeway for the claim by Northpointe. Even under this lax threshold, the analysis by Northpointe fails.

5 RELATED WORK

There are a plethora of fairness criteria, and subtle changes in their definition can change the implications on decision-making [7]. Practitioners need support when selecting, designing, and guaranteeing fairness for deployed machine learning algorithms. Prior work on fairness has helped develop nuanced notions and algorithms to help train more 'fair' machine learning models. These include group fairness measures such as inter alia, minimizing disparate impact [6, 15], maximizing the equality of opportunity [22] In contrast with group fairness notions, causal notions of fairness [31] and individualized notions of fairness [13] provide alternative statistical mechanisms for understanding discriminatory behaviors of automated decision systems. Thomas et al. [38] proposed the Seldonian Framework as a generic mechanism for model users to design algorithms that help train machine learning models that can regulate them against undesirable behaviors. Yan and Zhang [45] propose a query-efficient framework to audit an unknown function chosen from a known hypothesis class of decision-making functions.

We focus on the problem of detecting and diagnosing whether systems designed under any framework follow any prescribed regulatory constraints supported within the grammar of AVOIR. That is, we are agnostic to the framework; instead, we are interested in testing the adherence of models to specified criteria. We use a probabilistic framework to verify this behavior. Alternative frameworks such as the AI Fairness 360 [5] provide mechanisms to quantify fairness uncertainty, though they are restricted to pre-supported metrics. Uncertainty quantification [20, 21] is an alternative mechanism to provide adaptive guarantees. However, existing work is designed for commonly used outcome metrics, such as accuracy and F1-score, rather than for fairness metrics. Justicia [19] optimizes uncertainty for fairness metrics estimates using stochastic SAT solvers but can only be applied to a limited class of tree-based classification algorithms.

Machine learning testing [47] is an avenue that can expose undesired behavior and improve the trustworthiness of machine learning systems. Prior work on fairness testing is most closely related to AVOIR. Fairness testing [18] provides a notion of causal fairness and generates tests to check the fairness of a given decision-making procedure. Given a specific definition of fairness, Fairtest [39] and Verifair (VF) [4] build a comprehensive framework for investigating fairness in data-driven pipelines. Fairness-aware Programming (FP) [2] combined the two demands of machine learning testing and fairness auditing to make fairness a first-class concern in programming. Fairness-aware programming applies a runtime monitoring system for a decision-making procedure with respect to an initially stated fairness specification. The overall failure probability of an assertion is computed as the sum of the failure probabilities of each constituting sub-expression (using the union bound). FP does not provide any specific mechanism for splitting uncertainty, and Verifair splits it equally across all constituent *elementary subexpressions*. Thus, assertion bounds for subexpressions in both FP and VF are split inefficiently compared to AVOIR.

6 CONCLUSION & FUTURE WORK

We presented the AVOIR framework to easily define and monitor fairness specifications online and aid in the refinement of specifications. AVOIR is easy to integrate within modern database systems but can also serve as a standalone system evaluating whether black-box machine learning models meet specific fairness criteria on specific datasets (including both structured and unstructured data) as described in our case studies. AVOIR extends the grammar from Fairness Aware Programming [2] with operations that enhance expressiveness. In addition, we derive probabilistic guarantees that improve the confidence with which specification violations are reported. Through case studies, we demonstrate that AVOIR can provide users with insights and context that contribute directly to refinement decisions. To understand the robustness of AVOIR, we evaluated it along two dimensions: the data/ML model used and changing parameters (thresholds, fairness definitions). We demonstrated the robustness of the data/model used by evaluating three datasets of varying domains and types (criminal justice - COMPAS, text classification - RateMyProfs, census data - Adult Income). For robustness to the thresholds, we used varying failure probability levels (0.05, 0.1, 0.15) in our case studies. Note that any probability thresholds over these values for the corresponding studies would converge in fewer iterations, while lower thresholds would require additional data samples. Our framework builds the foundation for further improvements in fairness specification, auditing, and verification workflows. Although contextual information from AVOIR makes decisions more straightforward, it is not always clear how to alter a specification in light of a violation and its relevant context.

