Individual Consistency eXplorer (ICX): An Interactive Dashboard for the Exploration of Individual Fairness

Madeleine Waller[a,*], **Odinaldo Rodrigues**[a] **and Oana Cocarascu**[a]

[a]King's College London

Abstract. We present ICX (**I**ndividual **C**onsistency e**X**plorer), an interactive dashboard designed to support stakeholders in exploring individual fairness notions within algorithmic decision-making systems. ICX focuses on a set of metrics based on the consistency score, a key measure of individual fairness, by allowing the visualisation of how the classification of an individual compares with that of similar individuals. Stakeholders can define and fine-tune the notion of similarity according to domain-specific criteria, and examine individual-level views that highlight comparable individuals and their classification outcomes. ICX empowers non-technical users, such as policy-makers, advocates, and auditors to interrogate, analyse and interpret fairness at the individual level, making algorithmic decision-making more transparent and accountable.

1 Introduction

Fair outcomes in algorithmic decision-making that significantly impact human lives are essential, particularly when machine learning drives these decisions. Algorithmic decision-making systems (ADMS) can unintentionally produce discriminatory outcomes that disproportionately affect certain demographic groups. For example, Amazon's hiring algorithm demonstrated gender bias by disadvantaging women due to historical hiring patterns that underrepresented female candidates [5]. The emergence of the field of algorithmic fairness addresses concerns about equity in ADMS [14], which have become prevalent across different domains such as criminal justice [13], healthcare [3], and finance [6].

Discrimination is typically assessed through legal frameworks and societal normative standards [4, 16, 20, 7]. Algorithmic fairness research translates these traditional fairness definitions into formal metrics that can measure unfairness of ADMS [17, 11]. These metrics primarily target binary classifiers using tabular data containing protected characteristics (e.g., race, gender, age, disability status) [2, 10] and evaluate fairness at both group and individual levels.

In this work, we present ICX[1] a dashboard for evaluating aspects of individual fairness in ADMS.[2] It allows users to examine whether an individual receives a decision that is compatible with the decisions of similar individuals regardless of their protected attributes. This approach ensures equitable treatment at the individual level rather than merely balancing outcomes across an entire dataset, as in e.g., group fairness [10]. Indeed, previous work has shown that the perceived fairness of individual decisions can be affected by several factors, including the notion of similarity between individuals, the number of similar individuals considered, and the proportion of similar individuals whose decision we wish an individual's decision to agree with [18]. ICX allows alternatives and caveats to be explored effectively by stakeholders (e.g. developers, auditors, and potentially affected individuals) to evaluate aspects of individual fairness in ADMS.[3] Specifically, it allows users to:

1. Explore and operate on a tabular dataset of individuals provided with their corresponding binary classifications;
2. Define how similarity between individuals is measured, by configuring categorisation of attributes and how distances between attribute values are computed;
3. Compute and visualise five individual fairness metrics that summarise the consistency of classifications across the dataset; and
4. Inspect attributes of specific individuals and of those individuals most similar to them, to explore variations in attribute values and allow "like-for-like" comparisons of classifications.

By allowing the exploration of individual fairness and associated metrics to show how similarly people are treated compared to their peers, ICX can be used to identify and question unfair algorithmic decisions.

ICX differs from existing tools addressing individual fairness as follows. It builds upon the consistency metric used to assess individual fairness in IBM's AI Fairness 360 (AIF360) [2], an open-source Python library, by providing an interactive user interface, additional consistency-based metrics and exploration of specific individuals, motivated by findings that the standard consistency metric may not fully capture the nuances of individual fairness [18]. Google's What-If Tool (WIT) [19] provides an interactive visual tool with some support for exploring individual fairness. Whilst WIT addresses individual fairness through a counterfactual fairness perspective [12], we focus on the notion of consistency.

2 Individual Fairness Notions

Central to our considerations is the notion of similarity between individuals and the consistency of an individual's classification decision with respect to the decisions of those most similar to them.

