Neurocomputing 481 (2022) 33–45

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Modeling implicit bias with fuzzy cognitive maps

Gonzalo Nápoles [a,*], Isel Grau [b], Leonardo Concepción [c,d], Lisa Koutsoviti Koumeri [c], João Paulo Papa [e]

[a] Department of Cognitive Science & Artificial Intelligence, Tilburg University, The Netherlands
[b] Information Systems Group, Eindhoven University of Technology, The Netherlands
[c] Business Informatics Research Group, Hasselt University, Belgium
[d] Department of Computer Science, Central University of Las Villas, Cuba
[e] Department of Computing, São Paulo State University, Brazil

ARTICLE INFO

Article history:
Received 2 October 2021
Revised 12 December 2021
Accepted 18 January 2022
Available online 22 January 2022
Communicated by Zidong Wang

Keywords:
Fairness
Implicit bias
Fuzzy cognitive maps
Convergence analysis

ABSTRACT

This paper presents a Fuzzy Cognitive Map model to quantify implicit bias in structured datasets where features can be numeric or discrete. In our proposal, problem features are mapped to neural concepts that are initially activated by experts when running what-if simulations, whereas weights connecting the neural concepts represent absolute correlation/association patterns between features. In addition, we introduce a new reasoning mechanism equipped with a normalization-like transfer function that prevents neurons from saturating. Another advantage of this new reasoning mechanism is that it can easily be controlled by regulating nonlinearity when updating neurons' activation values in each iteration. Finally, we study the convergence of our model and derive analytical conditions concerning the existence and unicity of fixed-point attractors.

1. Introduction

Automated or algorithmic decision-making is nowadays used by organizations worldwide partly because machines seem neutral and unbiased [40,39]. Such an assumption is rather naive since bias can be encoded into historical data by capturing the discriminatory beliefs of people involved in the data generation process [27]. Artificial Intelligence (AI) algorithms then use the contaminated data to facilitate life-changing decisions such as accepting loan applications or suggesting appropriate medical treatments. Therefore, detecting bias in data used by Machine Learning (ML) systems is of paramount importance to the welfare of both society and individuals [17].

Bias in decision-making tasks can be expressed implicitly or explicitly [27]. Direct discrimination (or explicit bias) occurs when decisions are influenced by sensitive or protected features like gender, race, or marital status. Indirect discrimination (or implicit bias) occurs when decisions are influenced by non-sensitive features that strongly correlate with sensitive ones, thus resulting in non-favorable outcomes towards underprivileged groups. Implicit bias is also referred to in the literature as unconscious bias [21]. It expresses unintentional forms of discrimination that infect decision-making structurally and systematically and remains hard to address during data collection, labeling, or at an algorithmic level [21].

Available literature dedicated to discovering implicit bias is scarce compared to respective literature regarding explicit bias mainly because the former is harder to uncover [39,34]. We identify three main paradigms to discover and measure indirect discrimination: (i) statistical tests and correlation [39], (ii) causal models [26], and (iii) classification rules [17]. We sum up the limitations of these approaches that are relevant to our proposed model. Firstly, several methods require discretization of numeric features, which might differ per case and/or distort the information encoded in the data. Secondly, some approaches are still dependent on black-box machine learning models whose outputs are dependent on data pre-processing or the train-test split. Thirdly, all approaches consider interactions among features or feature category pairs selected by experts. However, experts might misjudge the impact of feature categories on the decision outcomes [16]. Finally, all approaches do not account for every possible interaction among the rest of the concepts within the system.

In previous work, we proposed a bias quantification measure based on granular computing, called fuzzy-rough uncertainty (FRU), for quantifying explicit bias [24]. FRU relies on fuzzy-rough sets to measure the extent to which the exclusion of protected features from a structured dataset changes the boundary regions. Positive changes mean that the boundary regions expanded themselves, which can be considered a proxy for explicit

* Corresponding author.
E-mail address: g.r.napoles@uvt.nl (G. Nápoles).

https://doi.org/10.1016/j.neucom.2022.01.070
0925-2312/

bias. FRU values measured after removing protected features from the data are not absolute and need to be interpreted concerning the respective FRU values of unprotected features. Extending this concept to detect implicit bias in [31], we hypothesize that a strong correlation[1] between a protected feature and unprotected ones coupled with a low FRU value of the protected feature in question can reveal which unprotected features might be (to some extent) implicitly biased.

Table 1 shows the possible scenarios concerning implicit and explicit bias, as discussed in [31]. A limitation of such a four-scenario approach is that one would need to analyze all possible pairwise feature combinations to study the implicit bias patterns in the data, and their interpretation might be too complex. Moreover, such a pairwise analysis would not suffice when bias is implicitly encoded by more than two features simultaneously or arises from feedback loops. To address this limitation, we propose a solution that considers interactions among all combinations of variables simultaneously.

In this paper, we introduce a model based on Fuzzy Cognitive Maps (FCMs) [23] to quantify the amount of bias towards protected features given initial conditions related to unprotected ones. In our FCM model, neural concepts denote protected or unprotected problem features, and weights comprise correlation patterns encoding the implicit bias. To run what-if simulations with this model, we activate neurons denoting (selected) unprotected features, execute the reasoning mechanism, and measure the amount of bias collected by neurons denoting protected features. Other contributions of our research concern theoretical challenges about FCM's reasoning and convergence that have remained unsolved until now. Firstly, we couple the quasi nonlinear reasoning rule proposed in [32] with a normalization-like transfer function that promotes competition among the neurons in each iteration. Such a quasi nonlinear reasoning rule involves a parameter controlling the nonlinearity degree of the model. Secondly, we analytically study the convergence properties of our model when it comes to the existence and uniqueness of the fixed point. Overall, we derive conditions guaranteeing (i) that the fixed point exists and is unique, or (ii) that the fixed point cannot be unique. The former is useful when modeling invariant situations that do not depend on the initial conditions, while the latter is useful when running what-if simulations. Section 3 further motivates these theoretical contributions after having introduced the FCM formalism.

The rest of this paper is organized as follows. Section 2 revisits the literature devoted to detecting and quantifying indirect (implicit) bias in structured datasets. Section 3 presents the theoretical background concerning fuzzy cognitive modeling. Section 4 introduces the proposed FCM-based model and the new reasoning mechanism, while Section 5 elaborates on the convergence properties of our reasoning mechanism and derives analytical conditions related to fixed-point attractors. Section 6 presents the simulation results using a well-known case study, and Section 7 states concluding remarks.

2. Related works on bias detection and quantification

In this section, we briefly introduce the main paradigms used to discover indirect discrimination (or implicit bias) in the literature and mention their limitations.

The most popular approach to discover implicit bias comes from the field of statistics using regression models or correlation coefficients [39,35]. However, such methods are not easily generalizable [39], and higher-order interactions between many features and the target are hard to interpret. The authors in [4] also mention that state-of-the-art methods are mainly correlation-based while stressing out that correlation could be spurious and does not necessarily imply causation. On the one hand, statistical methods can be translated to algorithmic discrimination measures and optimization constraints [39]. On the other hand, one might wonder whether modeling implicit bias must have causal semantics.

Table 1
Scenarios relating correlation and FRU values.

	Large FRU value	Small FRU value
Strong correlation	Explicit & Implicit bias	Implicit bias
Weak correlation	Explicit bias	Safe scenario

A second approach to discover implicit bias relies on generating classification rules learned from the data as reported in [17]. In such a work, the authors use predetermined discriminatory items (feature category pairs) and the concepts of support and confidence of classification rules to quantify and correct implicit and explicit bias. Their measure computes the correlation between predefined and non-discriminatory items to identify redlining or biased rules. This approach has two potential limitations. First, the decision variable and protected features must be binary or nominal, requiring numeric features to be discretized. Second, manually predetermining the sensitive feature categories might be risky as human experts might misjudge their impact on the decision outcomes [16].

The authors in [1] introduce a method to quantify indirect influence using black-box prediction models. First, they estimate the information content of a feature by predicting it from the remaining features. Then, they modify or obscure the feature so that it can no longer be predicted from the data and, finally, analyze the difference between the original feature and the modified one. Possible limitations of this approach are their measure's dependency on the data distribution, the need to discretize numeric attributes, and the black-box manner in which predictions are made. The authors applied their suggested method to the German Credit dataset [12] and concluded that implicit bias is evident in the features *credit amount*, *checking status* and *existing credits*.

A third approach to discover implicit bias is based on causal networks [26]. Causal models assume that implicit bias is transmitted as causal effects along paths that pass through proxy attributes [38]. Zhang et al. [38] attempt to discover implicit bias in historical data by coupling causal networks and probabilistic inference. They use the path-specific effects as a proxy to model bias. The authors in [4] also use probabilistic causal theory to detect implicit bias using the German Credit dataset. They examine the existence of implicit bias between three protected features (*gender*, *age* and *foreign worker*) and five unprotected features (*employed since* being the only one that is also considered in our approach), but report that discrimination patterns are less palpable partly because this dataset contains less proper causal relations. Their conclusion states that automatically exploring the whole space of possible discrimination patterns is an ongoing research challenge.

Causal paths are also explored by Chiappa in [6] combined with the notion of counterfactual fairness in order to compute indirect causal effects. In order to eliminate unfair information, the authors apply a correction that is suitable to complex nonlinear models, and its main limitation is that the data-generation process needs to be provided. In their numerical simulations using the German Credit dataset, they conclude that there is evidence of implicit bias between the features *housing* and *gender* which aligns with our findings.

In conclusion, the causal approach considers one-way interactions among a limited number of features or feature categories, whereas an FCM-based approach using a square correlation matrix

[1] For simplicity purposes, the term correlation simultaneously represents three different measures chosen based on variable type (numeric or nominal) described in Section 4.2: Pearson correlation, strength of association and R-squared coefficient of determination.

as weights accounts for interactions among all features simultaneously. This is also the case with the classification rule-based approach. Moreover, both approaches use the target variable during the process of measuring indirect bias in contrast to our FCM-based approach, where information flow among variables is modeled without being affected by the possibly partial decision feature.

