Fairness Shields: Safeguarding against Biased Decision Makers

Filip Cano[1], Thomas A. Henzinger[2], Bettina Könighofer[1], Konstantin Kueffner[2], Kaushik Mallik[3]*

[1]Graz University of Technology
[2]Institute of Science and Technology Austria (ISTA)
[3]IMDEA Software Institute
filipcano95@gmail.com, {tah, konstantin.kueffner}@ist.ac.at, bettina.koenighofer@tugraz.at, kaushik.mallik@imdea.org

Abstract

As AI-based decision-makers increasingly influence human lives, it is a growing concern that their decisions may be unfair or biased with respect to people's protected attributes, such as gender and race. Most existing bias prevention measures provide probabilistic fairness guarantees in the long run, and it is possible that the decisions are biased on any decision sequence of fixed length. We introduce *fairness shielding*, where a symbolic decision-maker—the fairness shield—continuously monitors the sequence of decisions of another deployed black-box decision-maker, and makes interventions so that a given fairness criterion is met while the total intervention costs are minimized. We present four different algorithms for computing fairness shields, among which one guarantees fairness over fixed horizons, and three guarantee fairness periodically after fixed intervals. Given a distribution over future decisions and their intervention costs, our algorithms solve different instances of bounded-horizon optimal control problems with different levels of computational costs and optimality guarantees. Our empirical evaluation demonstrates the effectiveness of these shields in ensuring fairness while maintaining cost efficiency across various scenarios.

1 Introduction

With the increasing popularity of machine learning (ML) in human-centric decision-making tasks, including banking (Liu et al. 2018) and college admissions (Oneto et al. 2020), it is a growing concern that the decision-makers often show biases based on protected attributes of individuals, like gender and race (Dressel et al. 2018; Obermeyer et al. 2019; Scheuerman et al. 2019). Therefore, mitigating biases in ML decision-makers is an important problem and an active area of research in AI.

A majority of existing bias prevention techniques use *design-time* interventions, like pre-processing the training dataset (Kamiran et al. 2012; Calders et al. 2013), tailoring the loss function used for training (Agarwal et al. 2018; Berk et al. 2017), or post-processing the decisions using a statically calibrated output function (Hardt et al. 2016; Caton et al. 2020). We propose *fairness shielding*, the first *run-time* intervention procedure for safeguarding the fairness of already deployed decision-makers.

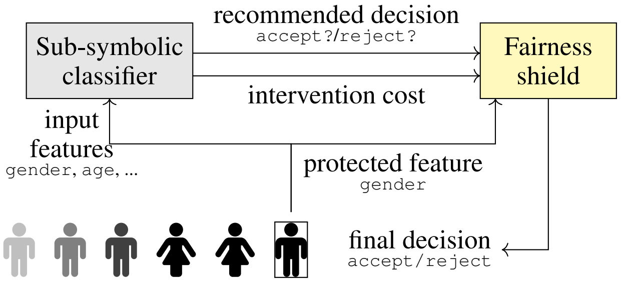


Figure 1: The operational diagram of fairness shields.

Fairness shields consider fairness from the *sequential* decision-making standpoint, where decisions are made on individuals appearing sequentially over a period of time, and each decision may be influenced by those made in the past. While the classical fairness literature typically evaluated fairness in a history-independent manner, the sequential setting, with its history-dependence, has been shown to better capture real-world decision-making problems (Zhang et al. 2021). Among the works on fairness in the sequential setting, most prior works are aimed at achieving fairness in the long run (Hu et al. 2022). Recently, the *bounded-horizon* and the *periodic* variants have been proposed, which are often more realistic, as regulatory bodies usually assess fairness after bounded time or at regular intervals, such as yearly or quarterly (Alamdari et al. 2024). Our fairness shields guarantee bounded-horizon and periodic fairness of deployed unknown decision-makers by monitoring each decision and minimally intervening if necessary. Fairness is guaranteed on *all* runs of the system, whereas existing algorithms guarantee fairness only *on average*, leaving individual runs prone to exhibit biases (Alamdari et al. 2024).

The basic functionality of fairness shields is depicted in Fig. 1. We assume a bounded-horizon or periodic fairness property is given, with a known time horizon or period, respectively. For each individual appearing in sequence, the fairness shield observes the protected attribute, the classifier's recommendation, and the cost of changing that recommendation to a different value, where the cost is assumed to be either provided by the decision-maker or pre-specified as constant. The shield then makes the final decision, ensur-

[*] Part of the research was done when the author was at ISTA.

ing that the given fairness criteria will be fulfilled while the associated total intervention cost is minimized.

Example 1 (Running example - Bilingual team)**.** *Consider the task of building a customer service team in a bilingual country where both languages, A and B, hold official status. To ensure high-quality service, it is essential to maintain a balanced representation of competent speakers of both languages. To achieve this, the company enforces a policy requiring that the difference between the number of employees proficient in each language must not exceed* 20% *of the total team size. The hiring process operates within a bounded time horizon, with a fixed number of T candidates to be screened. Candidates apply sequentially, and decisions about each applicant must be made before considering future candidates. An automated decision-making system, such as an ML model, screens applicants, which may or may not have been designed with fairness considerations. A fairness shield can be deployed, which will monitor and intervene in the decisions at runtime to guarantee that the final team is linguistically balanced as required while keeping the deviations from the decision-maker's at a minimum.*

Computation of fairness shields. Fairness shields are computed by solving bounded-horizon optimal control problems, which incorporate a *hard fairness constraint* and a *soft cost constraint* designed to discourage interventions. For the hard fairness constraint, we consider the empirical variants of standard group fairness properties, like demographic parity and equal opportunity. We require that the *empirical bias remains below a given threshold* with the bias being measured either at the end of the horizon or periodically.

For the soft cost constraint, it is assumed that the shield receives a separate cost penalty for each decision modification. The shield is then required to minimize the total expected future cost, either over the entire horizon or within each period. The definition of cost is subjective and varies by application. In our experiments, we consider constant costs that simply discourage a high *number* of interventions. Future works will include more fine-grained cost models, such as costs proportionally varying with the classifier's confidence, thereby discouraging interventions on high-confidence decisions and possibly resulting in higher final utilities.

For shield computation, we assume that the distribution over future decisions (of the classifier) and costs are known, either from the knowledge of the model or *learned* from queries. Note that even if the distribution is learned and imprecise, as long as it shares the same support as the true distribution, the fairness guarantees provided by the shield remain unaffected; only the cost-optimality may be compromised. Fairness shields are computed through dynamic programming. While the straightforward approach would require exponential time and memory, we present an efficient abstraction for the dynamic programming algorithm that reduces the complexity to only *polynomial time*.

Types of fairness shields. We propose four types of shields: (i) FinHzn, (ii) Static-Fair, (iii) Static-BW, and (iv) Dynamic shields. FinHzn is specific to the bounded-horizon problem, ensuring fairness in every run while being cost-effective. The other three are suited for the periodic setting, guaranteeing fairness under mild assumptions on how rarely individuals from each group will appear in a period (formalized in Sec. 4). Static-Fair and Static-BW reuse a statically computed FinHzn shield for each period, while Dynamic shields require online re-computation of shields at the start of each period.

Experiments. We empirically demonstrate the effectiveness of fairness shielding on various ML classifiers trained on well-known datasets. While unshielded classifiers often show biases, their shielded counterparts are fair in *every* run in the bounded-horizon setting and in most runs in the periodic setting (fairness may not be guaranteed in runs that don't meet the rareness assumption). In most cases the shielded classifiers exhibit a slightly lower classification accuracy as their unshielded counterparts. This discrepancy is more pronounced under stricter fairness conditions and less pronounced, if the classifier was already trained to be fair.

