INFORMATION SYSTEMS RESEARCH

Explainable AI (XAI) and Set Operation: A Dual Approach to Decoding Fairness in Decision Analysis

Journal:	*Information Systems Research*
Manuscript ID	ISR-2024-1540
Manuscript Type:	Research Article
Manuscript Category:	Regular Issue
Keywords:	Human-Computer Interaction, Explainable AI (XAI), Group Fairness, Individual Fairness, Knowledge Representation, Probation Information System

Abstract:	Human decision-making processes are inherently subjective. Identifying the source of perceived differences poses significant challenges, particularly when decisions are based on structured and comparable representations. This paper introduces a systematic dual process for decision factor analysis to decompose bias through decoding fairness. The proposed fairness methodology can be used as an information-supporting tool for discovering decision analysis patterns related to a broader concept of fairness, both at the group and individual levels. We utilize set-theoretical operations for the group fairness decision factors to describe group differences and similarities uniformly. At the individual level, a CatBoost machine learning algorithm and SHAP values are employed to restore explainable algorithmic rationale, systematically identifying impactful variables to detect individual differences. Extensive numerical experiments are conducted with a recent dataset of over 100,000 criminal justice records from 51 judicial circuits within the State of Georgia Department of Community Supervision (DCS). From the empirical results, we identified the key feature for decomposing bias in each circuit. Comparing the pattern of decomposed bias across all judicial circuits, we provide the central authorities key insights that could used to improve their policymaking in a fair way.

SCHOLARONE™
Manuscripts

Submitted to *Information Systems Research*
manuscript (Please, provide the manuscript number!)

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Explainable AI (XAI) and Set Operation: A Dual Approach to Decoding Fairness in Decision Analysis

(Authors' names blinded for peer review)

Human decision-making processes are inherently subjective. Identifying the source of perceived differences poses significant challenges, particularly when decisions are based on structured and comparable representations. This paper introduces a systematic dual process for decision factor analysis to decompose bias through decoding fairness. The proposed fairness methodology can be used as an information-supporting tool for discovering decision analysis patterns related to a broader concept of fairness, both at the group and individual levels. We utilize set-theoretical operations for the group fairness decision factors to describe group differences and similarities uniformly. At the individual level, a CatBoost machine learning algorithm and SHAP values are employed to restore explainable algorithmic rationale, systematically identifying impactful variables to detect individual differences. Extensive numerical experiments are conducted with a recent dataset of over 100,000 criminal justice records from 51 judicial circuits within the State of Georgia Department of Community Supervision (DCS). From the empirical results, we identified the key feature for decomposing bias in each circuit. Comparing the pattern of decomposed bias across all judicial circuits, we provide the central authorities key insights that could used to improve their policymaking in a fair way.

Key words: Human-Computer Interaction, Explainable AI (XAI), Group Fairness, Individual Fairness, Knowledge Representation, Probation Information System

1. Introduction

In this research, we address high-stakes decision-making scenarios where fairness is a primary concern. Fundamentally, achieving fairness is complex and often requires more than a single-step solution. A key component of this iterative process is the decomposition of biases in the informa-

tion system, enabling continuous human-computer interaction (HCI) that progressively improves decision-making and promotes fairness.

1.1. Challenges and motivations

Yet, challenges persist in this approach to decomposing bias and supporting fair decision-making. First, the group of data users may primarily consist of non-technical professionals, whose lack of technical expertise may create barriers to effective human-computer interaction. This can lead to job burnout from technology-related stress (Ragu-Nathan et al. 2008, Nastjuk et al. 2024) and insufficiency in interpreting obtained information (Bera et al. 2014, Rudin 2019, Bauer et al. 2021). Additionally, bias is decomposed after the evaluation of fairness, and we need to consider the possibility of differing interpretations at the group and individual levels, where fairness may be achieved when groups are differentiated based on a single attribute, but not when two protected attributes or individual-level considerations are taken into account. Moreover, while fairness is typically evaluated within a single decision-making unit, we observe a surging need that a broader context of fairness should be considered when multiple decision-making units are involved.

1.2. High-stakes decision making: Enhanced Supervision Program (ESP) skills

For example, a high-stakes decision-making scenario in the criminal justice system involves supervisors in Georgia deciding whether to apply Enhanced Supervision Program (ESP) skills to individual supervisee. The goal of applying ESP skills is to improve supervisor-supervisee interactions and help determine if a supervisee can be granted unsupervised status (Department of Community Supervision 2020). Fairness concerns would arise if protected attributes such as race were relevant in differentiating these decisions, highlighting the need for bias decomposition.

However, while supervisors have increasing autonomy in using criminal justice information systems, they are typically less required to have a technical background (ZipRecruiter 2024), which makes them more susceptible to lacking the expertise needed to navigate complex datasets and reliant on digital tools for fair decision-making. Simultaneously, the information system containing ESP decisions encompasses multiple features, such as race, sex, and age, necessitating an approach

that represents and evaluates fairness at various levels of data granularity, including broader categories (e.g., male vs. female) and more specific subgroups (e.g., young Black female vs. young White female). Further, probation services are intended to follow consistent policies; however, variations in staffing levels and workloads across districts could result in differing implementation (United States Courts 2003). In states with multiple judicial circuits, it is essential to evaluate fairness within and across circuits, comparing the decomposed bias patterns to inform central policy adjustments where necessary. Therefore, a comparable framework representing fairness must be established within fairness evaluation tools to decompose bias efficiently and clearly.

1.3. Methodology and empirical application

To tackle these challenges, we incorporate technical assistance in the proposed fairness module designed to decompose the bias underlying decisions, supporting the needs of data users to fully leverage all available data across various fairness scenarios.

We first provide conceptual modeling support (Wand and Weber 2002, Ong Jr and Jabbari Sabegh 2019) to assist in understanding the abstract representations of fairness at the group level. This includes frameworks that identify key components and relationships by encoding the fairness context into a tree diagram, employing visualization tools for diagramming, and using performance indicators to decode fairness and decomposite bias in a uniform and comparable structure. Next, we introduce AI Coaching (Schmidt et al. 2020, Bauer et al. 2023, Lu and Zhang 2024) by leveraging machine learning to assess fairness at the individual level. We gain insights by applying Explainable AI (XAI) techniques to capture the underlying algorithmic reasoning by considering all instances – individuals – in the database. Rather than predicting ESP skills usage based on features defining similar individuals, we focus on demonstrating how each feature contributes to an individual's probability of receiving ESP skill usage to detect hidden bias.

We requested confidential data from 50 judicial circuits and one virtual circuit in Georgia, encompassing 103,717 criminal justice outcomes for November 2022. This information system includes ESP decisions for all individuals with active probation cases on November 30, 2022, excluding those

in custody, administrative, or warrant status. In the given decision-centric information system, for each judicial circuit, we obtain 5 features that could be used to define protected attributes including subpopulation features – `Race`, `Sex`, `Age`, `Education` – and a baseline factor `Risk_Score` in evaluating fairness in the decision variable `ESP Usage` as shown in Table 1.

In the conceptual modelling approach, a tree diagram provides a hierarchical representation of groups based on the intersection of the above features to generate performance indicators for `ESP Usage`. In the AI coaching approach, a SHAP-adapted decision tree machine learning architecture is used to decompose bias arising from the algorithmic reasoning between the independent variables (subpopulation features with the baseline factor: `Race`, `Sex`, `Age`, `Education`, `Risk_Score`) and the dependent outcome variable (`ESP usage`).

Aligning with the broader requirement of fairness, we treat each judicial circuit, delineated by its geographic location, as a decentralized unit. Managerial benefits can be derived by supporting supervisors in each circuit to evaluate their decision-making processes for decomposing bias. Assessing the fairness across the 51 circuits, valuable insights can be provided to central authorities, such as policymakers, enabling them to reconsider and develop fair policies that could apply uniformly across all circuits. Our numerical results suggest that some of the features may be biased at both group and individual levels, for example, the `Risk_Score` feature. For continuous improvement, policymakers should examine the key factors in the methodology used to derive the `Risk_Score`, as it might be associated with factors such as unbiased components. For example, factors such as `Race` and `Sex` may appear unbiased at the individual level, yet their effects may be used in determining the value of `Risk_Score` that lead to differing probabilities of ESP usage.

1.4. Paper organization

The structure of the paper is as follows: Section 2 reviews the literature and key concepts for the proposed fairness module. Sections 3 and 4 introduce a dual bias decomposition approach: Section 3 covers group-level fairness decomposition, while Section 4 details an AI coaching system for individual-level fairness evaluation. Both sections outline a methodology to represent fairness and performance indicators decomposite bias. Section 5 reports empirical results from Georgia data, and Section 6 highlights the relevance of the dual approach in managerial implications.

Table 1 Georgia DCS supervisee data on November 30th, 2022

Data Type	Supervisee Data	Explanation
Outcome	ESP Use	A binary variable indicates whether an officer used at least 1 Enhanced Supervision Program (ESP) Skill during an interaction with the individual for the entire month of November
Subpopulation	Sex	Self-reported sex: Female, Male
	Race	Self-reported race: Asian, Black, Hispanic, Native American, Pacific Islander, Racially Mixed, White, Other, Unknown
	Age	Difference in years between date of birth and 11/30/22: 15-97 or Unknown
	Edu	Highest self-reported years of education: 1- 21 or Unknown
Evaluation	Risk_Score	Risk Score: 1 - 10 or Unknown
Location	Judicial Circuit	50 judicial circuits and one interstate judicial circuit

For details of risk score calculation, see Georgia Department of Community Supervision Unified Risk Assessment 2017.

2. Literature Review and Motivations

In proposing our dual approach, we revisit the concepts of fairness and bias to outline the need for a structured representation in encoding and decoding fairness during bias decomposition. This structure is designed to ensure that the decomposed bias is reliable, expressive, and adequate to support data users' needs in navigating complex datasets for continuous human-computer interaction.

2.1. Fairness

Fairness is introduced to ensure equal treatment of comparable individuals and groups (Dwork and Ilvento 2018a), which metrics could be broadly classified into statistical metrics, similarity-based metrics, and metrics derived from causal reasoning (Verma and Rubin 2018).

Statistical measurements for fairness underscore equal treatment across different groups varied on protected attributes (e.g., statistical parity and disparate impact (Chouldechova 2017)). These metrics could be further expanded by incorporating (i) fixed non-protected attributes (e.g., conditional statistical parity (Corbett-Davies et al. 2017)), (ii) fixed binary prediction outcome as a condition (e.g., equal opportunity and equalized odds (Hardt et al. 2016, Zafar et al. 2017, Berk et al. 2021)), or (iii) fixed prediction probability score as a condition (e.g., predictive parity and calibration, balance for positive class and negative class (Kleinberg et al. 2016, Fu et al. 2020)).

Similarity-based and causal-based metrics are more practical for scenarios where individual fairness is enforced. In such cases, fairness is evaluated among individuals instead of comparing aggregated group levels. In the similarity-based measures (Dwork et al. 2012, Dwork and Ilvento

2018b), individual fairness mandates that similar individuals—regardless of whether clustering is based on predefined protected attributes or on features not predetermined—should receive similar "distribution on outcomes (Dwork et al. 2020)." From causal reasoning (Kusner et al. 2017, Nabi and Shpitser 2018), individual fairness dictates that the treatment of an individual should not exhibit a causal relationship with protected attributes.

2.2. Bias decomposition for fairness

When setting fairness as the ultimate goal, the initial step starts from bias decomposition: quantifying the deviations from fairness (Kordzadeh and Ghasemaghaei 2021).

