Identifying Reasons for Bias: An Argumentation-Based Approach

Madeleine Waller, Odinaldo Rodrigues, Oana Cocarascu

King's College London
{madeleine.waller, odinaldo.rodrigues, oana.cocarascu}@kcl.ac.uk

Abstract

As algorithmic decision-making systems become more prevalent in society, ensuring the fairness of these systems is becoming increasingly important. Whilst there has been substantial research in building fair algorithmic decision-making systems, the majority of these methods require access to the training data, including personal characteristics, and are not transparent regarding which individuals are classified unfairly. In this paper, we propose a novel model-agnostic argumentation-based method to determine why an individual is classified differently in comparison to similar individuals. Our method uses a quantitative argumentation framework to represent attribute-value pairs of an individual and of those similar to them, and uses a well-known semantics to identify the attribute-value pairs in the individual contributing most to their different classification. We evaluate our method on two datasets commonly used in the fairness literature and illustrate its effectiveness in the identification of bias.

1 Introduction

As machine learning (ML) algorithms are increasingly used in decision-making systems with high impact on individuals, there is a need to ensure not only that the decisions made are fair, but also to explain the decision to the individual affected. A system is considered to be fair if it does not discriminate based on protected personal characteristics such as race, sex, religion, etc. There have been instances where decision-making systems discriminated against individuals in domains such as criminal justice (The Partnership on AI 2019), recruitment (Tilmes 2022), and social services (Gillingham 2019). An analysis of COMPAS (Northpointe 2019), a popular tool used in the US to predict whether criminals will re-offend, found that black defendants were identified incorrectly as re-offending at a higher rate than white defendants (Larson et al. 2016).

Research on fairness has received increased attention in recent years and several fairness metrics have been developed to quantify the fairness of a system (see Mehrabi et al. (2022) for an overview on bias and fairness in machine learning). These metrics can be classified into group fairness (i.e. detecting bias across different values of a protected attribute, e.g., male and female individuals (Garg, Villasenor, and Foggo 2020)) and individual fairness (i.e. detecting bias for an individual compared to similar individuals (Mukherjee et al. 2020)). Whilst several notions of evaluating fairness have been proposed in the literature, there is no agreement as to which fairness metric to apply in which scenario (Verma and Rubin 2018). Furthermore, interpreting the meaning of the values returned by a metric is not always intuitive; for example, simply reporting the percentage level of fairness of a system (e.g., 80%) may not give full confidence for stakeholders in the system. Most existing group metrics also require the specification of protected attributes and can only detect unwanted bias with respect to one binary protected attribute (Waller, Rodrigues, and Cocarascu 2023b). Finally, quantifying fairness requires full access to the training data and protected attributes in order to measure the difference in positive classifications across protected groups. In reality, this data may not be available and it may be difficult to pre-define protected attributes before deploying the system (Haeri and Zweig 2020).

Identifying a link between the input data and the final decision is key towards providing a fair and transparent explanation (Hamon et al. 2022). Computational argumentation has long been seen as a means for explaining reasoning (see Vassiliades, Bassiliades, and Patkos (2021); Cyras et al. (2021) for an overview). Specifically, abstract Argumentation Frameworks (AFs) were proposed as a way to represent and reason with conflicting information (Dung 1995). Several types of semantics have been proposed to evaluate the acceptability of arguments in AFs (Baroni et al. 2015) and their extensions. AFs have been used for a variety of applications such as decision-making systems (Amgoud and Prade 2009; Brarda, Tamargo, and García 2021), recommender systems (Cocarascu, Rago, and Toni 2019; Rago, Cocarascu, and Toni 2018), knowledge-based systems (Kökciyan et al. 2020), and planning and scheduling systems (Cyras et al. 2019). However, until now they have not been explored in relation to individual fairness in decision-making systems.

In this paper, we propose a novel argumentation-based approach for identifying bias in relation to individual fairness which does not require access to labelled data, the training algorithm, or the specification of protected attributes before deployment. We focus on individual fairness as individuals impacted by decisions will mostly be concerned about their personal treatment rather than any group. We move away from quantifying fairness using existing group fairness metrics and offer a transparent representation of the reasons for a classification from which an explanation can be extracted.

We use a quantitative argumentation framework to represent the arguments of similar individuals that reason why the queried individual received a classification. Reasons are differences in the values of attributes in the queried individual in relation to those of similar individuals with different classifications. The strength of attacks between attribute-value pairs is calculated as the proportion of similar individuals with particular characteristics and the overall evaluation is done using the weighted h-Categorizer semantics (Amgoud, Doder, and Vesic 2022) which calculates the final weights of arguments. As a result, final weights correspond to the attribute-value pairs that contribute most to the negative classification of a queried individual compared to similar individuals.

2 Background

In this section, we provide the core background on fairness in ML and argumentation on which our method relies.

Fairness in ML

Fairness in ML involves ensuring decision-making ML systems are fair for individuals and for different groups defined by protected personal characteristics. Various metrics have been proposed to quantify fairness in decision-making systems (Mehrabi et al. 2022). Individual fairness metrics determine whether similar individuals receive the same classifications (Dwork et al. 2012). For example, Zemel et al. (2013) counts the pairs of individuals receiving the same classification, where the notion of similarity depends on the context of the application. Group fairness metrics have been defined as the difference in the number of positive and negative classifications across two protected groups (Mehrabi et al. 2022). For example, demographic parity is calculated as the proportion of positive classifications for the protected group divided by the proportion for the non-protected group (Calders, Kamiran, and Pechenizkiy 2009).

