Argumentative Debates for Transparent Bias Detection

Hamed Ayoobi [1], Nico Potyka[2], Anna Rapberger[3,4], Francesca Toni[4]

[1]University of Groningen, Netherlands
[2]Cardiff University, United Kingdom
[3]Technical University Dortmund, Germany
[4]Imperial College London, United Kingdom
h.ayoobi@umcg.nl, potykan@cardiff.ac.uk, anna.rapberger@tu-dortmund.de, f.toni@imperial.ac.uk

Abstract

As the use of AI in society grows, addressing emerging biases is essential to prevent systematic discrimination. Several bias detection methods have been proposed, but, with few exceptions, these tend to ignore transparency. Instead, interpretability and explainability are core requirements for algorithmic fairness, even more so than for other algorithmic solutions, given the human-oriented nature of fairness. We present ABIDE (Argumentative BIas detection by DEbate), a novel framework that structures *bias detection* transparently as debate, guided by an underlying argument graph as understood in (formal and computational) *argumentation*. The arguments are about the success chances of groups in local *neighbourhoods* and the significance of these neighbourhoods. We evaluate ABIDE experimentally and demonstrate its strengths in performance against an argumentative baseline.

1 Introduction

As the use of AI in society grows, addressing its potential unfairness is becoming increasingly crucial. For example, an unfair AI model that supports decision-making in healthcare, discriminating against a certain sub-population, would be extremely harmful. In particular, it is essential to address any potential biases against specific groups/individuals in data and AI models trained thereupon (Mehrabi et al. 2022; Caton and Haas 2024). Various notions of fairness have been proposed in the literature (Mehrabi et al. 2022; Caton and Haas 2024; Waller et al. 2024). Amongst these notions, *statistical (or demographic) parity* (Corbett-Davies et al. 2017) defines fairness as an equal probability of being classified with the desirable (positive) label in different groups.

Existing solutions to identify unfairness tend to focus on optimising fairness and ignore the need for transparency for roots of (un)fairness (Caton and Haas 2024). However, transparency, afforded by interpretability and explainability, is crucial for the trustworthy detection and mitigation of bias. While considerable efforts have been made towards explainability of machine learning models, e.g. in (Ayoobi et al. 2023b; Dejl et al. 2025b), algorithmic fairness has received limited attention (exceptions include (Grabowicz, Perello, and Mishra 2022; Waller, Rodrigues, and Cocarascu 2024)).

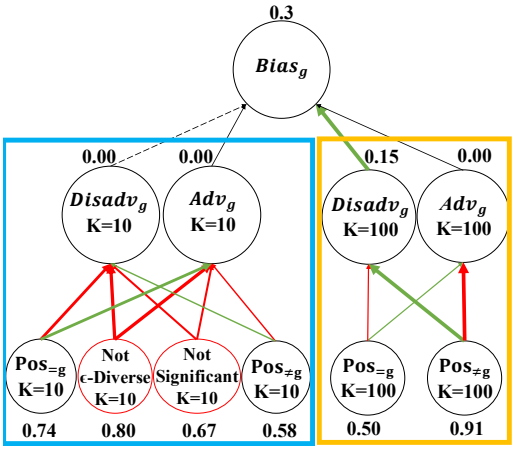

Figure 1: A QBAF generated by ABIDE for the COMPAS dataset (with protected feature $X_p = race$ and protected group $g = African\text{-}American$) for $K = 10, 100$ neighbourhoods (critical question arguments with zero strength omitted). Dashed/solid edges are supports/atttacks. Green/red edges are supports/attacks (from arguments) with nonzero strength, and black edges are (from arguments) with zero strength. Edge width reflects the strength of arguments. (Strengths below/above nodes. Details in Section 6.)

We propose ABIDE (Argumentative Bias Detection), a novel transparent method to detect bias through argumentative *debates* about the presence of bias against individuals, based on values of *protected features* for the individuals and others in their *neighbourhoods*. Debates recently emerged as a powerful mechanism to address various challenges in contemporary AI, such as AI safety (Brown-Cohen, Irving, and Piliouras 2024), and to improve LLM performance (Khan et al. 2024). In much of this literature, debates are unstructured and inform other entities that decide the debates' outcome. Thus, while these debates provide a justification for the outcomes, they are not faithfully explained by the debates.

ABIDE is designed to guide debates about the presence of bias. To structure debates, we rely upon *argument schemes* (with critical questions) (Walton, Reed, and Macagno 2008; Macagno, Walton, and Reed 2017), from formal argumentation. To decide the debates' outcomes, we compute argu-

ments' strength with *gradual semantics* (Baroni, Rago, and Toni 2018) from computational argumentation. In summary:

1. We propose a neighbourhood-based notion of fairness for individual bias detection, adapting statistical parity.
2. We define novel argument schemes about the presence of bias based on (properties of) neighbourhoods.
3. We develop a mapping of these schemes into Quantitative Bipolar Argumentation Frameworks (QBAFs) (Baroni, Rago, and Toni 2018), enabling structured and modular reasoning for transparent bias detection (see Figure 1).
4. We formally connect properties of gradual semantics for QBAFs to desirable properties of bias detection.
5. We conduct empirical evaluations with synthetically biased models, models trained on real-world datasets, and ChatGPT-4o, consistently outperforming an argumentative baseline (Waller, Rodrigues, and Cocarascu 2024).
6. We show how our approach can empower debate-based bias detection in human-agent and multi-agent scenarios.

The proofs of all results and additional material can be found in the accompanying (Ayoobi et al. 2025a). The code is available at https://github.com/hamed-ayoobi/ABIDE.

2 Related Work

Fairness in Machine Learning. Various methods for identifying, measuring and mitigating bias have been proposed (Mehrabi et al. 2022; Caton and Haas 2024), but they often neglect transparency (Waller et al. 2024). One exception is (Grabowicz, Perello, and Mishra 2022), which uses explanations to unearth and rectify bias, but relies on feature importance, rather than debates, for transparency.

Bias can be defined in different ways, e.g. by comparing the success probability of the protected group to the one of unprotected individuals or the general population (*statistical parity*), or by comparing the risk of misclassification of the protected to the one of the non-protected group (*predictive equality*) (Corbett-Davies et al. 2017). ABIDE is based on a local notion of statistical parity. That is, we identify neighbourhoods where groups are locally advantaged/disadvantaged with respect to success probability. The decision whether a bias exists is then based on the evidence provided by these neighbourhoods in a transparent way.

Debate and AI. Some approaches rely upon debate protocols between (two) AI models trained as players in zero-sum games, e.g. towards AI safety (Irving, Christiano, and Amodei 2018; Brown-Cohen, Irving, and Piliouras 2024) or for post-hoc explaining image classifiers (Kori, Glocker, and Toni 2024). (Thauvin et al. 2024) propose a transparent model based on debate protocols between agents drawing arguments in the form of visual features. (Khan et al. 2024) use debate to resolve disagreement between (two) strong LLMs to inform and improve decisions by weaker models (LLMs or humans). In all these settings, debates are between two pre-determined entities (e.g. models or agents) and they inform decisions made by other entities (e.g. a verifier or a weaker model). Instead, in our approach, debates may be in the "mind" of a single entity or across two or more entities, as our approach focuses on providing guidance for structuring and formally evaluating (with gradual semantics) debates.