2.1 Defining Similar Individuals

Evaluating similarity of individuals requires several key design choices that must be specified within a given context: (1) how to com-

* Corresponding Author. Email: madeleine.waller@kcl.ac.uk

[1] https://github.com/maddiewaller/Individual-Consistency-eXplorer

[2] https://individual-consistency-explorer.streamlit.app

[3] We also provide a Python package for ICX, available at https://pypi.org/project/icx/, that can be downloaded and run offline on local data, thereby mitigating concerns related to data privacy or protection.

pute distances between values of the attributes of two individuals; (2) how to aggregate these into an overall distance between individuals; and (3) how to select the set of individuals closest to an individual according to the aggregated distances, i.e. the *nearest neighbours*.

ICX enables users to define their own notion of similarity based on criteria (1)–(3). For numerical attributes, users can retain them as continuous values, discretise them, or treat them as categorical. Categorical attributes can be treated as nominal (unordered) or assigned a numerical ordering. These choices determine how distances between attribute values are calculated: categorical values have distance 0 when they are identical, or 1 otherwise; while numerical and ordered categorical attributes yield distances on a continuous scale between 0 and 1 (via normalisation). The overall distance between individuals x and y is computed using the Gower distance:

$$D(x,y) = \sum_{i=1}^{p} d_i(x,y), \text{where}$$
$$d_i(x,y) = \begin{cases} 0, & \text{if } v(x,z_i) = v(y,z_i) \\ \frac{|v(x,z_i) - v(y,z_i)|}{range(z_i)}, & \text{if } z_i \text{ is numerical} \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

In the above, p is the number of attributes in the dataset and $v(w, z_i)$ is the value of the individual w's ith attribute. Notice that the Gower distance equates to the Hamming distance when all attribute values are treated as nominal categorical.

Finally, users can define how many neighbours to compare with, i.e., the number k of individuals with minimal distance to a given individual (neighbours are chosen arbitrarily if there are ties).

2.2 Measuring Consistency

The *consistency score* in Definition 1 measures the overall level of disparity in a dataset E in the classification f of individuals with respect to their most similar individuals.

Definition 1 (Consistency score [21])**.** *Let $nbr_\sigma(x,k)$ be the set of k individuals in E closest to individual x, according to some notion of similarity σ. The consistency score of E w.r.t. f is defined as*

$$C^\sigma(E) = 1 - \frac{1}{|E|} \sum_{x \in D} |f(x) - \tfrac{1}{k} \sum_{y \in nbr_\sigma(x,k)} f(y)|$$

Although the consistency score is commonly used to measure individual fairness in algorithms, it has some limitations. As shown by [18], this score remains relatively stable across different groupings because it only captures average fairness levels. This average-based approach hides important information about individuals experiencing significant unfairness, specifically those whose outcomes differ greatly from similar people. These outliers are our main concern in real-world applications. To address this, the notion of an *individual consistency score* was introduced. The individual consistency score quantifies how similar an individual's classification is to similar individuals' classifications.

Definition 2 (Individual consistency score [18])**.** *The individual consistency score $c^\sigma(x)$ computes the proportion of x's most similar individuals (cf. σ), with the same classification as x's.*

$$c^\sigma(x) = 1 - |f(x) - \frac{1}{k} \sum_{y \in nbr_\sigma(x,k)} f(y)|$$

Previous work [18] showed that Definition 2 and Definition 1 can be obtained from each other. Given that Definition 2 also allows the analysis of the decisions at the individual level, it can used to define the more nuanced measures of individual consistency at the dataset level proposed in [18] (Definitions 3–6).

2.3 Dataset Metrics based on Individual Consistency

Given a dataset E provided alongside the classifications of its individuals, ICX calculates and displays E's overall consistency score (cf. Definition 1), alongside each individual's individual consistency score (cf. Definition 2), and the values of the four metrics given in definitions 3–6. We now briefly introduce the new metrics. In the dashboard, the notion of similarity σ can be specified according to the criteria described in Section 2.1, using a given number k of most similar neighbours, and the special threshold parameter $\delta \in [0,1]$ (see below).