Bias detection methods are used to assess the fairness of a system. Mitigation methods attempt to make a system fairer with respect to some metric(s) by targeting different stages of the ML system including the pre-processing of the training data (Kamiran and Calders 2011), the training algorithm (Hu et al. 2020), or in the classifications (Lohia et al. 2019). The scenarios in which each method may be applied are not usually specified (Weinberg 2022; Wachter, Mittelstadt, and Russell 2021) and there is usually no consideration into the legality of the use of protected attributes for the identification of bias (Haeri and Zweig 2020).

Most existing bias detection and mitigation methods focus on group fairness and do not consider individual fairness. Satisfying individual fairness in a decision-making system is necessary but not sufficient to ensure overall fairness (Fleisher 2021). This is because individuals in a protected group could all be given negative outcomes and this would conflict with the notion of group fairness despite individual fairness being satisfied (Chakraborty, Peng, and Menzies 2020).

Computational Argumentation

Argumentation theory has been explored as a way of representing arguments and forms of reasoning (Dung 1995; Baroni, Caminada, and Giacomin 2011; Atkinson et al. 2017). An Abstract Argumentation Framework (AAF) is a tuple $\langle A, R\rangle$, where A is a set of arguments and $R \subseteq A \times A$ is an attack relation between them. Arguments are atomic entities whose content is not specified, the focus being on acceptability criteria for the arguments based on the attack relationship, resulting in alternative semantics (Efstathiou 2011; Baroni, Caminada, and Giacomin 2011).

Weighted (or Quantitative) Argumentation Frameworks (WAFs) augment AAFs with *weights*. For our method, we are interested in WAFs that define weights for both arguments and attacks.

Definition 1. [Weighted Argumentation Graph (Amgoud and Doder 2019)] $G = \langle A, \sigma, R, \pi\rangle$, where A is a non-empty finite set of arguments, $R \subseteq A \times A$, $\sigma : A \mapsto [0, 1]$, and $\pi : R \mapsto [0, 1]$.

A and R define arguments and attacks as before, but an argument $a \in A$ is given an initial weight $\sigma(a)$, and attacks (a, b) between arguments are given a strength $\pi(a, b)$. Semantics for WAFs define how the final weights of the arguments are calculated given G. We use the strength of attacks to express the relative representation of "votes" by similar individuals, inspired by ideas from Eğilmez, Martins, and Leite (2014); Gabbay and Rodrigues (2014). WAFs may have cycles, and applicable semantics include the Trust-based semantics (da Costa Pereira, Tettamanzi, and Villata 2011), the Simple Product (Leite and Martins 2011), Weighted Max-based *(Mbs)*, Weighted Cardinality-based *(Wbs)* and Weighted h-Categorizer (*Hbs*) (Amgoud, Doder, and Vesic 2022). These last three were originally developed for frameworks with attack and support relations (Amgoud et al. 2017) and have also been extended to account for strengths of attacks (Amgoud and Doder 2019). *Mbs* favours the strength over quantity of attacks, whereas *Cbs* favours quantity over strength. *Hbs* considers both (Amgoud and Doder 2019).

Hbs is based on the h-Categorizer semantics originally developed for non-weighted graphs (Besnard and Hunter 2001; Pu et al. 2014). *Hbs* defines an infinite sequence of weights for arguments $s(a)^{(1)}, s(a)^{(2)}, \ldots$ such that

$$s(a)^{(1)} = \sigma(a) \tag{1}$$

$$s(a)^{(n+1)} = \frac{\sigma(a)}{1 + \sum_{b \in Att_a} \pi((b,a)) \cdot s(b)^{(n)}}, \text{ for } n > 0 \tag{2}$$

Notice that the denominator of the fraction on the right-hand side of Equation 2 is greater than 1 for any *attacked* argument a provided the strength of the attack is non-null. This means the weight of such arguments decrease in proportion to the quantity as well as the strength of the attacks. The sequence $\{s(a)^{(n)}\}_{n=1}^{+\infty}$ is infinite. However, Amgoud, Doder, and Vesic (2022, Theorem 17) showed that for every $a \in A$, it converges and the final weight of an argument a is defined as $\lim_{n\to+\infty} s(a)^{(n)}$. We will describe how to approximate these final weights in Section 3.

3 Identifying Reasons for Bias Using Argumentation

In this section we describe our method. Given a *queried* individual's classification, we aim to ensure it is fair with respect to similar individuals. For simplification, we only consider queried individuals that are classified *negatively*, although our method equally applies to positive classifications. Our approach requires a set of unlabelled individuals, which we obtain from the test data (see Section 4).

Let E be the set of unlabelled individuals with attributes $Z = \langle z_1, z_2, ..., z_p \rangle$ and corresponding domains $\langle D_1, D_2, ..., D_p \rangle$. Let $v : E \times Z \to \cup_{i=1}^{p} D_i$ be the function that given an individual e and attribute z_i, returns the value of the attribute z_i for e.

Defining Similar Individuals

To represent individual fairness, we must first define what it means for individuals to be "similar". Several definitions of similarity have been explored in the literature in relation to individual fairness. For example, Zemel et al. (2013) proposed the use of k-nearest neighbours (KNN) while Dwork et al. (2012) proposed context-specific similarity definitions whereby certain attributes are considered more important in the evaluation of how similar individuals are to one another.