To assist in these decisions, we are currently examining mechanisms that suggest edits that are likely to achieve the desired intent of a model developer. We plan to extend this work to provide intelligent specification refinement suggestions and support distributed machine learning settings. In addition to improving the usability of our tools for making fairness specification refinements, we also envision a more scalable framework. Our case studies looked at a single model with respect to a single dataset. However, real-world deployment of machine learning often contains many clients with models and datasets that may evolve and drift over time. We also expect to examine efficient monitoring of machine learning behavior for a fairness specification in a distributed context, enabling horizontal scalability. We believe techniques such as decoupling the observation of data and reporting results from monitoring the results are promising and can lead to the desired scalability.

ACKNOWLEDGMENTS

The authors acknowledge support from National Science Foundation (NSF) grants #1949037 and #2112471 (AI-EDGE). Any opinions and findings are those of the author(s) and do not necessarily reflect the views of the granting agencies.

REFERENCES

[1] Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V Nori. 2017. Fairsquare: probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages* 1, OOPSLA (2017), 1–30.

[2] Aws Albarghouthi and Samuel Vinitsky. 2019. Fairness-Aware Programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 211–219. https://doi.org/10.1145/3287560.3287588

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[4] Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. 2019. Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages* 3, OOPSLA (2019), 1–27.

[5] 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. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

[6] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[7] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, and Daniele Regoli. 2021. The zoo of fairness metrics in machine learning. *arXiv preprint arXiv:2106.00467* (2021).

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

[9] US Congress. 2023. H.R.3369 - AI Accountability Act. https://www.congress.gov/bill/118th-congress/house-bill/3369/text

[10] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 864–876.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* 7, 4 (2016).

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[14] The U.S. EEOC. 1979. Uniform guidelines on employee selection procedures. (March 1979).

[15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[16] Avi Feller, Emma Pierson, Sam Corbett-Davies, and Sharad Goel. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post* 17 (2016).

[17] Gemma Galdon Clavell, Mariano Martín Zamorano, Carlos Castillo, Oliver Smith, and Aleksandar Matic. 2020. *Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization*. Association for Computing Machinery, 265–271.

[18] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*. 498–510.

[19] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. 2021. Justicia: a stochastic SAT approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 7554–7563.

[20] Soumya Ghosh, Q Vera Liao, Karthikeyan Natesan Ramamurthy, Jiri Navratil, Prasanna Sattigeri, Kush R Varshney, and Yunfeng Zhang. 2021. Uncertainty Quantification 360: A Holistic Toolkit for Quantifying and Communicating the Uncertainty of AI. *arXiv preprint arXiv:2106.01410* (2021).

[21] Tony Ginart, Martin Jinye Zhang, and James Zou. 2022. MLDemon: Deployment Monitoring for Machine Learning Systems. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3962–3997.

[22] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[23] William E Hart, Jean-Paul Watson, and David L Woodruff. 2011. Pyomo: modeling and solving mathematical programs in Python. *Mathematical Programming Computation* 3, 3 (2011), 219–260.

[24] Dennis Hirsch, Tim Bartley, Arvind Chandrasekaran, Srinivasan Parthasarathy, Piers Turner, Devon Norris, Keir Lamont, and Christina Drummond. 2020. Corporate Data Ethics: Data Governance Transformations for the Age of Advanced Analytics and AI (Final Report). In *Appeared at the Privacy Law Scholars Conference (also available as SSRN Abstract: 3828239)*.

[25] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. 2021. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* 49, 2 (2021), 1055–1080.

[26] Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can i trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. *Advances in Neural Information Processing Systems* 33 (2020), 18600–18612.

[27] B Justice Srikrishna. 2018. A free and fair digital economy: Protecting privacy, empowering Indians.