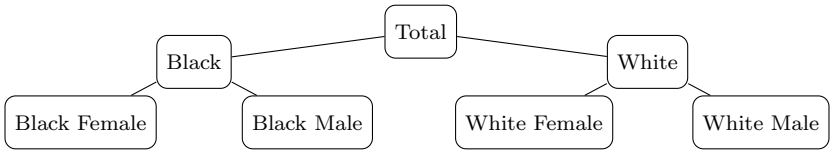
Assume a targeted outcome (variable $\mathbf{a}$) across disjoint groups defined by protected attributes (variable $\mathbf{b}$), optionally conditioned on fixed prediction factors (variable $\mathbf{z}$). The bias metrics evaluate the degree of dissimilarities: from the group level perspectives, captured by some function $f(p(\mathbf{a} \mid \mathbf{b} = b_i, \mathbf{z}), p(\mathbf{a} \mid \mathbf{b} = b_j, \mathbf{z}))$, where groups differ by protected attributes $\mathbf{b}$; from the individual level perspectives, captured by $g(u, v)$, when u and v are both in the same cluster $\mathbf{C}$ that $u \overset{\mathbf{C}}{\sim} v$. At the group level, we identify the feature $\mathbf{b}$ that exhibits deviations from fairness when decomposing bias. At the individual level, we identify the feature used to define the cluster $\mathbf{C}$ if similar individuals within the same cluster are treated differently.

2.3. Bias decomposition in hierarchical data

When decomposing bias, it's important to account for data organized hierarchically. For example, probation information systems often contain data on subpopulation features like race and sex, which could be organized hierarchically with varying data granularity.

Past research has highlighted potential discrepancies in fairness results derived from aggregated versus disaggregated data (Bickel et al. 1975, Binns 2020, Gohar and Cheng 2023). Those are relevant to Simpson's Paradox (Simpson 1951), as the consideration of the hierarchical structure can shift the fairness conclusion and alter the observed pattern of decomposed bias when the groups are combined (or disaggregated). It cautions against potential discrepancies when comparing fairness results derived from coarse and fine (partially aggregated) data, with implications extending to

Figure 1 Sample hierarchy



include disaggregated and individual-level data. For example, consider the structure in Figure 1, fairness might appear to be achieved when comparing the Black and White groups, but this may not hold true when comparing the Black female group with the White female group.

2.4. Structured representation of fairness for bias decomposition

Given the presence of multiple features in the information system, the proposed approach must account for the interplay between aggregated and disaggregated data while providing a consistent and structured representation of fairness for comparison across all decentralized decision units. Fundamentally, bias decomposition would lack reliability and effectiveness without an appropriate fairness representation, as it relies on a well-defined fairness structure.

Yet, developing an effective fairness representation for bias decomposition poses a significant challenge, as it requires structured methodologies to address complex decision-making problems (Covaliu and Oliver 1995, Frini et al. 2017). In information retrieval systems, representation aims to organize objects for processing and presenting information (Orłowska and Pawlak 1984, Burton-Jones and Grange 2013, Bhatia 2019). The expressive power of a representation is evaluated by its ability to distinguish elements based on their characteristics (Orłowska and Pawlak 1984, Kolaitis 2005) at two levels: the "knowledge level," which examines conceptual modeling, and the "symbol level," which addresses computational data structures (Brachman and Levesque 2004).

2.5. Representation at the knowledge level

Tree diagrams, characterized with node-edge structure, have been widely used in applied combinatorics (Keller and Trotter 2017) and probability space field (Lyons and Peres 2016). These diagrams are often utilized in conceptual modeling to represent objects based on set relationships, particularly in instances involving hierarchical orders.

In a specific branch of the tree, the child nodes represent disjoint subsets that share the same defining feature set ($\mathbf{z}$) found in the parent nodes, but differ in the feature ($\mathbf{b}$) – which can be considered a potential protected attribute. The sets encoded by the nodes could represent different group fairness contexts, while the leaf nodes represent distinct groups characterized by the intersection of features from the separating conditions of their respective preceding nodes.

When two nodes branch at the same level, it can construct the context of fairness considering the intersection of features. By adding hierarchical layers or expanding the branches, tree diagrams can be used to achieve the desired level of data granularity in conceptual modeling. However, according to the set theory, disconnected nodes are incomparable, as each forms distinct partially ordered sets with their preceding nodes. This necessitates additional mapping functions to infer comparable relationships. Inefficient management of the complexities of varied levels of intersectionality by these functions and the tree construction can compromise the expressiveness of a fairness module at the "knowledge level."

2.6. Representation at the symbol level

The decision tree, a variant of tree diagrams, can be used in algorithmic representation (Quinlan 1986). It is a valuable tool for organizing information by providing "Clarity and Conciseness," "Context Sensitivity," and "Flexibility," particularly when dealing with categorical data (Quinlan 1990). The main goal is to represent computational data structures within a tree, where decision trees use predefined algorithms to categorize objects based on their labels.

When used within a single decision unit, a decision tree is well-suited for representing fairness if it doesn't allow feature reusability. However, concerns may arise regarding decision trees with a top-down approach due to their limited formalism (Quinlan 1990). The classic algorithm divides the next branch on the feature $\mathcal{N}_t$ by considering the split space defined by a single attribute: the parent node feature at $\mathcal{N}_{t-1}$. This hierarchical structure, based on the condition derived from $H(\mathcal{N}_t|\mathcal{N}_{t-1})$, leads to non-comparability across different decision units, as each unit may generate a tree with a different hierarchical ordering of splitting features. In cases requiring a comparison of multiple decision-making units, this context sensitivity might reduce the effectiveness of representing fairness at the "symbol level" for a comparable representation of group-level performance indicators.

2.7. Proposed contributions

Based on the above, alternative or revised methods based on tree structures should be proposed to address the identified drawbacks of efficient representation.

In Section 3, we propose an alternative tree structure model—a bottom-up tree transducer—to address the group fairness context, in which the performance indicators are the separating features used to achieve a desired level of intersectional fairness. We adopt an outcome-oriented approach that integrates set inclusion with probabilistic representation to derive an abstraction of fairness, addressing issues at the knowledge level. Our method does numerical reconstruction across varying levels of the hierarchy, providing an expressive and uniform representation of group fairness with guarantees at the symbolic level for empowered users.

In Section 4, we propose a SHAP-adapted decision tree method (XAI) to compare multiple decision-making units while mitigating individual biases. The decision tree enhances the interpretation of individual fairness due to its computational data structure. Each individual can be represented through a unified mapping function—a function of feature importance derived from decision tree —within a given decision unit, thereby facilitating a consistent performance indicator at the individual level.

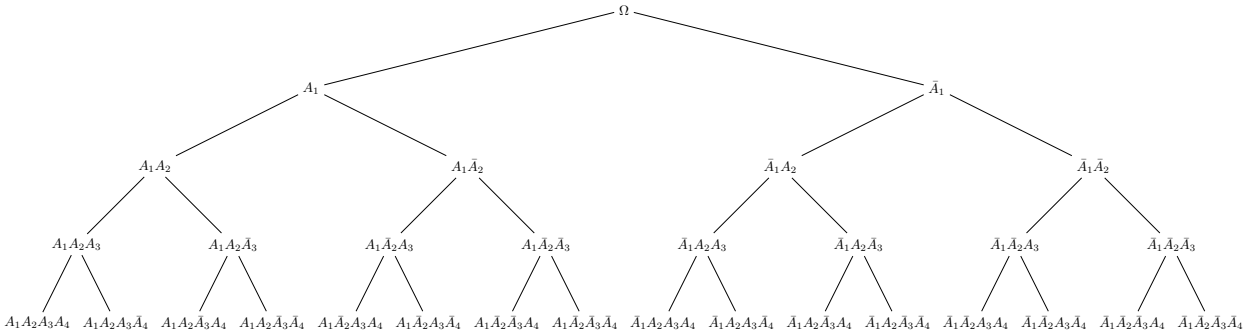
3. Fairness Methodology: Group Level

We propose a framework for group-level fairness analysis, including constructs for representing fairness in the information retrieval architecture (Sections 3.1, 3.2) and rules for generating performance indicators (Section 3.3, 3.4) to decompose bias in decentralized systems.

3.1. Fairness encoding: a uniform structure

Suppose that sets A_i, $i = 1, 2, 3, 4$ represent females, low-risk, high education, and married groups, respectively. Then, related fairness could be evaluated for (1) group A_1, assessing whether females receive equal treatment relative to other sex groups, and (2) group A_1A_2, the intersection of A_1 and A_2, evaluating whether low-risk females receive equal treatment compared to other sex-risk group intersections. In Figure 2, a unified tree structure describes the above-mentioned example

Figure 2 Description of subdivision into disjoint sets



with $n = 4$. On the bottom, $2^4 = 16$ groups assess disaggregated group fairness; right above the bottom level, $2^3 = 8$ groups evaluate fairness at a finer partially aggregated level related to sex, risk level, and education intersection.

Group fairness requires comparing disjoint groups at the same granularity. Concerns about discrepancies emerge when extending to a different data granularity level. For instance, group A_1A_2 is considered fair if no disparities against $A_1\bar{A}_2$, $\bar{A}_1A_2$, and $\bar{A}_1\bar{A}_2$. In contrast, group $A_1A_2A_3$ is assessed as unfair if discrepancies arise, for example, with $A_1A_2\bar{A}_3$. As discussed in Section 2.3, group fairness can be framed by positioning the protected groups and the comparison groups at the same level within a tree structure. Simply put, group fairness can be assessed based on comparisons of nodes at the same level of the tree structure as in Figure 2.

In more general cases, let $A_1, \ldots, A_n$ be given features in a dataset Ω. We can subdivide Ω into 2^n disjoint subsets by taking intersections of the original events and their complementary events $(A_1A_2\cdots A_n, \ldots, \bar{A}_1\bar{A}_2\cdots\bar{A}_n)$. Positioning the 2^n disjoint subsets as the base nodes and fixing the order of sequential aggregation, the bottom-up tree encodes a unified architecture that addresses fairness across varying data granularities. In this research, we treat the level-number increase in the top-down approach, with the level number at the bottom level being n.

Let us consider indices of given n different categorical features A_i for a more systematic mathematical form. Then we let $V = [n] = \{1, 2, \ldots, n\}$, a set of integers (or indices), together with $\mathbb{B} = \{0, 1\}$. For subdivision into disjoint ones, the following binary vector would be useful for a

more compact mathematical form: $\mathbf{b} = (b_1, \ldots, b_n) \in \mathbb{B}^n$, for any subset $S \subseteq V$, where

$$b_k = \begin{cases} 1 & \text{if } k \in S \subseteq V = [n] \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Introducing the binary outcome variable $\mathbf{a}$, suppose that we're given six categorical features $A_1, \ldots, A_6$. If we're interested in $p(\mathbf{a} = 1 \mid A_1\bar{A}_2A_3A_4\bar{A}_5\bar{A}_6)$ then we can reformulate it as $p(\mathbf{a} = 1 \mid \mathbf{b} = b)$, where $b = (1, 0, 1, 1, 0, 0)$. In case of $n = 3$ (i.e., $\mathbb{B}^3$), there are $2^3 = 8$ possible cases (therefore 8 distinct binary vectors $\mathbf{b} = (b_1, b_2, b_3) \in \mathbb{B}^3$):

$$A_1A_2A_3, \bar{A}_1A_2A_3, A_1\bar{A}_2A_3, A_1A_2\bar{A}_3, \bar{A}_1\bar{A}_2A_3, \bar{A}_1A_2\bar{A}_3, A_1\bar{A}_2\bar{A}_3, \bar{A}_1\bar{A}_2\bar{A}_3, \tag{2}$$

equivalent to $(1,1,1), (0,1,1), (1,0,1), (1,1,0), (0,0,1), (0,1,0), (1,0,0), (0,0,0)$.