We identify an individual e's similar individuals as the nearest neighbours in $E/\{e\}$ using a Ball Tree clustering algorithm (Leibe, Mikolajczyk, and Schiele 2006) using the Hamming distance. We assume all values to be categorical, converting numerical attributes such as *age* to categorical attributes by grouping them appropriately (see Section 4).[1]

Hence, the distance $d(e_1, e_2)$ between individuals e_1 and e_2 is given by

$$d(e_1, e_2) = \sum_{i=1}^{p} d_i, \text{ where } d_i = \begin{cases} 0, \text{ if } v(e_1, z_i) = v(e_2, z_i) \\ 1, \text{ if } v(e_1, z_i) \neq v(e_2, z_i) \end{cases}$$

The distance between individuals with the same attribute values is 0 and the distance between two individuals increases in proportion to the number of attributes with different values in the individuals.

$Sim_k(e)$ will denote the set of k individuals most similar to the individual e, i.e. the individuals $\{e_1, \ldots, e_k\} \subseteq E/\{e\}$ minimising $\sum_{i=1}^{k} d(e, e_i)$. We assume that $|E/\{e\}| \geq k$ and that if there are more than k individuals minimising $\sum_{i=1}^{k} d(e, e_i)$, we can choose arbitrarily between them.

In Section 4 we discuss the effect of using different values of k. In what follows, we present a working example to illustrate our method using $k = 5$ similar individuals.

Example 1. *Table 1 shows six individuals from the commonly used Adult dataset (Becker and Kohavi 1996), restricted to three attributes and their classifications. The top row represents the queried individual e and the other rows represent e's similar individuals as described in this section. The classification $+$ represents that the individual's income is predicted to exceed \$50K/year, and $-$ otherwise.*

workclass	**education**	**race**	**Classification**
Local-gov	*Bachelors*	*Black*	$-$
Private	*Bachelors*	*White*	$+$
Local-gov	*HS-grad*	*White*	$+$
Local-gov	*Bachelors*	*White*	$+$
Private	*Masters*	*White*	$+$
Local-gov	*Masters*	*White*	$+$

Table 1: Sample data with queried individual (grey) and five similar individuals.

Intuitively, Example 1 shows that given the queried individual in the top row and its similar individuals, $(race, Black)$ is the attribute-value pair contributing the most to the negative classification of the queried individual, since all combinations of values of the other attributes in the similar individuals lead to a positive classification. Our aim is to create a mechanism by which this type of bias can be identified, focusing on *why* an individual has been classified negatively (and which attributes contribute most to this), as opposed to *how many* individuals have been treated unfairly which is the aim of most existing bias detection methods (Waller, Rodrigues, and Cocarascu 2023a).

Constructing the Argumentation Graph

Let f be a binary classifier $f : E \to Y$ which takes an individual $e \in E$ as input and outputs a classification $f(y) \in \{+, -\}$. Recall that $v(e, z)$ is the value of the attribute z for e. We use e_0 to denote the queried individual.

We show how to construct the weighted argumentation graph $\langle A, \sigma, R, \pi \rangle$, which will be used to detect the attribute-value pairs that contribute the most to the classification of e_0. In all that follows, we assume that the set of attributes is $Z = \{z_1, \ldots, z_p\}$.

We define the set of arguments A as the unique attribute-value pairs $(attribute, value)$ representing an association $attribute = value$ found in $\{e_0\} \cup Sim_k(e_0)$.

Definition 2. [Set of arguments] Let e_0 be the queried individual and $Sim_k(e_0)$ the set with the k individuals most similar to e_0 according to some similarity measure Sim. The set of arguments A is defined as follows.

$$A = \bigcup_{i=0}^{k} \bigcup_{j=1}^{p} \{(z_j, v(e_k, z_j))\}$$

At the outset, we have no prior information about which attribute-value pairs contribute the most to the negative classification of the queried individual, so the initial weight of all arguments is set to 1.[2]

[1]The distance between individuals could also be calculated using another metric that would consider the difference between numerical values and/or the relative distance between ordered categorical attributes such as $education\text{-}level$.

[2]It is left for future work to explore whether the initial weight of an argument could represent prior knowledge such as prevalence of attribute values in the dataset or existing group fairness metrics for a deployed model.

Definition 3. [Initial argument weights] For all $a \in A$, $\sigma(a) = 1$.

We now define the attack relations and strengths using the frequency of occurrence of those pairs amongst the similar individuals. One way of looking at this is that individuals "argue" for the potential reasons of the negative classification of the queried individual, which can be the presence of one, or multiple attribute-value pairs. If a similar positively-classified individual has a value for an attribute z different to the value of the negatively-classified queried individual, then the similar individual attacks that attribute-value pair in the queried individual. This will in turn reduce the initial weight of the pair. Specifically *all* attribute-value pairs of similar individuals e_i will attack the value of the attribute z in the queried individual e_0 if the values of z differ in e_0 and e_i. This accounts for the possibility that e_0's negative classification could be due to combinations of multiple attribute-value pairs in the similar individuals.

Definition 4. [Attack relationship] Given a binary classifier $f : E \to \{+, -\}$, a queried individual e_0, and the set of similar individuals $Sim_k(e_0)$:

$$R = \bigcup_{i=1}^{k} \bigcup_{j=1}^{p} \bigcup_{l=1}^{p} \{((z_l, v(e_i, z_l)), (z_j, v(e_0, z_j))) \mid f(e_i) \neq f(e_0) \text{ and } v(e_0, z_j) \neq v(e_i, z_j)\}$$

Notice that by Definition 4 the targets of attacks can only be arguments of the type $(z_j, v(e_0, z_j))$ (i.e. towards the queried individual's attribute-value pairs) and that attacks from the attribute-value pairs of a similar individual e_i only occur towards an argument $(z_j, v(z_j, e_0))$ when the values of the attribute z_j differ in e_0 and e_i (i.e. $(v(e_0, z_j) \neq v(e_i, z_j))$. In addition, we also want the strength of any potential attacks to reflect the *proportion* of similar individuals with a particular differing attribute-value pair. This approach is inspired by voting, with related ideas explored in detail in Eğilmez, Martins, and Leite (2014); Gabbay and Rodrigues (2014). The following definition captures the intuition.