Argumentation and Bias. (Walton, Reed, and Macagno 2008; Macagno, Walton, and Reed 2017) propose an argument scheme to capture *arguments from bias*[1], allowing to draw the conclusion that an arguer is unlikely to have taken both sides of an issue into account if the arguer is biased, subject to addressing critical questions regarding (i) evidence for the arguer's bias and (ii) the need to take multiple sides for that issue. Our new argument scheme below is orthogonal to this and can be used in combination with it, by addressing the first critical question using properties of neighbourhoods to collect evidence for the bias of the arguer (a classifier).

(Waller, Rodrigues, and Cocarascu 2024) were the first to use gradual semantics for QBAFs to characterise bias. While we were inspired by them, our approach differs in a simpler QBAF structure (with fewer nodes and relations), the use of neighbourhoods and their properties (for more robust reasoning on bias), and the empowering of formal links between properties of semantics and of bias detection.

Argumentation and Debate. Our approach follows the general idea of using argumentation towards explainability, e.g. as in (Ayoobi et al. 2019, 2021, 2022, 2023a, 2025b; Dejl et al. 2025a). Several have pointed to the connection between argumentation and debate, e.g. (Panisson, McBurney, and Bordini 2021; de Tarlé, Bonzon, and Maudet 2022) in general, and (Rago et al. 2025) for product recommendation. While we can leverage on these works towards debate generation, our approach is orthogonal to them.

3 Preliminaries

Classification Problems. We consider variables (features) $X_1, \ldots, X_k$ with associated domains $D_1, \ldots, D_k$ and a class variable Y with associated domain $\mathcal{C}$ of class labels. We use X_p to denote a *protected feature* of interest and g to denote its value for the protected group. Thus, $X_p = g$ identifies the protected group and $X_p \neq g$ the remaining individuals.

For $\mathcal{D} = \bigtimes_{i=1}^{k} D_i$, vectors $\mathbf{x} \in \mathcal{D}$ are the *inputs* of the classification problem. We consider *(binary) classifiers* $\mathrm{c} : \mathcal{D} \to \mathcal{C} = \{0, 1\}$, and regard 1 as a desirable (e.g., acceptance of a loan application) and 0 as a negative outcome.

Notation 1. Capital letters refer to variables, lowercase letters to values of variables. Bold capital letters refer to sequences of variables, bold lowercase letters to variable assignments. For example, if $\mathbf{V} = (X_1, X_2, X_3)$, then $\mathbf{v}$ refers to an assignment (x_1, x_2, x_3) with $x_i \in D_i$ for $i \in \{1, 2, 3\}$. For each $\mathbf{x} \in \mathcal{D}$ and sequence of variables $\mathbf{V} = (X_{i_1}, \ldots, X_{i_m})$, we let $\mathbf{x}|_{\mathbf{V}}$ denote the projection of $\mathbf{x}$ onto $\mathbf{V}$. For example, if $\mathbf{x} = (a, b, c, d)$ and $\mathbf{V} = (X_2, X_4)$, then $\mathbf{x}|_{\mathbf{V}} = (b, d)$. Finally, we let $\mathbf{X}$ denote the sequence $(X_1, \ldots, X_k)$ of all variables.

QBAFs. Quantitative bipolar argumentation frameworks (QBAFs) (Baroni, Rago, and Toni 2018) are quadruples $\mathcal{Q} = (\mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau)$ consisting of a finite set of *arguments* $\mathcal{A}$, disjoint binary relations of *attack* $\mathcal{R}^- \subseteq \mathcal{A} \times \mathcal{A}$ and *support* $\mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$ and a *base score function* $\tau : \mathcal{A} \to [0, 1]$. QBAFs

[1] www.rationaleonline.com/map/9gy9gd/argument-from-bias

can be seen as graphs with arguments as nodes and elements of the attack and support relations as edges. For $a \in \mathcal{A}$, $a^{att} = \{b \in A \mid (b,a) \in \mathcal{R}^-\}$ and $a^{sup} = \{b \in A \mid (b,a) \in \mathcal{R}^+\}$ denote the set of all *attackers* and *supporters*, resp., of a.

To assess the acceptability of arguments in a QBAF $\mathcal{Q}$, (gradual) semantics can be given by means of a *strength function* $\sigma_{\mathcal{Q}} : \mathcal{A} \to [0,1]$, often defined by an iterative procedure that initializes strength values with the base scores and then repeatedly updates the strength of arguments based on the strength of their attackers and supporters. For acyclic QBAFs, this procedure always converges and is equivalent to a linear-time algorithm that first computes a topological ordering of the arguments and then updates each argument only once following the order (Potyka 2019, Proposition 3.1).

A QBAF semantics is called *modular* if the update function can be decomposed into an *aggregation function* $\mathrm{agg}(A,S)$ that aggregates the strength values A of attackers and S of supporters, and an *influence function* $\mathrm{infl}(b,a)$ that adapts the base score b based on the aggregate a (Mossakowski and Neuhaus 2018). Most QBAF semantics are modular. Examples include the DF-QuAD (Rago et al. 2016), Euler-based (Amgoud and Ben-Naim 2017) and quadratic energy (Potyka 2018) semantics. Aggregation functions include *product* $\mathrm{agg}(A,S) = \prod_{a \in A}(1-a) - \prod_{s \in S}(1-s)$ used in DF-QuAD and *sum* $\mathrm{agg}(A,S) = \sum_{s \in S} s - \sum_{a \in A} a$ used in Euler-based and quadratic energy semantics. The influence functions of DF-QuAD and quadratic energy both take the form $\mathrm{infl}(b,a) = b - b \cdot f(-s) + (1-b) \cdot f(s)$, where $f(x) = \max\{0,x\}$ for DF-QuAD and $f(x) = \frac{\max\{0,x\}}{1+\max\{0,x\}}$ for quadratic energy.

QBAF Properties. QBAFs are often compared on properties (Leite and Martins 2011; Amgoud and Ben-Naim 2016; Amgoud and Ben-Naim 2017; Baroni, Rago, and Toni 2018), many of which are tied to properties of aggregation and influence functions (Mossakowski and Neuhaus 2018). We will use properties of *(strict) monotonicity* and *balance* of these functions from (Potyka and Booth 2024b) and recap them in (Ayoobi et al. 2025a). As shown in (Potyka and Booth 2024a), product and sum aggregation satisfy balance and monotonicity, sum satisfies strict monotonicity; the influence functions of DF-QuAD and quadratic energy satisfy all properties.

4 Neighbourhoods and Their Properties

In this section, we will formalize the idea of *local bias*. Intuitively, a local bias against a reference individual exists if other similar individuals are advantaged by the classifier. Formally, these similar individuals can be represented by a *neighbourhood* surrounding the reference individual.

Definition 1. A *neighbourhood* is a set $\mathcal{N} \subseteq \mathcal{D}$. A *neighbourhood of a point* $\mathbf{x} \in \mathcal{D}$ is a set $\mathcal{N}$ such that $\mathbf{x} \in \mathcal{N}$.

If we were allowed to choose neighbourhoods arbitrarily, we may always find one that indicates presence and one that indicates absence of bias. We thus present some desirable properties of neighbourhoods. To begin with, a neighbourhood that contains only few individuals may not be reliable as they may be exceptions. The larger the neighbourhood, the larger our confidence that it is representative of the domain.