Using (1), any target human groups with multiple features can be expressed by a binary vector. For simplicity of mathematical expressions, we present $p_i = p(a \mid b_i)$ for $p(\mathbf{a} = 1 \mid \mathbf{b} = b_i)$, where $b_i \in \mathbb{B}^n$. Also, for the counterpart of b_i in terms of sets, let us use B_i. For example, (2) can be written as

$$\begin{aligned} &B_1 = A_1A_2A_3, B_2 = \bar{A}_1A_2A_3, B_3 = A_1\bar{A}_2A_3, B_4 = A_1A_2\bar{A}_3, \\ &B_5 = \bar{A}_1\bar{A}_2A_3, B_6 = \bar{A}_1A_2\bar{A}_3, B_7 = A_1\bar{A}_2\bar{A}_3, B_8 = \bar{A}_1\bar{A}_2\bar{A}_3. \end{aligned} \tag{3}$$

3.2. Bottom-up tree transducer with numerical reconstruction

Due to the commutative property of set intersections, the bottom nodes (at the highest level number) remain unchanged, while the nodes at the intermediate level number may differ.

In the tree diagram (Figure 2), fairness can be assessed by considering sex (A_1 vs. $\bar{A}_1$), intersectional fairness of sex and risk score (A_1A_2 vs. $\bar{A}_1A_2$), and at other granularity levels. Even though intersectional fairness could not be directly inferred when examining sex and education levels (A_1A_3 vs. $\bar{A}_1A_3$), a bottom-up weighted tree transducer can be used to reorder nodes based on set aggregations (Seidl 1994, Bozapalidis 1999, Borchardt and Vogler 2003, Maletti 2005).

When the most disaggregated groups are placed as the bottom nodes of the tree, intersectional fairness can be efficiently assessed at various levels of data granularity using a bottom-up approach. The reason is as follows. Let B be the union of arbitrary disjoint sets B_i, $i = 1, \ldots, N$:

$$B = \bigcup_{i=1}^{N} B_i, B_i \cap B_j = \emptyset, 1 \leq i \neq j \leq N. \tag{4}$$

Then, the conditional probability of E given B can be written up as:

$$P(E \mid B) = \frac{P(EB)}{P(B)} = \frac{P(E \cap (\bigcup_{i=1}^{N} B_i)}{P(B)} = \frac{P(\bigcup_{i=1}^{N} EB_i)}{P(\bigcup_{i=1}^{N} B_i)} = \frac{\sum_{i=1}^{N} P(EB_i)}{\sum_{i=1}^{N} P(B_i)} = \sum_{i=1}^{N} \frac{P(EB_i)}{P(B_i)} = \sum_{i=1}^{N} P(E \mid B_i). \tag{5}$$

From equation (5), we observe that once $P(E \mid B_i), B_i \cap B_j = \emptyset, 1 \leq i \neq j \leq N$ are obtained, a conditional probability of E given any union of B_i's can be obtained by efficient union operation (i.e., addition) since B_i's are disjoint.

Algorithm 1 Aggregation using a bottom-up approach for tree construction

Require: Let k denote a level of the tree. Let $p_i^{(k)} = p(\mathbf{a} = 1 \mid B_i^k) = p(a \mid b_i^k)$, for $i = 1, \ldots, N$, and B_i^k's be mutually exclusive categorical features. Suppose that p_i's are the node values at the same level k of a tree: $(p_1^{(k)}, \ldots, p_N^{(k)})$.

for node values at the level $k-1$ (one level above level k) **do**

$$(p_1^{(k-1)}, \ldots, p_J^{(k-1)}) = (p_1^{(k)}, \ldots, p_N^{(k)}) \begin{pmatrix} M_{11} & M_{12} & \ldots & M_{1J} \\ M_{21} & M_{22} & \ldots & M_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ M_{N1} & M_{N2} & \ldots & M_{NJ} \end{pmatrix}, \tag{6}$$

where $M_{ij} = 1$ for features i's are added together for node j (otherwise, 0); and J denotes the number of nodes at level $k-1$.

end for

When designing the knowledge representation system, two factors are key: (i) efficiently structuring information during modeling and (ii) querying it during retrieval. Fairness is evaluated between any two nodes (subsets) at the same tree level, with efficiency enhanced by the disjoint property. Once bottom-level indicators are obtained, Algorithm 1 can efficiently encode the entire tree (any order, any number of nodes) recursively.

In conclusion, the tree structure can efficiently represent fairness conceptually (bottom-up approach) and computationally (weighted transducers). It addresses the comparability using a uniform tree structure that can be used to reconstruct the fairness for varying data granularity.

3.3. Systematic deviation and fairness indicator

Let p_i and p_j denote probabilities given a target group and another group to compare with (say, groups i and j) conditioned on groups B_i, B_j. That is, $p_i = p(\mathbf{a} = 1 \mid B_i) = p(a \mid b_i)$ and $p_j = p(\mathbf{a} = 1 \mid B_j) = p(a \mid b_j)$, where $\mathbf{a}$ is a binary outcome variable. We can think of two metrics for deriving the deviations: $p_i - p_j$ (statistical parity) and p_i/p_j (disparate impact). Fairness between group i and j is indicated by $p_i - p_j$ being closer to 0 and p_i/p_j being closer to 1.

In a broader context, systematic deviations can be evaluated within specific application scenarios. For example, including a central utility measurement $\bar{\mathbf{u}}$ is to constrain the variation for all groups i in $f(\mathbf{u}_{z,\mathbf{b}=b^{(i)}}, \bar{\mathbf{u}})$ (Chen and Hooker 2022). In scenarios where decentralized groups adhere to a unified policy, group-level fairness can be addressed by introducing a central measurement. For the mapping function that connects the concept of fairness with entailment relation on incomparable sets, we present two functions for effective comparison using a central benchmark.

Definition 1 (Indicator for a target group) *Suppose that the given target outcome is favorable. For group B_i, we calculate the following by introducing the average performance, $p(a)$, derived from all decentralized units as a central measurement:*

$$g_i = \begin{cases} 1 & \text{if } p(a \mid b_i) > p(a) \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Note that when the difference between $p(a \mid b_i)$ and $p(a)$ is negligible, a relative difference function in the following Definition 2 can be more suitable.

Definition 2 (Relative difference) *As in Definition 1, we assume the target outcome is favorable.*

$$d_i = \frac{p(a \mid b_i) - p(a)}{p(a)}, \tag{8}$$

which is negative when discrimination is against group b_i, and vice versa. (8) can be converted to a percent difference by multiplying it by 100%.

Remark 1 (Relative difference vs. statistical parity and disparate impact). Statistical parity and disparate impact are probably two of the most widely used metrics for parity calculation. The relative difference can be considered a combination of statistical parity and disparate impact since the numerator of (8), $p(a \mid b_i) - p(a)$ is equivalent to the statistical parity. If we rewrite (8) as $d_i = \frac{p(a \mid b_i)}{p(a)} - 1$, then it is a disparate impact minus 1.

3.4. Bias decomposition: Conceptual modeling support

Conceptual modeling uses a process-oriented approach to show how bottom-up tree structures reveal human reasoning. However, explaining the abstract representation complicates understanding for non-technical data users. To address this, we decode the fairness script with an outcome-oriented approach after encoding the fairness methodology using the bottom-up tree transducer.

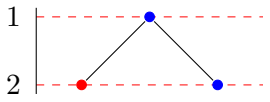
For our analyses on group level fairness throughout this paper, we use the relative difference measurement of (8) as base functions for effective comparison. See Figure 3 where the node colors represent the values of the relative difference function of (8), together with the value of n of $b_i \in \mathbb{B}^n$ (on the vertical axis).

Example 1 (Dimension of $\mathbb{B}^n$ as a layer). As mentioned above, when $n = 1$ it is a single feature for conditions in $p(a \mid b_i^{(1)})$, where $b_i^{(1)} \in \mathbb{B}^1$. When $n = 2$, it is pairwise intersections of given features. For the general case, when $n = k$, it is k-tuples of features: $p(a \mid b_i^{(k)})$, where $b_i^{(k)} \in \mathbb{B}^k$. In Figure 3, there are two decision units that we want to compare in terms of the relative difference. Suppose that $b^{(2)} = (b_1, b_2)$, where $b_1 = 1$ if a person has a high school degree (otherwise, 0); $b_2 = 1$ if a person is a female (otherwise, 0).

Let the blue and red nodes in Figure 3 denote positive and negative values of the relative difference of (8), respectively. We observe that all blue nodes are at level 1 of the figures, inferring both circuits give the treatment (a special program) to people at a higher probability than the average of the entire organization, whether the applicants have high school degrees or not. At level 2, Circuit 1 is more likely to put males into the program among parolees without high school degrees. Related numerical works are presented in Section 5.1, where a recent dataset of over 100,000 criminal justice records from 51 judicial circuits within the State of Georgia are analyzed.

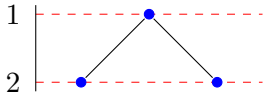
Figure 3 Two color characterization graph for effective comparison

(a) Judicial Circuit 1



Note. All nodes are blue except one bottom left.

(b) Judicial Circuit 2



Note. All nodes are blue.

4. Fairness Methdology: Individual Level

Group level fairness can be assessed using the tree diagram with set-theoretic operation and probabilistic reasoning (Section 3). To evaluate individual-level fairness, we would need to take a closer look at each of the individual observations by applying decision tree. In this section, we leverage insights from an explainable machine learning (XAI) framework to derive algorithmic decision-making patterns, thereby generating statements for decomposing bias at the individual level.

4.1. Machine learning and decision tree

Machine Learning (ML), an integral facet of Artificial Intelligence (AI), facilitates logical inference by generating algorithmic rasoning (Jordan and Mitchell 2015, Athey and Imbens 2019). For the sake of completeness, let us briefly present some notions. Given the input data $\mathbf{X} \in \mathbb{R}^{m \times d}$, a machine learning model f is trained by minimizing the loss $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}; \theta)$ between the prediction $\hat{\mathbf{y}} = f(\mathbf{X}; \theta)$ and the ground truth $\mathbf{y} \in \mathbb{R}^m$, with m: the number of observations; d: the number of features; and θ: model parameters (Mitchell and Mitchell 1997). For classification tasks, a threshold τ is introduced that a targeted label would be assigned if $\hat{y} \geq \tau$.

In training f, a decision tree is an efficient ML architecture. It partitions data into subsets for subsequent predictions, with each branch representing a rule (e.g., "Age ≥ 10") that directs instances to different nodes. This process, applied recursively to maximize information gain, continues until reaching a leaf node with the final prediction (Quinlan 1986). Unlike the tree structures discussed in Section 3, which use set inclusion or exclusion to represent relationships among groups, decision trees branch on decision rules and are evaluated based on overall individual performance.

After constructing the model architecture, the parameters θ in $f(\mathbf{X};\theta)$ must be optimized to minimize the prediction loss. For a binary classification task, the ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (TPR), $\frac{TP}{TP+FN}$, against the False Positive Rate (FPR), $\frac{FP}{FP+TN}$, with the Area Under the Curve (AUC) serving as a measure to quantify performance. Generally, an AUC of 1 indicates a perfect model, an AUC of 0.5 corresponds to random guessing, and an AUC of 0.7 suggests a moderate level of performance (Fawcett 2006, Mandrekar 2010).