The Low Individual Consistency Count ($LICC$) metric measures the absolute number of individuals whose individual consistency score is below the threshold δ. $LICC$ provides an *absolute* measure of the total number of individuals in a dataset affected by controversial decisions.

Definition 3 (Low Individual Consistency Count ($LICC$))**.** *$LICC$ provides the number of individuals whose individual consistency score is less than the threshold δ.*

$$LICC^\sigma_\delta(E) = |\{x \in E : c^\sigma(x) < \delta\}|$$

We can also measure the *proportion* of individuals in the datasets whose decisions are deemed 'acceptable', i.e., with an individual consistency score above a certain threshold using the Proportional Consistency Score (PCS).

Definition 4 (Proportional Consistency Score (PCS))**.** *$PCS^{\sigma,k}_\delta(E)$ quantifies the proportion of individuals in E with an individual consistency score greater than or equal to δ.*

$$PCS^{\sigma,k}_\delta(E) = \frac{|\{x \in E : c^{\sigma,k}(x) \geq \delta\}|}{|E|}$$

PCS measures the proportion of acceptable decisions (i.e., above the threshold δ), but not how *close* to full agreement those decisions were. It allows each individual to contribute to the score with $1/|E|$ as long as their individual consistency score meets the threshold, regardless of how close to the threshold the score is (or how far from full agreement, i.e., 1, the score is). The Balanced Conditioned Consistency Score BCC allows us to capture this.

Definition 5 (Balanced Conditioned Consistency Score (BCC))**.** *BCC gives the sum of the individual consistency scores above or equal the threshold δ divided by the total number of individuals.*

$$BCC^{\sigma,k}_\delta(E) = \frac{1}{|E|} \sum_{x \in E} v(x), \textit{where } v(x) = \begin{cases} c^\sigma(x), & \textit{if } c^\sigma(x) \geq \delta \\ 0, & \textit{otherwise} \end{cases}$$

It is possible to replace 0 with a suitable *penalty*, effectively bringing down the overall BCC score for each controversial decision made. To mirror the positive contribution of each individual, it is suggested that the penalty is a value in $[-1, 0]$.

Definition 6 (BCC with penalty)**.** *BCC with penalty adds a penalty $p \in [-1,0]$ for each individual consistency score below δ.*

$$BCC^{\sigma,k}_{\delta,p}(E) = \frac{1}{|E|} \sum_{x \in E} v(x), \textit{where } v(x) = \begin{cases} c^\sigma(x), & \textit{if } c^\sigma(x) \geq \delta \\ p, & \textit{otherwise} \end{cases}$$

3 Inspecting Specific Individuals

In ICX, individuals whose individual consistency scores (cf. Definition 2) fall below a user-specified threshold δ are automatically

highlighted in the dataset view. By clicking on such an individual, the individual's attribute values and individual consistency score can be inspected alongside the attribute values of the nearest neighbours, their distances from the individual and their classifications. This enables a more fine-grained assessment of the consistency of the classifications. This interactive inspection helps users identify individuals who may have been treated unfairly and explore potential reasons behind such discrepancies.

Example Consider a company that uses a machine learning model to assist in screening applicants for a job. Each applicant is described by attributes such as *education level*, *years of experience*, and *university attended*. The model produces a binary outcome: whether the applicant is **shortlisted** or **not shortlisted** for the job.

To evaluate the fairness of this decision-making process, a compliance officer performs an individual fairness audit using ICX. The officer uploads a dataset of 5,000 historical applications along with the model's decisions. A subset of this dataset is shown in Figure 1.

	id	Education Level	Years of Experience	University Attended	Classification
☐	0	Masters	5	University A	0
☐	1	Masters	5	University B	1
☐	2	Bachelors	5	University B	1
☐	3	Masters	6	University A	1
☐	4	Masters	4	University A	1
☐	5	PhD	5	University A	1

Figure 1. Example dataset (image taken from ICX).

To define *similarity*, the officer configures ICX to treat *education level* as an ordered categorical variable (Bachelors$<$Masters$<$PhD); *years of experience* as a numerical variable; and *university attended* as a nominal categorical variable.