Definition 5. [Strength of attacks] Take $(a, b) \in R$, where $a = (z_1, v_1)$ and $b = (z_2, v_2)$.

$$\pi((a, b)) = \frac{|\{e_i \mid v(e_i, z_1) = v_1 \text{ and } v(e_0, z_2) \neq v(e_i, z_2)\}|}{k}$$

Since $(a, b) \in R$, there are between 1 and k individuals for which the condition in Definition 5 holds. Therefore $\frac{1}{k} \leq \pi((a, b)) \leq 1$.

Example 2. *Take the attribute* workclass *in Table 1 for which two similar individuals have the value $Private$, different to the queried individual. There is an attack relation from all attributes in these two similar individuals to $(workclass, \mathit{Local\text{-}gov})$. The values of attribute* education *for these individuals are $Bachelors$ and $Masters$ therefore an attack is added from $(education, Bachelors)$ and $(education, Masters)$ to $(workclass, \mathit{Local\text{-}gov})$, both with strength $\frac{1}{5}$ according to Definition 5. $(race, White)$ also attacks $(workclass, \mathit{Local\text{-}gov})$ but both individuals have that attribute value so the strength of the attack is $\frac{2}{5}$. Similarly for $(workclass, Private)$ to $(workclass, \mathit{Local\text{-}gov})$.*

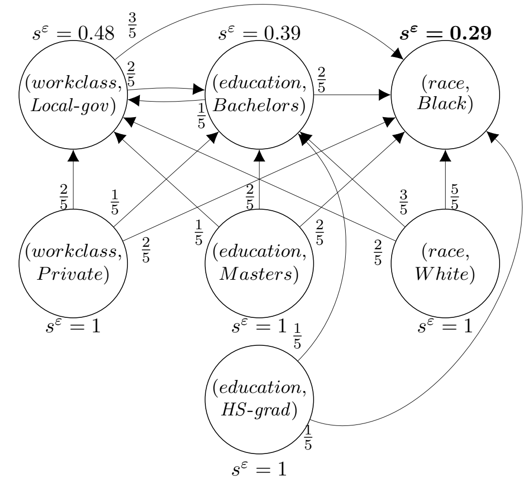

Figure 1: The weighted argumentation graph constructed from Table 1 with the final weights s^ε of all arguments.

Figure 1 shows the weighted argumentation graph constructed from the data in Table 1, with the strengths of attacks between arguments, and the final weights of all arguments, which will be discussed next.

Calculating Final Weights of Arguments

We now need to calculate the arguments' final weights, using an appropriate well-defined semantics. Good candidates for semantics are those that aggregate attacks in an "argumentation-friendly" way (see Gabbay and Rodrigues (2015) for a discussion). At the very least, the semantics needs to be able to deal with potential cycles in the graph and reduce initial weights in proportion to the strength of attacks on arguments. We chose the *Hbs* semantics not only because it satisfies these requirements, but it is also efficient to compute and provided good results with the datasets used in our experiments. Further properties of this semantics have been explored in Amgoud and Doder (2019).[3]

Recall Equation 2. We aim to approximate $\lim_{n\to+\infty} s(a)^{(n)}$ by successively calculating the values of the sequence until the difference between successive values of all arguments is within a desired convergence threshold ε, i.e. $s(a)^{(n+1)} - s(a)^{(n)} < \varepsilon$, for all $a \in A$. By an abuse of notation we will denote this value $s(a)^\varepsilon$ and call it the "final" weight of the argument a (up to ε). Section 4 describes the particular choice of ε used in the experiments.

Having computed all final weights, the weakest arguments, i.e. those with the smallest final weights, correspond to the attribute-value pairs of the queried individual that

[3]We also ran experiments using the Quadratic Energy Model semantics (Potyka 2018) (without supports) and it identified the same weakest arguments for all individuals.

most contribute to its negative classification. These are given as the explanation Exp for the classification: $Exp = \{a \in A \mid s(a)^{\varepsilon} = \min_{b \in A}\{s(b)^{\varepsilon}\}\}$.

The approximated final weights calculated using *Hbs* for the sample data in Table 1 are shown in Figure 1. It is easy to see that $(race, Black)$ is the weakest argument in the graph, supporting our intuition in Example 1. We now show general properties of the final weights.