3. Fuzzy cognitive maps

FCMs were introduced in [23] as a knowledge-based method for modeling complex systems and running what-if simulations. Such interpretable recurrent neural networks consist of neural concepts and weighted connections. Neural concepts represent variables, entities, or states related to the physical system under investigation. The signed weight associated with each connection denotes the strength of the causality (or simply correlation) between the corresponding variables. Weights are quantified in the $[-1, 1]$ interval, while the neurons' activation values can take values in either $[0, 1]$ or $[-1, 1]$ depending on the nonlinear transfer function attached to each neural unit.

In each iteration, FCM-based models update the activation values of neural concepts based on the neurons' states in the previous iteration and the weight matrix. Overall, the recurrent reasoning process allows several times to propagate a given initial condition across the whole network, which might be necessary to model scenarios with feedback loops.

Eq. (1) shows the reasoning rule to compute the activation value of the i-th neural concept in the $(t+1)$-th iteration for a given initial condition provided by the domain expert when running what-if simulations:

$$a_i^{(t+1)} = f\left(\sum_{j=1}^{M} a_j^{(t)} w_{ij}\right), \tag{1}$$

where M is the number of neurons and w_{ij} is the weight connecting the C_i and C_j neurons, while $f(\cdot)$ is the transfer function used to keep the neurons' activation values within the allowed activation interval. The sigmoid function below is often used as a transfer function in neural systems:

$$f(x) = \frac{1}{1 + e^{-\lambda x}}, \tag{2}$$

such that $\lambda > 0$ is the inclination parameter to control the function steepness. Larger values of λ cause the sigmoid function to resemble a binary activator. It is worth mentioning that we can use other bounded, continuous transfer functions such as the hyperbolic tangent function, but we need to re-evaluate how to interpret the simulation results.

The reasoning rule stops when either (i) the model converges to a fixed point or (ii) a maximal number of iterations T is reached. Overall, we have three possible states:

- **Fixed point** $\left(\exists t_\alpha \in \{1, \ldots, (T-1)\} : a_i^{(t+1)} = a_i^{(t)}, \forall i, \forall t \geqslant t_\alpha\right)$: the FCM produces the same state vector after t_α, thus $a_i^{(t_\alpha)} = a_i^{(t_\alpha+1)} = a_i^{(t_\alpha+2)} = \ldots = a_i^{(T)}$. If the fixed point is unique, the FCM model will produce the same state vector regardless of the initial conditions. The unique fixed-point attractor allows modeling invariant situations rather than performing what-if simulations.
- **Limit cycle** $\left(\exists t_\alpha, P, j \in \{1, \ldots, (T-1)\} : a_i^{(t+P)} = a_i^{(t)}, \forall i, \forall t \geqslant t_\alpha\right)$: the FCM produces the same state vector periodically after the period P, thus $a_i^{(t_\alpha)} = a_i^{(t_\alpha+P)} = a_i^{(t_\alpha+2P)} = \ldots = a_i^{(t_\alpha+jP)}$, where $t_\alpha + jP \leqslant T$.
- **Chaos**: the FCM produces different state vectors for successive iterations with no clear pattern.

The convergence of FCM-based models is of paramount importance when running what-if simulations. However, we have noticed that most published papers using FCMs to model complex systems rarely touch upon the practical implications of having multiple or unique fixed points, or cyclic patterns. There could be several reasons for such a gap. Firstly, the nonlinearity of FCM-based models operating as closed systems makes them difficult to control. Secondly, the most prominent theoretical results reported in the literature [5,20,22] are primarily devoted to deriving sufficient conditions ensuring the existence and uniqueness of the fixed point. These conditions might be quite strict and difficult to fulfill in reality. Moreover, they have limited usability if we want to perform what-if simulations where the outputs should not remain static.

Another issue that emerges when using bounded transfer functions and large, densely connected FCM-based models is that the neuron's activation values often locate either in the lower or upper boundary of the activation interval. Hence, the neurons tend to *saturate* when receiving a large negative or positive information flow. One way to address this drawback is to modify the reasoning rule to promote competition among the neurons in each iteration. Finally, some transfer functions such as the sigmoid function can distort the simulation results by activating neurons that are not supposed to be active given the information flow they receive.

In the following subsection, we overcome these drawbacks and propose a sound FCM-based model to perform simulations concerning the impact of the implicit bias over protected features in pattern classification problems.

4. Modeling implicit bias using cognitive modeling

This section presents our FCM-based simulation model to quantify the amount of implicit bias towards protected features given initial conditions imposed by (selected) unprotected features. Firstly, we further discuss the problem of quantifying implicit bias and introduce the intuition of our model. Secondly, we describe how to build the FCM structure using the correlation/association patterns between problem features. Thirdly, we introduce a rescaled transfer function coupled with a parametric quasi nonlinear reasoning rule to prevent saturation and convergence issues from happening.

4.1. Motivating our FCM-based model

One of the difficulties in quantifying implicit bias is that protected features can be affected by seemingly unrelated features through hidden patterns. For example, the postal code might influence whether or not an applicant gets a loan because people living in more expensive neighborhoods might be more likely to receive the loan. Such hidden patterns result from complex dependencies, correlations, associations, causalities, and feedback loops that are difficult to detect by inspecting the data. This suggests that implicit bias should be modeled as a *complex system* involving feedback loops and non-trivial dependencies. Note that having feedback loops justifies using recurrent neural networks even when the target problem does not involve sequences or an explicit time component.

In our paper, modeling implicit bias means building an FCM-based model to quantify the bias amount for protected features given initial conditions associated with unprotected ones. The model will enable performing what-if simulations with the form: "What is the amount of bias attached to the protected feature $y \in Y$ if x_1 is $a_1^{(0)}, x_2$ is $a_2^{(0)}, \ldots$, and x_n is $a_n^{(0)}$, such that Y is the set

of protected features and X is the set of unprotected ones. The initial conditions $a_1^{(0)}, a_2^{(0)}, \ldots, a_n^{(0)}$ should be determined by domain experts when running the simulations. To perform such simulations, we need to ensure that the fixed point is not unique. At the same time, we might want to quantify the amount of bias towards a protected feature regardless of these initial conditions, just relying on the weight matrix. By doing so, we need to ensure that the fixed point is unique to prevent that the FCM-based model converges to a different fixed point as the initial conditions change.

4.2. Building the correlation network

The first step towards modeling implicit bias using cognitive mapping is to determine the network structure (that is to say, the concepts and the weight matrix). Each problem feature is denoted with a neural concept in our neural reasoning system regardless of whether the feature is nominal or continuous. The initial activation values of these neurons represent the extent to which the corresponding feature is active in the model. Initial activation values are defined by the experts when running what-if simulations. In contrast, the neurons' activation values resulting from the recurrent reasoning process (see the following subsection) represent the amount of implicit bias given the initial condition provided by the domain expert.

The weights connecting the neurons represent correlation patterns to be derived from the dataset being analyzed. The intuition behind these weights is that if a protected feature correlates with an unprotected one, then part of the bias will be encoded in the unprotected feature. Hence, capturing the information that arrives at a neuron representing a protected feature given an initial condition is reliable for detecting implicit bias. More importantly, our approach detects implicit bias considering non-trivial relationships since the interaction among variables is modeled as a complex system.

The weight matrix is obtained using three measures that are conceptually similar (i.e. based on Pearson's correlation coefficient) in an effort to preserve consistency. Let C_i and C_j be two neural concepts in the network. If both C_i and C_j represent numeric features, then w_{ij} will be given by the Pearson correlation coefficient [33] (even though features in German Credit dataset do not meet the assumptions of normality, linearity or homoscedasticity). If both C_i and C_j represent nominal features, then w_{ij} will denote the association strength between these variables as computed by the Cramér's V [10] which is based on Pearson's chi-squared statistic [3]. If C_i represents a nominal feature and C_j represents a numeric feature, then w_{ij} is the R-squared coefficient of determination [29] which is the square of the Pearson's correlation. This measure quantifies the percentage of variation in the numeric feature that is explained by the nominal one coupled with an F-test of joint significance [25]. Notice that the weight matrix is symmetric such that $w_{ij} = w_{ji}$. Finally, we compute the absolute values of these weights since we are not interested in the correlation signs but their intensity. It is important to state that in this paper, the term correlation simultaneously represents all the three measures described above for simplicity.

4.3. Quasi nonlinear reasoning rule

As mentioned earlier, the reasoning rule in Eq. (1) often converges to unique fixed-point attractors or suffers from saturation problems. The former could be useful when modeling invariant situations, but it is less attractive when performing what-if simulations with some generalization. For example, if we want to design an FCM-based model to solve a regression-like problem and the model converges to a unique fixed-point attractor, the predicted value will be the same regardless of the inputs. Moreover, when having dense FCM-based models, neurons often produce saturated activation values even when their inputs are significantly different. In this paper, we couple the quasi nonlinear reasoning rule recently proposed in [32] with a re-scaled transfer function that performs a normalization operation. The resulting reasoning model solves the issues mentioned above while producing relative activation values, thus making the model more intuitive and easy to control.