2 Shielding Fairness

Data-driven classifier. Suppose we are given a population of individuals partitioned into groups a and b, where $\mathcal{G} = \{a, b\}$ are the *protected features*, like race, gender, or language. Consider a data-driven classifier that at each step samples one individual from the population, and outputs a *recommended decision* from the set $\mathbb{B} = \{1, 0\}$ along with an intervention cost from the finite set $\mathbb{C} \subset \mathbb{R}_{\geq 0}$. As convention, decisions "1" and "0" will correspond to "accept" and "reject," respectively. We assume that the sampling and classification process gives rise to a given *input distribution* $\theta \in \Delta(\mathcal{X})$, where the set $\mathcal{X} := (\mathcal{G} \times \mathbb{B} \times \mathbb{C})$ is called the *input space*. The non-protected features of individuals are hidden from the input space because they are irrelevant for shielding. We will assume that θ is given; in the extended version of this paper (Cano et al. 2024b) we discuss extensions to cases when θ is to be estimated from data.

Example 2 (Continuation of Ex. 1)**.** *In the bilingual team example, an individual is represented by a tuple $(g, z) \in \mathcal{G} \times \mathcal{Z}$, where $\mathcal{G} = \{a, b\}$ denotes the language in which the candidate is proficient, and $\mathcal{Z}$ encompasses all non-protected features relevant to evaluating a candidate's suitability for the job, such as years of experience, relevant education, and so on. For simplicity, we assume that a candidate is proficient in only one of the two languages.*

The company uses a classifier $f\colon \mathcal{G} \times \mathcal{Z} \to \mathbb{B} \times \mathbb{C}$, which outputs a preliminary decision for each candidate (accept or reject) along with a cost associated with altering that decision. The cost reflects the classifier's confidence: candidates who are clearly good or bad incur a high cost for decision changes, while borderline candidates can have their decisions reversed at a lower cost.

Shields. A *shield* is a symbolic decision-maker—independent from the classifier—and selects the *final decision* from the *output space* $\mathcal{Y} := \mathbb{B}$ after observing a given input from $\mathcal{X}$, and possibly accounting for past inputs and outputs. Formally, a shield is a function $\pi\colon (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Y}$, and its bounded-horizon variants are functions of the form $(\mathcal{X} \times \mathcal{Y})^{\leq t} \times \mathcal{X} \to \mathcal{Y}$, for a given t. We will write Π and Π^t to respectively denote the set of all shields and the set of bounded-horizon shields with horizon t. The *concatenation* of a sequence of shields $\pi_1, \pi_2, \ldots \in \Pi^t$ is a shield π, such that for every trace τ, if τ can be decomposed as $\tau\tau'$ with $|\tau| = jt$ for some j and $\tau' < t$, then $\pi(\tau, x) \coloneqq \pi_{j+1}(\tau', x)$.

Shielded sequential decision making. We consider the sequential setting where inputs are sampled from θ one at a time, and the shield π needs to produce an output without seeing the inputs from the future. Formally, at every time $i = 1, 2, \ldots$, let $x_i = (g_i, r_i, c_i)$ be the input appearing with the probability $\theta(x_i) > 0$, and let the shield's output be $y_i = \pi((x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), x_i)$. The resulting finite sequence $\tau = (x_1, y_1), \ldots, (x_t, y_t)$ is called a *trace* induced by θ and π, and the integer t is called the *length* of the trace, denoted as $|\tau|$; the notation $\mathtt{FT}^t_{\theta,\pi}$ will denote the set of every such trace. For every t, the probability distribution θ and the shield π induce a probability distribution $\mathbb{P}(\cdot; \theta, \pi)$ over the set $(\mathcal{X} \times \mathcal{Y})^t$ as follows. For every trace $\tau \in (\mathcal{X} \times \mathcal{Y})^t$, if $\tau \in \mathtt{FT}^t_{\theta,\pi}$ then $\mathbb{P}(\tau; \theta, \pi) \coloneqq \prod_{i=1}^{t} \theta(x_i)$ and otherwise $\mathbb{P}(\tau; \theta, \pi) \coloneqq 0$. Given a prefix τ, the probability of observing the trace $\tau\tau'$, for some $\tau' \in (\mathcal{X} \times \mathcal{Y})^*$, is $\mathbb{P}(\tau' \mid \tau; \theta, \pi) = \mathbb{P}(\tau\tau'; \theta, \pi) / \mathbb{P}(\tau; \theta, \pi)$. (The statistical dependence of τ' on τ is due to π's history-dependence.)

Cost. Let $\tau = (x_1, y_1), \ldots, (x_t, y_t)$ be a trace of length t, where $x_i = (g_i, r_i, c_i)$. At time i, the shield pays the cost c_i if its output y_i is different from the recommended decision r_i. The *total* (intervention) *cost incurred by the shield* on τ up to a given time $s \leq t$ is $cost(\tau; s) \coloneqq \sum_{i=1}^{s} c_i \cdot \mathbf{1}\{r_i \neq y_i\}$. The cost incurred up to time t (the length of τ) is simply written as $cost(\tau)$, instead of $cost(\tau; t)$.

For a given time horizon t, we define the expected value of cost after time t as $\mathbb{E}[cost; \theta, \pi, t] \coloneqq \sum_{\tau \in (\mathcal{X} \times \mathcal{Y})^t} cost(\tau) \cdot \mathbb{P}(\tau; \theta, \pi)$, and if additionally a prefix τ is given, the conditional expected cost after time t (from the end of τ) is $\mathbb{E}[cost \mid \tau; \theta, \pi, t] \coloneqq \sum_{\tau' \in (\mathcal{X} \times \mathcal{Y})^t} cost(\tau') \cdot \mathbb{P}(\tau' \mid \tau; \theta, \pi)$.

Example 3 (Continuation of Ex. 2)**.** *The shield π is an element external to the classifier. It takes the language group of the candidate and the classifier's recommendation as inputs and has the authority to issue a final accept/reject decision. If the shield's decision differs from the classifier's, the incurred cost is as specified by the classifier. The shield's inputs are the features of candidates, the classifier's decisions, and the costs, and the input distribution is assumed to be known in advance.*

Note that, from the shield's perspective, the distribution of non-protected features is unimportant, as these features are already processed by the data-driven classifier and summarized into a single cost value. By sampling individuals from the candidate pool and processing them through both f and π, we obtain a trace τ that records the individuals and their decisions. This trace encapsulates the results of the hiring process, including the linguistic distribution of hired candidates and the total cost incurred by the shield.

Fairness. We model (group) *fairness properties* as functions that map every finite trace to a real-valued *bias* level through intermediate statistics. A *statistic* μ maps each finite trace τ to the values of a finite set of counters, represented as a vector in $\mathbb{N}^p$, where p is the number of counters. The *welfare* for group $g \in \{a, b\}$ is a function $\mathtt{WF}^g \colon \mathbb{N}^p \to \mathbb{R}$. When μ is irrelevant or clear, we will write $\mathtt{WF}^g(\tau)$ instead of $\mathtt{WF}^g(\mu(\tau))$. A fairness property φ is an aggregation function mapping $(\mathtt{WF}^a(\tau), \mathtt{WF}^b(\tau))$ to a real-valued *bias*. Tab. 1 summarize how existing fairness properties, namely demographic parity (DP) (Dwork et al. 2012), disparate impact (DI) (Feldman et al. 2015), and equal opportunity (EqOpp) (Hardt et al. 2016) can be cast into this form.