Definition 2. Let $N \in \mathbb{N}$. A neighbourhood $\mathcal{N} \subseteq \mathcal{D}$ is called *N-significant* if $|\mathcal{N}| \geq N$.

Without further restrictions, an adversarial agent could define a neighbourhood by picking individuals systematically to demonstrate the (non-)existence of a bias. To prevent this, one reasonable assumption is that when we pick two individuals from a sample, we cannot leave out individuals between them. Mathematically, *betweenness* can be described by *convexity*.

Definition 3. Let $S \subseteq \mathcal{D}$ be a sample. A neighbourhood $\mathcal{N} \subseteq S$ is called *S-objective* if $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{N}$ and there is an $\mathbf{x}_3 \in S$ that is a convex combination of $\mathbf{x}_1$, $\mathbf{x}_2$, then $\mathbf{x}_3 \in \mathcal{N}$.

Formally, an input $\mathbf{x} \in \mathbb{R}^n$ is a convex combination of $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ if $\mathbf{x} = \lambda \cdot \mathbf{x}_1 + (1-\lambda) \cdot \mathbf{x}_2$ for some $\lambda \in [0,1]$. Discrete ordinal features such as Boolean ($\{0,1\}$) or qualitative descriptions like small (0), medium (1), large (2) can be mapped to integers to match the definition. For nominal features such as color or marital status, we demand that they are equal in $\mathbf{x}_1$ and $\mathbf{x}_2$ (and thus in $\mathbf{x}$).

The next proposition explains that S-objectivity is always satisfied when we use an *ϵ-neighbourhood* $\mathcal{N}_{\mathbf{x}} = \{\mathbf{x}' \in S \mid \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon\}$ of a point $\mathbf{x}$. It can be defined w.r.t. any common distance measuresuch as Euclidean, Manhattan and the Hamming distance and their weighted variants.

Proposition 1. *If $\mathcal{N}_{\mathbf{x}}$ is the ϵ-neighbourhood of a point $\mathbf{x} \in S$ with respect to the distance induced by a seminorm[2] $\|.\|$, then $\mathcal{N}$ is S-objective.*

For the remaining properties, we introduce some notation.

Definition 4. Given a finite neighbourhood $\mathcal{N}$ and a partial feature assignment $\mathbf{v}$ to a sequence of variables $\mathbf{V}$, the *local probability of $\mathbf{v}$ in $\mathcal{N}$* is $P_{\mathcal{N}}(\mathbf{v}) = \frac{|\{\mathbf{x} \in \mathcal{N} \mid (\mathbf{x}|\mathbf{v}) = \mathbf{v}\}|}{|\mathcal{N}|}$. The *local success probability* for $\mathbf{v}$ is defined as $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid \mathbf{v}) = \frac{|\{\mathbf{x} \in \mathcal{N} \mid \mathbf{x}|_{\mathbf{V}} = \mathbf{v}, c(\mathbf{x}) = 1\}|}{|\{\mathbf{x} \in \mathcal{N} \mid \mathbf{x}|_{\mathbf{V}} = \mathbf{v}\}|}$.

Ideally, our neighbourhoods contain good representations of both the protected and non-protected groups. For example, even for a strongly biased model, we may be able to find one protected individual that is treated similar to non-protected ones even though the model strongly discriminates against the non-protected individuals in most cases. Similarly, if it contained only one non-protected individual, this may simply be an outlier. Since protected groups are often underrepresented, we should not expect balanced neighbourhoods. However, if almost all elements in a neighbourhood belong to one group, the neighbourhood may not provide very strong evidence. To quantify neighbourhood diversity, we use entropy.

Definition 5. The *entropy* $H_{\mathcal{N}}(X_p)$ of the protected feature X_p in $\mathcal{N}$ is $H_{\mathcal{N}}(X_p) = -P_{\mathcal{N}}(X_p = g) \cdot \log P_{\mathcal{N}}(X_p = g) - P_{\mathcal{N}}(X_p \neq g) \cdot \log P_{\mathcal{N}}(X_p \neq g)$.

The entropy takes its maximum 1 when $\mathcal{N}$ is maximally diverse ($P_{\mathcal{N}}(X_p = g) = P_{\mathcal{N}}(X_p \neq g) = 0.5$), its minimum 0 when it is minimally so ($P_{\mathcal{N}}(X_p = g) = 1$ or $P_{\mathcal{N}}(X_p \neq g) = 1$), and is monotonically decreasing in between.

Definition 6. Let $\epsilon \in [0,1]$. A neighbourhood $\mathcal{N} \subseteq \mathcal{D}$ is called *ϵ-diverse* if $H_{\mathcal{N}}(g) = \epsilon$.

[2] A seminorm is a function $\|.\| : \mathcal{D} \to \mathcal{D}$ that satisfies subadditivity (triangle inequality) and absolute homogeneity.

Major premise	Generally, if, in $\mathcal{N}$, $X_p \neq g$ leads to a positive decision and $X_p = g$ leads to a negative decision, then $X_p = g$ is disadvantaged in $\mathcal{N}$ and $X_p \neq g$ is advantaged in $\mathcal{N}$
Minor premise	In the case of the chosen $\mathcal{N}$, $X_p = g$ leads to a negative decision.
Minor premise	In the case of a chosen $\mathcal{N}$, $X_p \neq g$ leads to a positive decision.
Conclusion	Thus, $X_p = g$ is disadvantaged in $\mathcal{N}$ and $X_p \neq g$ is advantaged in $\mathcal{N}$.

Figure 2: Argument scheme for neighbourhood $\mathcal{N}$. $X_p = g$ identifies the protected, potentially disadvantaged group.

Major premise	If $X_p{=}g$ is disadvantaged in $\mathcal{N}_1, \ldots, \mathcal{N}_m$, then there is a bias against $X_p = g$.
Minor premise	In the case of chosen $\mathcal{N}_1, \ldots, \mathcal{N}_m$, $X_p{=}g$ is disadvantaged in $\mathcal{N}_1, \ldots, \mathcal{N}_m$.
Conclusion	Thus, there is a bias against $X_p = g$.

Figure 3: Argument scheme for combining multiple neighbourhoods $\mathcal{N}_1, \ldots, \mathcal{N}_m$.

Finally, we define that a neighbourhood is biased if the local conditional probability of the positive outcome is significantly larger for the non-protected than the protected group.

Definition 7. A neighbourhood $\mathcal{N} \subseteq \mathcal{D}$ is called *ϵ-biased against* $X_p = g$ if $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g) - P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g) = \epsilon$ and $\epsilon > 0$.

5 Argument Schemes for Detecting Local Bias

ABIDE makes use of properties of neighbourhoods to drive argumentative debates about the presence of bias against individuals with $X_p = g$. To shape these debates, we define novel argument schemes with accompanying critical questions. In the next section, we will turn both into QBAFs.

Figure 2 gives our argument scheme for detecting local bias against the protected feature taking the value of interest ($X_p = g$). The critical questions address the overall quality (properties) of the neighbourhood: ***CQ1. Is $\mathcal{N}$ N-significant? CQ2. Is $\mathcal{N}$ S-objective? CQ3. Is $\mathcal{N}$ ϵ-diverse?*** When the critical questions are instantiated they give attacks against the conclusion of the (instantiated) argument scheme.