Let $\mathbf{A}^{(t)} = \left(a_1^{(t)}, \ldots, a_i^{(t)}, \ldots, a_M^{(t)}\right)$ be the activation vector produced by an FCM such that $a_i^{(t)}$ is the activation value of the i-th neuron at the iteration t. Similarly, let $\bar{\mathbf{A}}^{(t)} = \left(\bar{a}_1^{(t)}, \ldots, \bar{a}_i^{(t)}, \ldots, \bar{a}_M^{(t)}\right)$ be the raw activation vector where $\bar{a}_i^{(t)}$ is the raw activation value of the i-th neuron in the current iteration. In other words, the vector $\bar{\mathbf{A}}^{(t)}$ is the argument of the transfer function and can be calculated as $\bar{\mathbf{A}}^{(t)} = \mathbf{A}^{(t)}\mathbf{W}$ where $\mathbf{W}_{M\times M}$ is the weight matrix defining the interaction among the neurons. The quasi nonlinear reasoning rule using the re-scaled transfer function can be computed as follows:

$$a_i^{(t+1)} = \underbrace{\phi f\left(\sum_{j=1}^{M} a_j^{(t)} w_{ij}\right)}_{\text{nonlinear component}} + \underbrace{(1-\phi)a_i^{(0)}}_{\text{linear component}}, \tag{3}$$

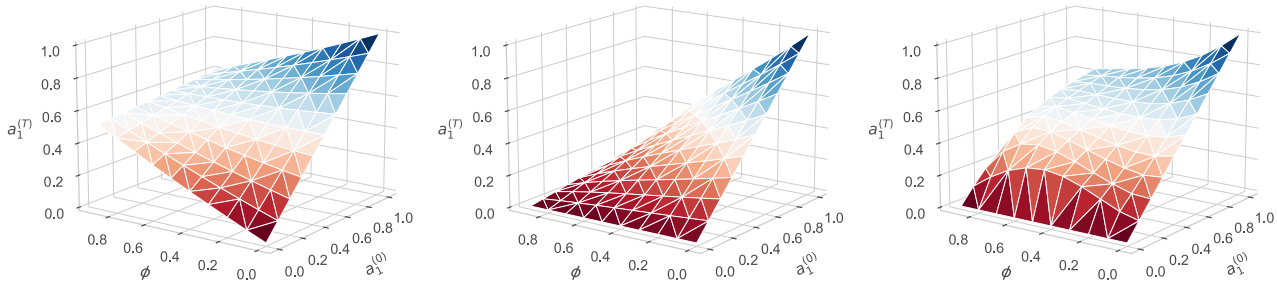
where

$$f\left(\bar{a}_i^{(t+1)}\right) = \frac{\bar{a}_i^{(t+1)}}{\left\|\bar{\mathbf{A}}^{(t+1)}\right\|_2}, \tag{4}$$

such that $||\cdot||_2$ stands for the Euclidean norm, and $0 \leqslant \phi \leqslant 1$ controls the nonlinearity of the reasoning rule. When $\phi = 1$, the model performs as a closed system where the activation value of a neuron depends on the activation values of connected neurons in the previous iteration. When $0 < \phi < 1$, we add a linear component to the reasoning rule devoted to preserving the initial activation values of neurons when updating their activation values in the current iteration. When $\phi = 0$, the model narrows down to a linear regression where the initial activation values of neurons act as regressors. It should be mentioned that, when the denominator of the nonlinear component is zero and $\phi \neq 0$, then the neuron's activation value will be given by $(1-\phi)a_i^{(0)}$. Two reasons move this rule: (i) we need to avoid indetermination issues in the denominator, and (ii) we want to prevent situations in which the neurons receiving no information flow at all get arbitrarily activated by the transfer function. For example, the sigmoid transfer function produces 0.5 when its argument is zero, leading to misleading simulation results.

Let us simulate to illustrate the behavior of the sigmoid function, the hyperbolic tangent function, and the new re-scaled transfer function. In this experiment, we couple these transfer functions with the quasi nonlinear reasoning rule introduced in Eq. (3). Moreover, we use an FCM model comprised of two neurons with weights $w_{12} = 0.8, w_{21} = 0.5$, and zeros in the main diagonal. Fig. 1 illustrates the activation values of the first neuron when activating it with random numbers in the $[0, 1]$ interval and without activating the second one. The simulation results show that the sigmoid neuron is activated when its input is zero, while the hyperbolic tangent and the re-scaled transfer functions keep the neuron inactive. It can also be noticed that the surface generated with the re-scaled function is smoother than the hyperbolic tangent.

Another attractive property of the quasi nonlinear reasoning rule using the re-scaled transfer function is that activation values are not absolute but relative. In the model, the neurons' activation values depend not on the neurons' activation values in the previous iteration, the weights, and the initial conditions but the neurons'



(a) Sigmoid activation function (b) Hyperbolic activation function (c) Re-scaled activation function

Fig. 1. Final activation value (after performing $T = 30$ iterations) of a single neuron using the quasi nonlinear reasoning rule with different transfer functions. In this simulation, we vary the nonlinearity degree and the raw activation value of the first neuron without activating the second one.

activation values in the current iteration. Overall, dividing each raw activation value by the Euclidean norm of the activation vector ensures that activation values are in the $[0, 1]$ interval while promoting competition among the neurons. In the following section, we will mathematically study the convergence properties of the quasi nonlinear reasoning rule and the proposed re-scaled transfer function.

5. Convergence analysis

As mentioned in Section 3, FCM-based models can reach three possible states when it comes to their dynamic behavior: (i) fixed-point, (ii) limit cycle, or (iii) chaos. Likewise, these states apply to our model. In this section, we aim for convergence to diverse fixed-point attractors and to elude cycles, chaos, and unique fixed-point attractors.

A few FCM-based models have been mathematically studied in the literature with regard to their convergence properties. The existence and uniqueness of fixed points of sigmoid FCMs was firstly discussed by Boutalis et al. [5] when the transfer function is given by $f(x) = 1/(1 + e^{-x})$. Other theoretical results about this topic are the ones reported in Knight et al. [22], and Harmati et al. [19]. As a generalization of the stability and instability properties of FCM-based models, in [30] we introduced the definitions of E-stability and E-instability, and four sufficient conditions to fulfill these properties.

In [7], the authors analyzed the behavior of FCMs from the perspective of their state spaces. More specifically, the research estimate bounds for the activation values of each neuron. Moreover, it was analyzed the coverage (i.e., the proportion of the induced activation space that is achievable by the neurons' activation values) and the proximity (i.e., the mean relative distance of the neurons' activation values at the borders of the feasible activation space) of neural concepts. The main theoretical result shows that the state space of any FCM equipped with an F-transfer function [7] shrinks infinitely with no guarantees of reaching a fixed point, but it does converge to the limit state space. Finally, in [8] it was proven that in an FCM-based model, a sigmoid neuron with self-feedback and no other incoming connection will always converge to a unique fixed-point regardless of its initial stimulus.

As a further addition to this research line, we will study the convergence features of the proposed reasoning mechanism. First, we establish the requirements for the appearance of cycles while examining some special situations. Then, we analyze the model conditions to avoid unique fixed-point attractors and the conditions to reach them.

5.1. Basic notions and definitions

Before presenting the main theoretical results devoted to the convergence of our model, let us re-write Eqs. (3) and (4) using the following matrix-like notation:

$$\mathbf{A}^{(t+1)} = \phi f\left(\mathbf{A}^{(t)}\mathbf{W}\right) + (1-\phi)\mathbf{A}^{(0)} \tag{5}$$

such that $f(.) : \mathbb{R}^M \rightarrow \mathbb{R}^M$is defined as follows:

$$f(\mathbf{X}) = \begin{cases} \frac{\mathbf{X}}{||\mathbf{X}||_2} & \text{if} X \neq \overrightarrow{0} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

In our FCM-based model, we need to ensure that activation values lie in the $[0, 1]$ interval. Next, we will prove that $A^{(t+1)} \in [0, 1]^M$ by mathematical induction.

Proof. The fact that $\mathbf{A}^{(0)} \in [0, 1]^M$ can be considered a restriction of the proposed neural reasoning model. Let us assume that $\mathbf{A}^{(t)} \in [0, 1]^M$ and prove that $\mathbf{A}^{(t+1)} \in [0, 1]^M$.

Let us define that $\mathbf{A}^{(t)}\mathbf{W} = \mathbf{V}$ such that $\mathbf{V} = (v_1, v_2, \ldots, v_M) \in \mathbb{R}^M$. Having said that $\mathbf{A}^{(t)} \in [0, 1]^M$ and that the elements of $\mathbf{W}$ are in $[0, 1]$, the vector–matrix product produces non-negative components. Then, we will demonstrate that $f(\mathbf{V}) \in [0, 1]^M$ and that there are two cases. On the one hand, having $\mathbf{V} = \overrightarrow{0}$ implies that $f(\mathbf{V}) = \overrightarrow{0}$. On the other hand, having that $\mathbf{V} \neq \overrightarrow{0}$, we will demonstrate that $\frac{\mathbf{V}}{||\mathbf{V}||_2} \in [0, 1]^M$. Since we are using the Euclidean norm, we have that:

$$\frac{\mathbf{V}}{||\mathbf{V}||_2} = \frac{(v_1, v_2, \ldots, v_M)}{\sqrt{\sum_{i=1}^{M} v_i^2}} = \frac{(v_1, v_2, \ldots, v_M)}{\sqrt{v_1^2 + v_2^2 + \ldots + v_M^2}}$$

It can be verified that, for every v_i, the inequality

$$0 \leqslant \frac{v_i}{\sqrt{\sum_{i=1}^{M} v_i^2}} \leqslant 1$$

holds and $\frac{\mathbf{V}}{||\mathbf{V}||_2} \in [0, 1]^M$. Hence, we have that $\mathbf{A}^{(t+1)} = \phi f(\mathbf{V}) + (1-\phi)\mathbf{A}^{(0)}$, with $f(\mathbf{V}), \mathbf{A}^{(0)} \in [0, 1]^M$. Since $0 \leqslant \phi \leqslant 1$, the right-hand side yields a non-negative vector that reaches its maximum (component-wise) when $f(\mathbf{V}) = \mathbf{A}^{(0)} = (1, 1, \ldots, 1)$. Such a value for $\phi f(\mathbf{V}) + (1-\phi)\mathbf{A}^{(0)}$ is equal to $\phi(1, 1, \ldots, 1) + (1-\phi)(1, 1, \ldots, 1) = (1, 1, \ldots, 1)$. Finally, we obtain that $\mathbf{A}^{(t+1)} \in [0, 1]^M$ and this concludes our proof. □

5.2. Cyclic behavior analysis

As we mentioned earlier, the evasion of cycles is desired and we need to know when our system falls into them. Generally speaking, a cycle is reached when:

$\mathbf{A}^{(t+v)} = \mathbf{A}^{(t)}$ for any $v \in \mathbb{N} : v > 0$.