Name	Counters	$\mathtt{WF}^g$	φ
DP	n_a, n_{a1}, n_b, n_{b1}	n_{g1}/n_g	$\lvert\mathtt{WF}^a(\tau) - \mathtt{WF}^b(\tau)\rvert$
DI	n_a, n_{a1}, n_b, n_{b1}	n_{g1}/n_g	$\left\lvert\mathtt{WF}^a(\tau) \div \mathtt{WF}^b(\tau)\right\rvert$
EqOpp	$n'_a, n'_{a1}, n'_b, n'_{b1}$	n'_{g1}/n'_g	$\lvert\mathtt{WF}^a(\tau) - \mathtt{WF}^b(\tau)\rvert$

Table 1: Empirical variants of fairness properties: For $g \in \{a, b\}$, the counters n_g and n_{g1} represent the total numbers of individuals from group g who appeared and were accepted, respectively. Counters n'_g and n'_{g1} denote the total numbers of appeared and accepted individuals whose ground truth labels are "1." If a welfare value is undefined due to a null denominator, we set $\varphi = 0$.

Estimating EqOpp requires the ground truth labels of the individuals be revealed after the shield has made its decisions on them. To accommodate ground truth, we introduce the set $\mathcal{Z} = \{0, 1\}$, such that traces are of the form $\tau = (x_1, y_1, z_1), \ldots, (x_t, y_t, z_t) \in (\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^*$, where each z_i is the ground truth label of the i-th individual. The shield is adapted to $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})^* \times \mathcal{X} \to \mathcal{Y}$, where the set $\mathcal{Z}$ is treated as another input space and the probability distribution $P(\mathcal{Z} = z_i \mid \mathcal{X} = x_i)$ is assumed to be available.

Example 4 (Continuation of Ex. 3)**.** *In the bilingual team example, the welfare of a linguistic group g is defined as the fraction of the team proficient in language g, which is the empirical variant of DP. A more nuanced interpretation considers the welfare of group g as the fraction of accepted candidates among those proficient in language g. This measure accounts for the possibility that the linguistic distribution of the population may not be evenly split. If one language is more prevalent in the target population, the hired team should proportionally include more members proficient in that language. The fairness property derived from this measure corresponds to an empirical version variant of EqOpp.*

Bounded-horizon fairness shields. From now on, we use the convention that θ is the input distribution, φ is the fairness property, and κ is the *bias threshold*. Let T be a given time horizon. The set $\Pi^{\theta,T}_{\mathtt{fair}}$ of *fairness shields over time T* is the set of every shield that fulfills $\varphi(\cdot) \leq \kappa$ after time T, i.e., $\Pi^{\theta,T}_{\mathtt{fair}} \coloneqq \{\pi \in \Pi^T \mid \forall \tau \in \mathtt{FT}^T_{\theta,\pi} \,.\, \varphi(\tau) \leq \kappa\}$. We now define optimal bounded-horizon fairness shields as below.

Definition 1 (FinHzn shields)**.** Let $T > 0$ be the time horizon. A FinHzn shield is the one that solves:

$$\pi^* := \arg \min_{\pi \in \Pi^{\theta,T}_{\mathtt{fair}}} \mathbb{E}[cost; \theta, \pi, T]. \tag{1}$$

Periodic fairness shields. FinHzn shields stipulate that fairness be satisfied at the end of the given horizon. However, in many situations, it may be desirable to ensure fairness not only at the end of the horizon but also at intermediate points occurring at regular intervals. For instance, a human resources department required to maintain a fair distribution of employees over the course of a quarter might also need to ensure a similar property for every yearly revision. This type of fairness is referred to as *periodic fairness* in the literature (Alamdari et al. 2024). For this class of fairness properties, we define the set of T-periodic fairness shields as $\Pi_{\mathtt{fair\text{-}per}} := \{\pi \in \Pi \mid \forall m \in \mathbb{N} \,.\, \forall \tau \in \mathtt{FT}^{mT}_{\theta,\pi} \,.\, \varphi(\tau) \le \kappa\}$.

Definition 2 (Optimal T-periodic fairness shield)**.** Let $T > 0$ be the time period. An *optimal T-periodic fairness shield* is given by:

$$\pi^* := \arg \min_{\pi \in \Pi_{\mathtt{fair\text{-}per}}} \sup_{\substack{m \in \mathbb{N} \\ \tau \in \mathtt{FT}^{mT}_{\theta,\pi}}} \mathbb{E}[cost \mid \tau; \theta, \pi, T]. \tag{2}$$

Eq. (2) requires fairness at each mT-th time (measured from the beginning), and minimizes the maximum expected cost over each period. The existence of this minimum remains an open question. In Sec. 4, we propose three "best-effort" approaches to compute periodically fair shields (under mild assumptions) that are as cost-optimal as possible.

3 Algorithm for FinHzn Shield Synthesis

We present our algorithm for synthesizing FinHzn shields as defined in Def. 1. A FinHzn shield π^* computes an output $y = \pi^*(\tau, x)$ for every trace τ and every input x. Our synthesis algorithm builds π^* recursively for traces of increasing length, using an auxiliary *value function* $v(\tau)$ that represents the minimal expected cost conditioned on traces with prefix τ. To define $v(\tau)$, we generalize fairness shields with the condition that a certain trace has already occurred. Given a time horizon t and a trace τ (length can differ from t), the set of *fairness shields over time t after τ* is defined as: $\Pi^{\theta,t|\tau}_{\mathtt{fair}} := \{\pi \in \Pi^t \mid \forall \tau' \in (\mathcal{X} \times \mathcal{Y})^t \,.\, \tau\tau' \in \mathtt{FT}^{|\tau|+t}_{\theta,\pi} \implies \varphi(\tau\tau') \le \kappa\}$. Then $v(\tau)$ is given by:

$$v(\tau) := \min_{\pi \in \Pi^{\theta,(T-|\tau|)|\tau}_{\mathtt{fair}}} \mathbb{E}[cost \mid \tau; \theta, \pi, T - |\tau|].$$

For every trace τ and every input $x \in \mathcal{X}$, the optimal value of the shield is $\pi^*(\tau, x) = \arg\min_{y \in \mathcal{Y}} v(\tau, (x, y))$.

In Sec. 3.1, we present a recursive dynamic programming for computing $v(\tau)$, whose complexity grows exponentially with the length of τ. In Sec. 3.2, we present an efficient solution using only the p counters defining the fairness property, thus solving the synthesis problem in $\mathcal{O}(T^p \cdot |\mathcal{X}|)$-time. From now on, we present the main ideas and refer the reader to (Cano et al. 2024b) for detailed proofs of all results.