To prevent overfitting and enhance generalization, machine learning involves splitting the dataset $\mathbf{X}$ into a training set $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$ for model development and a test set $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$ for evaluation. In addition, cross-validation offers a robust evaluation by dividing the dataset into multiple subsets, enhancing the reliability of performance assessments. For example, in leave-one-out cross-validation, the model is trained on all subsets except one and tested on the excluded subset in each iteration, providing a comprehensive assessment across all data points.

4.2. Explainable AI (XAI) and feature importance

When training machine learning models, we often encounter opaque models that are difficult for humans to interpret due to their complex architectures in $f(\mathbf{X};\theta)$, making trust challenging (Rudin 2019, Hickey et al. 2021, Grabowicz et al. 2022). As a solution, Explainable AI (XAI) has gained significant attention from both industry and academia, aiming to address the lack of transparency, especially for high-stake decision-making (Adadi and Berrada 2018, Arrieta et al. 2020).

Regarding explainability, decision trees do not necessitate post-hoc analysis, making it one of the most widely used methods with added visualizability. Nevertheless, it is recommended to implement techniques such as feature relevance when multiple tree structures are involved (Arrieta et al. 2020). The rationale is that, despite the challenges in comparing varied hierarchical structures across multiple units, analyzing feature relevance or feature importance can yield more structured and insightful conclusions. This approach provides a comparable structure with importance scores that can be standardized, aiding in identifying similarities and outliers across multiple decision units.

For decision trees, metrics for feature importance include (i) Mean Decrease in Impurity (MDI) (Breiman 2001); (ii) Permutation Importance (PI) (Altmann et al. 2010); and (iii) SHAP (SHapley Additive exPlanations) importance (Lundberg and Lee 2017a,b). MDI calculates the average

reduction in impurity (e.g., Gini impurity) within the training data; PI evaluates the performance decrease when a single feature is permuted using the testing data; and SHAP, an alternative to PI, assesses each feature's contribution to the prediction (Christoph 2020).

However, the selection of feature importance metrics should consider both data structure and model architecture. For instance, Gini impurity becomes insufficient when non-binary categorical features are involved (Strobl et al. 2007). Considering the fairness context typically involves multi-categorical subpopulation features, we apply SHAP importance.

Remark 2 (Adoption and Impact of SHAP importance). Introduced by Lundberg and Lee in 2017, SHAP (SHapley Additive exPlanations), a game-theoretic method for explaining machine learning outputs, is widely recognized for their effectiveness in explaining complex models in academia and industry. As of August 2024, the method has been cited over 25,877 times on Google Scholar, highlighting its strong theoretical foundation and practical relevance. Originally derived from Shapley values (Shapley et al. 1953), SHAP importance assesses the change based on the coalition function $\mathcal{V}$ given the feature space in S by removing a given feature g from a set F (Lundberg and Lee 2017b):

$$\Phi_g = \sum_{F \subseteq S \setminus \{g\}} \frac{|F|!(|S| - |F| - 1)!}{|S|!} \left[\mathcal{V}(F \cup \{g\}) - \mathcal{V}(F)\right]. \tag{9}$$

For a more suitable model architecture, we also consider the model's accuracy and numerical reliability, acknowledging that various off-the-shelf packages for decision tree models require high precision and accuracy (Hill et al. 2024). In handling categorical datasets, for this paper, we employ the CatBoost algorithm regarding its foundation in the gradient boosting algorithm on decision trees (CatBoost 2018, Prokhorenkova et al. 2018, Dorogush et al. 2018).

4.3. SHAP in representing fairness and decomposing bias

To the authors' best knowledge, no method has yet been proposed using SHAP values to explain fairness at the individual level. By accounting for interactions between features among all possible combinations, SHAP values can be useful as a standardized method for measuring each feature's contribution to predictions, enabling the assessment of individual fairness.

As presented in Section 2.1, similar individuals ($u \overset{C}{\sim} v$) should receive identical predictions (Dwork et al. 2012, Dwork and Ilvento 2018b, Dwork et al. 2020). When evaluated under the same threshold τ, we can write:

$$\mathbb{1}(\hat{y}(u) \geq \tau) = \mathbb{1}(\hat{y}(v) \geq \tau) \tag{10}$$

where $\mathbb{1}(\cdot)$ denotes the binary indicator function applied to the model predictions $\hat{y}(u)$ and $\hat{y}(v)$.

Due to possible truncation errors, a relaxed condition could be generated considering the fairness with *acceptable threshold* ϵ (some small positive constant):

$$|\hat{y}(u) - \hat{y}(v)| \leq \epsilon. \tag{11}$$

Given a trained model $\mathcal{M}$, we aim to evaluate feature importance for similar observations in datasets $\mathbf{X} \in \mathbb{R}^{n \times d}$ to assess individual-level fairness. Let the SHAP value of the j-th feature for the i-th observation be denoted as $\Phi_{(i,j)}^{\mathcal{M}}$, where $i \in [n] = \{1, \ldots, n\}$ and $j \in [d] = \{1, \ldots, d\}$ correspond to the indices of the rows (samples) and columns (features), respectively, within the dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$. Then, $\Phi_{(i,j)}^{\mathcal{M}} = a$ indicates that for i-th individual, the j-th feature contributes an increment of a to the predicted outcome $\hat{y}_i^{\mathcal{M}}$.

According to the additive nature of Shapley values (Shapley et al. 1953), the SHAP importance of all input features will sum to the difference between the baseline model and model $\mathcal{M}$ when all qualifying features are presented (Lundberg and Lee 2017b):

$$\sum_{j \in [d]} \Phi_{(i,j)}^{\mathcal{M}} = \hat{y}_i^{\mathcal{M}} - E(\hat{y}), \tag{12}$$

where $E(\hat{y})$ denotes the model's average prediction across all data points in the baseline model.

In the context of individual fairness, similar individuals—specifically, the u-th and v-th observations in a given database ($X_u \overset{C}{\sim} X_v$)—should exhibit comparable feature contributions ($u, v \in [n], u \neq v$). And Equations (11) and (12) can be reformulated as:

$$|\hat{y}_u^{\mathcal{M}} - \hat{y}_v^{\mathcal{M}}| = |(\hat{y}_u^{\mathcal{M}} - E(\hat{y})) - (\hat{y}_v^{\mathcal{M}} - E(\hat{y}))| = |\sum_{j \in [d]} \Phi_{(u,j)}^{\mathcal{M}} - \sum_{j \in [d]} \Phi_{(v,j)}^{\mathcal{M}}| \leq \epsilon, \tag{13}$$

Strict individual fairness, we propose as *Augmented Individual Fairness*, can be assessed by evaluating whether, for similar individuals, all features contribute similarly:

$$|\Phi^{\mathcal{M}}_{(u,j)} - \Phi^{\mathcal{M}}_{(v,j)}| \leq \epsilon, \forall j \in [d]. \tag{14}$$

Remark 3 (Augmented Individual Fairness). Inequalities (13) and (14) can be related to the triangular inequality. This establishes a strict requirement based on the fairness requirement that "no strong features are used to differentiate individuals (Hickey et al. 2021, Grabowicz et al. 2022)." For similar individuals, the most significant feature contributing to the prediction should be identical. Let $j^{(k)}_u$ and $j^{(k)}_v$ denote the feature corresponding to the k-th largest SHAP values for individuals u and v, respectively. Ideally, $j^{(1)}_u$ should be the same as $j^{(1)}_v$, with $j^{(1)}_u = \arg\max_j(\Phi^{\mathcal{M}}_{(u,j)})$ and $j^{(1)}_v = \arg\max_j(\Phi^{\mathcal{M}}_{(v,j)})$, where $j \in [d]$. Following an inductive process, the second most important feature should be identical and similar in magnitude. This allows for a pairwise comparison that supports *Augmented* individual fairness, as presented in Inequality (14) and Example 2.

Remark 4 (Predefined Protected Attributes). Individual fairness can be assessed when protected attributes A are predefined in clustering similar individuals $X_u \overset{C_A}{\sim} X_v$ (Section 2.1). Let j_A denote the indices corresponding to the feature set A, where $|A| = d_A$. Then Inequality (14) can be updated to: $|\Phi^{\mathcal{M}}_{(u,j_A)} - \Phi^{\mathcal{M}}_{(v,j_A)}| \leq \epsilon^{'}, \forall j_A \in [d_A], A \subseteq S$.

In this research, we propose a performance indicator $\mathbf{D}(\Phi^{\mathcal{M}}_j)$ to evaluate the variations in the feature importance for the $j-$th feature:

$$\mathbf{D}(\Phi^{\mathcal{M}}_j) = \frac{R(\Phi^{\mathcal{M}}_j)}{f(\Phi^{\mathcal{M}}_j)}, j \in [d], \tag{15}$$

in which $R(\Phi^{\mathcal{M}}_j)$ denotes the range of the feature importance for all data points given feature j: $R(\Phi^{\mathcal{M}}_j) = \max(\Phi^{\mathcal{M}}_{(i,j)}) - \min(\Phi^{\mathcal{M}}_{(i,j)}), \forall i \in [n]$. To capture the differing interpretations of positive and negative contributions, we assess the sign importance in the denominator $f(\Phi^{\mathcal{M}}_j)$:

$$f(\Phi^{\mathcal{M}}_j) = \max\left(\frac{|\{i \mid \Phi^{\mathcal{M}}_{(i,j)} \geq 0\}|}{n}, \frac{|\{i \mid \Phi^{\mathcal{M}}_{(i,j)} \leq 0\}|}{n}\right), \forall i \in [n], j \in [d], \tag{16}$$

that fairness is indicated by $\mathbf{D}(\Phi^{\mathcal{M}}_j)$ being closer to 0.

Remark 5 (Evaluation of Simultaneous Positive and Negative Contributions). The denominator in Equation (15) addresses the imbalance between positive and negative contributions, with the worst-case scenario being 50% of observations evaluated positive contributions, i.e., $\min f(\Phi_j^{\mathcal{M}}) = 0.5$. For instance, consider the ranges $[-0.1, 0.1]$ and $[0.1, 0.3]$. Both ranges have a magnitude of 0.2. However, the first range includes negative and positive contributions, while the second includes positive contributions. Despite having the same magnitude, the first range may be perceived as less fair than the second due to simultaneous positive and negative influences.

4.4. Augmented individual fairness and performance indicator

Following the proposed *Augmented Individual Fairness* defined in Equation (14) with the performance indicator proposed in Equation (15), it is feasible to construct a SHAP-adapted decision tree to address individual-level fairness, as outlined in Algorithm 2.

Algorithm 2 Generation of Performance Indicators in SHAP-Adapted Decision Tree

Require: Dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$, t fold cross-Validation, clustering rules R.

Ensure: Decision tree model $\mathcal{M}$, SHAP importance $\Phi_{(i,j)}^{\mathcal{M}}$, and performance indicators $\mathbf{D}(\Phi_j^{\mathcal{M}})$.

1: For a given decision unit, split $\mathbf{X}$ into t folds.
2: **for** each fold $k = 1$ to t **do**
3: Train CatBoost model $\mathcal{M}^k$ on $\mathbf{X}_{train}^k$ and evaluate on $\mathbf{X}_{test}^k$ until desired AUC is achieved.
4: Compute SHAP values $\Phi_{(i,j)}^{\mathcal{M}^k}$.
5: **end for**
6: Apply clustering rules R to $\mathbf{X}$ to form intermediate groups $\mathcal{C}_R$.
7: Compute $\mathbf{D}(\Phi_j^{\mathcal{M}})$ using the averaged SHAP values for observations $(\Phi_{(i',j)}^{\mathcal{M}^k})$ within $\mathcal{C}_R$.