[28] Moniba Keymanesh, Tanya Berger-Wolf, Micha Elsner, and Srinivasan Parthasarathy. 2021. Fairness-aware summarization for justified decision-making. *arXiv preprint arXiv:2107.06243* (2021).

[29] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[30] Ron Kohavi. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.. In *Kdd*, Vol. 96. 202–207.

[31] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[32] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[33] E.F. Moore. 1956. Gedanken-experiments on sequential machines. *Automata Studies, Princeton University Press* 129-153 (1956).

[34] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2, 1 (1996), 25–42.

[35] Central Digital Office and Data. 2021. Algorithmic transparency standard. https://www.gov.uk/government/collections/algorithmic-transparency-standard

[36] European Parliament. 2018. 2018 reform of EU data protection rules. *European Commission* (2018).

[37] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 56–67.

[38] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (2019), 999–1004.

[39] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering Unwarranted Associations in Data-Driven Applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS P)*. 401–416. https://doi.org/10.1109/EuroSP.2017.29

[40] Richard Vanderford. 2022. New York's Landmark AI Bias Law Prompts Uncertainty. *WSJ* (2022). https://www.wsj.com/articles/new-yorks-landmark-ai-bias-law-prompts-uncertainty-11663752602

[41] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[42] Andreas Wächter and Lorenz T Biegler. 2006. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming* 106, 1 (2006), 25–57.

[43] Anne L Washington. 2018. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ* 17 (2018), 131.

[44] Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H Kennedy, and Aaditya Ramdas. 2021. Time-uniform central limit theory, asymptotic confidence sequences, and anytime-valid causal inference. *arXiv preprint arXiv:2103.06476* (2021).

[45] Tom Yan and Chicheng Zhang. 2022. Active fairness auditing. In *International Conference on Machine Learning*. PMLR, 24929–24962.

[46] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.

[47] Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine Learning Testing: Survey, Landscapes and Horizons. *IEEE Transactions on Software Engineering* (2020), 1–1. https://doi.org/10.1109/TSE.2019.2962027

[48] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. 2016. Adaptive concentration inequalities for sequential decision problems. *Advances in Neural Information Processing Systems* 29 (2016).

$$\frac{X : \left(\overline{\mathbb{E}}[X], \varepsilon_X, \delta_X\right), Y : \left(\overline{\mathbb{E}}[Y], \varepsilon_Y, \delta_Y\right)}{X \pm Y : \left(\overline{\mathbb{E}}[X] \pm \overline{\mathbb{E}}[Y], \varepsilon_X + \varepsilon_Y, \delta_X + \delta_Y\right)}$$

$$\frac{X : \left(\overline{\mathbb{E}}[X], \varepsilon_X, \delta_X\right), Y : \left(\overline{\mathbb{E}}[Y], \varepsilon_Y, \delta_Y\right)}{X \times Y : (\overline{\mathbb{E}}[X]\overline{\mathbb{E}}[Y], \varepsilon_X\varepsilon_Y + \overline{\mathbb{E}}[X]\varepsilon_Y + \overline{\mathbb{E}}[Y]\varepsilon_X, \delta_X + \delta_Y)}$$

$$\frac{X : \left(\overline{\mathbb{E}}, \varepsilon, \delta\right), \overline{\mathbb{E}} - \varepsilon > 0}{X^{-1} : \left(\overline{\mathbb{E}}^{-1}, \frac{\varepsilon}{\overline{\mathbb{E}}(\overline{\mathbb{E}}-\varepsilon)}, \delta\right)} \text{(Inverse)}$$

$$\frac{X : \left(\overline{\mathbb{E}}, \varepsilon, \delta\right)}{X^{-1} : \left(\overline{\mathbb{E}}^{-1}, \frac{\varepsilon}{\overline{\mathbb{E}}(\overline{\mathbb{E}}-\varepsilon)}, \delta\right), \{\overline{\mathbb{E}} - \varepsilon > 0\}} \text{(Inverse C)}$$