Proposition 1. *Let $a = (z_i, v) \in A$. If $s((z_i, v))^{\varepsilon} < 1$, then $v(e_0, z_i) = v$, where e_0 is the queried individual.*

Proof. According to Definition 3, $s((z_i, v))^{(1)} = \sigma((z_i, v)) = 1$. According to Equation 2, argument weights can only decrease and only when an argument is attacked. According to Definition 4, attacks are only defined towards attribute-value pairs of the queried individual. If $s((z_i, v))^{\varepsilon} < 1$, then (z_i, v) must have been attacked, and hence (z_i, v) is in the queried individual. □

Proposition 2. *Assume the queried individual e_0 has the same classification as all of its similar individuals. Let a_i be the attribute-value pair (z_i, v_i), such that $v_i = v(e_0, z_i)$, then $s(a_i)^{\varepsilon} = \sigma(a_i) = 1$.*

Proof. Let the argument a_j be an attribute-value pair (z_j, v_j), such that $v_j = v(e_0, z_j)$. According to Definition 4, an attack is added into a_j, only when the classification of e_0 is different to the classification of a similar individual e_i. Since the classification of e_0 is the same as that of all of its similar individuals, there will be no attacks into any such a_j, and hence according to Equation 2, $s(a_j)^{n+1} = s(a_j)^1 = s(a_j)^{\varepsilon}$, for all $1 \leq j \leq p, n \geq 0$. □

Proposition 2 shows that when the classification of the queried individual e_0 matches the classifications of its similar individuals, no attribute-value pair is singled-out in e_0.

Proposition 3. *Assume there is only one attribute z for which all positively-classified similar individuals have a different value than the queried individual and all other attributes have the same values for all individuals. Let a be the attribute-value pair $(z, v(e_0, z))$, then $s(a)^{\varepsilon} < 1, s(a)^{\varepsilon} = \min_{b \in A}\{s(b)^{\varepsilon})\}\}$ and for all $b \in \{A \setminus a\}, s(b)^{\varepsilon} = 1$.*

Proof. According to Definition 4, $(z, v(e_0, z))$ is attacked by all attribute values in all positively-classified similar individuals. Since all other attribute-value pairs are the same, then $(z, v(e_0, z))$ is the only argument that is attacked.

According to Equation 2, $s(a)^{n+1}$ only decreases if $\{b \mid (b, a) \in R\} \neq \emptyset$ thus only $(z, v(e_0, z))$ can have final weight less than 1 (all other final weights remaining unchanged, i.e. equal to 1). Hence, $s(a)^{\varepsilon}$ is the minimum value and identified as the only possible explanation for the negative classification. □

4 Evaluation

We conduct our analysis with experiments on the Adult (Becker and Kohavi 1996) and the Bank Marketing (Moro, Rita, and Cortez 2012) datasets, commonly used in the fairness literature. Further, we introduce bias to the Adult dataset to illustrate our method identifies bias as expected.

	Attribute value	+ labels	− labels	% of − labels
sex	Male	9539	20,988	69%
	Female	1669	13,026	89%
race	White	10,207	28,696	74%
	Black	534	3694	87%
	AsianPacIslander	369	934	72%
	AmerIndianEskimo	53	382	88%
	Other	45	308	87%

Table 2: Prevalence of positive and negative labels for different protected attribute values in the Adult dataset.

Experiments on Real Data

We cannot easily compare our results to existing fairness metric values or bias detection methods as we focus on the reasons for a classification, as described in Section 3, which provides the basis for *explaining* a classification. We envision our method being employed to identify possible bias in an individual's classification (see Example 1), rather than for a group of individuals. Nonetheless, we offer some results to showcase the efficacy of our approach.

To be able to show meaningful results, we first provide a qualitative data analysis of the Adult and Bank Marketing datasets, with a focus on their protected attributes, even though our method does not require their specification. We then provide the results of our experiments.[4].

Qualitative data analysis The original Adult dataset has 48,842 instances. We pre-process to remove instances with null values to result in 45,222 instances. For this dataset, the attributes sex and $race$ are identified as protected throughout the literature (Le Quy et al. 2022). Table 2 shows the prevalence of positive and negative labels for the values of these attributes. A negative (positive) label represents the fact that an individual's income is below (above) $50K, respectively. The percentage of negative labels is the proportion of individuals in a group with a negative label out of all individuals in that group. Table 2 shows there is a greater percentage of negative labels for females than males and for non-white individuals (except $AsiaPacIslander$) compared to white individuals.

The Bank Marketing dataset has 45,211 instances and the attributes age and $marital$ are identified as protected. We pre-process the dataset to categorise the attribute age into two groups corresponding to the protected groups identified in the literature (Le Quy et al. 2022). Specifically, converting the attribute value of age to $YoungOrOld$ where it is less than 25 or greater than 60 and $MidAge$ otherwise.

It is not as clear what a positive label is for this dataset. We define a positive (negative) label as representing individuals that have not (have) subscribed to a term deposit. As previously mentioned, we could easily adapt our method to identify bias with respect to a positive classification. Table 3 shows the prevalence of positive and negative labels for the values of these attributes.

For the attribute age, individuals with value

[4]The code is available here: https://github.com/maddiewaller/IdentifyingReasonsForBias

	Attribute value	+ labels	− labels	% of − labels
age	MidAge	38,634	4580	11%
	YoungOrOld	1288	709	36%
marital	Married	24,459	2755	10%
	Single	10,878	1912	15%
	Divorced	4585	622	12%

Table 3: Prevalence of positive and negative labels for different protected attribute values in the Bank Marketing dataset.

$YoungOrOld$ are considered in the protected group, whereas the individuals with value $MidAge$ are not. For the attribute $marital$, non-married individuals with the values of $single$ or $divorced$ are considered in the protected group, whereas $married$ individuals are not. Table 3 shows that there is a greater percentage of negative labels for young or old individuals than mid-age individuals, and for single and divorced individuals compared to married individuals, however the percentage of negative labels does not vary greatly for $marital$.

Setup For our experiments, we pre-process the Adult dataset as before by removing null values. Additionally we remove the attributes $fnlwgt$[5] and *education-num*[6]. We then apply the same numerical attribute categorisation as outlined by Le Quy et al. (2022) to both datasets. This aligns with our similarity definition in Section 3 and ensures meaningful argument representations.