3.1 Recursive Computation of $v(\tau)$

Base case. Let T be the time horizon and τ be a trace of length T. Since the horizon is reached if $\varphi(\tau) \le \kappa$ then the expected cost is zero because fairness is already satisfied and no more cost needs to be incurred, whereas if $\varphi(\tau) > \kappa$, the expected cost is infinite, because no matter what cost is paid fairness can no longer be achieved. Formally,

$$v(\tau) = \begin{cases} 0 & \varphi(\tau) \le \kappa, \\ \infty & \text{otherwise.} \end{cases} \tag{3}$$

Recursive case. Let τ be a trace of length smaller than T. The probability of the next input being $x = (g, r, c)$ is $\theta(x)$, and the shield decides to output y that either agrees with the recommendation r (the case $y = r$) or differs from it (the case $y \neq r$)—whichever minimizes the expected cost. When $y = r$, then the trace becomes $(\tau, (x, y = r))$. Therefore, no cost is incurred and the total cost remains the same as $v(\tau, (x, y = r))$. When $y \neq r$, the trace becomes $(\tau, (x, y \neq r))$. Thus, the incurred cost is c and the new total cost becomes $c + v(\tau, (x, y = r))$. Therefore

$$v(\tau) = \sum_{x=(g,r,c) \in \mathcal{X}} \theta(x) \cdot \min \left\{ \begin{matrix} v(\tau, (x, y = r)), \\ v(\tau, (x, y \neq r)) + c \end{matrix} \right\}. \tag{4}$$

Eqs. (3) and (4) can be used to recursively compute $v(\tau)$ for every τ of length up to T, and the time and space complexity of this procedure is $\mathcal{O}(|\mathcal{X} \times \mathcal{Y}|^T)$. The correctness of Eq. (4) is formally proven in (Cano et al. 2024b).

Example 5 (Continuation of Ex. 4)**.** *Consider the task of hiring a linguistically balanced team with a horizon of $T = 50$ candidates and a target demographic parity property φ with a threshold $\kappa = 0.2$. By the end of the process, a trace τ of $|\tau| = 50$ candidates must satisfy $\varphi(\tau) < \kappa$.*

Consider the following situation. Suppose τ' be the trace obtained after observing the first 48 candidates, i.e., $|\tau'| = 48$, and just two more candidates are going to be observed before the horizon ends. In τ', 24 candidates have been observed for each language proficiency group among A and B, and among them 12 from group A and 17 from group B have been accepted, resulting in $\varphi(\tau') = |12/24 - 17/24| = 0.208 > \kappa$. This temporary violation of DP is allowed since the process is ongoing.

Suppose a new candidate $x = (g, r, c)$ appears, with $g = B$. The classifier tentatively accepts x and informs the shield that reversing this decision would incur a cost c. If the shield accepts x, the shield will be forced to reject the next candidate proficient in B or accept the next candidate proficient in A, regardless of the cost. Conversely, if the shield rejects x, it incurs an immediate cost of c but balances the languages to a point where intervention will not be required for the next decision.

The shield must therefore weigh its options: either incur a known cost c now by rejecting x or risk an unknown future cost c' by accepting x. If the candidate is exceptionally qualified, the shield might choose to accept x, accepting the potential risk of rejecting another well-qualified candidate proficient in B in the next round.

3.2 Efficient Recursive Computation of $v(\tau)$

We now present an efficient recursive procedure for computing FinHzn shields that runs in polynomial time and space. The key observation is that φ is a fairness property that depends on τ through a statistic that uses p counters. Consequently, $v(\tau)$ in Eq. (3) and Eq. (4) depend only on counter values, not on exact traces. This allows us to define our dynamic programming algorithm over the set of counter values taken by the statistic μ. Let $R_{\mu,T} \subseteq \mathbb{N}^p$ be the set of values the statistic μ can take from traces of length at most T. We have the following complexity result.

Theorem 1. *Solving the bounded-horizon shield-synthesis problem requires $\mathcal{O}(|R_{\mu,T}| \cdot |\mathcal{X}|)$-time and $\mathcal{O}(|R_{\mu,T}| \cdot |\mathcal{X}|)$-space.*

Most fairness properties, e.g., DP and EqOpp, have a range of $R_{\mu,T} = [0,T]^p$, where p is the number of counters ($p = 4$ for DP, and $p = 5$ for EqOpp), making the complexity polynomial in the length of the time horizon.

4 Algorithms for Periodic Shield Synthesis

We present algorithms for computing periodic fairness shields for a broad subclass of group fairness properties, termed *difference of ratios* (DoR) properties. A statistic μ is *single-counter* if it maps every trace τ to a single counter value, i.e., $\mu(\tau) \in \mathbb{N}$, and *additive* if $\mu(\tau\tau') = \mu(\tau) + \mu(\tau')$ for any traces τ and τ'. A group fairness property φ is DoR if (a) for each group g, $\mathtt{WF}^g(\tau) = \mathtt{num}^g(\tau)/\mathtt{den}^g(\tau)$, where $\mathtt{num}^g(\tau)$ and $\mathtt{den}^g(\tau)$ are additive single-counter statistics, and (b) $\varphi(\tau) = |\mathtt{WF}^a(\tau) - \mathtt{WF}^b(\tau)|$. Many fairness properties, including DP and EqOpp, are DoR, though DI is not because it violates the condition (b). For DoR fairness properties, we propose two approaches for constructing periodic fairness shields: *static* and *dynamic*, and we explore their respective strengths and weaknesses.

4.1 Periodic Shielding: The Static Approach

In the static approach, a periodic shield is obtained by *concatenating infinitely many identical copies of a statically computed bounded-horizon shield* π, synthesized with the time period T as the horizon. We present two ways of computing π so that its infinite concatenation is T-periodic fair.

Approach I: Static-Fair shields.

Definition 3 (Static-Fair shields)**.** A shield is called Static-Fair if it is the concatenation of infinite copies of a FinHzn shield (from Def. 1).

Unfortunately, Static-Fair shields do not always satisfy periodic fairness. Consider a trace $\tau = \tau_1 \dots \tau_m$ for an arbitrary $m > 0$, generated by a Static-Fair shield, such that each segment τ_i is of length T. It follows from the property of FinHzn shields that $\varphi(\tau_i) \leq \kappa$ for each individual i. However, T-periodic fairness may be violated because $\varphi(\tau)$ need not be bounded by κ. A counter-example for DP is below; for additional examples see (Cano et al. 2024b).

Example 6. *Consider DP with $0 < \kappa < 1 - 2/T$. Suppose τ_1 and τ_2 are traces of length T such that for τ_1, $n_a = 1$,$n_b = T-1$, and $n_{a1} = n_{b1} = 0$, and for τ_2, $n_a = n_b = T$, $n_{a1} = T$, and $n_{b1} = 1$. Then $\varphi(\tau_1) = \varphi(\tau_2) = 0$ (fair), but $\varphi(\tau_1\tau_2) = |(T-1)/T - 1/T| = 1 - 2/T > \kappa$ (biased).*

An important feature of these counter-examples is the excessive skewness of appearance rates across the two groups. We show that Static-Fair shields are T-periodic fair if the appearance rates of the two groups are equal at every period.

Definition 4 (Balanced traces)**.** Let $\mu^a, \mu^b \colon (\mathcal{X} \times \mathcal{Y})^* \to \mathbb{N}$ be a pair of group-dependent (single-counter) statistics, $T > 0$ be a given time horizon, and $N \leq T/2$ be a given integer. A trace τ of length T is *N-balanced with respect to* μ^a and μ^b if both $\mu^a(\tau) \geq N$ and $\mu^b(\tau) \geq N$; the set of all such traces is written $\mathtt{BT}^T(\mu^a, \mu^b, N)$.