Example 2. Given three decision tree models, $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, each trained on separate decentralized decision units to a desired performance of AUC = 0.85 with feature set $S = \{\texttt{Sex}, \texttt{Risk\_Score}, \texttt{Age}\}$. Consider two similar individuals $(\text{Female}, 5, 30)$ and $(\text{Female}, 4, 28)$ obtained through a clustering rule. The SHAP values, as assessed by the three models, are listed in Table 2.

Table 2 SHAP values

		Φ_1	Φ_2	Φ_3			Φ_1	Φ_2	Φ_3			Φ_1	Φ_2	Φ_3
	u	0.50	0.30	0.10		u	0.50	0.30	0.03		u	0.50	-0.20	-0.10
$\mathcal{M}_1$	v	0.49	0.31	0.11	$\mathcal{M}_2$	v	-0.51	0.25	-0.02	$\mathcal{M}_3$	v	-0.01	-0.39	0.60

Note that SHAP, as a permutation indicator, is calculated on the testing dataset after a model is trained. In this example, assuming we are evaluating two observations with $n = 2, d = 3$.

Comparing SHAP values, the range in decision unit 1 is $(R(\Phi_1^{\mathcal{M}_1}), R(\Phi_2^{\mathcal{M}_1}), R(\Phi_3^{\mathcal{M}_1})) = (0.50 - 0.49,\ 0.31 - 0.30,\ 0.11 - 0.10) = (0.01, 0.01, 0.01)$. Thus, individual-level fairness is attained within decision unit 1, which is also supported by the small value derived the performance indicator $(\mathbf{D}(\Phi_1^{\mathcal{M}_1}), \mathbf{D}(\Phi_2^{\mathcal{M}_1}), \mathbf{D}(\Phi_3^{\mathcal{M}_1})) = (0.01, 0.01, 0.01)$. In contrast, bias is present within decision unit 2, where the differences are $(R(\Phi_1^{\mathcal{M}_2}), R(\Phi_2^{\mathcal{M}_2}), R(\Phi_3^{\mathcal{M}_2})) = (1.01, 0.05, 0.05)$, with the calculated metric $(\mathbf{D}(\Phi_1^{\mathcal{M}_2}), \mathbf{D}(\Phi_2^{\mathcal{M}_2}), \mathbf{D}(\Phi_3^{\mathcal{M}_2}) = (2.02, 0.05, 0.10)$. This deviation primarily arises from features `Sex` and `Age`, as the performance indicator $\mathbf{D}(\Phi_1^{\mathcal{M}_2})$ and $\mathbf{D}(\Phi_3^{\mathcal{M}_2})$ considerably surpasses zero. Specifically, bias is evident in the feature `Sex`, which could be utilized to differentiate outcomes for similar individuals, indicating a fairness issue at the individual level.

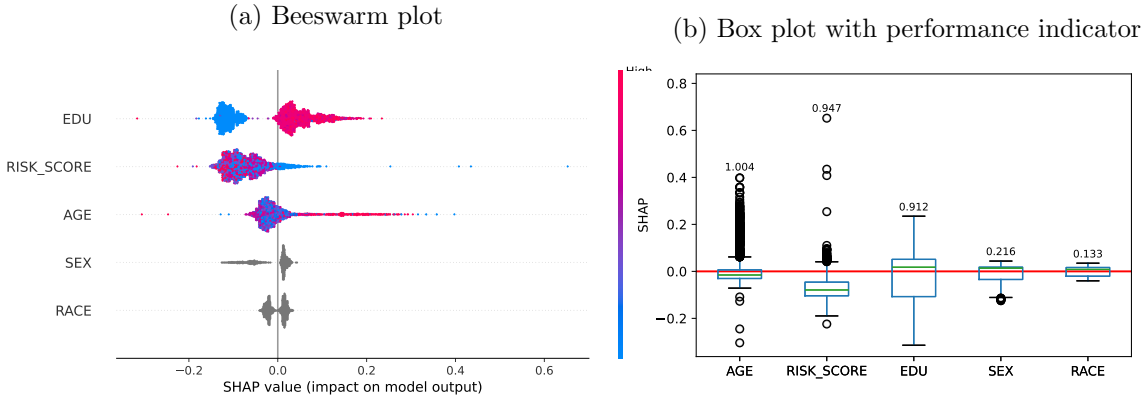
The results from decision unit 3 demonstrate the necessity of augmented individual fairness. Although the aggregated contribution difference, $\left|\sum_{j=1}^{3} \Phi_{(1,j)}^{\mathcal{M}_3} - \sum_{j=1}^{3} \Phi_{(2,j)}^{\mathcal{M}_3}\right| = |(0.50 + (-0.20) + (-0.10)) - (-0.01 + (-0.39) + (0.60))| = |0.20 - 0.20| = 0.00$, is zero, the individual feature contributions, i.e., the pairwise comparisons show significant variation with $(R(\Phi_1^{\mathcal{M}_1}), R(\Phi_2^{\mathcal{M}_1}), R(\Phi_3^{\mathcal{M}_1})) = (0.51, 0.19, 0.70)$ and $(\mathbf{D}(\Phi_1^{\mathcal{M}_3}), \mathbf{D}(\Phi_2^{\mathcal{M}_3}), \mathbf{D}(\Phi_3^{\mathcal{M}_3}) = (1.02, 0.19, 1.40)$. This implies that modifying the decision-making hierarchy could lead to differential treatment of the two individuals, which could incur intersectional fairness concerns. Specifically, features `Sex` and `Age` can differentiate between individuals, and removing either can alter prediction outcomes.

4.5. Bias decomposition: AI coaching

We utilize a beeswarm plot (Figure 4a) along with an accompanying box plot annotated with performance indicators (Figure 4b) to enhance information interaction in AI coaching.

The beeswarm plot (Figure 4a) is a type of plot used to visualize SHAP values. Such plots use scatter points that are color-coded by feature values. The spread of points along the x-axis displays

Figure 4 AI coaching with visual presentation

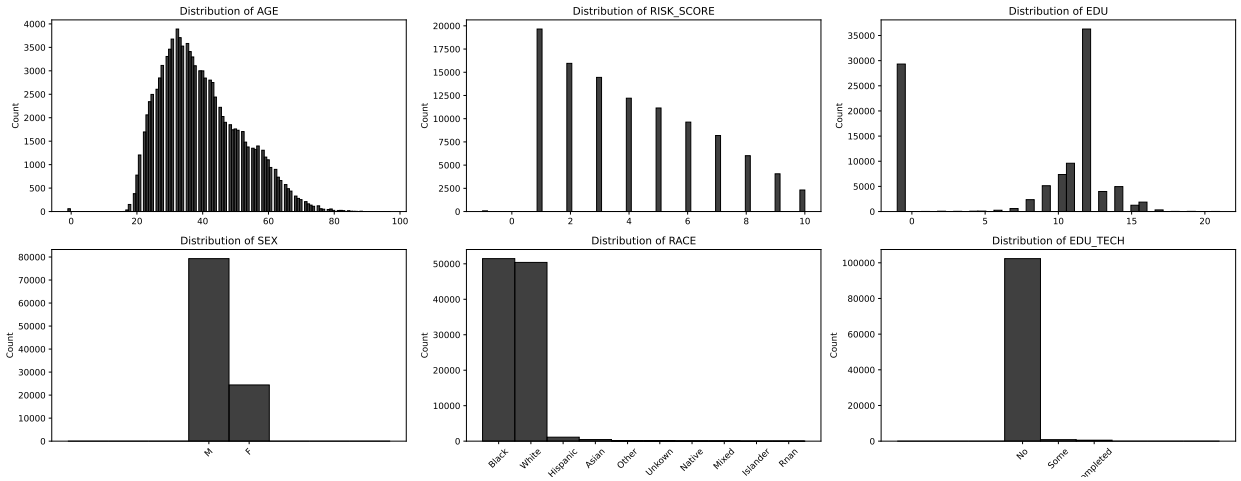


the distribution of derived SHAP values for the feature on the y-axis. Along the y-axis, features are ranked in ascending order to list the most important features at the top. In our approach, we assess individual fairness and detect potential bias by examining the color and density of scatter points, focusing on the distribution of similar colored points corresponding to each feature.

It should be noted that in the beeswarm plot, the scales of the x- and y-axis may vary depending on the range of SHAP values, which can distort the magnitude of variation in the visualization if we are comparing two beeswarm plots; nevertheless, the performance indicator remains a reliable measure for comparison. Figure 4b uses boxplots to display the distribution of SHAP values in Figure 4a across each feature. The box plot is arranged in descending order along the x-axis based on the calculated values of the corresponding performance indicator, $\mathbf{D}(\Phi^{\mathcal{M}})$, with the performance indicator annotated on top.

Providing a fictionary example as illustrated in Figure 4, consider that similar individuals are defined based on a clustering rule R applied to the feature set $S = \{\texttt{Risk\_Score}, \texttt{Edu}, \texttt{Age}, \texttt{Sex}, \texttt{Race}\}$. In this context, individuals are classified as comparable if they have similiar `Risk_Score`, `Edu`, `Age`, and identical `Sex` and `Race`. From the beewarm plot, using SHAP values, the most important feature is `Edu`, as it is positioned at the top in Figure 4a. However, given that the performance indicator $\mathbf{D}(\Phi_2^{\mathcal{M}}) = 0.912$, it demonstrates substantial variation in the model's contribution across instances clustered by the given feature, inferring significant inconsistent treatment within similar individuals. Similar conclusions can be drawn for the features `Risk_Score` and `Age`, where individual fairness may be compromised.

Figure 5 Histogram of data distribution at the state level



Note. `Sex`: Female (24,419), Male (79,296). `Race`: Black (51,461), White (50,394), Hispanic (1,119), Asian (450), Other (96), Unknown (84), Native American (55), Racially Mixed (52), Pacific Islander (3), with 1 missing value. `Age` has 60 missing values. `Edu` has 29,323 missing values. `Risk_Score` has 73 missing values. Missing values for `Sex` ,`Age`, and `Risk_Score` (all under 1%) are coded as -1. Due to over 28% missing values for `Edu`, this feature is also encoded as -1 for missing values and converted into a categorical variable.

5. Empirical Application: State of Georgia's Department of Community Supervision (DCS)

In this section, we present our numerical results on bias decomposition at both the group and individual levels. We then integrate these findings to deepen our insights using the dual approach by combining results from both levels.

5.1. Numerical results at the group level

To represent fairness at the group level and decomposite bias, we define disjoint groups based on the encoded data presented in Table 3.