It should be noticed that when $v = 1$, we fall into a special kind of cycle: *convergence*. If we expand this definition using the proposed reasoning rule, we have:

$$\phi f\left(\mathbf{A}^{(t+v-1)}\mathbf{W}\right) + (1-\phi)\mathbf{A}^{(0)} = \phi f\left(\mathbf{A}^{(t-1)}\mathbf{W}\right) + (1-\phi)\mathbf{A}^{(0)},$$

such that

$$\phi f\left(\mathbf{A}^{(t+v-1)}\mathbf{W}\right) = \phi f\left(\mathbf{A}^{(t-1)}\mathbf{W}\right).$$

At this point, we can put aside the case when $\phi = 0$, since it implies that $\mathbf{A}^{(t+1)} = \mathbf{A}^{(t)}, \forall t$, so convergence is attained. Simplifying on ϕ leads to the following:

$$f\left(\mathbf{A}^{(t+v-1)}\mathbf{W}\right) = f\left(\mathbf{A}^{(t-1)}\mathbf{W}\right). \tag{7}$$

Firstly, let us analyze the cases when $\mathbf{A}^{(t+v-1)}\mathbf{W} = \overrightarrow{0}$ and $\mathbf{A}^{(t-1)}\mathbf{W} = \overrightarrow{0}$. In general, it holds that, if $\mathbf{A}^{(t-1)}\mathbf{W} = \overrightarrow{0}$, then $\mathbf{A}^{(t)} = (1-\phi)\mathbf{A}^{(0)}$ and $\mathbf{A}^{(t+1)} = \mathbf{A}^{(1)}$. Such a cycle is found by simplifying the following expression:

$$\mathbf{A}^{(t+1)} = \phi f\left((1-\phi)\mathbf{A}^{(0)}\mathbf{W}\right) + (1-\phi)\mathbf{A}^{(0)}.$$

In contrast, if $\mathbf{A}^{(0)}W \neq \overrightarrow{0}$ and $\phi < 1$ (otherwise, the cycle would have appeared earlier), then we have:

$$\mathbf{A}^{(t+1)} = \phi \frac{(1-\phi)\mathbf{A}^{(0)}\mathbf{W}}{||(1-\phi)\mathbf{A}^{(0)}\mathbf{W}||_2} + (1-\phi)\mathbf{A}^{(0)}.$$

Since the norm of a vector multiplied by a scalar is equal to the scalar multiplied by the norm, we have that

$$\mathbf{A}^{(t+1)} = \phi \frac{(1-\phi)\mathbf{A}^{(0)}W}{(1-\phi)||\mathbf{A}^{(0)}W||_2} + (1-\phi)\mathbf{A}^{(0)},$$

and

$$\mathbf{A}^{(t+1)} = \phi \frac{\mathbf{A}^{(0)}\mathbf{W}}{||\mathbf{A}^{(0)}W||_2} + (1-\phi)\mathbf{A}^{(0)} = \mathbf{A}^{(1)}.$$

Next, we can use the nonzero branch of Eq. (6) to develop Eq. (7) as follows:

$$\frac{A^{(t+v-1)}W}{||A^{(t+v-1)}W||_2} = \frac{A^{(t-1)}W}{||A^{(t-1)}W||_2}.$$

Finally, we obtain that:

$$\left(\frac{\mathbf{A}^{(t+v-1)}}{||\mathbf{A}^{(t+v-1)}\mathbf{W}||_2} - \frac{\mathbf{A}^{(t-1)}}{||\mathbf{A}^{(t-1)}\mathbf{W}||_2}\right)\mathbf{W} = 0.$$

The satisfaction of this equality is difficult but still possible. This situation may occur: the vector resulting as the difference of the fractions is perpendicular to every column of $\mathbf{W}$ or such a difference is equal to $\overrightarrow{0}$. The former is less likely to happen, while the latter has a slightly higher probability of occurrence. This more probable case turns out when

$$\frac{\mathbf{A}^{(t+v-1)}}{||\mathbf{A}^{(t+v-1)}\mathbf{W}||_2} = \frac{\mathbf{A}^{(t-1)}}{||\mathbf{A}^{(t-1)}\mathbf{W}||_2},$$

and

$$\mathbf{A}^{(t+v-1)} = \frac{||\mathbf{A}^{(t+v-1)}\mathbf{W}||_2}{||\mathbf{A}^{(t-1)}\mathbf{W}||_2}\mathbf{A}^{(t-1)}.$$

The above means that $\mathbf{A}^{(t+v-1)}$ is equal to $\mathbf{A}^{(t-1)}$ multiplied by a scalar factor, which is the quotient of the Euclidean norms. Beyond the mathematical meaning, it implies that if an activation vector is multiple of another for the same initial stimulus, then the cyclic behavior appears.

5.3. Unique fixed-points analysis

When performing what-if simulations, it is desirable not to converge to unique fixed-point attractors. In these situations, reaching the same output regardless of the initial conditions must be avoided. Next, we analyze how the model parameters influence the system stability properties. To do that, we split the task into two cases: $0 \leqslant \phi < 1$ and $\phi = 1$.

Case 1: $0 \leqslant \phi < 1$. Our goal is to prove that there is no unique fixed-point attractor for all initial stimuli. Furthermore, in the process we found other interesting results.

Aiming at obtaining a more general result, let us define $g(.) : \mathbb{R}^M \to \mathbb{R}^M$ as a continuous transfer function for our model. Also, the model's intrinsic restrictions for $\mathbf{W}$ (symmetry and weights in $[0, 1]$) and $\mathbf{A}^{(t)}$ (activation values in $[0, 1]$) are not required for our demonstration. This means that we will use real matrices and activation vectors whose only requirement is that the product $\mathbf{A}^{(t)}\mathbf{W}$ must be defined. Thus, the generic reasoning rule can be rewritten once again as follows:

$$\mathbf{A}^{(t+1)} = \phi g\left(\mathbf{A}^{(t)}\mathbf{W}\right) + (1-\phi)\mathbf{A}^{(0)} \tag{8}$$

Theorem 1. *(Injective convergence). In an FCM model using Eq. (8), when $0 \leqslant \phi < 1$, there are not two different initial stimuli leading to the same fixed-point attractor.*

Proof. Let $\mathbf{A}_i^{(0)}, \mathbf{A}_j^{(0)} \in \mathbb{R}^M$ denote two unequal initial activation vectors and also assume that both lead to the same fixed-point attractor $\mathbf{V}$. Leaning on the mathematical analysis, we define that $\mathbf{V} = \lim_{t\to\infty}\mathbf{A}_i^{(t)} = \lim_{t\to\infty}\mathbf{A}_j^{(t)}$.

The activation vectors $\mathbf{A}_i^{(0)}$ and $\mathbf{A}_j^{(0)}$ can be inserted in the Eq. (8), producing the following formulas:

$$\mathbf{A}_i^{(t+1)} = \phi g\left(\mathbf{A}_i^{(t)}\mathbf{W}\right) + (1-\phi)\mathbf{A}_i^{(0)},$$

and

$$\mathbf{A}_j^{(t+1)} = \phi g\left(\mathbf{A}_j^{(t)}\mathbf{W}\right) + (1-\phi)\mathbf{A}_j^{(0)}.$$

Also, after applying the limit when $t \to \infty$, we have that:

$$\lim_{t\to\infty}\mathbf{A}_i^{(t+1)} = \phi \lim_{t\to\infty} g\left(\mathbf{A}_i^{(t)}W\right) + (1-\phi)\mathbf{A}_i^{(0)},$$

and

$$\lim_{t\to\infty}\mathbf{A}_j^{(t+1)} = \phi \lim_{t\to\infty} g\left(\mathbf{A}_j^{(t)}\mathbf{W}\right) + (1-\phi)\mathbf{A}_j^{(0)}.$$

Given the continuity of $g(.)$ and using the basic properties of limit operations, it yields:

$$\lim_{t\to\infty}\mathbf{A}_i^{(t+1)} = \phi g\left(\lim_{t\to\infty}\mathbf{A}_i^{(t)}\mathbf{W}\right) + (1-\phi)\mathbf{A}_i^{(0)},$$

and

$$\lim_{t\to\infty}\mathbf{A}_j^{(t+1)} = \phi g\left(\lim_{t\to\infty}\mathbf{A}_j^{(t)}\mathbf{W}\right) + (1-\phi)\mathbf{A}_j^{(0)}.$$

Now, substituting all limits by its value and solving on $\mathbf{A}_i^{(0)}$ and $\mathbf{A}_j^{(0)}$ respectively, we obtain:

$$\mathbf{A}_i^{(0)} = \frac{\mathbf{V} - \phi g(\mathbf{VW})}{1-\phi},$$

and

$$\mathbf{A}_j^{(0)} = \frac{\mathbf{V} - \phi g(\mathbf{VW})}{1-\phi}.$$

Note that the formulas above are not undefined since $\phi \neq 1$. At this point, the assumption $\mathbf{A}_i^{(0)} \neq \mathbf{A}_j^{(0)}$ leads us to $\mathbf{A}_i^{(0)} = \mathbf{A}_j^{(0)}$. Therefore, we reach a contradiction and the proof by *reductio ad absurdum* is completed. □

Theorem 1 brings a straightforward consequence, which is portrayed in the following corollary.

Corollary 1.1. *(Relaxed injective convergence). In an FCM model using Eq. (8), when $0 \leqslant \phi < 1$, there is no unique fixed-point attractor for all initial stimuli.*

Theorem 1 and Corollary 1.1 are valid when using the Eq. (8), but such a general equation differs from Eq. (5). Hence, we need to analyze the continuity of $f(.)$.