Theorem 2 (Conditional correctness of Static-Fair shields)**.** *Let φ be a DoR fairness property. Consider a Static-Fair shield π, and let $\tau = \tau_1 \dots \tau_m \in \mathtt{FT}^{mT}_{\theta,\pi}$ be a trace such that $|\tau_i| = T$ for all $i \leq m$. If $\mathtt{den}^a(\tau_i) = \mathtt{den}^b(\tau_i)$ for every $i \leq m$, then the fairness property $\varphi(\tau) \leq \kappa$ is guaranteed.*

While the condition in Thm. 2 appears conservative, we show in (Cano et al. 2024b) that it is in fact tight for the worst-case satisfaction of DP, in the sense that for every κ, there exist m and $\lfloor (T-1)/2 \rfloor$-balanced traces $\tau_1, \dots, \tau_m$ such that $\varphi_{\mathtt{DP}}(\tau_i) \leq \kappa$ for each i, but $\varphi_{\mathtt{DP}}(\tau_1 \dots \tau_m) > \kappa$. However, these are worst-case scenarios and are "uninteresting." In our experiments, Static-Fair shields fulfill periodic fairness in a majority of cases even if the condition of Thm. 2 is violated.

Approach II: Static-BW shields. When the condition of Thm. 2 is violated, Static-Fair shields cannot guarantee fairness as the bound on the bias is not closed under concatenation of traces (see Ex. 6). A stronger property that is closed under concatenation is when a bound is imposed on each group's welfare. Let l, u be constants with $0 \leq l < u \leq 1$. A trace τ has *bounded welfare* (BW) if for each group $g \in \mathcal{G}$, $\mathtt{WF}^g(\tau) = \mathtt{num}^g(\tau)/\mathtt{den}^g(\tau)$ belongs to $[l, u]$. The pair (l, u) will be called *welfare bounds*. We show that BW is closed under trace concatenations, which depends on the additive property of $\mathtt{num}^g$ and $\mathtt{den}^g$.

Lemma 3. *Let (l, u) be given welfare bounds, and $\mathtt{WF}^g(\cdot) \equiv \mathtt{num}^g(\cdot)/\mathtt{den}^g(\cdot)$ for additive $\mathtt{num}^g, \mathtt{den}^g$. For a trace $\tau = \tau_1 \dots \tau_m$, if for each i, $\mathtt{WF}^g(\tau_i) \in [l, u]$, then $\mathtt{WF}^g(\tau) \in [l, u]$.*

For DoR properties, BW implies fairness when $u - l \leq \kappa$. Combining this with Lem. 3, we infer that if π is a bounded-horizon shield that fulfills BW on every trace τ of length T for welfare bounds (l, u) with $u - l \leq \kappa$, then the concatenation of infinite copies of π would be a T-periodic fairness shield. The natural course of action for computing shields that fulfill BW is to mimic Def. 1, replacing the condition on φ with a condition on welfare. However, if we define the set of BW-fulfilling shields as $\Pi^{\theta,T}_{\mathtt{BW}} := \{\pi \in \Pi \mid \forall \tau \in \mathtt{FT}^T_{\theta,\pi} \,.\, \forall g \in \{a, b\} \,.\, l \leq \mathtt{WF}^g(\tau) \leq u\}$, the set $\Pi^{\theta,T}_{\mathtt{BW}}$ can be empty for some T, l, u. Following is an example.

Example 7. *Suppose $\mathtt{WF}^g(\tau) = n_{g1}/n_g$, where n_{g1} and n_g are the total numbers of accepted and appeared individuals from group g (as in DP). Suppose $T = 2, l = 0.2, u = 0.4$. It is easy to see that no matter what the shield does, for every τ of length 2, $\mathtt{WF}^g(\tau) \in \{0, 0.5, 1\}$. Therefore, $\Pi^2_{[0.2,0.4]} = \emptyset$.*

The emptiness of $\Pi_{\mathtt{BW}}^{\theta,T}$ is due to a large disparity between the appearance rates of individuals from the two groups, which occurs for shorter time horizons and for datasets where one group has significantly lesser representation than the other group. To circumvent this technical inconvenience, we make the following assumption on observed traces.

Assumption 1. *Let l, u be welfare bounds, and $\tau = \tau_1 \ldots \tau_m \in \mathtt{FT}_{\theta,\pi}^{mT}$ be a trace with $|\tau_i| = T$ for each i. Every τ_i is N-balanced w.r.t. $\mathtt{den}^a$ and $\mathtt{den}^b$ for $N = \lceil 1/(u-l) \rceil$.*

Assump. 1 may be reasonable depending on l, u, T, and the input distribution θ. Intuitively, for a larger T and a smaller skew of appearance probabilities for individuals between the two groups, the probability of fulfilling Assump. 1 is larger (for a given finite m). In in (Cano et al. 2024b) we quantify this probability as the probability of a sample from a binomial distribution lying between N and $T - N$.

Definition 5 (Static-BW shields)**.** Let l, u be given welfare bounds, and T be a given time period. A Static-BW shield is the concatenation of infinite copies of the shield π^* solving

$$\pi^* = \arg \min_{\pi \in \Pi_{\mathtt{BW}}^{\theta,T,N}} \mathbb{E}[cost; \theta, \pi, T], \tag{5}$$

where $N = \lceil 1/(u-l) \rceil$, and

$$\Pi_{\mathtt{BW}}^{\theta,T,N} := \{\pi \in \Pi \mid \forall \tau \in \mathtt{FT}_{\theta,\pi}^{T} \cap \mathtt{BT}_N^T \\ \forall g \in \{a,b\} \;.\; l \leq \mathtt{WF}^g(\tau) \leq u\}.$$

In in (Cano et al. 2024b) we provide a constructive proof showing that $\Pi_{\mathtt{BW}}^{\theta,T,N}$ is indeed non-empty when Assump. 1 is fulfilled. This result guarantees that the optimization problem in (5) is feasible, and thus Static-BW shields are well-defined. Intuitively, we obtain a "best-effort" solution for π^*: when a trace satisfies Assump. 1, π^* guarantees that τ satisfies BW with minimum expected cost. Otherwise, π^* has no BW requirement, and thus for traces that violate Assump. 1, the shield will incur zero cost by never intervening.

Synthesis of Static-BW shields follows the same approach as in Sec. 3 with Eq. (3) replaced by:

$$v(\tau) = \begin{cases} 0 & \tau \notin \mathtt{BT}_N^T \vee \bigwedge_{g \in \{a,b\}} \mathtt{WF}^a(\tau) \in [l,u], \\ \infty & \text{otherwise.} \end{cases}$$

We summarize the fairness guarantee below.

Theorem 4 (Conditional correctness of Static-BW shields)**.** *Let φ be a DoR fairness property. Let l, u be welfare bounds such that $u - l \leq \kappa$. For a given Static-BW shield π, let $\tau = \tau_1 \ldots \tau_m \in \mathtt{FT}_{\theta,\pi}^{mT}$ be a trace with $|\tau_i| = T$ for each $i \leq m$. If Assump. 1 holds, then the fairness property $\varphi(\tau) \leq \kappa$ is guaranteed.*

4.2 Periodic Shielding: The Dynamic Approach

While the static approaches repeatedly use one statically computed bounded-horizon shield, the dynamic approach recomputes a new bounded-horizon shield at the beginning of each period, and thereby adjusts its future decisions based on the past biases. We formalize this below.