As introduced in Section 3.4, we employ coloration to visualize the represented fairness and use the color contradiction to highlight the decomposition of bias. Specifically, (1) a treemap (abbreviated as "TM" in the figure) is used to help comprehend hierarchically encoded fairness, and (2) a refined binary coloration (abbreviated as "BC" in the figure) is applied to uniformly reveal the 1-0 (blue-red) contradictions. For example, in Figure 6a, a treemap with colormap illustrates the performance indicator for each subgroup in the Atlanta circuit. At the circuit level, Atlanta

Table 3 Data coding for group level analysis

Data Type	Supervisee Data	Data Coding (ordered by group counts)
Outcome (a)	ESP Use	Indicates whether an officer used at least 1 Enhanced Supervision Program (ESP) Skill during an interaction with the individual for the entire month of November
Subpopulation	Sex	Self-reported sex (S-M = Male, S-F = Female)
	Race	Self-reported race (R-B = Black, R-W = White)
	Age	Difference in years between the date of birth and 11/30/22 (A-S: 40 years of age or older, A-Y: Under 40 years old or any other case)
	Edu	Highest self-reported education level (E-H = High education level: 12 years or more, E-L = Low education level: all other cases)
Evaluation	Risk_Score	Risk Score (RISK-H = High risk score: 5 above, RISK-L = Low risk score: all other cases)
Location	Judicial Circuit	50 judicial circuits and one interstate judicial circuit

Note: We value inclusively and respect the diverse identities of all individuals. However, we presently lack other subcategories for `Sex`, `Race`, and `Judicial Circuit` in the data collection process due to their unavailability (less than 10 headcount). Minor groups (based on their group counts) are excluded, enhancing the presentation of indicators at the group level. Notably, Black and White constitute over 98% of Georgia's population, leading to a possible exclusion of other racial subgroups to avoid data bias in group comparison using set inclusions. Data coding criteria for other features are based on the median value of each feature, as illustrated in Figure 5. In the given Georgia dataset, the benchmark in Equation (8) is $p(a) = 0.088$, denoting the proportion of applying ESP usage at the state level.

has $d = -0.16$ (red), shown as a red dot at level 1 in Figure 6e. At level 2, the RISK-L group has $d = -0.41$ (red), and the RISK-H group has $d = 0.23$ (blue), represented by corresponding dots (red and blue) at level 2 in Figure 6e.

Remark 6. [Favorable vs. Unfavorable Outcome] In addressing the fairness concerns related to evaluating ESP treatment outcomes, it is important to note that in Equation (8), the treatment is assumed to be favorable, with the fairness being attained with a positive performance indicator. Rather than taking a stance on whether ESP skills are inherently positive or negative, we focus on unraveling bias by detecting contradictions (1 vs. 0) in performance indicators across disjoint groups differentiated by protected attributes, as these inconsistencies can lead to imbalanced treatment.

Empirical analysis # 1: At the group level, can bias be decomposed in `Risk_Score`?

For this analysis, we evaluate if bias could be decomposed in `Risk_Score` at the group level in each circuit. Here, level 1 corresponds to the aggregated group by circuit. level 2 represents risk subgroups, specifically the low-risk score group (RISK-L) and the high-risk score group (RISK-H) according to the rules in Table (3). As defined in Section 3.4, level 0 refers to the state level.

Figure 6a and 6e illustrate contradictions in coloration among level 2 nodes, demonstrating that bias can be decomposed in the `Risk_Score` in the given circuit. The distinct colorations for RISK-L and RISK-H indicate different treatment across these groups showing the exhibition of bias. Figure 6b and 6f illustrate consistency in coloration among level 2 nodes, demonstrating bias would not be composed by the `Risk_Score` in the provided circuit. Due to page limits, we select circuits with large group counts and the most representative patterns for illustration.

Empirical analysis # 2: At the group level, can bias be decomposed in `Race`?

In this analysis, we fix the hierarchical order as `Circuit` and `Race`, i.e., level 2 represents race subgroups. We do observe `Race` could exhibit bias at the group level for certain circuits as shown color contradiction at level 2 (e.g., Figure 7c and 7g). The proportion of circuits where bias can be decomposed in `Race` (10/51) is smaller than that observed for `Risk_Score` (19/51). However, concerns could arise that bias might be exaggerated at the group level due to the class imbalance, as seen in the Treemap (Figure 7a, 7b, 7c, and 7d), where the S-W rectangle is smaller than the S-B rectangle, with the rectangle area representing the group size. Further investigation is needed to assess the influence of `Race` at the individual level.

Empirical analysis # 3: At the group level, can bias be decomposed in `Sex`?

For this analysis, we use level 2 to represent `Sex`. We observe that `Sex` may exhibit bias at the group level in certain circuits, as indicated by the color inconsistencies at level 2 in Figures 8c and 8g. Similar concerns may arise when analyzing `Sex`, as we also observed group imbalances.

Empirical analysis # 4: At the group level, can patterns and outliers be identified?

Comparing Figure 6, 7, 8, we can observe some circuits may have distinct patterns in the decomposed bias. For example, in the Cherokee circuit, bias is decomposed into `Race` and `Sex` but not into `Risk_Score`. In contrast, the other circuits listed do not show bias decomposed into `Race` and `Sex`. Central policymakers should reassess the case presented in the Cherokee circuit to determine whether the current policy is applicable to this specific context.

By adding more hierarchical layers (levels), Table 4 presents performance indicators across 51 circuits, organized by the hierarchical structure of `Circuit`, `Risk_Score`, and `Race`. The order

Figure 6 **Contradiction and consistency between subgroups at level 2 (`Risk_Score`) within a circuit.**

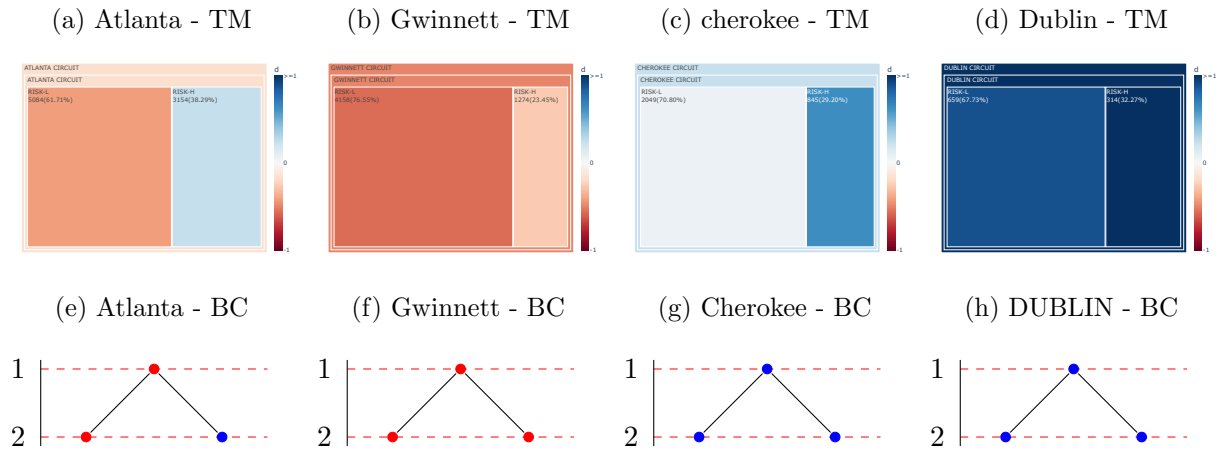


Figure 7 **Contradiction and consistency between subgroups at level 2 (`Race`) within a circuit.**

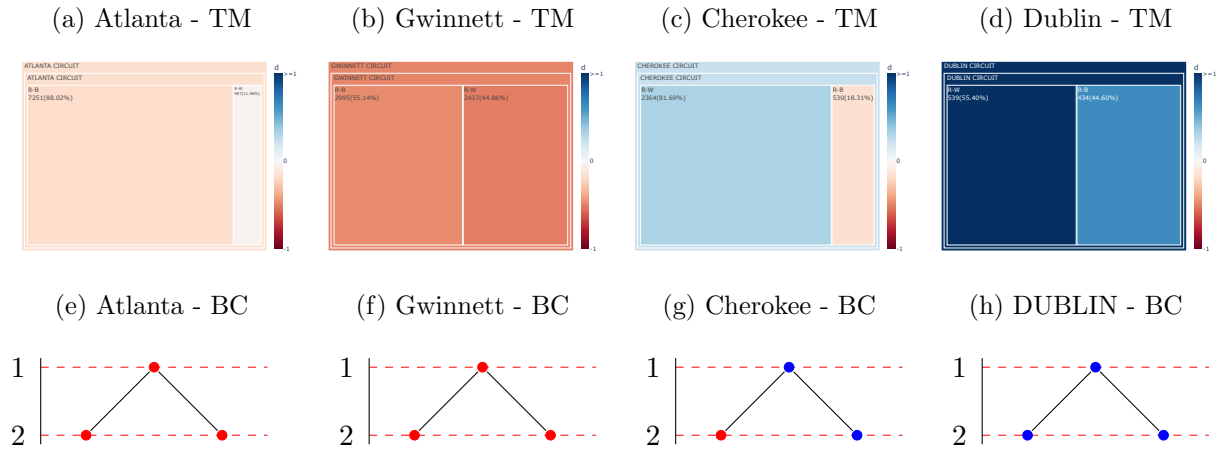


Figure 8 **Contradiction and consistency between subgroups at level 2 (`Sex`) within a circuit.**

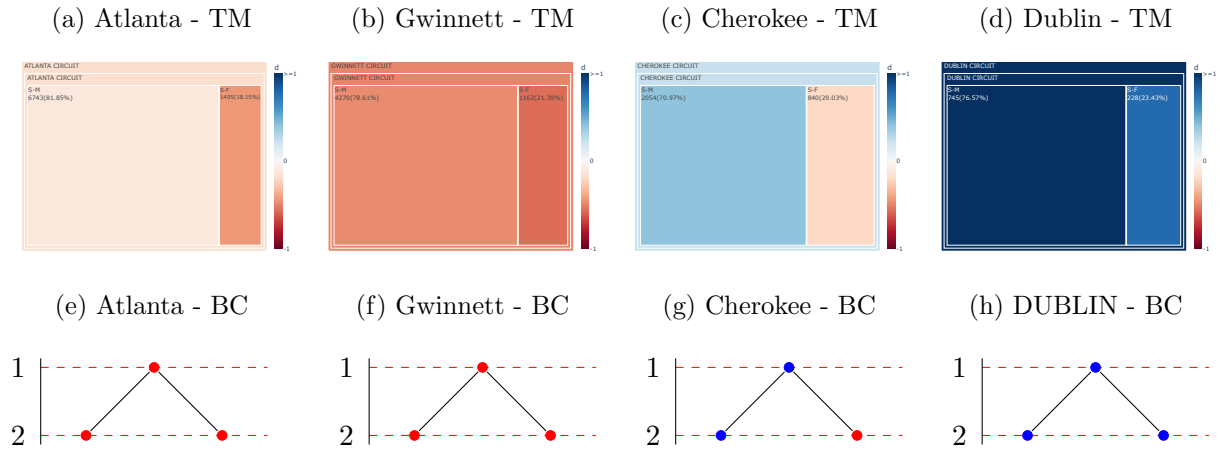


Table 4 **Level 1:** `Circuit`**, level 2:** `Risk_Score`**, level 3:** `Race`

	Patterns on d	Circuits
Pattern 1	0000000	Chattahoochee, Cordele, Enotah, Gwinnett, Ogeechee
Pattern 2	0000100	Atlantic, Southwestern
Pattern 3	0000001	Brunswick, Coweta, Flint
Pattern 4	0001000	Bell-forsyth
Pattern 5	0010001	Augusta, Cobb, Lookout mountain, Northeastern
Pattern 6	0010010	Houston, South georgia
Pattern 7	0010011	Atlanta, Blue ridge, Conasauga, Douglas, Eastern, Oconee, Pataula, Stone mountain, Western
Pattern 8	1101100	Interstate compact
Pattern 9	1011001	Appalachian, Tallapoosa
Pattern 10	1010011	Alcovy, Columbia, Dougherty, Piedmont
Pattern 11	1010111	Paulding
Pattern 12	1111001	Alapaha
Pattern 13	1110111	Cherokee, Northern, Ocmulgee, Rome, Waycross
Pattern 14	1111111	Clayton, Dublin, Griffin, Macon, Middle, Mountain, Rockdale, Southern, Tifton, Toombs, Towaliga

Note that the second column lists the binarized performance indicators for the subgroups, organized according to the level order (breadth-first traversal) as depicted in the treemap in Figure (6a). For example, the pattern 0010011 indicates that a given circuit has an overall ESP assignment below the state average, below-average assignment for RISK-L subgroups, above-average assignment for RISK-H subgroups, below-average for RISK-L R-B individuals, below-average for RISK-L R-W individuals, above-average for RISK-H R-B individuals, and above-average for RISK-H R-W individuals, consistent with the pattern observed in the Atlanta circuits.

follows the level sequence (breadth-first traversal) of the encoded tree structure (as noted in Table 4). We observe four circuits emerge as unique ones with distinct patterns. For example, Bell-Forsyth is the only circuit where ESP usage is (1) below the statewide average for the circuit overall, (2) below the statewide average for the RISK-L group, (3) below average for the RISK-H group, (4) above average for RISK-L and R-B, (5) below average for RISK-L and R-W, (6) below the statewide average for subgroups of RISK-H and R-B, and (7) below average for RISK-H and R-W.