On the one hand, the first branch concerning $\frac{\mathbf{x}}{||\mathbf{x}||_2}$ is continuous since it is a elementary function that never gets undefined when $\mathbf{X} \neq \overrightarrow{0}$. On the other hand, $f\left(\overrightarrow{0}\right) = \overrightarrow{0}$, and for this branch the continuity also requires that:

$$\lim_{\mathbf{x}\to\overrightarrow{0}} f(\mathbf{X}) = \overrightarrow{0}.$$

Next, we will prove that the previous expression does not hold by using the limit on $\mathscr{D} = \{(u,u,\ldots,u) \in \mathbb{R}^M : u \neq 0\}$, such that $\mathscr{D} \subset Domf(.)$ and $Domf(.)$ is the domain of $f(.)$. Then, if $\mathbf{X} \in \mathscr{D}$ we have that the following:

$$f(X) = \frac{(u,u,\ldots,u)}{\sqrt{u^2+u^2+\ldots+u^2}} = \frac{(u,u,\ldots,u)}{u\sqrt{M}}.$$

Now, the limit on $\mathscr{D}$ is calculated as follows:

$$\lim_{\mathbf{x}\to\overrightarrow{0}} f(\mathbf{X}) = \lim_{u\to 0}\frac{(u,u,\ldots,u)}{u\sqrt{M}} = \left(\frac{1}{\sqrt{M}},\frac{1}{\sqrt{M}},\ldots,\frac{1}{\sqrt{M}}\right) \neq \overrightarrow{0},$$

which implies that $f(.)$ is not continuous at $\overrightarrow{0}$.

As we know from the cyclic behavior analysis in Section 5.2, if after some initial stimulus we get $\mathbf{A}^{(t)}\mathbf{W} = \overrightarrow{0}$ for some t, then our model will exhibit cyclic behavior. Similarly, if $f(.)$ is evaluated on the discontinuity point $\overrightarrow{0}$ during the inference process, a cycle will be reached. Since the Theorem 1 and the Corollary 1.1 conclude on the nonexistence of stimuli (more than one) leading to a fixed-point attractor, then the cyclic behavior is consistent with such a conclusion.

Finally, the theorem and the corollary are also valid using the Eq. (5) instead of Eq. (8). The following theorem and corollary summarize the results exposed.

Theorem 2. *In an FCM model using the Eq. (5), when $0 \leqslant \phi < 1$ there is no couple of different initial stimuli leading to the same fixed-point attractor.*

Corollary 2.1. *In the model described by the reasoning rule in the Eq. (5), when $0 \leqslant \phi < 1$ there is no unique fixed-point attractor for all initial stimuli.*

Summarizing, the obtained results for the case $0 \leqslant \phi < 1$ allow performing what-if simulations without the concerns of converging to a unique fixed point.

Case 2: $\phi = 1$. We establish the equivalence between our reasoning rule and the *power iteration method* [28] by using matrix properties. This leads us to interesting conclusions about the convergence of our model. For the case in point, the reasoning rule is reduced to the following expression:

$$\mathbf{A}^{(t+1)} = f\left(\mathbf{A}^{(t)}\mathbf{W}\right). \tag{9}$$

We could exclude the case when $\mathbf{A}^{(t)}\mathbf{W} = \overrightarrow{0}$ for some t, for it causes the model to converge to $\overrightarrow{0}$. Therefore, we rewrite the previous formula as follows:

$$\mathbf{A}^{(t+1)} = \frac{\mathbf{A}^{(t)}\mathbf{W}}{||\,\mathbf{A}^{(t)}\mathbf{W}\,|\,|_2}. \tag{10}$$

Note that the dimensions of $\mathbf{A}^{(t)}$ and $\mathbf{W}$ are 1xM and MxM, respectively. By transposing the elements in Eq. (10), we change the order of the vectors, not the results. After applying the property of the transpose of a matrix product and the fact that the Euclidean norm of a matrix does not change after transposition, we obtain the following:

$$\mathbf{A}^{(t+1)^\top} = \frac{\mathbf{W}^\top A^{(t)^\top}}{||\,\mathbf{W}^\top\mathbf{A}^{(t)^\top}\,|\,|_2}. \tag{11}$$

Given that the weight matrix W derived from the correlation patterns is symmetric, it yields that:

$$\mathbf{A}^{(t+1)^\top} = \frac{\mathbf{W}\mathbf{A}^{(t)^\top}}{||\,\mathbf{W}\mathbf{A}^{(t)^\top}\,|\,|_2}. \tag{12}$$

Eq. (12) matches with the power iteration method [28]. Given a diagonalizable matrix $\mathbf{W}$, the algorithm will produce a number λ, which is the greatest (in absolute value) eigenvalue of $\mathbf{W}$, and a nonzero vector $\mathbf{v}$, which is the corresponding eigenvector of λ, that is, $\mathbf{Wv} = \lambda\mathbf{v}$. The algorithm is also known as the power method or the Von Mises iteration [28] and starts with a vector $\mathbf{b}_0$, which may be an approximation to the dominant eigenvector or just a random vector. In every iteration, the vector $\mathbf{b}_t$ is multiplied by $\mathbf{W}$ and normalized. The algorithm is described by the recurrence relation:

$$\mathbf{b}_{t+1} = \frac{\mathbf{W}\mathbf{b}_t}{||\,\mathbf{W}\mathbf{b}_t\,||}. \tag{13}$$

If we assume that W has an eigenvalue that is strictly greater in magnitude than its other eigenvalues and that the starting vector $\mathbf{b}_0$ has a nonzero component in the direction of an eigenvector associated with the dominant eigenvalue, then the subsequence $(\mathbf{b}_t)$ converges to an eigenvector associated with the dominant eigenvalue. Such an eigenvector always exists and it is unique provided that these two assumptions are fulfilled; otherwise, the sequence might not converge.

In order to match the terminology used in the recurrence relation of the power iteration with our terms, we establish that $\mathbf{b}_t = \mathbf{A}^{(t)^\top}$. Moreover, the symmetry of W implies that it is diagonalizable and there are M real eigenvalues. Summarizing, the case $\phi = 1$ allows modeling invariant situations in which the model needs to converge to the same fixed-point attractor regardless of the initial conditions.

6. Numerical simulations

In this section, we conduct numerical simulations using three scenarios involving protected features reported in the literature. Also, we provide experimental evidence further supporting the main theoretical results of this research.

6.1. German credit dataset

Aiming at testing our FCM-based model, we use the German Credit dataset [12] that is widely used in the context of fairness in machine learning and classifies loan applicants as good or bad credit risks. Features are treated either as numeric or nominal following the data description at the UCI Machine Learning Repository [12]. Ordinal variables are treated as nominal as it is not clear how to order the categories within some variables. Treatment of ordinal variables should be decided by domain experts who are familiar with the data-generating procedure and the decision making rationale. It is worth mentioning that features *age*, *foreign worker* and *gender* are considered to be protected following similar practices in the literature *gender*, *foreign worker* and *age* [2,4,11,36,37,15].

During pre-processing steps, (i) the nominal features *sex* & *marital status* are re-coded to include information only related to gender following [2]'s analysis and (ii) numeric features are normalized in the $[0, 1]$ interval. The FCM's weight matrix is computed as described in Section 4.2 where each entry represents the absolute correlation between every pair of variables.

Table 2 shows the features associated with the German Credit dataset, the absolute correlation values between the problem features and each protected feature (highlighted in gray) and the FRU values as reported in [31], which provide insight into the explicit bias patterns (i.e., how the boundary regions change when suppressing a given feature). The star symbol is used to denote significant correlation results.

Fig. 2a provides an overview of the correlation patterns in an attempt to visualize all possible pathways through which implicit bias related to the sensitive features can be propagated within the system. In this chord chart, we highlighted the correlation between each unprotected feature and the protected ones. However, our FCM model takes into account all interactions when quantifying implicit bias.

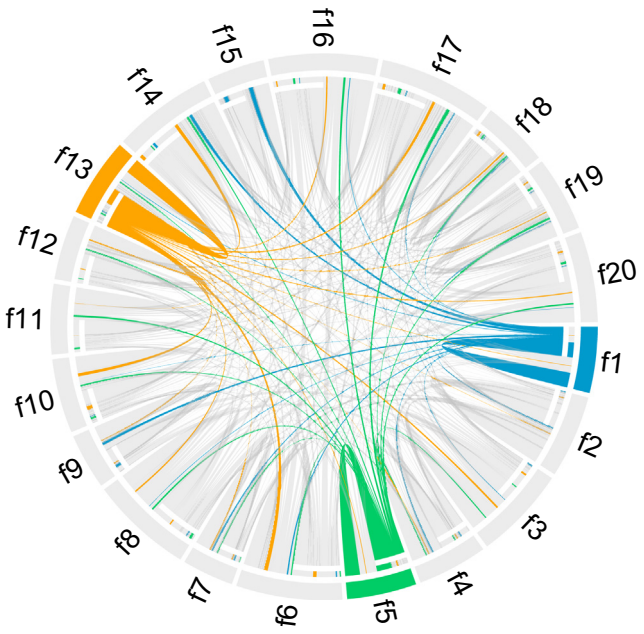


Fig. 2. Pairwise correlations between features in the German Credit dataset. The colored chords show the direct interactions between protected and unprotected features, while gray chords denote the interaction among unprotected features. Overall, this graph shows the complex interactions among the features when modeling implicit bias.

6.2. Results and discussion

Leveraging the simulation capability of the proposed FCM-based model, we perform a what-if analysis in three different scenarios. Moreover, we will illustrate how the nonlinear parameter in Eq. (3) impacts the results and discuss how it can be used as a powerful tool to perform richer what-if simulations that can be useful for decision makers.

Table 2
Correlation[1] values between protected and unprotected features, and FRU values that provide information about the explicit bias.

ID	Features	Correlation with			FRU
		Age	Foreign worker	Gender	
f1	Age	1.0*	0	0.03*	0.11
f2	Credit amount	0.03	0	0.01*	0.07
f3	Credit history	0.03*	0.07	0.12*	0.26
f4	Months	0.04	0.02*	0.01	0.11
f5	Foreign worker	0.0	1.0*	0.04	0.09
f6	Housing	0.09*	0.07	0.23*	0.17
f7	Installment rate	0.06	0.01*	0.01	0.2
f8	Job	0.03*	0.1*	0.09*	0.23
f9	Existing credits	0.15*	0.0	0.01*	0.23
f10	People liable	0.01*	0.07*	0.2*	0.14
f11	Other debtors	0.0	0.12*	0.01	0.15
f12	Other installment	0.0	0.04	0.05	0.23
f13	Gender	0.03*	0.04	1.0*	0.22
f14	Employment since	0.17*	0.08	0.22*	0.36
f15	Residence since	0.27*	0.0	0.0	0.19
f16	Property	0.05*	0.14*	0.09*	0.34
f17	Purpose	0.03*	0.17*	0.15*	0.37
f18	Savings account	0.01	0.04	0.07	0.32
f19	Checking account	0.01	0.08	0.03	0.39
f20	Telephone	0.02*	0.1*	0.07*	0.22

[1] Pearson correlation is used between numeric features and Cramér's V is used between nominal features such that the star indicates significant p-value ($p < 0.05$). The R-squared coefficient of determination is used between numeric and nominal features with the star indicating a significant F-statistic (larger than the critical value).