Definition 6 (Dynamic shields)**.** Suppose we are given a parameterized set of *available* shields $\Pi'(\tau) \subseteq \Pi$ where the parameter τ ranges over all finite traces. A Dynamic shield π is the concatenation of a sequence of shields $\pi_1, \pi_2, \ldots$ such that for every trace $\tau \in \mathtt{FT}_{\theta,\pi}^{mT}$ with $m \geq 0$, for every $\tau' \in (\mathcal{X} \times \mathcal{Y})^{<T}$, and for every input $x \in \mathcal{X}$, $\pi(\tau\tau', x) = \pi_{m+1}(\tau', x)$ where

$$\pi_{m+1} = \arg \min_{\pi' \in \Pi'(\tau)} \mathbb{E}[cost \mid \tau; \theta, \pi', T]. \tag{6}$$

The set $\Pi'(\tau)$ restricts the available set of shields that can be used for the next period for the given history τ. A naïve attempt for $\Pi'(\tau)$ would be to choose $\Pi'(\tau) = \Pi_{\mathtt{fair}}^{\theta,T|\tau}$ for every τ, so that fairness is guaranteed at the end of the current period. However, there exist histories for which $\Pi_{\mathtt{fair}}^{\theta,T|\tau}$ would be empty, implying that Eq. (6) would not have a feasible solution for some τ, and the Dynamic shield would exhibit undefined behaviors. To circumvent this technical inconvenience, we make the following mild assumption on the set of allowed histories, requiring $\Pi'(\tau)$ to fulfill fairness only if τ fulfills this assumption.

Assumption 2. *For a given trace $\tau \in \mathtt{FT}_{\theta,\pi}^{jT}$ with $j > 0$, every valid suffix τ' of length t (i.e., $\tau' \in \{\tau'' \in (\mathcal{X} \times \mathcal{Y})^T \mid \tau\tau'' \in \mathtt{FT}_{\theta,\pi}^{(j+1)T}\}$) fulfills:*

$$\frac{1}{\mathtt{den}^a(\tau\tau')} + \frac{1}{\mathtt{den}^b(\tau\tau')} \leq \kappa + \varphi(\tau).$$

The set of shields $\Pi'(\cdot)$ available to the Dynamic shield in Def. 6 is then defined as:

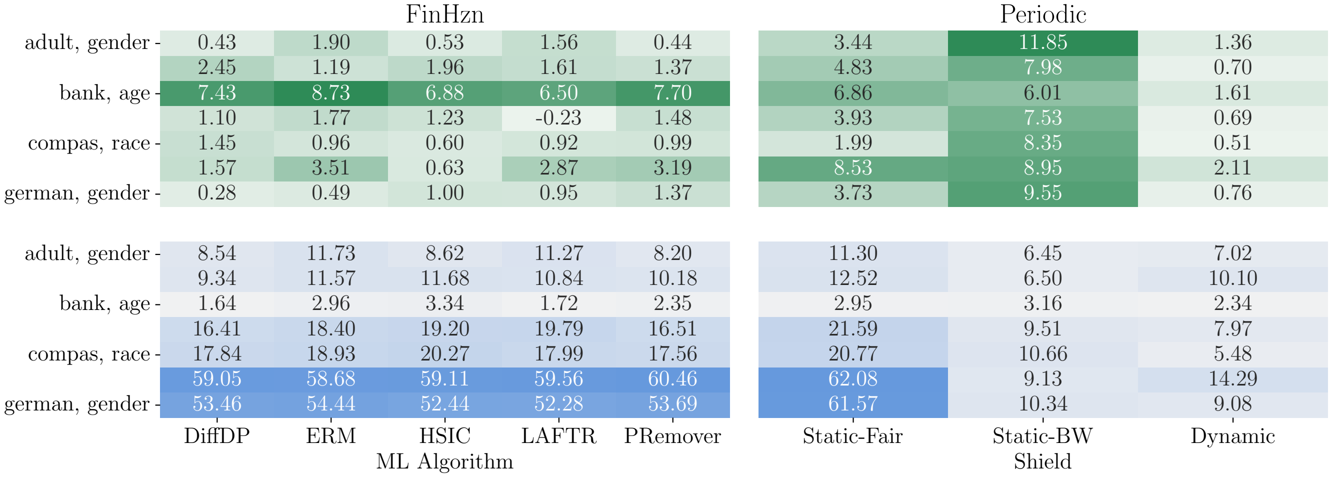
$$\Pi'(\tau) = \Pi_{\mathtt{fair\text{-}dyn}}^{\theta,T}(\tau) := \begin{cases} \Pi_{\mathtt{fair}}^{\theta,T|\tau} & \tau \text{ fulfills Assump. 2,} \\ \Pi & \text{otherwise.} \end{cases}$$

We prove that $\Pi_{\mathtt{fair}}^{\theta,T|\tau}$ is non-empty whenever τ fulfills Assump. 2 – see (Cano et al. 2024b)–, implying that $\Pi_{\mathtt{fair\text{-}dyn}}^{\theta,T}(\tau)$ is non-empty for every τ. Technically, this guarantees that the optimization problem in (6) is feasible and π_{m+1} always exists and Dynamic shields are well-defined. Intuitively, we obtain a "best-effort" solution: If Assump. 2 is fulfilled then π_{m+1} is in $\Pi_{\mathtt{fair}}^{\theta,T|\tau}$ and achieves fairness for the minimum expected cost. Otherwise, π_{m+1} can be any shield in Π that only optimizes for the expected cost; in particular, π_{m+1} will be the trivial shield that never intervenes (has zero cost).

Synthesis of Dynamic shields involves computing the sequence of shields $\pi_1, \pi_2, \ldots$, which are to be concatenated. We outline the algorithm below.

1. Generate a FinHzn shield (Def. 1) π for the property φ and the horizon T. Set $\pi_1 := \pi$.
2. For $i \geq 1$, let π be the concatenation of the shields $\pi_1, \ldots, \pi_i$, and let $\tau \in \mathtt{FT}_{\theta,\pi}^{iT}$ be the generated trace. Compute π_{i+1} that uses the same approach as in Sec. 3 with Eq. (3) being replaced by:

$$v(\tau') = \begin{cases} 0 & \varphi(\tau\tau') \leq \kappa \\ \infty & \text{otherwise.} \end{cases}$$



	FinHzn					Periodic		
	DiffDP	ERM	HSIC	LAFTR	PRemover	Static-Fair	Static-BW	Dynamic
	ML Algorithm					Shield		
adult, gender	0.43	1.90	0.53	1.56	0.44	3.44	11.85	1.36
	2.45	1.19	1.96	1.61	1.37	4.83	7.98	0.70
bank, age	7.43	8.73	6.88	6.50	7.70	6.86	6.01	1.61
	1.10	1.77	1.23	-0.23	1.48	3.93	7.53	0.69
compas, race	1.45	0.96	0.60	0.92	0.99	1.99	8.35	0.51
	1.57	3.51	0.63	2.87	3.19	8.53	8.95	2.11
german, gender	0.28	0.49	1.00	0.95	1.37	3.73	9.55	0.76
adult, gender	8.54	11.73	8.62	11.27	8.20	11.30	6.45	7.02
	9.34	11.57	11.68	10.84	10.18	12.52	6.50	10.10
bank, age	1.64	2.96	3.34	1.72	2.35	2.95	3.16	2.34
	16.41	18.40	19.20	19.79	16.51	21.59	9.51	7.97
compas, race	17.84	18.93	20.27	17.99	17.56	20.77	10.66	5.48
	59.05	58.68	59.11	59.56	60.46	62.08	9.13	14.29
german, gender	53.46	54.44	52.44	52.28	53.69	61.57	10.34	9.08

Table 2: Utility loss (in %) incurred by FinHzn shields for different ML models (left) and by periodic shields on the ERM model (right) for the fairness properties DP (top, green) and EqOpp (bottom, blue). Lighter colors indicate smaller utility loss.