$$\frac{X : \left(\overline{\mathbb{E}}, \varepsilon, \delta\right), \overline{\mathbb{E}} - \varepsilon > c}{X > c : (T, \delta)} \text{(True)} \quad \frac{X : \left(\overline{\mathbb{E}}, \varepsilon, \delta\right), \overline{\mathbb{E}} + \varepsilon < c}{X < c : (F, \delta)} \text{(False)}$$

$$\frac{X : \left(\overline{\mathbb{E}}, \varepsilon, \delta\right)}{X > c : (T, \delta), \{\overline{\mathbb{E}} - \varepsilon > c\}} \text{(True C)}$$

$$\frac{X : \left(\overline{\mathbb{E}}, \varepsilon, \delta\right)}{X < c : (T, \delta), \{\overline{\mathbb{E}} + \varepsilon < c\}} \text{(False C)}$$

$$\frac{\psi_1 : (\mathbb{B}_1, \delta_1), \psi_2 : (\mathbb{B}_2, \delta_2)}{\psi_1 \wedge \psi_2 : (\mathbb{B}_1 \wedge \mathbb{B}_2, \delta_1 + \delta_2)} \text{(and)} \frac{\psi_1 : (\mathbb{B}_1, \delta_1), \psi_2 : (\mathbb{B}_2, \delta_2)}{\psi_1 \vee \psi_2 : (\mathbb{B}_1 \vee \mathbb{B}_2, \delta_1 + \delta_2)} \text{(or)}$$

$$\frac{\psi_1 : (\mathbb{B}_1, \delta_1), \{C_{11,\ldots,1k}\}, \psi_2 : (\mathbb{B}_2, \delta_2), \{C_{21,\ldots,2m}\}}{\psi_1 \wedge \psi_2 : (\mathbb{B}_1 \wedge \mathbb{B}_2, \delta_1 + \delta_2), \{C_{11,\ldots,1k}, C_{21,\ldots,2m}\}} \text{(and C)}$$

$$\frac{\psi_1 : (\mathbb{B}_1, \delta_1), \{C_{11,\ldots,1k}\}, \psi_2 : (\mathbb{B}_2, \delta_2)}{\psi_1 \vee \psi_2 : (\mathbb{B}_1 \vee \mathbb{B}_2, \delta_1 + \delta_2), \{C_{11,\ldots,1k}\} \vee \{C_{21,\ldots,2m}\}} \text{(or C)}$$

Figure 7: Inference rules used to guarantees for expressions.The inference rules for each compound expression build on the union bound, triangle inequality, and structural induction approach described by [4]. C: Constraint.

A INFERENCE RULES

In Figure 7, we provide the rules used to determining the constraints and guarantees for a specification. We represent $X \odot Y : (E, \varepsilon, \delta) \equiv \Pr(|\mathbb{E}[X] \odot \mathbb{E}[Y] - E| \geq \varepsilon) \leq \delta$ where $\odot$ represents a binary operator. Constraints are represented in $\{\}$. The proof of correctness for each inference rule starts from the assumptions above the horizontal line and derives the assertions below. These proofs use ideas similar to those in [4]. We reproduce the proofs in Appendix A.1 here for completeness. Note that the assertions in the base case (elementary subexpressions) can be arrived at by applying AIN.

A.1 Inference rules with Constraints

In Section 3.3 we provided the proofs for $X \pm Y$, $X > c$. In the following text, we provide the remaining proofs.