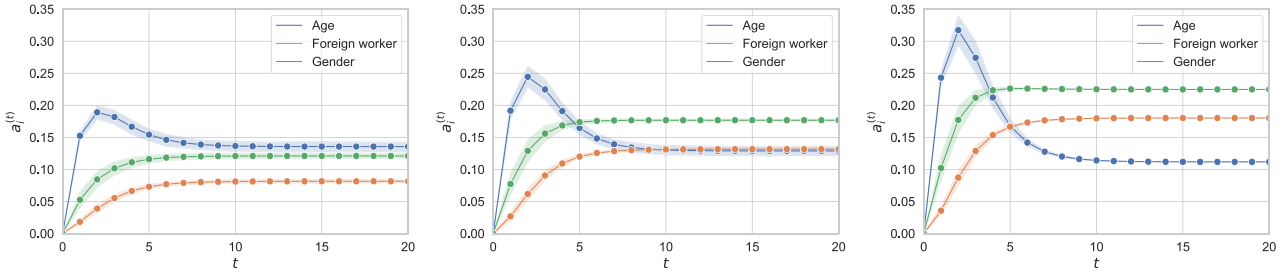
As mentioned, these what-if simulations can be performed by (i) assigning initial values to selected concepts denoting unprotected features, (ii) executing the inference mechanism of the FCM model, and (iii) observing the resulting activation values of concepts denoting protected features as a way to quantify implicit bias. The scenarios to be investigated and the structure of initial activation vectors are detailed below. We generate 20 activation vectors for each scenario where concepts denoting selected unprotected features are randomly initialized, whereas the remaining ones are set to zero.

- **Scenario 1**. We activate three neural concepts representing unprotected features having large absolute correlation with the protected feature *age*. These unprotected features are *existing credits* (f9), *employment since* (f14) and *residence since* (f15), which encode the set of initial activation vectors of the form: $A^{(0)} = (0, 0, 0, 0, 0, 0, 0, 0, f9, 0, 0, 0, 0, f14, f15, 0, 0, 0, 0, 0)$ where f9, f14 and f15 are uniform random values in $(0, 1]$.
- **Scenario 2**. We activate three neural concepts representing unprotected features having large absolute correlation with the protected feature *foreign worker*. These unprotected features are *other debtors* (f11), *property* (f16) and *purpose* (f17), which encode the set of initial activation vectors of the form: $A_1^{(0)} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, f11, 0, 0, 0, 0, f16, f17, 0, 0, 0)$ where f11, f16 and f17 are uniform random values in $(0, 1]$.
- **Scenario 3**. We activate four concepts representing unprotected features having large absolute correlation with the protected feature *gender*. These unprotected features are *housing* (f6), *people liable* (f10), *employment since* (f14) and *purpose* (f17), which encode the set of initial activation vectors of the form: $A_1^{(0)} = (0, 0, 0, 0, 0, f6, 0, 0, 0, f10, 0, 0, 0, f14, 0, 0, f17, 0, 0, 0)$ where f6, f10, f14 and f17 are uniform random values in $(0, 1]$.

According to our definition of implicit bias, one would expect each scenario to be more biased towards the protected feature that correlates with the unprotected ones that were activated. Let us analyze the simulation results. Figs. 3–5 show the activation values of neurons representing protected features for $T = 20$ and different ϕ values for the three scenarios, respectively. As expected, in the third scenario, the activation values of the neuron denoting *gender* are consistently larger than the ones for *age* or *foreign worker*, when varying the ϕ parameter. In the first and second scenarios, however, we see that the model reports more bias towards *gender* than *age* or *foreign worker*, as ϕ approaches one. In other words, the bias against *gender* increases as the neurons' activation values are determined by the interaction among the neurons and the correlation patterns encoded in the weight matrix.

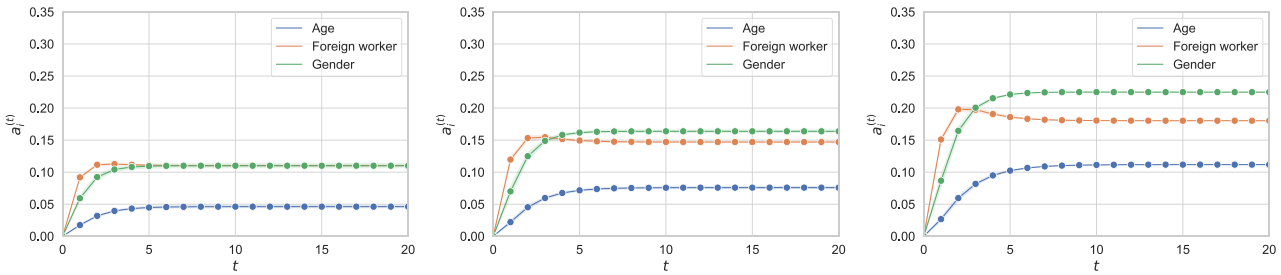
The reader can notice that the final state of our FCM-based model matches in all scenarios when $\phi = 1$. This means that the map will produce the same output regardless of the initial activation vector, thus indicating the presence of a unique fixed-point attractor with larger bias towards gender. This result is consistent with our previous research devoted to quantifying explicit bias using fuzzy-rough sets [24]. Moreover, these simulations provide experimental evidence about the uniqueness of the fixed-point attractor when the nonlinear component in Eq. (3) is entirely suppressed.

Now, let us suppose that we want to build a decision-making model concerning the first scenario in which the bias towards *gender* must be reduced at least 50% compared to the implicit bias encoded in the original data. Using the quasi nonlinear reasoning rule, we can conclude that we can rely on up to 60% of the correlation patterns in the data ($\phi = 0.6$) when activating the unprotected features *existing credits*, *employment since* and *residence since*. More explicitly, we can see in Fig. 3 (c) that the amount of bias concerning gender is about 0.22 when using $\phi = 1$. Such an amount goes



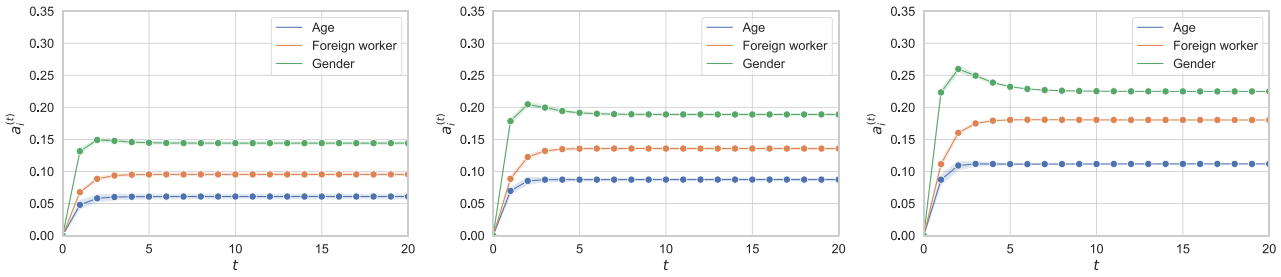
(a) $\phi = 0.6$ (b) $\phi = 0.8$ (c) $\phi = 1.0$

Fig. 3. Activation values of neurons denoting protected features for the first scenario. The unprotected features *existing credits*, *employment since* and *residence since*, which have the largest correlation with the protected feature *age*, are randomly initialized when running the simulations.



(a) $\phi = 0.6$ (b) $\phi = 0.8$ (c) $\phi = 1.0$

Fig. 4. Activation values of neurons denoting protected features for the second scenario. The unprotected features *other debtors*, *property* and *purpose*, which have the largest correlation with the protected feature *foreign worker*, are randomly initialized when running the simulations.



(a) $\phi = 0.6$ (b) $\phi = 0.8$ (c) $\phi = 1.0$

Fig. 5. Activation values of neurons denoting protected features for the third scenario. The unprotected features *housing*, *people liable*, *employment since* and *purpose*, which have the largest correlation with the protected feature *gender*, are randomly initialized when running the simulations.

down to 0.11 (50%) as shown in Fig. 3 (a) when using $\phi = 0.6$, which means that the model is based on 60% of the correlation patterns and 40% on the initial conditions. This suggests that our proposal allows making fairer decisions and reusing the historical data that might contain relevant pieces of knowledge.

Finally, we compare the differences in the fixed points produced by the FCM model when suppressing the linear component in Eq. (3) and using different transfer functions. Fig. 6 shows the results for the second scenario. We selected a single scenario to keep the simulation simple but the same idea can be extrapolated to the remaining scenarios.

Observe that the neurons' activation values are more similar among them (and closer to one) when using either the sigmoid or hyperbolic tangent function, which would suggest that neurons are in risk of saturation. This situation will worsen if we increase the slope of these transfer functions. Overall, saturated neurons do not allow making reliable decisions since their activation values offer little possibilities to discriminate among the options. In contrast, the activation values produced by our re-scaled transfer function allow for better separation between the decisions while being far from one. The latter indicates that there are other active neurons in that iteration, otherwise their activation values would be closer to one.