We summarize the fairness guarantee below.

Theorem 5 (Conditional correctness of Dynamic shields)**.** *Let φ be a DoR fairness property. Let π be a Dynamic shield that uses $\Pi^{\theta,T}_{\texttt{fair-dyn}}(\cdot)$ as the set of available shields. Let $\tau = \tau_1 \dots \tau_m \in \mathtt{FT}^{mT}_{\theta,\pi}$ be a trace with $|\tau_i| = T$ for each $i \leq m$. Suppose for every $i \leq m$, $\tau_1 \dots \tau_i$ fulfills Assump. 2. Then the fairness property $\varphi(\tau) \leq \kappa$ is guaranteed.*

5 Experiments

Experimental setup. We performed our experiments on the datasets Adult (Becker et al. 1996), COMPAS (Kirchner et al. 2016), German Credit (Hofmann 1994), and Bank Marketing (Moro et al. 2012). The protected attributes include race, gender, and age. We synthesized shields to ensure DP and EqOpp with thresholds $\kappa \in \{0.05, 0.1, 0.15, 0.2\}$.

5.1 FinHzn Shields

The ML models were trained with DiffDP (Chuang et al. 2021), HSIC (Pérez-Suay et al. 2017), LAFTR (Madras et al. 2018), and PRemover (Kamishima et al. 2012). As a baseline, we also trained a classifier using empirical risk minimization (ERM). For all models and datasets, FinHzn shields were synthesized with $T = 100$ for DP and $T = 75$ for EqOpp. Shield synthesis took about 1 second and 30 MB for DP, and 1.5 seconds and 1.3 GB for EqOpp. A more detailed report on resource usage can be found in (Cano et al. 2024b). We compared model performances—with and without shielding—across 30 simulated runs.

		Recomp.	Assump. satisfied	Fairness satisfied
DP	Static-Fair	no	0.0%	95.71%
	Static-BW	no	43.8%	83.1%
	Dynamic	yes	100%	100%
EqOpp	Static-Fair	no	0.0%	100%
	Static-BW	no	4.1%	56.4%
	Dynamic	yes	49.8%	100%

Table 3: Comparison of different types of fairness shields.

Fairness. Unshielded ML models violated bounded-horizon fairness in 44% of the cases for DP and in 65% for EqOpp. Shielded models were fair at the horizons, empirically validating the effectiveness of FinHzn shields. Detailed results are reported in (Cano et al. 2024b).

Utility loss. Classification utility is measured using classification accuracy. Note that interventions by the fairness shield can reduce this utility. We define *utility loss* as the difference in utility between unshielded and shielded runs. Tab. 2 shows the average utility loss across all simulations for a threshold of 0.1. We can observe that the median utility loss is smaller when the classifier is trained to be fair, as fewer interventions are needed. In general, utility loss increases as the bias threshold κ decreases, with more pronounced differences between classifiers for smaller κ.

5.2 Periodic Shielding

ML models were trained using the ERM algorithm across all datasets. We synthesized Static-Fair, Static-BW, and Dynamic shields with $T = 50$ for DP and EqOpp, and simulated them for 10 periods. Shield synthesis took about 1 second and 30 MB for DP, and 1.5 seconds and 1.3 GB for EqOpp. We compared the models' performances—with and without shielding—across 20 simulated runs.

Fairness. In Tab. 3, we present the rates of assumption and fairness satisfaction across all datasets and runs. The assumption for Static-Fair (see Thm.2) never met, and the assumption for Static-BW (see Thm. 4) is also often violated. Nevertheless, both Static-Fair and Static-BW still perform well as heuristics, with many runs satisfying the fairness constraint. Dynamic shields outperform both.

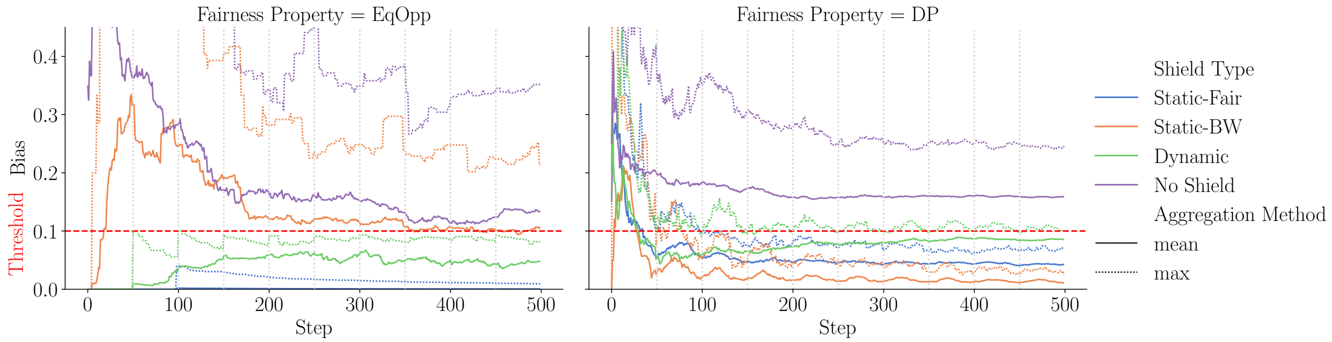


Figure 2: Variations of bias over time for the ERM classifier on the Adult dataset with and without periodic shielding.

Utility loss. In Tab. 2, we report the average utility loss across all simulations for each shield for the bias threshold of 0.1. In general, if the assumptions are satisfied, Dynamic shields incur the least loss and Static-BW shields incur the most, which is due to their stricter BW objectives. However, an assumption violation forces both Dynamic and Static-BW shields to go inactive incurring no additional utility loss. Therefore, the low utility loss of Static-BW shields in EqOpp can be explained by the frequent assumption violations.

6 Related Work

Existing work on fairness in AI focuses on how to *specify*, *design*, and *verify* fair decision-making systems. Specification involves quantifying fairness across groups (Feldman et al. 2015; Hardt et al. 2016) and individuals (Dwork et al. 2012). Design approaches ensure that decision-makers meet fairness objectives (Hardt et al. 2016; Gordaliza et al. 2019; Zafar et al. 2019; Agarwal et al. 2018; Wen et al. 2021). Verification includes static (Albarghouthi et al. 2017; Bastani et al. 2019; Sun et al. 2021; Ghosh et al. 2021; Meyer et al. 2021; Li et al. 2023) and runtime (Albarghouthi et al. 2019; Henzinger et al. 2023a,b) methods to assess fairness. Our shields are *verified* by *design*, and act as trusted third-party intervention mechanisms, ensuring fairness without requiring knowledge of the AI-based classifier.

Traditionally, fairness is defined using the classifier's output distribution. However, this can lead to biases over short horizons (Alamdari et al. 2024). To address this, we adopt the recently proposed bounded-horizon fairness properties (Alamdari et al. 2024), ensuring decisions remain empirically fair over a bounded horizon. To the best of our knowledge, our work is the first to provide algorithmic support for guaranteeing bounded-horizon fairness properties.

Sequential decision-making problems have been extensively studied under the umbrella of optimal stopping problems (Shiryaev 2007; Bandini et al. 2018; Ankirchner et al. 2019; Bayraktar et al. 2024; Palmer et al. 2017; Källblad 2022). These works focus on designing policies that approximate those that have perfect foresight about the future. However, statistical properties like fairness are not addressed by existing algorithms in this literature.