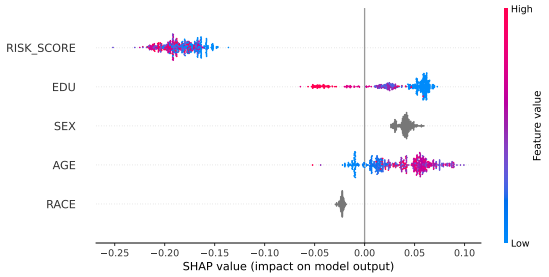
5.2. Numerical results at the individual level

In this section, we use Beeswarm plots to visualize SHAP values and apply the performance indicator $\mathbf{D}(\Phi_j^{\mathcal{M}})$ (in Equation (15)) for decomposing bias. We employ three-fold cross-validation with the Catboost algorithm to train decision trees using `Risk_Score`, `Race`, `Sex`, `Age`, and `Edu` as independent variables and `ESP usage` as the dependent variable, achieving an AUC of 0.70.

Remark 7 (Performance indicator at the individual level). In Remark 6, we use 1/0 contradiction in performance indicators d to decompose bias at the group level. Bias occurs when two groups exhibit d with different signs—one positive and one negative—revealing inconsistency.

Figure 9 **Atlanta circuit: RISK-L, R-B, S-M, A-S, EDU-H**

(a) Bee swarm plot



(b) Box plot with performance indicator

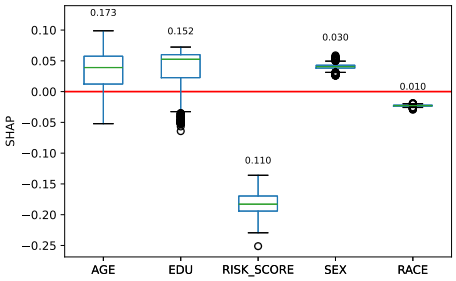
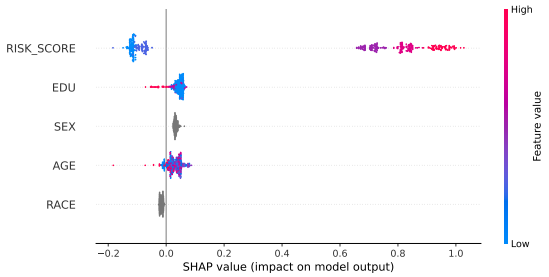
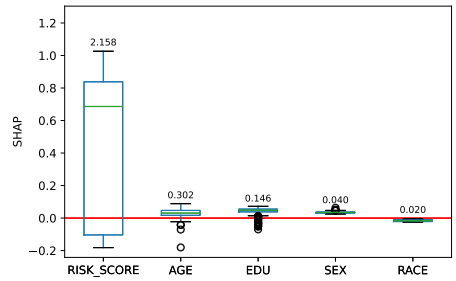


Figure 10 **Atlanta circuit: RISK-H, R-B, S-M, A-S, EDU-H**

(a) Bee swarm plot



(b) Box plot with performance indicator



At the individual level, the proposed performance indicator $\mathbf{D}(\Phi_j^{\mathcal{M}})$ (Equation 15) compares the magnitude of feature contributions to decompose bias. A wide range of feature contributions indicates treatment inconsistencies. Therefore, a larger value of $\mathbf{D}(\Phi_j^{\mathcal{M}})$ can reveal bias. And we use the 5% rule that bias would not be decomposed in a feature j if $\mathbf{D}(\Phi_j^{\mathcal{M}}) \leq 5\%$. Note that due to scaling issues, the Beeswarm plot may display an exaggerated range of variation, while the box plot provides a clearer representation of the absolute variation, allowing for better interpretation.

Due to page limits, we mainly assess individual-level fairness for the Atlanta circuit, the largest circuit in Georgia. In Section 5.1, we identified the circuit Atlanta has bias decomposed in `Risk_Score` but not in `Race` and `Sex`. We further decompose bias at the individual level (by defining similar individuals) to identify key insights for local supervisors within this circuit.

Let us focus on the protected feature set of $S = \{\texttt{Sex}, \texttt{Race}, \texttt{Risk\_Score}\}$. Please note that `Age` and `Edu` could also be considered protected features. However, illustrating individual-level fairness

for these variables may require extensive discussion, as it would involve defining a significant number of clusters of similar individuals.

By defining clusters of similar individuals according to the data coding rules outlined in Section 5.1, we can generate two clusters for the Atlanta circuit (1) Cluster *A1*: low-risk Black males over 40 with high education (Figure 9), and (2) Cluster *A2*: high-risk Black males over 40 with high education (Figure 10).

Empirical analysis # 5: At the individual level, can bias be decomposed in `Sex` and `Race`?

In Figure 9a, similar individuals in *A1* exhibit tightly clustered points for `Sex` and `Race`, indicating treatment are not differentiated based on the two features, which is also justified with their performance indicators being small ($\mathbf{D}(\Phi_1^{\mathcal{M}}) = 0.030, \mathbf{D}(\Phi_2^{\mathcal{M}}) = 0.010$) in Figure 9b. Note, the features `Sex` and `Race` have identical categorical values; all scatter points are colored in grey. In Figure 10a, similar individuals in *A2* exhibit tightly clustered points of the same color for `Sex` and `Race` ($\mathbf{D}(\Phi_1^{\mathcal{M}}) = 0.040$, $\mathbf{D}(\Phi_2^{\mathcal{M}}) = 0.020$), suggesting similar treatment and individual fairness if protected features are `Sex` and `Race`, i.e., no bias decomposed in `Sex` and `Race`.

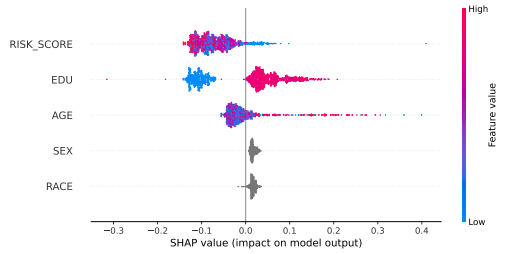
Empirical analysis # 6: At the individual level, can bias be decomposed in `Risk_Score`?

For similar individuals in the cluster *A1*, `Risk_Score` appears to be a biased feature, as indicated by the moderate indicator value of $\mathbf{D}(\Phi_3^{\mathcal{M}}) = 0.110 > 5\%$ (Figure 9b). For the second cluster *A2*, the performance indicator of $\mathbf{D}(\Phi_3^{\mathcal{M}}) = 2.158$ (Figure 10b) suggests a substantial presence of bias. In predicting binary categorical features, scores typically range from 0 to 1; thus, a performance indicator of 2.158 suggests significant variation supporting the conclusion of decomposed bias.

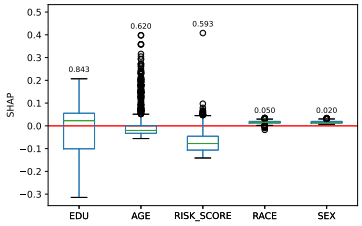
While `Risk_Score` shows a high bias indicator in the second cluster, the clustering of similarly color-coded points (Figure 10a) indicates that a more refined risk-based cluster could merit further investigation. Individuals with `Risk_Scores` of 5, 6, and 7 (blue dots in Figure 10b corresponding to feature `Risk_Score`) show negative contributions for `ESP Usage` compared to those with scores of 8 and above (purple and red dots in Figure 10b) show positive contribution. However, individuals with a `Risk_Score` of 8 and above still exhibit variation in their contribution to the predicted probability of `ESP Usage`, as the updated performance indicator shows a level of $\mathbf{D}(\Phi_5^{\mathcal{M}}) = 0.450$.

Figure 11 Gwinnett circuit: Two clusters varied on `Sex`

(a) RISK-L, R-B, S-M: Bee swarm plot

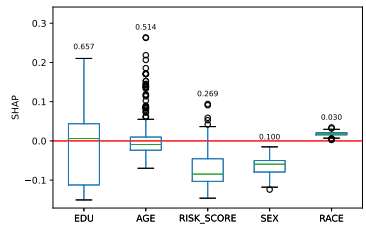


(b) RISK-L, R-B, S-M: Box plot



(c) RISK-L, R-B, S-F: Bee swarm plot

(d) RISK-L, R-B, S-F: Box plot



5.3. Dual Process: Potential hidden bias

By integrating results from both the group and individual levels, our dual approach could be applied to detect potential hidden bias. In this research, we define hidden bias as the one that cannot be directly observed through evaluations of fairness at either the group or individual level. Instead, it requires a synthesis of findings from both levels to uncover such hidden bias.

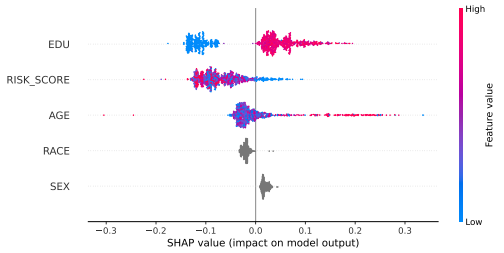
Empirical analysis # 7: Can hidden bias be decomposed in `Sex`?

In this section, we target the Gwinnett circuit, which, as noted in Section 5.1, does not exhibit bias decomposition in `Sex`, `Race`, or `Risk_Score` at the group level. Upon introducing two clusters for the Gwinnett circuit — Cluster *G1*: low-risk Black males (Figure 11a and 11b) and Cluster *G2*: low-risk Black females (Figure 11c and 11d) — we observe some intriguing patterns that could be related to decompose the hidden bias.

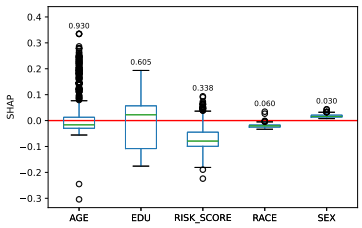
This evaluation could help us detect hidden bias by coupling the results we derived at the group level versus the results we derived at the individual level. In Section 5.1, we find for the Gwinnett circuit, no group-level bias is observed in `Sex` as shown in Figure 8f. At the individual level, following the process we used in analysis #5, we find that fairness would be attained when

Figure 12 Gwinnett circuit: Two clusters varied on `Race`

(a) RISK-L, R-W, S-M: Bee swarm plot



(b) RISK-L, R-W, S-M: Box plot



evaluating similar Black males, and also for the case involving similar Black females. Yet, we find positive contributions from `Sex` for Black males (box above the red zero-line in Figure 11b), but negative contributions from `Sex` for Black females (box below the red zero-line in Figure 11d). This can suggest a hidden bias in `Sex`, where despite both males and females receiving similar `ESP Usage` at the group level, the decision makers treat those two sex subgroups differently. Applying this approach to cluster *A1* and *A2*—as seen in Figures 9b and 10b—yields a similar conclusion for the Atlanta circuit: hidden bias can be decomposed in `Race`.