We can also combine arguments drawn from different neighbourhoods, via the argument scheme in Figure 3. For brevity, we omit to consider critical questions for this scheme.

6 Arguing About Bias with ABIDE

ABIDE empowers debates on bias, in three steps.

1. Build a *local bias-QBAF* $Q_{\mathcal{N}}$ (based on Figure 2).
2. Add critical questions (CQ1-CQ3).
3. Build a *global bias-QBAF* $\mathcal{Q}_{\mathbf{g}}$ (based on Figure 3).

We assume a fixed modular gradual semantics σ with aggregation function agg and influence function infl and omit the QBAF subscript when applying σ.

Arguing about Bias in a single Neighbourhood. The *conclusion* of the argument scheme in Figure 2 amounts to two arguments, Disadv_g/Adv_g, that express that individuals with $X_p = g$ are, resp., disadvantaged/advantaged. Their base score is set to 0. The *minor premises* give rise to two further arguments $\mathrm{Pos}_{=g}$ and $\mathrm{Pos}_{\neq g}$. Their base score corresponds to the group's local success probability in the neighbourhood. The *major premise* informs about the connection between these arguments: $\mathrm{Pos}_{=g}$ attacks/supports Disadv_g/Adv_g, while $\mathrm{Pos}_{\neq g}$ supports/attacks Disadv_g/Adv_g.

Definition 8. The *local bias-QBAF for g w.r.t. neighbourhood $\mathcal{N}$* is the QBAF $Q_{\mathcal{N}} = (\mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau)$ with
$\mathcal{A} = \{\mathrm{Disadv}_g, \mathrm{Adv}_g, \mathrm{Pos}_{=g}, \mathrm{Pos}_{\neq g}\}$,
$\mathcal{R}^- = \{(\mathrm{Pos}_{=g}, \mathrm{Disadv}_g), (\mathrm{Pos}_{\neq g}, \mathrm{Adv}_g)\}$,
$\mathcal{R}^+ = \{(\mathrm{Pos}_{=g}, \mathrm{Adv}_g), (\mathrm{Pos}_{\neq g}, \mathrm{Disadv}_g)\}$, and

$$\tau(a) = \begin{cases} P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g), & a = \mathrm{Pos}_{\neq g} \\ P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g), & a = \mathrm{Pos}_{=g}. \\ 0, & else. \end{cases}$$

Figure 1 shows this QBAF (yellow, right box). Intuitively, the base score of Disadv_g/Adv_g is 0 as our default assumption is that individuals with $X_p{=}g$ are not disadvantaged/advantaged. The base scores of $\mathrm{Pos}_{=g}$/$\mathrm{Pos}_{\neq g}$ are the local success probabilities of the groups $X_p{=}g$/ $X_p{\neq}g$.

We show that our choice of QBAF is sensible when choosing the semantics appropriately. If $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g)$ and $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g)$ are equal, then there is neither reason to accept that individuals with $X_p = g$ are disadvantaged nor that they are advantaged. However, as $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g)$ becomes larger than $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g)$, we should gradually accept Disadv_g. Conversely, as $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g)$ becomes smaller than $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g)$, we should gradually accept Adv_g. We can guarantee this behaviour by choosing aggregation and influence functions with certain properties.

Proposition 2. *If* agg *and* infl *satisfy*

1. *monotonicity, then* $\sigma(\mathrm{Adv}_g){=}0$ *or* $\sigma(\mathrm{Disadv}_g){=}0$,
2. *balance, then if* $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g) = P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g)$, *then* $\sigma(\mathrm{Adv}_g) = \sigma(\mathrm{Disadv}_g) = 0$,
3. *strict monotonicity, then if* $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g) > P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g)$, *then* $\sigma(\mathrm{Disadv}_g) > 0$,
4. *strict monotonicity, then if* $P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g) < P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g)$, *then* $\sigma(\mathrm{Adv}_g) > 0$.

Item 1 states that we always fully reject at least one of Adv_g and Disadv_g. We fully reject both if both groups have the same local success probability (item 2). Items 3 and 4 explain that if the protected/non-protected group is at a disadvantage, the strength of Disadv_g/Adv_g will be above 0.

As the inequality between the protected and non-protected groups increases, the strength of Disadv_g/Adv_g should change accordingly. To study this situation, we compare the strength values in two independent local bias-QBAFs, connecting again properties of aggregation and influence functions with the expected behaviour:

Proposition 3. *Let $\mathcal{Q}$ be a QBAF composed of two independent local bias-QBAFs $Q_{\mathcal{N}^1}, Q_{\mathcal{N}^2}$ If* agg *and* infl *satisfy*

1. *balance, then, if $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p \neq g) = P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p \neq g)$ and $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p = g) = P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p = g)$, we have $\sigma(\mathrm{Adv}_g^1) = \sigma(\mathrm{Adv}_g^2)$ and $\sigma(\mathrm{Disadv}_g^1) = \sigma(\mathrm{Disadv}_g^2)$;*
2. *monotonicity, then if $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p \neq g) \geq P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p \neq g)$ and $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p = g) \leq P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p = g)$, we have $\sigma(\mathrm{Adv}_g^1) \leq \sigma(\mathrm{Adv}_g^2)$ and $\sigma(\mathrm{Disadv}_g^1) \geq \sigma(\mathrm{Disadv}_g^2)$;*
3. *strict monotonicity, then the guarantees of item 2 hold, and, if additionally $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p \neq g) > P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p \neq g)$ or $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p = g) < P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p = g)$, then $\sigma(\mathrm{Adv}_g^1) < \sigma(\mathrm{Adv}_g^2)$ and $\sigma(\mathrm{Disadv}_g^1) > \sigma(\mathrm{Disadv}_g^2)$;*
4. *monotonicity, then if $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p \neq g) \leq P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p \neq g)$ and $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p = g) \geq P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p = g)$, we have $\sigma(\mathrm{Adv}_g^1) \geq \sigma(\mathrm{Adv}_g^2)$ and $\sigma(\mathrm{Disadv}_g^1) \leq \sigma(\mathrm{Disadv}_g^2)$;*
5. *strict monotonicity, then the guarantees of item 4 hold, and, if additionally $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p \neq g) < P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p \neq g)$ or $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p = g) > P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p = g)$, then $\sigma(\mathrm{Adv}_g^1) > \sigma(\mathrm{Adv}_g^2)$ and $\sigma(\mathrm{Disadv}_g^1) < \sigma(\mathrm{Disadv}_g^2)$.*

Item 1 states that if we find the same inequality in two neighbourhoods, we will also have the same strengths. Items 2, 3 deal with the case where the protected group is disadvantaged. Item 2 is a generalization of item 1 and guarantees that if the non-protected/protected group in $\mathcal{N}_1$ is at least/most as successful as the non-protected/protected group in $\mathcal{N}_2$, then the strength of Adv_g/ Disadv_g in $\mathcal{N}_1$ will be at least/most as large as in $\mathcal{N}_2$. Item 3 guarantees that there will be a strict difference in the strength values if there is a strict difference between the probabilities. Items 4, 5 give symmetrical guarantees when the protected group is advantaged.