Product Starting with ϕ_X, ϕ_Y First, from union bound, both of these hold true with probability at least $1 - \delta_X - \delta_Y$. Then,

$$|\mathbb{E}[X]| = |\overline{\mathbb{E}}[X] - \overline{\mathbb{E}}[X] + \mathbb{E}[X]|$$

$$\leq ||\overline{\mathbb{E}}[X]| + |\overline{\mathbb{E}}[X] + \mathbb{E}[X]| \leq ||\overline{\mathbb{E}}[X]| + \varepsilon_X$$

$$\begin{aligned}
|\overline{\mathbb{E}}[X]\overline{\mathbb{E}}[Y] - \mathbb{E}[XY]| &= |\overline{\mathbb{E}}[X]\overline{\mathbb{E}}[Y] - \mathbb{E}[X]\mathbb{E}[Y]| \\
&= |\overline{\mathbb{E}}[X](\overline{\mathbb{E}}[Y] - \mathbb{E}[Y]) + \mathbb{E}[Y](\overline{\mathbb{E}}[X] \ - \mathbb{E}[X])| \\
&\leq |\overline{\mathbb{E}}[X]||(\overline{\mathbb{E}}[Y] - \mathbb{E}[Y])| + |\mathbb{E}[Y]||(\overline{\mathbb{E}}[X] \ - \mathbb{E}[X])| \\
&\leq |\overline{\mathbb{E}}[X]|\varepsilon_Y + |\mathbb{E}[Y]|\varepsilon_X \\
&\leq |\overline{\mathbb{E}}[X]|\varepsilon_Y + (|\overline{\mathbb{E}}[Y]| + \varepsilon_Y)\varepsilon_X \\
&= |\overline{\mathbb{E}}[X]|\varepsilon_Y + |\overline{\mathbb{E}}[Y]|\varepsilon_X + \varepsilon_X\varepsilon_Y
\end{aligned}$$

where the first step follows as X, Y are Bernoulli r.vs. Therefore, $X \times Y : (\overline{\mathbb{E}}[X]\overline{\mathbb{E}}[Y], \varepsilon_X\varepsilon_Y + \overline{\mathbb{E}}[X]\varepsilon_Y + \overline{\mathbb{E}}[Y]\varepsilon_X, \delta_X + \delta_Y)$

Inverse/Inverse C Assume $X : \left(\overline{\mathbb{E}}, \varepsilon, \delta\right)$ and $\overline{\mathbb{E}} - \varepsilon > 0$. In the constrained case, we start with only the prior assumption. Then,

$$\begin{aligned}
|\mathbb{E}[X]| &= |\mathbb{E}[X] - \overline{\mathbb{E}}[X] + \overline{\mathbb{E}}[X]| \\
&\leq |\mathbb{E}[X] - \overline{\mathbb{E}}[X]| + |\overline{\mathbb{E}}[X]| \leq \varepsilon_X + |\overline{\mathbb{E}}[X]|
\end{aligned}$$

i.e., $|\mathbb{E}[X]| \leq \varepsilon_X + |\overline{\mathbb{E}}[X]|$. Also,

$$\begin{aligned}
|\mathbb{E}[X]^{-1} - \overline{\mathbb{E}}[X]^{-1}| &= \left|\frac{\overline{\mathbb{E}}[X]^{-1} - \mathbb{E}[X]^{-1}}{\overline{\mathbb{E}}[X]\mathbb{E}[X]^{-1}}\right| \\
&\leq \frac{\varepsilon}{|\mathbb{E}[X]||\overline{\mathbb{E}}[X]|} \leq \frac{\varepsilon}{|\mathbb{E}[X]|(\mathbb{E}[X] - \varepsilon_X)}
\end{aligned}$$

VF adds $E[X] - \varepsilon_X > 0$ as a precondition; AVOIR as a post-constraint.

Boolean Operators. Starting from $\psi_1 : (b_1, \delta_1)$, $\psi_2 : (b_2, \delta_2)$, we can apply the union bound for $\psi_1 \wedge \psi_2$, $\psi_1 \vee \psi_2$ to derive the rules for and/or. Similarly, constraints follow the semantics specified by the rules as they also follow from the union bound.