To evaluate our method, we split our pre-processed datasets into training (80%) and test (20%) partitions and train a logistic regression classifier on the training sets. We chose logistic regression because it is the most common classification model used to evaluate bias mitigation methods (Hort et al. 2022). However, our method is model-agnostic and can be applied with any classifier.

We took as queried individuals every individual with a negative classification from our test sets amounting to 7,252 and 1,253 individuals for the Adult and Bank Marketing datasets, respectively. We then selected the 5 most similar individuals in the test data using a Ball-tree clustering algorithm as defined in Section 3.[7] This allowed us to collate groups of individuals such as those in Table 1 and use Definitions 2 – 4 to construct the weighted argumentation graph such as in Figure 1. We then calculated the strengths of the attacks according to Definition 5 and used the *Hbs* semantics (Equation 2) to find the weakest arguments in the graphs.

We tested varying convergence thresholds (ε) for approximating final weights, as in Equation 2. We chose $\varepsilon = 0.01$; smaller values did not impact weakest argument identification but increased computation time. Larger ε values led to imprecise weight approximations that could not reliably identify the weakest arguments. All approximated weights are rounded to two decimal places, treating two argument weights as equal if their rounded values match.

Dataset	Train	Test	Accuracy	$\mathbf{F_1}$
Adult	36177	9045	85%	66%
Bank Marketing	36169	9042	72%	82%

Table 4: Number of instances in the train and test set for the Adult and Bank Marketing datasets, with classifier performance statistics.

	Attribute value	Count (proportion)
sex	Male	130 (1.8%)
	Female	595 (8.2%)
race	White	86 (1.2%)
	Black	245 (3.4%)
	AsianPacIslander	64 (0.9%)
	AmerIndianEskimo	28 (0.4%)
	Other	21 (0.3%)

Table 5: Count (and proportion) of attribute values being the weakest argument in queried individuals (Adult).

As shown by Proposition 2, if all similar individuals have the same (negative) classification as a queried individual, the final weights of all arguments are 1 and hence the queried individual has been treated consistently with respect to the similar individuals. Otherwise the weakest arguments will correspond to the attribute-value pairs contributing the most to the negative classification.

Results on Adult dataset Table 5 shows the count and proportion where a protected attribute value was amongst the weakest arguments in the graph. Furthermore, 70% of the queried individuals are consistent with the similar individuals, meaning all similar individuals are also classified negatively. Our method identified 8.2% of queried individuals of which $(sex, Female)$ contributed the most to the negative classification, thus identifying bias against these individuals. As expected from the prevalence of negative labels for females versus males, this is greater than the proportion of $(sex, Male)$ contributing the most to the negative classification. This detects individuals who were given a negative classification unfairly based on the attribute values $Male$ or $Female$, highlighting a benefit of our method in not requiring to specify the protected group before deployment.

Similarly, Table 5 shows that $(race, Black)$ was identified as the attribute-value contributing the most to the negative classification in 159 more cases than the attribute-value $(race, White)$. Summing the proportions for all values of $race$ that are not equal to $White$, we obtain 5.0%, showing that there are a greater percentage of non-white individuals being negatively-classified due to their value of $race$ than white individuals.

Results on Bank Marketing dataset Table 6 shows the count and proportion of the weakest arguments for all the protected attribute values. Furthermore, 21% of the queried individuals are consistent with similar individuals, hence all similar individuals are also classified negatively.

Our method identifies 8.9% of queried individuals of

[5] $fnlwgt$ represents a weighting of how many individuals an instance represents and is not a feature of an individual (commonly removed from the dataset (Kamiran and Calders 2011)).

[6] *education-num* is equivalent to *education-level* which is already included in the dataset.

[7] In our experiments, the KD clustering algorithm had similar performance to Ball-tree clustering.

	Attribute value	Count (proportion)
age	MidAge	14 (1.1%)
	YoungOrOld	112 (8.9%)
marital	Married	59 (4.7%)
	Single	68 (5.4%)
	Divorced	51 (4.0%)

Table 6: Count (and proportion) of attribute values being the weakest argument in queried individuals (Bank Marketing).

which $(age, YoungOrOld)$ contributes the most to the negative classification, thus identifying bias against these individuals. As expected from the prevalence of negative labels for $MidAge$ versus $YoungOrOld$, this is less than the proportion of $(age, MidAge)$ contributing the most to the negative classification.

The counts of weakest arguments for various $marital$ values exhibit no significant difference, much like the proportion of negative labels for each group in the original dataset.

Experiments were run on a MacBook Pro with 32GB RAM and 10 CPU cores. Computing the final weights for all of the queried individuals for each dataset took under 2 minutes, showing our method is scalable to many queried individuals (see Amgoud, Doder, and Vesic (2022) for a discussion on the complexity of the *Hbs* semantics).

Experiments on Adapted Data

To provide evidence our method detects bias as expected, we add an attribute $bias\text{-}attr$ to the test set of the Adult dataset, which we first pre-process as previously described. The attribute $bias\text{-}attr$ takes value 0 or 1, each with probability 0.5. We then fix the classifier $f : E \to Y$ such that $f(e) = v(e, bias\text{-}attr)$.

We hypothesise that all queried individuals will either be consistent with their similar individuals, i.e. there is no difference in classifications between a queried individual and its similar individuals, or $bias\text{-}attr = 0$ will be identified as contributing the most to the negative classification of the queried individual.