By design, we treat evidence for being advantaged and disadvantaged equally. That is, in two neighbourhoods with symmetric evidence, the strength of Adv_g in one will be the strength of Disadv_g in the other:

Proposition 4. *For $Q_{\mathcal{N}^1}, Q_{\mathcal{N}^2}$ as in Prop. 3, if $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p \neq g) = P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p = g)$ and $P_{\mathcal{N}^1}(c(\mathbf{X}) = 1 \mid X_p = g) = P_{\mathcal{N}^2}(c(\mathbf{X}) = 1 \mid X_p \neq g)$, then $\sigma(\mathrm{Adv}_g^1) = \sigma(\mathrm{Disadv}_g^2)$ and $\sigma(\mathrm{Adv}_g^1) = \sigma(\mathrm{Disadv}_g^2)$.*

Finally, under DF-QuAD, the strength of the Adv_g and Disadv_g arguments has a particularly natural meaning:

Proposition 5. *For σ the DF-QuAD semantics, we have*

1. $\sigma(\mathrm{Adv}_g) = \max\{0, P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g) - P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g)\}$,
2. $\sigma(\mathrm{Disadv}_g) = \max\{0, P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p \neq g) - P_{\mathcal{N}}(c(\mathbf{X}) = 1 \mid X_p = g)\}$,
3. $\sigma(\mathrm{Disadv}_g) = \epsilon$ *iff $\mathcal{N}$ is ϵ-biased against $X_p = g$.*

Questioning Bias in a single neighbourhood. The significance of the evidence given by a neighbourhood can be debatable. To take account of this, we introduce additional arguments corresponding to the *critical questions*: s(ignificance) for CQ1, o(bjectivity) for CQ2 and d(iversity) for CQ3. By $CQ = \{s, o, d\}$, we denote the set of our novel arguments.

Intuitively, if the answer to these questions is negative, i.e., there is legitimate doubt regarding the significance, diversity, or objectivity of $\mathcal{N}$, then our confidence in detecting bias within $\mathcal{N}$ should decrease. Instead, if the answer is positive, then we can trust the bias detection outcome in $\mathcal{N}$. We thus treat the new arguments as attackers of Adv_g and Disadv_g.

The new arguments are unattacked and unsupported by other arguments, and thus their strength (for any σ) will be their base score. So, for s, the base score should be monotonically decreasing w.r.t. the size of $\mathcal{N}$ and eventually reach 0, so that small neighbourhoods provide strong attacks. For o, since objectivity is a binary criterion, an indicator function is most natural. For d, the attack strength should be maximal when the population consists of only one group, and approach 0 as the distribution becomes uniform: this behaviour can be captured by the entropy, as given in Section 4. However, since we cannot expect a uniform distribution for underrepresented groups, we should be able to rescale the entropy: our base score function for d should thus be monotonically decreasing. Formally, critical questions lead to the QBAF below:

Definition 9. Let $Q_{\mathcal{N}} = (\mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau)$ be a local bias-QBAF. The corresponding *local bias-QBAF with critical questions* is $\mathcal{Q}^c_{\mathcal{N}} = (\mathcal{A}', \mathcal{R}^{-\prime}, \mathcal{R}^+, \tau')$ with $\mathcal{A}' = \mathcal{A} \cup CQ$, $\mathcal{R}^{-\prime} = \mathcal{R}^- \cup \{(a, \mathrm{Adv}_g), (a, \mathrm{Disadv}_g) \mid a \in CQ\}$ and τ' is such that $\tau'(a) = \tau(a)$ for all $a \notin CQ$, and

- $\tau'(s) = f^s(|\mathcal{N}|)$ for a monotonically decreasing function $f^s : \mathbb{N} \to [0, 1]$ such that $f^s(1) = 1$, and $\lim_{x \to \alpha} f^s(x) = 0$ for some *significance threshold* $\alpha \in \mathbb{N} \cup \{\infty\}$;
- $\tau'(o) = f^o(\mathcal{N})$ where $f^o(\mathcal{N}) = 0$ if $\mathcal{N}$ is S-objective, and $f^o(\mathcal{N}) = \gamma$ else, where $\gamma \in (0, 1]$ is an *objectivity weight*;
- $\tau'(d) = f^d(H_{\mathcal{N}}(g))$ where for $f^d : [0, 1] \to [0, 1]$ is monotonically decreasing and satisfies $f^d(0) = 1$ and $f^d(\beta) = 0$ for some *entropy threshold* $\beta \in (0, 1]$.

We illustrate this QBAF in Figure 1 (in the blue, left box). In the experiments (Section 7), we use $f^s(x) = \frac{\max\{\alpha - x, 0\}}{\alpha}$ ($\alpha < \infty$) and $f^d(x) = 1 - \max\{\frac{x}{\beta}, 1\}$.

Similarly to Section 6, properties of σ guarantee that the new arguments behave as intended. We do not go into formal detail but observe that, if agg and infl satisfy balance, then $\mathrm{Adv}_g/\mathrm{Disadv}_g$ will remain unaffected by critical questions that do not apply (base score 0), and, if they satisfy monotonicity, then the strength of $\mathrm{Adv}_g/\mathrm{Disadv}_g$ will decrease monotonically w.r.t. the strength (i.e. base score) of s, o, d.

Arguing about Bias across neighbourhoods. Finally, we combine multiple local bias-QBAFs for different neighbourhoods $\mathcal{N}_1, \ldots, \mathcal{N}_m$, as per Figure 3. Their individual $\mathrm{Adv}_g/\mathrm{Disadv}_g$ arguments attack/support a *global bias* argument bias_g. We make the following assumptions: we have no prior information about biases of the classifier and, in the absence of information, we take that the classifier is unbiased. We capture this by setting the base score of bias_g to 0.

Definition 10. Let $Q^c_{\mathcal{N}_i} = (\mathcal{A}_i, \mathcal{R}_i{}^-, \mathcal{R}_i{}^+, \tau_i)$ be the local bias-QBAF with critical questions for g w.r.t. $\mathcal{N}_i \in$

$\{\mathcal{N}_1, \ldots, \mathcal{N}_m\}$. Then, the *global bias-QBAF for* g is the QBAF $\mathcal{Q}_{\mathbf{g}} = (\mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau)$ with

$\mathcal{A} = \{\text{bias}_g\} \cup \bigcup_{1 \leq i \leq m} \mathcal{A}_i$;

$\mathcal{R}^- = \{(\text{Adv}_g^i, \text{bias}_g) \mid 1 \leq i \leq m\} \cup \bigcup_{i=1}^m \mathcal{R}_i^-$;

$\mathcal{R}^+ = \{(\text{Disadv}_g^i, \text{bias}_g) \mid 1 \leq i \leq m\} \cup \bigcup_{i=1}^m \mathcal{R}_i^+$;

$\tau(a) = \tau_i(a)$ if $a \in \mathcal{A}_i$ and 0 if $a = \text{bias}_g$.

Figure 1 illustrates a global bias-QBAF, for $m = 2$.

Again, properties of σ can guarantee that local bias-QBAFs affect bias_g as expected and it will remain unaffected by fully rejected local bias arguments (strength 0) and its strength will increase/decrease monotonically w.r.t. local Disadv_g/Adv_g.