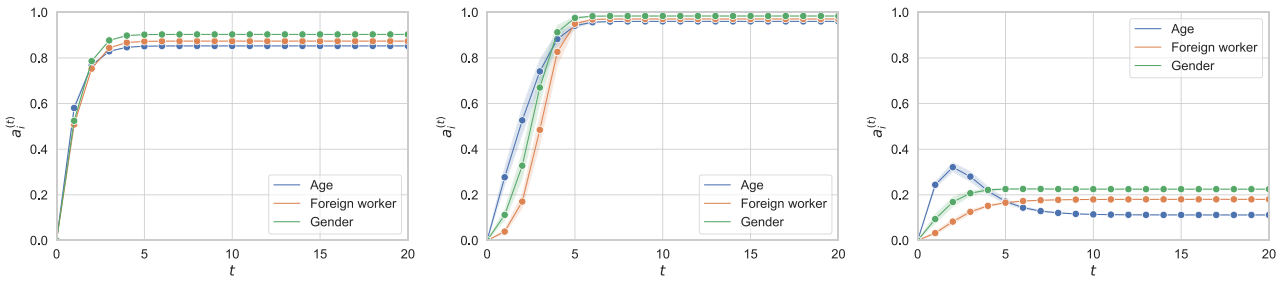
6.3. Comparison with state-of-the-art methods

Before diving into the comparison with other state-of-the-art techniques for detecting (implicit) bias, it is important to remark that our model is originally intended to work at the feature level. Similarly to the FRU measure [24,31], our proposal quantifies the bias towards interesting protected features. However, the majority of the approaches in the fairness literature [4,2,37,14] do this analysis at a group level, i.e., defining groups for the protected features. These methods usually discretize protected numerical features in two or more groups, where at least one is considered unprivileged. Although, the choice of thresholds for these groups can influence the results, even leading to masking the discrimination [9].

To make our model comparable with these methods, we replace the FCM's concepts denoting protected features by concepts symbolizing protected groups. For the protected features *age* and *gender*, we use the codification in [4] involving four possible values each. The groups for *age* are as follows: people younger than 30 years of age (*age_le_30*), people from 30 to 41 years old (*age_from_30_le_41*), people from 41 to 52 years old (*age_from_41_le_52*), and people older than 52 years of age (*age_gt_52*). The groups for *gender* have extra information on the civil status and categorize females independently of their civil status (*gender_female*), single males (*gender_male_single*), married or widowed males (*gender_male_mar_or_wid*), and divorced or separated males (*gender_male_div_or_sep*). For the protected feature *foreign worker*, we will have two groups representing the binary options. Numerical unprotected features will not be discretized to avoid losing information, therefore, the new FCM contains 27 concepts.

As for the nonlinearity degree, we choose $\phi = 1$ to guarantee the convergence to a unique fixed-point attractor. Nevertheless, we generate 20 initial activation vectors where randomly selected concepts denoting unprotected features are randomly initialized, whereas the remaining ones are set to zero. Fig. 7 shows the results of the simulation.

According to our FCM model, the concept with the highest activation value concerns young individuals (*age_le_30*). This result agrees with traditional statistical bias detection measures such as statistical parity difference [13], equal opportunity difference [18], average odds difference [18], and disparate impact [14], which signal high bias towards young individuals. Interestingly,



(a) Sigmoid activation function (b) Hyperbolic activation function (c) Re-scaled activation function

Fig. 6. Activation values of neurons denoting protected features for the first scenario, using different activation functions. Although the activation of *gender* is greater than the activation of *age* or *foreign worker* for all functions, the sigmoid and the hyperbolic tangent functions seem to be less sensitive to this difference. Meanwhile the proposed re-scaled activation function is able to capture the difference without saturating the neurons.

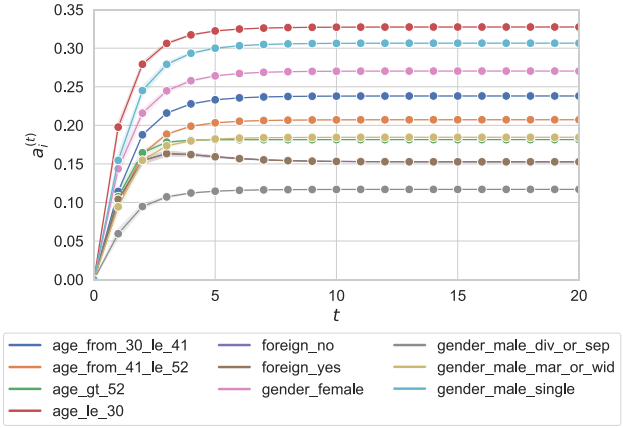


Fig. 7. Activation values for the neural concepts denoting protected groups. The proposed FCM model signals a higher bias against people under 30 years old and single males than the remaining groups.

when using the codification for *gender* that includes information on the civil status, we obtain that the single males (*gender_male_single*) is one of the most discriminated groups, as also concluded by [4]. Moreover, our model keeps indicating a strong bias against females and moderate bias towards middle age people. Less bias is detected towards foreign workers, older people, or males other than single.

It is worth further discussing the divergence between the results reported in [4] and the ones obtained with our model for the protected feature age. The mentioned study attributes more bias against people older than 52 years and less discrimination against young people. However, when looking at the distribution of good and bad credits for each of the groups associated with age (see Table 3), it is evident that younger applicants are less favored than older ones. More explicitly, the percentage of people younger than 30 years with respect to the total of bad credits (45%) is significantly larger that the same demographic for the good credit (33%). In contrast, the distribution of elderly people is very similar for obtaining a bad or a good credit, with 10% and 11%, respectively.

Before concluding the paper, let us summarize the main features of our proposal. First, the FCM model determines implicit bias in situations where pairwise analysis might lead to misleading results. Second, FCMs are able to handle both numeric and nominal features which minimizes data pre-processing and preserves information in its original state. Third, we suggest an interpretable way to rely on the data to a certain extent when taking decisions regarding to a sensitive feature. In practice, this feature allows decision makers to complement their data-driven decision making approach with alternative practices. Finally, the introduction of the nonlinear component expressed by the ϕ value and the re-scaled transfer function paves the way to perform richer what-if simulations.

Table 3
Distribution of good and bad credits per age group for the German credit dataset.

Age Groups	Bad Credit	Good Credit
le_30	137 (45%)	234 (33%)
from_30_le_41	91 (30%)	264 (38%)
from_41_le_52	42 (15%)	127 (18%)
gt_52	30 (10%)	75 (11%)
Total	300	700

7. Concluding remarks

This paper introduced an FCM-based model to quantify implicit bias in structured classification datasets where features are characterized as protected and unprotected ones. Our model uses correlation patterns as weights connecting the neural concepts since the intuition dictates that implicit bias can be quantified from the correlation between unprotected and protected features. At the same time, we solved relevant theoretical challenges concerning the controllability and convergence of FCM-based simulation models that had remained unsolved. Firstly, we employed a quasi nonlinear neural reasoning mechanism involving a parameter to control nonlinearity in the predictions. Secondly, we proposed a re-scaled transfer function that promotes competition among neurons in each iteration. Finally, we derived analytical conditions ensuring either that the model converges to a unique fixed point or that the unique fixed point does not exist. This means that we obtained an easily-controllable simulation model to either represent invariant situations or perform what-if simulations.

In our numerical simulations, an FCM-based model is built using absolute correlation values computed in a pairwise manner between all features of the German Credit dataset. Since this dataset has been widely examined on the basis of possibly being biased towards the features *age*, *foreign worker* and *gender*, we chose to investigate the flow of information from selected unprotected features to these three sensitive concepts. We built three scenarios: we used unprotected features that strongly correlate with each sensitive feature as activation neurons. The resulting activation values of neurons denoting a protected feature are conceptually equivalent to its relevance inside the system and are used to estimate the implicit bias flowing from the activated features towards them. In all scenarios, results show that activation values of the neuron representing *gender* are larger than the ones for *age* or *foreign worker* when $\phi = 1$ regardless of the initial activation vector, indicating that the system converged to a unique fixed point attractor. This denotes that the dataset is more biased against *gender* than *age* or *foreign worker*, a results consistent with our previous models using entirely different approaches [24,31]. A second finding arises when modeling bias that is implicitly reflected on *age* from highly correlated unprotected features: *age*'s activation values gradually become smaller than those of *gender* as we increasingly allow the neurons to interact without any constraints. This implies that one needs to examine different ϕ levels in order to select the optimal level.

Actually, we can formulate a constrained optimization problem to find out the maximal nonlinearity amount and activation values for given protected features such that we minimize the bias towards the protected features. The constraints can be used to ensure that the relevance of interesting features does not fall below a given threshold or that a desired order relationship among features is preserved. Such a research line will be the target of our future research efforts.

CRediT authorship contribution statement

Gonzalo Nápoles: Conceptualization, Methodology, Software, Formal analysis, Validation, Writing - original draft, Writing - review & editing, Supervision. **Isel Grau:** Investigation, Software, Formal analysis, Validation, Writing - original draft, Writing - review & editing, Visualization. **Leonardo Concepción:** Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Lisa Koutsoviti Koumeri:** Data curation, Software, Investigation, Writing - original draft, Writing - review & editing, Visualization. **João Paulo Papa:** Validation, Writing - original draft, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. Adler, C. Falk, S.A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, S. Venkatasubramanian, Auditing black-box models for indirect influence, Knowl. Inf. Syst. 54 (2018) 95–122, https://doi.org/10.1080/1369118X.2019.1573912.

[2] Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y., 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias https://arxiv.org/abs/1810.01943..

[3] K.J. Berry, J.E. Johnston, J.W. Mielke, The measurement of association: a permutation statistical approach, Springer, 2018.

[4] F. Bonchi, S. Hajian, B. Mishra, D. Ramazzotti, Exposing the probabilistic causal structure of discrimination, Int. J. Data Sci. Anal. 3 (2017) 1–21, https://doi.org/10.1007/s41060-016-0040-z.

[5] Y. Boutalis, T.L. Kottas, M. Christodoulou, Adaptive estimation of fuzzy cognitive maps with proven stability and parameter convergence, IEEE Trans. Fuzzy Syst. 17 (2009) 874–889, https://doi.org/10.1109/TFUZZ.2009.2017519.

[6] Chiappa, S., 2019. Path-specific counterfactual fairness, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7801–7808..