Results Our method correctly identifies that, for all negatively-classified queried individuals where at least one of the similar individuals is positively-classified, $bias\text{-}attr = 0$ is amongst the weakest arguments and therefore identified as contributing the most to the negative classification, i.e. $bias\text{-}attr = 0$ is amongst the weakest arguments in our constructed argumentation graph. This is the case for 56.4% of the queried individuals. The other 43.6% of individuals were consistent with similar individuals thus no bias was detected. This highlights a limitation of individual fairness — if all similar individuals have the same classification, there is no bias identified as they are treated the same, even if that they are all classified negatively due to a particular attribute value. By increasing the number of similar individuals we consider, we tend towards group fairness and the proportion of consistent queried individuals decreases. For example, running the experiments with *10 similar individuals*, we identify 76.8% of queried individuals for which the reason for the classification is $bias\text{-}attr = 0$ and 85.3% with *15 similar individuals*.

Our method correctly identifies the attribute value that contributes to the negative classification for all individuals that have at least one similar individual with a different classification. Ideally, dataset curation should prioritise minimising strong correlations between protected attributes and labels. This can be achieved through the implementation of bias mitigation techniques that emphasise group fairness. Assuming this precondition, our method excels in uncovering less obvious biases associated with individual fairness, which is often overlooked in practice.

5 Conclusion & Future Work

In this paper, we proposed a novel argumentation-based method for finding why an individual is classified differently from similar individuals, to be able to identify bias in relation to individual fairness. Our method is model-agnostic and does not require access to labelled data or the specification of protected characteristics.

We construct a quantitative argumentation framework based on the relationships between attribute-value pairs of a queried individual and its similar individuals. The argumentation framework can be used to extract an *explanation* for the classification of the queried individual compared to its neighbours, thus offering a transparent representation of the reasons for a classification, which is applicable to any machine learning classifier. We evaluated our method on the two most commonly used datasets in the fairness literature. Our results demonstrate a correlation between the attribute-value pairs detected by our method and the prevalence of protected attributes in the original dataset. In addition, we introduced synthetic bias into the Adult dataset and showed that our method correctly identified the attribute-value pairs artificially responsible for the classification.

Identifying the reasons for an individual's black-box binary classification is the first step to creating more transparent methods whose results can be better explained to users, hence increasing the trustworthiness of the classification.

There are multiple avenues for future work. Although we identified the reasons for bias in the datasets analysed, we would like to extend our experiments to cover other datasets and ensure that the current semantics is sufficiently fine-grained to identify potential more subtle causes for bias. In addition, we plan to employ alternative definitions of individual similarity and consider and compare the results obtained using different argumentation semantics. Future work also includes the development of templates that can be instantiated with more intelligible explanations in natural language rather than the current identification of attribute-value pairs, thus allowing better understanding of the explanations by end-users.

Acknowledgments

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence [grant number EP/S023356/1][8]. The first

[8]www.safeandtrustedai.org

author is an Affiliate of the King's Institute for Artificial Intelligence and additionally funded by The Alan Turing Institute's Enrichment Scheme.

References

Amgoud, L.; Ben-Naim, J.; Doder, D.; and Vesic, S. 2017. Acceptability Semantics for Weighted Argumentation Frameworks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia*, 56–62.

Amgoud, L.; and Doder, D. 2019. Gradual Semantics Accounting for Varied-Strength Attacks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada*, 1270–1278.

Amgoud, L.; Doder, D.; and Vesic, S. 2022. Evaluation of argument strength in attack graphs: Foundations and semantics. *Artificial Intelligence*, 302: 103607.

Amgoud, L.; and Prade, H. 2009. Using arguments for making and explaining decisions. *Artificial Intelligence*, 173(3-4): 413–436.

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards Artificial Argumentation. *AI Magazine*, 38(3): 25–36.

Baroni, P.; Caminada, M.; and Giacomin, M. 2011. An introduction to argumentation semantics. *Knowl. Eng. Rev.*, 26(4): 365–410.

Baroni, P.; Romano, M.; Toni, F.; Aurisicchio, M.; and Bertanza, G. 2015. Automatic evaluation of design alternatives with quantitative argumentation. *Argument Comput.*, 6(1): 24–49.

Becker, B.; and Kohavi, R. 1996. Adult dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20. Accessed: 2023-06-10.

Besnard, P.; and Hunter, A. 2001. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128(1-2): 203–235.

Brarda, M. E. B.; Tamargo, L. H.; and García, A. J. 2021. Using Argumentation to Obtain and Explain Results in a Decision Support System. *IEEE Intell. Syst.*, 36(2): 36–42.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building Classifiers with Independency Constraints. In *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA*, 13–18.

Chakraborty, J.; Peng, K.; and Menzies, T. 2020. Making Fair ML Software using Trustworthy Explanation. In *35th IEEE/ACM International Conference on Automated Software Engineering, ASE 2020, Melbourne, Australia*, 1229–1233.

Cocarascu, O.; Rago, A.; and Toni, F. 2019. Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada*, 1261–1269.

Cyras, K.; Letsios, D.; Misener, R.; and Toni, F. 2019. Argumentation for Explainable Scheduling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, 2752–2759.

Cyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada*, 4392–4399.

da Costa Pereira, C.; Tettamanzi, A.; and Villata, S. 2011. Changing One's Mind: Erase or Rewind? In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain*, 164–171.

Dung, P. M. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2): 321–358.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA*, 214–226. ACM.

Efstathiou, V. 2011. *Algorithms for computational argumentation in artificial intelligence*. Ph.D. thesis, University College London, UK.