A.2 Inferred Optimization Problem

For a given overall specification ψ, suppose $(\varepsilon_i, \delta_i)$, $i \in \{1, \ldots, n\}$ represents the concentration bounds associated with each constituent elementary subexpression. Using the inference rules, we can derive the overall $\delta_T = \sum_i \delta_i$, along with a set of (say) K constraints

$$g_k(\varepsilon_1, \ldots, \varepsilon_n, \overline{\mathbb{E}}[X_1], \ldots, \overline{\mathbb{E}}[X_n]) \leq \varepsilon_k$$
$$\text{where } \varepsilon_k = \left|c_k - \overline{\mathbb{E}}[f(\overline{\mathbb{E}}[X_1], \ldots, \overline{\mathbb{E}}[X_n])]\right|$$

denotes the maximum allowed margin for the k$^{\text{th}}$ subexpression of form `<ETerm> <comp-op> c`). The objective is to minimize the overall failure probability δ_T. The overall optimization problem can then be formulated as shown in 1, having n optimization variables δ_i and $2n + K$ constraints (bounds on δ_i provide the $2n$ constraints). A developer using AVOIR inputs a required acceptable upper bound of failure probability Δ. If the solution to the optimization problem $\delta^*_T = \sum_i \delta_i \leq \Delta$, then the optimization can conclude with the required confidence in the proved guarantee. At this point, the developer may choose to terminate AVOIR. However, using Corollary 4.1, they may continue to run and refine the estimates.

Metric Name	Definition/DSL
Statistical Parity [13]	$\Pr[R \mid S] = \Pr[R \mid \neg S]$
	$\mathbb{E}[r \mid s] / \mathbb{E}[r \mid !s] < c$
Predictive Parity [8]	$\Pr[Y \mid R, S] = \Pr[Y \mid R, S]$
	$\mathbb{E}[y \mid r, s] - \mathbb{E}[y \mid r, s] > c$
Equal Opportunity [22]	$\Pr[\neg R \mid Y, S] = \Pr[\neg R \mid Y, \neg S]$
	$\mathbb{E}[!r \mid y, s] - \mathbb{E}[!r \mid y, !s] < c$
Equalized Odds [22]	$\Pr[R \mid Y = i, S] = \Pr[R \mid Y = i, \neg S]$, $i = 0, 1$
	$(\mathbb{E}[r \mid y = 0, !s] - \mathbb{E}[r \mid y = 0, s] > c_0)$ & $(\mathbb{E}[r \mid y = 1, !s] - \mathbb{E}[r \mid y = 1, s] > c_1)$

Table 2: Examples of supported metrics.

B CONCENTRATION BOUNDS

Theorem 1 provides a mechanism for choosing the stopping time using arbitrary methods for a fixed δ. In general, any adaptive concentration inequality suffices; we use AIN_H However, we use confidence intervals to visualize the evolution of sub-expressions (and overall specification) over the sequence of observations. To do so, we require an additional result.

Theorem 4. *[48, Proposition 1, Lemma 1] Let $S_n = \sum_{i=1}^{n} X_i$ be a random walk from i.i.d. random variables $X_1, \ldots, X_t \sim D$. For any $\delta > 0$, $\Pr[S_\mathcal{T} \geq f(\mathcal{T})] \leq \delta$ for any stopping time $\mathcal{T}$ if and only if $\Pr\left[\exists n, S_t \geq f(t)\right] \leq \delta$*

Corollary 4.1. *For $\delta > 0$, $\Pr[|\overline{\mathbb{E}}_\mathcal{T}[X] - \mathbb{E}[X]| \leq \varepsilon(\delta, \mathcal{T})|] \geq 1 - \delta$ for any stopping time $\mathcal{T}$ if and only if*

$$\Pr\left[\forall t, |\overline{\mathbb{E}}_t[X] - \mathbb{E}[X]| \leq \varepsilon(\delta, t)|\right] \geq 1 - \delta$$

Corollary 4 follows directly from applying Theorem 4 to Theorem 1. Intuitively, Theorem 1 holds since we can choose an adversarial stopping rule for $\mathcal{T}$ that terminates as soon as the boundary for $\varepsilon(\delta, t)$ is crossed [48]. Thus, when we establish a bound with a stopping rule, the bound will hold prior to and after the stopping rule is enforced. Corollary 4.1 implies that once we choose an optimal bound for each subexpression, we can extend the bounds derived using Theorem 1 to following observations with continued guarantees for subexpressions.