(Cano et al. 2024a) proposed sequential decision making algorithms for the general class of finite-horizon statistical properties. They demonstrated that combining statistically indistinguishable traces in dynamic programming reduces computational costs without altering the output. We apply this idea in our FinHzn shields, where traces with identical counter values remain indistinguishable.

7 Discussion and Future Work

Static vs. dynamic shielding in the periodic setting. Static shields are computationally cheaper than Dynamic shields and have no runtime overhead, making them ideal for fast decision-making applications like online ad-delivery (Ali et al. 2019). However, they can't adjust decisions based on the actual history, leading to overly restrictive and frequent interventions—particularly in the long run. In contrast, Dynamic shields adapt to historical data, resulting in fewer interventions over time, making them suitable for applications like banking where decision-making can afford longer computation times (Liu et al. 2018).

On the feedback effect in sequential decision-making. Decisions that seem fair individually can introduce biases over time as the input distribution θ changes based on past actions (D'Amour et al. 2020; Sun 2023). Although we assumed a constant θ in this paper, our recursive synthesis algorithm from Sec. 3 could be adapted to handle trace-dependent θ by simply modifying Eq. 4. A detailed study of this adaptation is left for future work.

Fairness shields with humans in the loop. In applications where human experts make decisions with AI assistance, shields may not have final decision authority but can act as a runtime "fairness filter" to modify and de-bias the AI's outputs before presenting them to the human expert.

Other future directions. Valuable future work includes extending static and dynamic shields to broader classes of fairness properties beyond DoR. Additionally, a comparative study of fine-grained intervention costs would help identify cost models that minimize utility loss. Lastly, we plan to extend fairness shields to offer optimality guarantees when input distributions involve uncertainties, e.g., by replacing exact probabilities with intervals.

Acknowledgements

This work is partly supported by the European Research Council under Grant No.: ERC-2020-AdG 101020093. It is also partially supported by the State Government of Styria, Austria – Department Zukunftsfonds Steiermark.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, 60–69. PMLR.

Alamdari, P. A.; Klassen, T. Q.; Creager, E.; and Mcilraith, S. A. 2024. Remembering to Be Fair: Non-Markovian Fairness in Sequential Decision Making. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235, 906–920. PMLR.

Albarghouthi, A.; D'Antoni, L.; Drews, S.; and Nori, A. V. 2017. Fairsquare: probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA): 1–30.

Albarghouthi, A.; and Vinitsky, S. 2019. Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 211–219.

Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–30.

Ankirchner, S.; Klein, M.; and Kruse, T. 2019. A verification theorem for optimal stopping problems with expectation constraints. *Applied Mathematics & Optimization*, 79: 145–177.

Bandini, E.; Cosso, A.; Fuhrman, M.; and Pham, H. 2018. Backward SDEs for optimal control of partially observed path-dependent stochastic systems: a control randomization approach. *The Annals of Applied Probability*, 28(3): 1634–1678.

Bastani, O.; Zhang, X.; and Solar-Lezama, A. 2019. Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages*, 3(OOPSLA): 1–27.

Bayraktar, E.; and Yao, S. 2024. Optimal stopping with expectation constraints. *The Annals of Applied Probability*, 34(1B): 917–959.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Calders, T.; and Žliobaitė, I. 2013. Why unbiased computational processes can lead to discriminative decision procedures. *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases*, 43–57.

Cano, F.; Henzinger, T. A.; Könighofer, B.; Kueffner, K.; and Mallik, K. 2024a. Abstraction-Based Decision Making for Statistical Properties. In *International Conference on Formal Structures for Computation and Deduction (FSCD)*, volume 299 of *LIPIcs*, 2:1–2:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Cano, F.; Henzinger, T. A.; Könighofer, B.; Kueffner, K.; and Mallik, K. 2024b. Fairness Shields: Safeguarding against Biased Decision Makers. arXiv:2412.11994.

Caton, S.; and Haas, C. 2020. Fairness in machine learning: A survey. *ACM Computing Surveys*.

Chuang, C.; and Mroueh, Y. 2021. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

D'Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; and Halpern, Y. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 525–534.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1): eaao5580.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, 214–226. New York, NY, USA: ACM.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 259–268.

Ghosh, B.; Basu, D.; and Meel, K. S. 2021. Justicia: A stochastic SAT approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 7554–7563.

Gordaliza, P.; Del Barrio, E.; Fabrice, G.; and Loubes, J.-M. 2019. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning (ICML)*, 2357–2365. PMLR.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 3315–3323.

Henzinger, T. A.; Karimi, M.; Kueffner, K.; and Mallik, K. 2023a. Monitoring Algorithmic Fairness. In *Proceedings of the International Computer Aided Verification (CAV)*, 358–382. Springer-Verlag.

Henzinger, T. A.; Karimi, M.; Kueffner, K.; and Mallik, K. 2023b. Runtime Monitoring of Dynamic Fairness Properties. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 604–614. ACM.

Hofmann, H. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Hu, Y.; and Zhang, L. 2022. Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 9549–9557.

Källblad, S. 2022. A dynamic programming approach to distribution-constrained optimal stopping. *The Annals of Applied Probability*, 32(3): 1902–1928.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1): 1–33.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, (ECML PKDD)*, 35–50. Springer.

Kirchner, L.; Mattu, S.; Larson, J.; and Angwin, J. 2016. Machine Bias. *ProPublica*.

Li, Y.; Wang, J.; and Wang, C. 2023. Certifying the Fairness of KNN in the Presence of Dataset Bias. In *International Conference on Computer Aided Verification (CAV)*. Springer.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning (ICML)*, 3150–3158. PMLR.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning (ICML)*, 3384–3393. PMLR.

Meyer, A.; Albarghouthi, A.; and D'Antoni, L. 2021. Certifying Robustness to Programmable Data Bias in Decision Trees. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 26276–26288.

Moro, S.; Cortez, P.; and Rita, P. 2012. Bank Marketing. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K306.

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

Oneto, L.; and Chiappa, S. 2020. Fairness in machine learning. In *Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019)*, 155–196. Springer.

Palmer, A. Z.; and Vladimirsky, A. 2017. Optimal stopping with a probabilistic constraint. *Journal of Optimization Theory and Applications*, 175: 795–817.

Pérez-Suay, A.; Laparra, V.; Mateo-García, G.; Muñoz-Marí, J.; Gómez-Chova, L.; and Camps-Valls, G. 2017. Fair Kernel Learning. In Ceci, M.; Hollmén, J.; Todorovski, L.; Vens, C.; and Džeroski, S., eds., *Machine Learning and Knowledge Discovery in Databases (KDD)*, 339–355. Cham: Springer International Publishing.

Scheuerman, M. K.; Paul, J. M.; and Brubaker, J. R. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–33.

Shiryaev, A. N. 2007. *Optimal stopping rules*, volume 8. Springer Science & Business Media.

Sun, B.; Sun, J.; Dai, T.; and Zhang, L. 2021. Probabilistic verification of neural networks against group fairness. In *International Symposium on Formal Methods (FM)*, 83–102. Springer.

Sun, Y. 2023. *Algorithmic Fairness in Sequential Decision Making*. Ph.D. thesis, Massachusetts Institute of Technology.

Wen, M.; Bastani, O.; and Topcu, U. 2021. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1144–1152. PMLR.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1): 2737–2778.

Zhang, X.; and Liu, M. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*, 525–555. Springer.