Eğilmez, S.; Martins, J.; and Leite, J. 2014. Extending Social Abstract Argumentation with Votes on Attacks. In *Theory and Applications of Formal Argumentation*, 16–31. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-54373-9.

Fleisher, W. 2021. What's Fair about Individual Fairness? In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA*, 480–490.

Gabbay, D. M.; and Rodrigues, O. 2014. An Equational Approach to the Merging of Argumentation Networks. *Journal of Logic and Computation*, 24: 1253–1277.

Gabbay, D. M.; and Rodrigues, O. 2015. Equilibrium States in Numerical Argumentation Networks. *Logica Universalis*, 1–63.

Garg, P.; Villasenor, J. D.; and Foggo, V. 2020. Fairness Metrics: A Comparative Analysis. In *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA*, 3662–3666.

Gillingham, P. 2019. Decision support systems, social justice and algorithmic accountability in social work: A new challenge. *Practice*, 31(4): 277–290.

Haeri, M. A.; and Zweig, K. A. 2020. The Crucial Role of Sensitive Attributes in Fair Classification. In *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020, Canberra, Australia*, 2993–3002.

Hamon, R.; Junklewitz, H.; Sanchez, I.; Malgieri, G.; and de Hert, P. 2022. Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making. *IEEE Computational Intelligence Magazine*, 17(1): 72–85.

Hort, M.; Chen, Z.; Zhang, J. M.; Sarro, F.; and Harman, M. 2022. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. arXiv:2207.07068.

Hu, T.; Iosifidis, V.; Liao, W.; Zhang, H.; Yang, M. Y.; Ntoutsi, E.; and Rosenhahn, B. 2020. FairNN - Conjoint Learning of Fair Representations for Fair Decisions. In *Discovery Science - 23rd International Conference, DS 2020, Thessaloniki, Greece*, volume 12323 of *Lecture Notes in Computer Science*, 581–595.

Kamiran, F.; and Calders, T. 2011. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1): 1–33.

Kökciyan, N.; Parsons, S.; Sassoon, I.; Sklar, E.; and Modgil, S. 2020. An Argumentation-Based Approach to Generate Domain-Specific Explanations. In *Multi-Agent Systems and Agreement Technologies - 17th European Conference, EUMAS 2020, and 7th International Conference, AT 2020, Thessaloniki, Greece*, volume 12520 of *Lecture Notes in Computer Science*, 319–337.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed: 2023-09-12.

Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining Knowl. Discov.*, 12(3).

Leibe, B.; Mikolajczyk, K.; and Schiele, B. 2006. Efficient Clustering and Matching for Object Class Recognition. In *Proceedings of the British Machine Vision Conference 2006, Edinburgh, UK*, 789–798. British Machine Vision Association.

Leite, J.; and Martins, J. G. 2011. Social Abstract Argumentation. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain*, 2287–2292.

Lohia, P. K.; Ramamurthy, K. N.; Bhide, M.; Saha, D.; Varshney, K. R.; and Puri, R. 2019. Bias Mitigation Post-processing for Individual and Group Fairness. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, UK*, 2847–2851.

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

Moro, S.; Rita, P.; and Cortez, P. 2012. Bank Marketing, UCI Machine Learning Repository. https://doi.org/10.24432/C5K306. Accessed: 2023-06-10.

Mukherjee, D.; Yurochkin, M.; Banerjee, M.; and Sun, Y. 2020. Two Simple Ways to Learn Individual Fairness Metrics from Data. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 7097–7107. PMLR.

Northpointe. 2019. Practitioner's Guide to COMPAS Core. https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf. Accessed: 2023-09-12.

Potyka, N. 2018. Continuous Dynamical Systems for Weighted Bipolar Argumentation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona*, 148–157.

Pu, F.; Luo, J.; Zhang, Y.; and Luo, G. 2014. Argument Ranking with Categoriser Function. In *Knowledge Science, Engineering and Management - 7th International Conference, KSEM 2014, Sibiu, Romania*, volume 8793 of *Lecture Notes in Computer Science*, 290–301.

Rago, A.; Cocarascu, O.; and Toni, F. 2018. Argumentation-Based Recommendations: Fantastic Explanations and How to Find Them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden*, 1949–1955.

The Partnership on AI. 2019. Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System. https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/. Accessed: 2023-10-04.

Tilmes, N. 2022. Disability, fairness, and algorithmic bias in AI recruitment. *Ethics Inf. Technol.*, 24(2): 21.

Vassiliades, A.; Bassiliades, N.; and Patkos, T. 2021. Argumentation and explainable artificial intelligence: a survey. *Knowledge Engineering Review*, 36: e5.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden*, 1–7.

Wachter, S.; Mittelstadt, B. D.; and Russell, C. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Comput. Law Secur. Rev.*, 41: 105567.

Waller, M.; Rodrigues, O.; and Cocarascu, O. 2023a. Bias Mitigation Methods for Binary Classification Decision-Making Systems: Survey and Recommendations. arXiv:2305.20020.

Waller, M.; Rodrigues, O.; and Cocarascu, O. 2023b. Recommendations for Bias Mitigation Methods: Applicability and Legality. In *Aequitas 2023: Workshop on Fairness and Bias in AI, co-located with ECAI 2023, Kraków, Poland*.

Weinberg, L. 2022. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research (JAIR)*, 74: 75–109.

Zemel, R. S.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA3*, volume 28 of *JMLR Workshop and Conference Proceedings*, 325–333.