Proposition 6. *For* $\mathcal{Q}_{\mathbf{g}} = (\mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau)$ *a global bias-QBAF, let* S_d=$\{\text{Disadv}_g^i \mid i \in \{1, \ldots, m\}, \sigma(\text{Disadv}_g^i) > 0\}$ *and* $S_a = \{\text{Adv}_g^i \mid i \in \{1, \ldots, m\}, \sigma(\text{Adv}_g^i) > 0\}$. *If* agg *and* infl *satisfy balance and monotonicity, then 1.* $S_d \succeq S_a$ *implies* $\sigma(\text{bias}_g) \geq 0$*, and 2.* $S_a \succeq S_d$ *implies* $\sigma(\text{bias}_g) = 0$.

Note that item 1 of Proposition 2 implies that $S_d \cap S_a = \emptyset$.

7 Experiments

We conduct experiments with: 1. synthetically biased classifiers (two hand-crafted and one resulting from training on the Adult Census Income (ACI) dataset (Becker and Kohavi 1996), also used for testing); 2. trained classifiers (logistic regression, but our method applies to any classifier); 3. LLMs (ChatGPT-4o). For training (2) and for testing (2,3) we use the Bank Marketing (Moro, Rita, and Cortez 2014) and COMPAS (ProPublica 2016; Angwin et al. 2016) datasets. The datasets we choose are commonly used in the fairness literature (Räz 2024; Fawkes et al. 2024; Ehyaei, Farnadi, and Samadi 2024). We consider also LLMs as they are prone to societal biases due to their exposure to human-generated data (Ma et al. 2025; Xiang 2024). In all experiments, we compare ABIDE with the argumentative approach by (Waller, Rodrigues, and Cocarascu 2024) (*IRB* in short). We report results with quadratic energy (DF-QuAD gave similar results). We report results for 1000 randomly sampled individuals and their KNN-neighbourhoods (Cover and Hart 1967), drawn from the test sets of the chosen datasets. A sensitivity analysis evaluating the model's performance across varying parameters is provided in (Ayoobi et al. 2025a).

Experiments with Synthetically Biased Models. We developed three synthetic models to obtain an objective ground truth in the setting of the ACI dataset: two *globally biased models* (G1 and G2) and one *locally biased model* (L1). G1 and G2 are hand-crafted decision trees (see Figures 5, 6 in Appendix 3): the former predicts a negative label for all female individuals and a positive label for all others; the latter predicts a negative label for all black female individuals. For L1, we trained a Logistic Regression (LR) model on the ACI training set. In contrast to the global models, L1 operates at a neighbourhood level, changing the prediction to the negative label for female individuals if their neighbourhood exhibits an ϵ-bias against $gender = female$ in the originally trained model (see Figure 7 in (Ayoobi et al. 2025a)).

Bias in Single Neighbourhoods. Table 1 shows that our method consistently outperforms IRB with *G2* and *L1* (IRB yields all zero scores with *G2*, as it focuses on showing bias against individual features, rather than combinations thereof). For *G1*, the two approaches perform similarly. Overall, our method runs significantly faster due to a simpler QBAF with fewer nodes and relations[3].

Method	Model	K	Accuracy	Precision	Recall	F1-score	Runtime (s)
Our	G1	50	0.95	1.00	0.90	0.95	**3.87**
IRB			0.95	1.00	0.90	0.95	28.22
Our	G1	100	1.00	1.00	1.00	1.00	**3.88**
IRB			1.00	1.00	1.00	1.00	48.47
Our	G1	200	1.00	1.00	1.00	1.00	**3.82**
IRB			1.00	1.00	1.00	1.00	97.81
Our	G2	50	**0.96**	**1.00**	**0.96**	**0.98**	**0.34**
IRB			0.00	0.00	0.00	0.00	4.28
Our	G2	100	**1.00**	**1.00**	**1.00**	**1.00**	**0.35**
IRB			0.00	0.00	0.00	0.00	8.99
Our	G2	200	**1.00**	**1.00**	**1.00**	**1.00**	**0.38**
IRB			0.00	0.00	0.00	0.00	17.44
Our	L1	50	**1.00**	**1.00**	**1.00**	**1.00**	**5.00**
IRB			0.81	**1.00**	0.44	0.61	30.66
Our	L1	100	**1.00**	**1.00**	**1.00**	**1.00**	**4.96**
IRB			0.74	**1.00**	0.31	0.48	49.45
Our	L1	200	**1.00**	**1.00**	**1.00**	**1.00**	**5.07**
IRB			0.70	**1.00**	0.22	0.36	97.16

Table 1: The performance of our method (using $Q^c_{\mathcal{N}_i}$) versus IRB, with single neighbourhoods ($\mathcal{N}_i$) of different sizes (K) for synthetically biased classifiers. Best results in bold.

Approach	Model	Accuracy	Precision	Recall	F1-score
Our	G1	1.00	1.00	1.00	1.00
IRB		1.00	1.00	1.00	1.00
Our	G2	**1.00**	**1.00**	**1.00**	**1.00**
IRB		0.00	0.00	0.00	0.00
Our	L1	**1.00**	**1.00**	**1.00**	**1.00**
IRB		0.70	1.00	0.22	0.36

Table 2: The performance of our method (using $\mathcal{Q}_{\mathbf{g}}$ with $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ of sizes K=50,100,200, respectively) versus IRB (with the largest neighbourhood $\mathcal{N}_3$) for synthetically biased classifiers. Best results in bold.

Bias Across Neighbourhoods. We consider multiple neighbourhoods of varying sizes (50, 100, 200).[4] Table 2 shows that our method achieves perfect performance for all classifiers. IRB fails for G2 for the same reasons as before.

Experiments with Trained Models. We trained logistic regression models on Bank marketing and COMPAS. Here, we no longer have ground truth labels for bias and thus just compare the number of identified biases. Table 3a gives results drawn from COMPAS, showing that our method identifies substantially more biased cases against African-American individuals (77 vs. 2), in line with the literature (Angwin et al. 2016; Khademi and Honavar 2020), while detecting fewer or no biased cases for the other groups compared to IRB. Table 3b gives results for individuals identified as biased against

[3]The standard deviations of the metrics are reported in Appendix.

[4]We run experiments with other sizes, ranging from $K = 10$ to $K = 200$ in intervals of 10. See (Ayoobi et al. 2025a) for details.

Feature value		LR	
		IRB	Our
race	African-American	2	77
	Caucasian	1	1
	Hispanic	26	13
	Asian	3	0
	Native American	3	0
	Other	11	0
sex	Female	1	0

(a) COMPAS dataset

Feature value		LR	
		IRB	Our
age	MidAge	0	0
	YoungOrOld	20	78
marital	Married	0	4
	Single	4	11
	Divorced	19	10

(b) Bank Marketing dataset

Table 3: Count of feature values biased against in queried individuals using $\mathcal{Q}_{\mathbf{g}}$ with $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ of sizes K=50,100,200, resp., and IRB (with $\mathcal{N}_3$) and Logistic Regression (LR).

specific age and marital status groups in the Bank Marketing dataset. In particular, for "YoungOrOld" age (younger than 25 years of age or older than 60 years of age), "Married", and "Single" marital features, our method detects more biased cases than IRB. We show examples in (Ayoobi et al. 2025a).