Empirical analysis # 8: Can hidden bias be decomposed in `Race`?

A similar conclusion from Analysis #8 applies here, suggesting hidden bias could be decomposed in `Race` when introducing another cluster *G3* for low-risk White males in Gwinnett (Figure 12).

In Section 5.1, we observe that no significant group-level bias is present in the Gwinnett circuit (Figure 7f). At the individual level, fairness is generally upheld when evaluating similar Black and White males. It is possible to conclude that bias is not decomposed in `Race` for individuals in Cluster *G3*, as indicated by a performance indicator of $D(\Phi_2^{\mathcal{M}}) = 0.060$ (Figure 12b) close to 5%. However, the SHAP importance box plot reveals negative contributions for R-B individuals (Figure 11b), while contributions for R-W are positive (Figure 12b).

6. Concluding Remarks

6.1. Key insights

By summarizing our empirical analysis from Section 5, we provide supervisors in decentralized circuits and central policymakers with insights to support fairness in human-computer interactions, offering significant managerial benefits.

Supervisors in each judicial circuit can efficiently navigate the information system to decompose bias in ESP usage decisions, fostering informed and fair decision-making. Additionally, our dual process results highlight areas for central policymakers to reevaluate or adjust their policies.

Key insight # 1: Does the policy apply uniformly across all circuits?

This insight arises from comparing decomposed bias across judicial circuits, revealing that some circuits tend to avoid `ESP Usage`, while others apply it universally. For instance, Figures 6b, 7b and 8b show that in the given circuit, local officers generally tend to avoid `ESP Usage`, leading to all subgroups receiving treatment below the state average. In contrast, for the circuit shown in Figure 6d, 7d and 8d, local officers tend to favor `ESP Usage`, resulting in all subgroups receiving treatment above the state average. This discrepancy suggests a potential avenue for policymakers to systematically evaluate whether a circuit favors `ESP Usage` due to workforce limitations, resulting in a tendency of utilizing `ESP Usage` overall.

Key insight # 2: Does the policy fairly calculate `Risk_Score`?

From Table 1, `Risk_Score` is the only calculated feature. Assigned by state central policy, this score is derived from an encapsulated information module that supports decision-making for each supervisee. A key concern is the criteria used to determine `Risk_Score`, as values of 7 and 8 can significantly affect the probability of `ESP Usage`, as illustrated in Figure 10 of empirical analysis #7. This raises questions about the rationale for not adjusting an individual's score from 7 to 8 to improve the prediction for the targeted outcome (or vice versa). For central authority policymakers, this issue requires careful evaluation before implementing a standardized policy.

Key insight # 3: Can hidden bias in `Sex` and `Race` structured in `Risk_Score`?

However, the lack of transparency in calculating `Risk_Score` raises concerns about potential biases related to `Race` and `Sex`. As noted in Section 5.3, we identified hidden biases as systematic deviations in contributions to `ESP Usage` across racial and gender groups. Despite this, both group-level and individual-level assessments indicate fairness. Comparing Figures 11b and 11d, we observe that, apart from `Risk_Score` and `Sex`, other features display similar SHAP value distributions. This suggests that `Risk_Score` may be constructed with hidden bias from other features, resulting in fair assessment at both the group and individual levels for the other features.

6.2. Major contributions

Our study makes several key contributions to the field of human-computer interaction with the added requirement of unraveling bias.

Based on a comprehensive literature review of human-computer interaction, fairness, and information representation, we pioneered the use of tree structures to represent fairness at both group and individual levels, addressing the complexities of intersectional fairness across various data granularities. This dual approach proves practical in decentralized settings, offering technical support for individual users navigating complex human-computer interactions and aiding central policymakers. An additional achievement is the use of an explainable AI to reveal performance indicators related to individual fairness behind algorithmic reasoning. This innovation enhances transparency and interpretability in high-stake decision-making. While SHAP values have been used to design fair machine learning algorithms by constraining the use of strong features in reducing group-level bias (Hickey et al. 2021, Grabowicz et al. 2022), our work introduces a novel approach that addresses the transparency of individual fairness, marks a significant step forward in unraveling bias in decentralized systems.

In conclusion, we extend the fairness context to scenarios involving multiple decentralized units, addressing challenges and establishing a foundational approach for future research. This scheme can be applied to chain businesses, such as hotel brands, to ensure consistent service levels across locations with independent teams, providing managerial benefits through human-AI interactions.

References

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access* 6:52138–52160.

Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. *Bioinformatics* 26(10):1340–1347.

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, et al. (2020) Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion* 58:82–115.

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. *Annual Review of Economics* 11(1):685–725.

Bauer K, Hinz O, van der Aalst W, Weinhardt C (2021) Expl(ai)n it to me–explainable ai and information systems research. *Business & Information Systems Engineering* 63:79–82.

Bauer K, von Zahn M, Hinz O (2023) Expl (ai) ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research* 34(4):1582–1602.

Bera P, Burton-Jones A, Wand Y (2014) Research note—how semantics and pragmatics interact in understanding conceptual models. *Information Systems Research* 25(2):401–419.

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021) Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50(1):3–44.

Bhatia S (2019) Predicting risk perception: New insights from data science. *Management Science* 65(8):3800–3823.

Bickel PJ, Hammel EA, O'Connell JW (1975) Sex bias in graduate admissions: Data from berkeley. *Science* 187(4175):398–404.

Binns R (2020) On the apparent conflict between individual and group fairness. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 514–524.

Borchardt B, Vogler H (2003) Determinization of finite state weighted tree automata. *Journal of Automata, Languages and Combinatorics* 8(3):417–463.

Bozapalidis S (1999) Equational elements in additive algebras. *Theory of Computing Systems* 32(1):1–33.

Brachman R, Levesque H (2004) *Knowledge representation and reasoning* (Morgan Kaufmann).

Breiman L (2001) Random forests. *Machine learning* 45:5–32.

Burton-Jones A, Grange C (2013) From use to effective use: A representation theory perspective. *Information Systems Research* 24(3):632–658.

CatBoost (2018) Catboost is a high-performance open source library for gradient boosting on decision trees. URL `https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus`.

Chen VX, Hooker J (2022) Combining leximax fairness and efficiency in a mathematical programming model. *European Journal of Operational Research* 299(1):235–248.

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

Christoph M (2020) *Interpretable machine learning: A guide for making black box models explainable* (Leanpub).

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Covaliu Z, Oliver RM (1995) Representation and solution of decision problems using sequential decision diagrams. *Management science* 41(12):1860–1881.

Department of Community Supervision (2020) Georgia dcs adopts a person-centered framework to aims to refine its organizational model. URL `https://dcs.georgia.gov/press-releases/2020-11-20/georgia-dcs-adopts-person-centered-framework-aims-refine-its`.

Dorogush AV, Ershov V, Gulin A (2018) Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* .

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Dwork C, Ilvento C (2018a) Fairness under composition. *arXiv preprint arXiv:1806.06122* .

Dwork C, Ilvento C (2018b) Individual fairness under composition. *Proceedings of Fairness, Accountability, Transparency in Machine Learning* .

Dwork C, Ilvento C, Jagadeesan M (2020) Individual fairness in pipelines. *arXiv:2004.05167* .

Fawcett T (2006) An introduction to roc analysis. *Pattern recognition letters* 27(8):861–874.

Frini A, Guitouni A, Benaskeur A (2017) Solving dynamic multi-criteria resource-target allocation problem under uncertainty: A comparison of decomposition and myopic approaches. *International Journal of Information Technology & Decision Making* 16(06):1465–1496.

Fu R, Huang Y, Singh PV (2020) Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications. *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, 39–63 (INFORMS).

Gohar U, Cheng L (2023) A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. *arXiv preprint arXiv:2305.06969* .

Grabowicz PA, Perello N, Mishra A (2022) Marrying fairness and explainability in supervised learning. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1905–1916.

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29:3315–3323.

Hickey JM, Di Stefano PG, Vasileiou V (2021) Fairness by explicability and adversarial shap learning. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*, 174–190 (Springer).

Hill C, Du L, Johnson M, McCullough B (2024) Comparing programming languages for data analytics: Accuracy of estimation in python and r. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* e1531.

Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349(6245):255–260.

Keller MT, Trotter WT (2017) *Applied combinatorics* (Open Textbook Library).

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* .

Kolaitis PG (2005) Schema mappings, data exchange, and metadata management. *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 61–75.

Kordzadeh N, Ghasemaghaei M (2021) Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems* 1–22.

Kusner M, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4069–4079.

Lu T, Zhang Y (2024) 1+ 1¿ 2? information, humans, and machines. *Information Systems Research* .

Lundberg SM, Lee SI (2017a) Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:1706.06060* .

Lundberg SM, Lee SI (2017b) A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

Lyons R, Peres Y (2016) *Probability on Trees and Networks*, volume 42 of *Cambridge Series in Statistical and Probabilistic Mathematics* (Cambridge University Press, New York), ISBN 978-1-107-16015-6, URL `http://dx.doi.org/10.1017/9781316672815`, available at `https://rdlyons.pages.iu.edu/`.

Maletti A (2005) Hasse diagrams for classes of deterministic bottom-up tree-to-tree-series transformations. *Theoretical computer science* 339(2-3):200–240.

Mandrekar JN (2010) Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* 5(9):1315–1316.

Mitchell TM, Mitchell TM (1997) *Machine learning*, volume 1 (McGraw-hill New York).

Nabi R, Shpitser I (2018) Fair inference on outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Nastjuk I, Trang S, Grummeck-Braamt JV, Adam MT, Tarafdar M (2024) Integrating and synthesising technostress research: a meta-analysis on technostress creators, outcomes, and is usage contexts. *European Journal of Information Systems* 33(3):361–382.

Ong Jr D, Jabbari Sabegh M (2019) A review of problems and challenges of using multiple conceptual models. *Proceedings of the 27th European Conference on Information Systems*, 1–18 (Association for Information Systems).

Orłowska E, Pawlak Z (1984) Expressive power of knowledge representation systems. *International Journal of Man-Machine Studies* 20(5):485–500.

Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31.

Quinlan JR (1986) Induction of decision trees. *Machine learning* 1:81–106.

Quinlan JR (1990) Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics* 20(2):339–346.

Ragu-Nathan T, Tarafdar M, Ragu-Nathan BS, Tu Q (2008) The consequences of technostress for end users in organizations: Conceptual development and empirical validation. *Information Systems Research* 19(4):417–433.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1(5):206–215.

Schmidt P, Biessmann F, Teubner T (2020) Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29(4):260–278.

Seidl H (1994) Finite tree automata with cost functions. *Theoretical Computer Science* 126(1):113–142.

Shapley LS, et al. (1953) A value for n-person games. *Contributions to the Theory of Games* 2:307–317.

Simpson EH (1951) The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 13(2):238–241.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* 8:1–21.

United States Courts (2003) Probation and pretrial services - mission. URL `https://www.uscourts.gov/services-forms/probation-and-pretrial-services/probation-and-pretrial-services-mission`.

Verma S, Rubin J (2018) Fairness definitions explained. *Proceedings of the international workshop on software fairness*, 1–7.

Wand Y, Weber R (2002) Research commentary: information systems and conceptual modeling—a research agenda. *Information Systems Research* 13(4):363–376.

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017) Fairness constraints: Mechanisms for fair classification. *Artificial intelligence and statistics*, 962–970 (PMLR).

ZipRecruiter (2024) Probation officer must-have resume skills and keywords. URL `https://www.ziprecruiter.com/career/Probation-Officer/Resume-Keywords-and-Skills`.