B.1 Proof of Theorem 2 for Specifications

Consider any specification ψ_k. Let $\psi_k^t : (\hat{b}_{\psi_k}(t), \delta_{\psi_k}(t))$, where $\hat{b}_{\psi_k}(t) \subseteq \{T, F\}$ is the inferred value and $\delta_{\psi_k}(t)$ corresponds to the confidence for the assertion at time t. Let the *elementary* subexpressions involved be $X_{k_1}, \ldots, X_{k_D}$ corresponding to the index multiset $B_k = \{\{k_1, \ldots, k_D\}\}$. Denote b_{ψ_k} as the true value of ψ_k, and δ_{ψ_k} as the inferred threshold at stopping time $\mathcal{T}$. From INFER, we have

$$\hat{b}_k(t), \delta_{\psi_k}(t) = \text{INFER}(\phi^\text{t}_{\text{X}_{\text{k}_1}}, \ldots, \phi^\text{t}_{\text{X}_{\text{k}_\text{D}}}) \tag{9}$$

$$\begin{aligned}
&\Pr[\exists t \geq 1, b_k \notin \hat{b}_k(T)] \\
&\quad \leq \Pr\left[\bigcup_{i=1}^{D} \exists t \geq 1, \neg\phi^t_{X_{k_i}}\right] \text{ (From 9)} \\
&\quad \leq \sum_{i \in B_k} \Pr\left[\exists t \geq 1, \neg\phi^t_{X_{k_i}}\right] \text{ (union bound)} \\
&\quad = \sum_{i \in B_j} \Pr\left[\exists t \geq 1, |\overline{\mathbb{E}}_t[X_{k_i}] - \mathbb{E}_t[X_{k_i}| > \varepsilon_{X_{k_i}}(t)\right] \\
&\quad \leq \sum_{i \in B_j} \delta_{X_{k_i}} \text{(elementary subexpressions)} \\
&\quad \leq \delta_{\psi_k} \text{ (applying 8 for } t = \mathcal{T}\text{)}
\end{aligned}$$

Thus, $b_{\psi_k}(t)$ is a $1 - \delta_{\psi_k}$ confidence sequence for b_{ψ_k}

C TERMINATION CRITERION FOR AVOIR

Corollary 3.2. *Under mild conditions, AVOIR terminates in finite steps with an assertion over the required specification.*

Proof. We know that the stopping time $\mathcal{T} \leq \mathcal{T}^+$, the stopping time for AVOIR. Thus, AVOIR would terminate whenever Verifiar can. For completeness, we provide the conditions under which Verifair terminates. Note that $c \in \mathbb{R}$ corresponds to a constant threshold involved in specification, also presented in the grammar and bound proagation rules.

- For every subexpression C_k occurring in the specification such that it is involved in the inverse or inverse constr. rules (i.e., $\overline{\mathbb{E}}[C_k]^{-1}$), $\overline{\mathbb{E}}[C_k] \neq 0$, $C_k \neq 0$
- For every subexpression C_k such that it occurs a True/False type inequality (such as $C_k > c$), $\overline{\mathbb{E}}[C_k] \neq c$, $C_k \neq c$

□

D SUPPORTED METRICS

We provide a non-exhaustive list of statistical group-based fairness criteria and show an exact/approximate equivalent in the AVOIR DSL in Table 2. We use the notation from Table 1, assuming that the return value R is a Bernoulli r.v. We assume that the decision function f tracked by AVOIR as a signature that takes X, G, Y as input and produces S or d as output. Note that in their python implementation, = would be replaced by == and | by the given keyword.