Experiments with LLMs. We examine the widely used ChatGPT-4o. We followed the same protocol as in (Liu et al. 2024), prompting the model to predict the class label based on the individuals' feature values. Table 4 shows the results of our experiments. Since the model is not fine-tuned on COMPAS or Bank Marketing, the results reveal the presence of inherent biases within the model itself. In Table 4a, our method identifies a significantly higher number of biased instances against African Americans (129 cases) compared to zero by IRB. Similarly, the model demonstrates a higher degree of bias against females (6 instances). In Table 4b, our approach also uncovers bias which IRB fails to detect. These results suggest that our method may be more sensitive in uncovering latent bias patterns that IRB overlooks.

Feature value		LLM (GPT4)	
		IRB	Our
race	African-American	0	129
	Caucasian	0	2
	Hispanic	8	1
	Asian	2	0
	Native American	3	0
	Other	2	1
sex	Female	0	6

(a) COMPAS dataset

Feature value		LLM (GPT4)	
		IRB	Our
age	MidAge	0	2
	YoungOrOld	0	14
marital	Married	0	13
	Single	0	18
	Divorced	0	0

(b) Bank Marketing dataset

Table 4: Count of feature values being biased against in queried individuals using $\mathcal{Q}_{\mathbf{g}}$ with $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ of sizes K=50,100,200, resp., versus IRB (with $\mathcal{N}_3$).

8 Argumentative Debates

Our QBAFs can be used as debate-based explanations. The debates are naturally argumentative, by presenting evidence for and against bias, and for and against this evidence. These debates may take place in the "mind" of a *single agent*, taking two opposing role, of a *proponent* arguing that there is a bias and an *opponent* arguing that there is no such bias, in the spirit of dispute trees and dispute derivations (Thang, Dung, and Hung 2009; Cyras et al. 2017). As an example, for the QBAF in Figure 1, the proponent may state the top most argument in Figure 4 (left) and the opponent may then continue

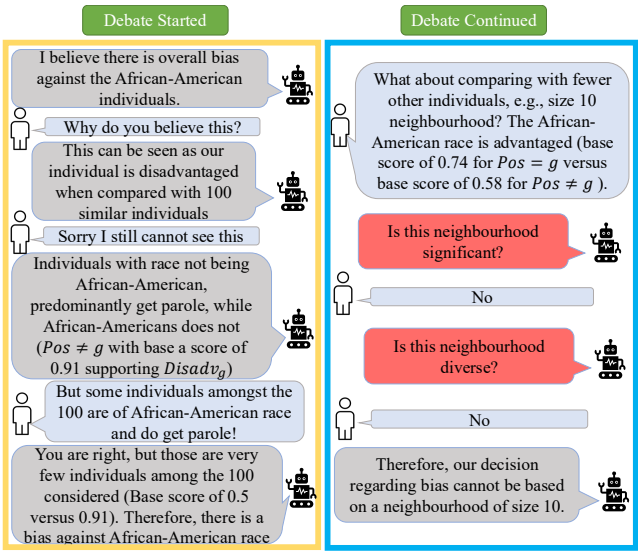

Figure 4: Argumentative Debate example (debates in yellow/blue boxes stem from corresponding boxes in Figure 1).

by stating: "But some individuals amongst the 100 are of African-American race and do get parole!" This behaviour can be obtained with conversational templates as in (Rago et al. 2025). Also, the QBAFs can be used as templates to drive debates amongst *two or more agents*. To do so, roles (for and against bias) or specific neighbourhoods could be associated with agents, with each using critical question arguments to criticise the neighbourhoods chosen by the others. Finally, the QBAFs can be used as the backbone for debates between *agents and humans* (as in Figure 4). When debates arise within or amongst agents, they could be generated and presented as explanations to users in one go, or incrementally, based on the user's prompts (as in Figure 4). This requires the ability of agents to match the prompts with specific parts of the underpinning QBAF, and generate replies based on relevant parts thereof. We leave this to future work.

9 Conclusion

We presented ABIDE, a novel model-agnostic approach for bias detection for analysing both local and global biases through argument schemes and quantitative bipolar argumentation. Our approach addresses a critical need for algorithmic fairness, by integrating transparency into bias detection, while also exhibiting desirable properties, linked to established properties for quantitative bipolar argumentation, and performance advantages over an argumentative baseline.

We discussed how ABIDE can support debates about bias. Future work will focus on the realisation thereof within and across agents, as well as between agents and humans, including with user studies. It will be especially interesting to explore how humans can contribute to the debates, specifically by contesting existing arguments and adding new arguments to the debate, in the spirit of (Leofante et al. 2024).

Acknowledgements

Ayoobi, Rapberger, and Toni were funded by the ERC under the EU's Horizon 2020 research and innovation programme (ADIX, grant number 101020934). Toni was also funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme.

References

Amgoud, L.; and Ben-Naim, J. 2016. Axiomatic Foundations of Acceptability Semantics. In *KR 2016*, 2–11. AAAI Press.

Amgoud, L.; and Ben-Naim, J. 2017. Evaluation of arguments in weighted bipolar graphs. In *ECSQARU 2017*, 25–35. Springer.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Accessed: 2025-05-06.

Ayoobi, H.; Cao, M.; Verbrugge, R.; and Verheij, B. 2019. Handling Unforeseen Failures Using Argumentation-Based Learning. In *CASE 2019*, 1699–1704. IEEE.

Ayoobi, H.; Cao, M.; Verbrugge, R.; and Verheij, B. 2021. Argue to Learn: Accelerated Argumentation-Based Learning. In Wani, M. A.; Sethi, I. K.; Shi, W.; Qu, G.; Raicu, D. S.; and Jin, R., eds., *ICMLA 2021*, 1118–1123. IEEE.

Ayoobi, H.; Cao, M.; Verbrugge, R.; and Verheij, B. 2022. Argumentation-Based Online Incremental Learning. *IEEE Trans Autom. Sci. Eng.*, 19(4): 3419–3433.

Ayoobi, H.; Kasaei, H.; Cao, M.; Verbrugge, R.; and Verheij, B. 2023a. Explain What You See: Open-Ended Segmentation and Recognition of Occluded 3D Objects. In *ICRA 2023*, 4960–4966. IEEE.

Ayoobi, H.; Potyka, N.; Rapberger, A.; and Toni, F. 2025a. Argumentative Debates for Transparent Bias Detection [Technical Report]. *CoRR*, abs/2508.04511.

Ayoobi, H.; Potyka, N.; Toni, F.; and and. 2023b. SpArX: Sparse Argumentative Explanations for Neural Networks. In *ECAI 2023*, 149–156. IOS Press.

Ayoobi, H.; Potyka, N.; Toni, F.; and and. 2025b. ProtoArgNet: Interpretable Image Classification with Super-Prototypes and Argumentation. In *AAAI-25*, 1791–1799. AAAI Press.

Baroni, P.; Rago, A.; and Toni, F. 2018. How Many Properties Do We Need for Gradual Argumentation? In *AAAI 2018*, 1736–1743. AAAI Press.

Becker, B.; and Kohavi, R. 1996. Adult. Accessed: 2025-05-06.