[7] L. Concepción, G. Nápoles, R. Falcon, K. Vanhoof, R. Bello, Unveiling the dynamic behavior of fuzzy cognitive maps, IEEE Trans. Fuzzy Syst. 29 (2021) 1252–1261, https://doi.org/10.1109/TFUZZ.2020.2973853.

[8] L. Concepción, G. Nápoles, I. Grau, W. Pedrycz, Fuzzy-rough cognitive networks: Theoretical analysis and simpler models, IEEE Trans. Cybern. (2020) 1–12, https://doi.org/10.1109/TCYB.2020.3022527.

[9] Corbett-Davies, S., Goel, S., 2018. The measure and mismeasure of fairness: A critical review of fair machine learning https://arxiv.org/abs/1808.00023..

[10] H. Cramér, Mathematical Methods of Statistics, Princeton University Press (2016), https://doi.org/10.1515/9781400883868.

[11] P. Delobelle, P. Temple, G. Perrouin, B. Frénay, P. Heymans, B. Berendt, Ethical adversaries: Towards mitigating unfairness with adversarial machine learning, ACM SIGKDD Explorations Newsletter 23 (2021), https://doi.org/10.1145/3468507.3468513.

[12] Dua, D., Graff, C., 2017. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)..

[13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Association for Computing Machinery, 2012, pp. 214–226, https://doi.org/10.1145/2090236.2090255.

[14] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2015, pp. 259–268.

[15] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D., 2019. A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 329–338. DOI: 10.1145/3287560.3287589..

[16] N. Grgić-Hlača, M.B. Zafar, K.P. Gummadi, A. Weller, Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[17] S. Hajian, J. Domingo-Ferrer, A methodology for direct and indirect discrimination prevention in data mining, IEEE Trans. Knowl. Data Eng. 25 (2012) 1445–1459, https://doi.org/10.1109/TKDE.2012.72.

[18] M. Hardt, E. Price, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates Inc, 2016.

[19] I. Harmati, L. Koczy, On the convergence of sigmoidal fuzzy grey cognitive maps, Int. J. Appl. Math. Comput. Sci. 29 (2019) 453–466, https://doi.org/10.2478/amcs-2019-0033.

[20] Harmati, I.Á., Kóczy, L.T., 2020. Improvements on the convergence and stability of fuzzy grey cognitive maps, in: Lesot, M.J., Vieira, S., Reformat, M.Z., Carvalho, J.P., Wilbik, A., Bouchon-Meunier, B., Yager, R.R. (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 509–523. DOI: 10.1007/978-3-030-50153-2_38..

[21] A.L. Hoffmann, Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, Inform., Commun. Soc. 22 (2019) 900–915.

[22] C.J. Knight, D.J. Lloyd, A.S. Penn, Linear and sigmoidal fuzzy cognitive maps: An analysis of fixed points, Appl. Soft Comput. 15 (2014) 193–202, https://doi.org/10.1016/j.asoc.2013.10.030.

[23] B. Kosko, Fuzzy cognitive maps, Int. J. Man Mach. Stud. 24 (1986) 65–75, https://doi.org/10.1016/s0020-7373(86)80040-2.

[24] Koutsoviti Koumeri, L., Nápoles, G., 2021. Bias quantification for protected featuresin pattern classification problems, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 1–10..

[25] Kramer, M., 2005. R2 statistics for mixed models, in: Proceedings of the Conference on Applied Statistics in Agriculture, pp. 148–160. DOI: 10.4148/2475-7772.1142..

[26] Makhlouf, K., Zhioua, S., Palamidessi, C., 2020. Survey on causal-based machine learning fairness notions https://arxiv.org/abs/2010.09553..

[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35, https://doi.org/10.1145/3457607.

[28] Mises, R., Pollaczek-Geiringer, H., Praktische verfahren der gleichungsauflösung. Zamm-zeitschrift Fur Angewandte Mathematik Und Mechanik 9, 152–164..

[29] N.J. Nagelkerke et al., A note on a general definition of the coefficient of determination, Biometrika 78 (1991) 691–692, https://doi.org/10.1093/biomet/78.3.691.

[30] G. Nápoles, L. Concepción, R. Falcon, R. Bello, K. Vanhoof, On the accuracy-convergence tradeoff in sigmoid fuzzy cognitive maps, IEEE Trans. Fuzzy Syst. 26 (2018) 2479–2484, https://doi.org/10.1109/TFUZZ.2017.2768327.

[31] Nápoles, G., Koumeri, L.K., 2021. A fuzzy-rough uncertainty measure to discover bias encoded explicitly or implicitly in features of structured pattern classification datasets. Submitted to Pattern Recogn. Lett. https://arxiv.org/abs/2108.09098..

[32] G. Nápoles, Y. Salgueiro, I. Grau, M.L. Espinosa, Recurrence-aware long-term cognitive network for explainable pattern classification, IEEE Trans. Cybern. (2021) (Submitted)https://arxiv.org/abs/2107.03423.

[33] M.J. Rovine, A. Von Eye, A 14th way to look at a correlation coefficient: Correlation as the proportion of matches, Am. Stat. 51 (1997) 42–46, https://doi.org/10.1080/00031305.1997.10473586.

[34] S. Ruggieri, S. Hajian, F. Kamiran, X. Zhang, Anti-discrimination analysis using privacy attack strategies, Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2014) 694–710, https://doi.org/10.1007/978-3-662-44851-9_44.

[35] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.P. Hubaux, M. Humbert, A. Juels, H. Lin, Fairtest: Discovering unwarranted associations in data-driven applications, in: 2017 IEEE European Symposium on Security and Privacy (EuroS&P), 2017, pp. 401–416, https://doi.org/10.1109/EuroSP.2017.29.

[36] Verma, S., Rubin, J., 2018. Fairness definitions explained, in: 2018 IEEE/ACM International Workshop on Software Fairness (FAIRWARE), pp. 1–7..

[37] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, International Conference on Machine Learning, PMLR (2013) 325–333.

[38] L. Zhang, Y. Wu, X. Wu, Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms, IEEE Trans. Knowl. Data Eng. 31 (2018) 2035–2050, https://doi.org/10.1109/TKDE.2018.2872988.

[39] I. Žliobaitė, Measuring discrimination in algorithmic decision making, Data Min. Knowl. Disc. 31 (2017) 1060–1089, https://doi.org/10.1007/s10618-017-0506-1.

[40] F.J. Zuiderveen Borgesius, Strengthening legal protection against discrimination by algorithms and artificial intelligence, Int. J. Human Rights 24 (2020) 1572–1593, https://doi.org/10.1145/3194770.3194776.

Prof. dr. Gonzalo Nápoles received his PhD degree from both Hasselt University (Belgium) and Maastricht University (The Netherlands) in 2017. Currently, he is an Assistant Professor at the Department of Cognitive Science & Artificial Intelligence, Tilburg University, The Netherlands. He has published several papers in peer-reviewed journals including IEEE Transactions on Neural Networks and Learning Systems, IEEE Transaction on Fuzzy Systems, IEEE Transaction on Cybernetics, Neural Networks, Neurocomputing, Neural Processing Letters, Knowledge-based Systems, Information Sciences, among others. Gonzalo was recipient of the Cuban Academy of Science Award twice (2013 and 2014), the highest scientific award in Cuba. He is the creator and senior developer of the FCM Expert software tool for neural cognitive modeling. His research interests include long-term and short-term reasoning, rough cognitive networks, neural networks and deep learning, learning systems and chaos theory.

Prof. dr. Isel Grau is an Assistant Professor in the Information Systems group at Eindhoven University of Technology (TU/e). She received her Ph.D. in Computer Science from the Vrije Universiteit Brussel (VUB), Belgium. Her research interests lie in the area of artificial intelligence, particularly recurrent neural networks, cognitive networks, (semi-)supervised classification, time-series analysis, data-driven decision making, and explainable AI. Her main research activity focuses on making machine learning black boxes more interpretable, formalizing prior knowledge from experts and using their feedback to improve AI models, identifying and correcting bias, and exploring these challenges in healthcare or business settings. She has over 45 peer-reviewed papers published at international conferences and journals. She has active collaboration with researchers from Vrije Universiteit Brussel, Queen's University Belfast, Hasselt University, Universidad de Talca, Warsaw University of Technology, Tilburg University, and industry partners.

Leonardo Concepción received his B.Sc. degree in Computer Science from the Central University of Las Villas (Cuba) in 2016. Currently, he is an Teacher Assistant at the Computer Science Department, Central University of Las Villas. He is currently a Ph.D. candidate at the Business Intelligence research group of Hasselt University (Belgium). He has published several papers in peer-reviewed journals and conference proceedings including IEEE Transaction on Fuzzy Systems, IEEE

Transaction on Cybernetics, the Lecture Notes in Artificial Intelligence series, among others. His research interests include long- and short-term reasoning, rough cognitive networks, neural networks and deep learning, learning systems, decision making, time series forecasting (classification), and chaos theory.

Lisa Koutsoviti Koumeri is a PhD student in the Business Informatics research group at Hasselt University (Belgium). She received an MSc in Data Science and Society at Tilburg University. Her research revolves around two topics: (1) bias quantification in tabular data for pattern classification problems using fuzzy-rough sets and (2) uni- and multivariate timeseries prediction using explainable models, namely LSTCN (Long Short Term Cognitive Networks), combined with timeseries augmentation.

Prof. dr. João Paulo Papa received his B.Sc. in Information Systems from the São Paulo State University, SP, Brazil. In 2005, he received his M.Sc. in Computer Science from the Federal University of São Carlos, SP, Brazil. In 2008, he received his Ph.D. in Computer Science from the University of Campinas, SP, Brazil. During 2008–2009, he had worked as a post-doctorate researcher at the same institute, and during 2014–2015 he worked as a visiting scholar at Harvard University. He has been an Associate Professor at the Computer Science Department, São Paulo, State University, since 2016, and his research interests include machine learning, pattern recognition, and image processing. He is also an IEEE Senior Member, the recipient of the Alexander von Humboldt fellowship, and the Brazilian representative of the International Association for Pattern Recognition.