Brown-Cohen, J.; Irving, G.; and Piliouras, G. 2024. Scalable AI Safety via Doubly-Efficient Debate. In *ICML 2024*. OpenReview.net.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 56(7): 166:1–166:38.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic Decision Making and the Cost of Fairness. In *SIGKDD 2017*, 797–806. ACM.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1): 21–27.

Cyras, K.; Fan, X.; Schulz, C.; and Toni, F. 2017. Assumption-based Argumentation: Disputes, Explanations, Preferences. *FLAP*, 4(8).

de Tarlé, L. D.; Bonzon, E.; and Maudet, N. 2022. Multiagent Dynamics of Gradual Argumentation Semantics. In *AAMAS 2022*, 363–371. IFAAMAS.

Dejl, A.; Ayoobi, H.; Cherrington, C.; Gardner, D.; Toni, F.; Pakzad-Shahabi, L.; and Williams, M. 2025a. ARGTUMOUR: INTEGRATING LARGE LANGUAGE MODELS AND COMPUTATIONAL ARGUMENTATION TO DISCUSS TREATMENT OPTIONS FOR HIGH-GRADE GLIOMA. *Neuro-Oncology*, 27(Supplement_2): ii26–ii26.

Dejl, A.; Zhang, D.; Ayoobi, H.; Williams, M.; and Toni, F. 2025b. Hidden Conflicts in Neural Networks and their Implications for Explainability. In *FAccT 2025*, 1498–1542. ACM.

Ehyaei, A.-R.; Farnadi, G.; and Samadi, S. 2024. Wasserstein Distributionally Robust Optimization through the Lens of Structural Causal Models and Individual Fairness. In *NeurIPS 2024*, 42430–42467. Curran Associates, Inc.

Fawkes, J.; Fishman, N.; Andrews, M.; and Lipton, Z. C. 2024. The Fragility of Fairness: Causal Sensitivity Analysis for Fair Machine Learning. In *NeurIPS 2024*, volume 37, 137105–137134. Curran Associates, Inc.

Grabowicz, P. A.; Perello, N.; and Mishra, A. 2022. Marrying Fairness and Explainability in Supervised Learning. In *FAccT '22*, 1905–1916. ACM.

Irving, G.; Christiano, P. F.; and Amodei, D. 2018. AI safety via debate. *CoRR*, abs/1805.00899.

Khademi, A.; and Honavar, V. 2020. Algorithmic Bias in Recidivism Prediction: A Causal Perspective (Student Abstract). In *AAAI 2020*, 13839–13840.

Khan, A.; Hughes, J.; Valentine, D.; Ruis, L.; Sachan, K.; Radhakrishnan, A.; Grefenstette, E.; Bowman, S. R.; Rocktäschel, T.; and Perez, E. 2024. Debating with More Persuasive LLMs Leads to More Truthful Answers. In *ICML 2024*. OpenReview.net.

Kori, A.; Glocker, B.; and Toni, F. 2024. Explaining Image Classifiers with Visual Debates. In *DS2024*, volume 15244 of *Lecture Notes in Computer Science*, 200–214. Springer.

Leite, J.; and Martins, J. G. 2011. Social Abstract Argumentation. In *IJCAI 2011*, 2287–2292. IJCAI/AAAI.

Leofante, F.; Ayoobi, H.; Dejl, A.; Freedman, G.; Gorur, D.; Jiang, J.; Paulino-Passos, G.; Rago, A.; Rapberger, A.; Russo, F.; Yin, X.; Zhang, D.; and Toni, F. 2024. Contestable AI Needs Computational Argumentation. In Marquis, P.; Ortiz, M.; and Pagnucco, M., eds., *KR2024*.

Liu, Y.; Gautam, S.; Ma, J.; and Lakkaraju, H. 2024. Confronting LLMs with Traditional ML: Rethinking the Fairness of Large Language Models in Tabular Classifications. In Duh, K.; Gomez, H.; and Bethard, S., eds., *NAACL 2024 (Volume 1: Long Papers)*, 3603–3620. Association for Computational Linguistics.

Ma, S.; Salinas, A.; Nyarko, J.; and Henderson, P. 2025. Breaking Down Bias: On The Limits of Generalizable Pruning Strategies. In *FAccT '25*, 2437–2450. ACM.

Macagno, F.; Walton, D.; and Reed, C. 2017. Argumentation Schemes. History, Classifications, and Computational Applications. *FLAP*, 4(8).

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6): 115:1–115:35.

Moro, S.; Rita, P.; and Cortez, P. 2014. Bank Marketing. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5K306.

Mossakowski, T.; and Neuhaus, F. 2018. Modular semantics and characteristics for bipolar weighted argumentation graphs. *arXiv preprint arXiv:1807.06685*.

Panisson, A. R.; McBurney, P.; and Bordini, R. H. 2021. A computational model of argumentation schemes for multi-agent systems. *Argument Comput.*, 12(3): 357–395.

Potyka, N. 2018. Continuous dynamical systems for weighted bipolar argumentation. In *KR 2018*, 148–157.

Potyka, N. 2019. Extending Modular Semantics for Bipolar Weighted Argumentation. In *AAMAS 2019*, 1722–1730. IFAAMAS.

Potyka, N.; and Booth, R. 2024a. Balancing Open-Mindedness and Conservativeness in Quantitative Bipolar Argumentation (and How to Prove Semantical from Functional Properties). In *KR 2024*, 597–607.

Potyka, N.; and Booth, R. 2024b. An Empirical Study of Quantitative Bipolar Argumentation Frameworks for Truth Discovery. In *COMMA 2024*, 205–216. IOS Press.

ProPublica. 2016. COMPAS Recidivism Racial Bias Dataset. https://www.kaggle.com/datasets/danofer/compass. Accessed: 2025-05-06.

Rago, A.; Cocarascu, O.; Oksanen, J.; and Toni, F. 2025. Argumentative review aggregation and dialogical explanations. *Artif. Intell.*, 340: 104291.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In *KR 2016*, 63–73.

Räz, T. 2024. Reliability Gaps Between Groups in COMPAS Dataset. In *FAccT '24*, 113–126. ACM.

Thang, P. M.; Dung, P. M.; and Hung, N. D. 2009. Towards a Common Framework for Dialectical Proof Procedures in Abstract Argumentation. *J. Log. Comput.*, 19(6): 1071–1109.

Thauvin, D.; Herbin, S.; Ouerdane, W.; and Hudelot, C. 2024. Interpretable Image Classification Through an Argumentative Dialog Between Encoders. In *ECAI 2024*, 3316–3323. IOS Press.

Waller, M.; Rodrigues, O.; and Cocarascu, O. 2024. Identifying Reasons for Bias: An Argumentation-Based Approach. In *AAAI 2024*, 21664–21672. AAAI Press.

Waller, M.; Rodrigues, O.; Lee, M. S. A.; and Cocarascu, O. 2024. Bias Mitigation Methods: Applicability, Legality, and Recommendations for Development. *J. Artif. Intell. Res.*, 81: 1043–1078.

Walton, D.; Reed, C.; and Macagno, F. 2008. *Argumentation Schemes*. Cambridge University Press.

Xiang, A. 2024. Fairness & Privacy in an Age of Generative AI. *Science and Technology Law Review*, 25(2).