ICX then computes the distances between applicants. For example, the distance between Applicant 1 and Applicant 2 is 1 as only the values of *university attended* differ. Calculating the distance between ordered categorical (*education level*) or numerical (*years of experience* requires the full range of possible values of that attribute as defined by the Gower distance (Section 2.1). Using these distances, ICX identifies similar individuals for each applicant and uses these to compute the *individual consistency scores* (Definition 2) for each.

The individual consistency scores are used to calculate five aggregate metrics of individual consistency which the officer can examine. Although the overall consistency score suggests reasonable fairness across the dataset, ICX highlights a subset of individuals whose consistency scores fall below a user-specified threshold ($\delta = 0.5$). These individuals are flagged as potential fairness concerns.

The officer selects one such applicant who was not shortlisted by the model and inspects their similar individuals. ICX reveals that the majority of the similar applicants were in fact shortlisted. This discrepancy prompts closer scrutiny, as it may indicate that the model is treating similar individuals inconsistently. The case is flagged for further investigation, potentially prompting a review of attribute handling, model retraining, or revision of the hiring criteria.

This example demonstrates how ICX can support stakeholders in identifying and investigating potential individual fairness violations. By enabling configurable similarity definitions and facilitating interactive inspection of individual cases, ICX provides actionable insights that bridge technical fairness metrics and real-world decision accountability.

4 Implementation

Web Framework. ICX was implemented using `streamlit`, a Python framework which provides a lightweight and interactive environment for building data-driven web applications. ICX runs as a standalone web app and requires no setup, making it easily accessible to both technical and non-technical stakeholders to explore. Streamlit-aligned packages were used to explore and manipulate the dashboard content.

Datasets and Data Handling. ICX includes three commonly used datasets in algorithmic fairness: Adult Census [1], COMPAS [15], and German Credit [9]. The datasets are drawn from the `aif360` library. Each dataset consists of attributes that are either numerical or categorical, and are processed dynamically based on user-specified configurations for defining similarity. This includes operations such as binning numerical attributes into custom ranges, or reordering categorical values to reflect ordinal relationships.

Similarity Computation The `gower` package computes the distances between all individuals, as specified in Section 2.1 which allows for the comparison of mixed-type attributes by normalising numerical differences and treating categorical mismatches as binary indicators. Once the pairwise Gower distance matrix is calculated, `scikit-learn`'s `NearestNeighbors` algorithm is applied with a precomputed distance metric to identify the k most similar individuals for each data point.

5 Conclusion

We presented ICX, an interactive dashboard that allows stakeholders to explore individual fairness notions within ADMS through the lens of consistency-based metrics. ICX supports customised definitions of similarity between individuals and computes the values of the four dataset metrics introduced in [18] besides the value of the consistency score of the dataset [21] and the values of the individual consistency scores of each individual (cf. Definition 2). Users can visualise how an individual's classification fares with respect to that of its nearest neighbours, allowing them to identify and investigate concerns about fairness at the individual level.

The legal aspects of individual fairness in ADMS remain unclear due to limited case law [8]. However, in the United Kingdom and European Union, individuals must prove discrimination by showing that a system disadvantages an individual compared to similar individuals [17]. This can be difficult, especially when algorithm details and training data are hidden. ICX equips non-technical users such as policymakers, auditors, and advocates with the ability to interrogate the model's behaviour, contributing to more transparent and accountable ADMS.

Future work includes ensuring ICX is scalable to large numbers of individuals and extending the Python package with new functions to support integration into fairness auditing workflows.

Acknowledgements

This work was supported by the UK Research and Innovation Centre for Doctoral Training in Safe and Trusted Artificial Intelligence[4] [grant number EP/S023356/1]. The authors are affiliates of the King's Institute for Artificial Intelligence.[5]

[4] https://www.safeandtrustedai.org
[5] https://www.kcl.ac.uk/ai

References

[1] B. Becker and R. Kohavi. Adult dataset. uci machine learning repository. https://doi.org/10.24432/C5XW20, 1996.

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

[3] M. Bergquist and B. Rolandsson. Exploring adm in clinical decision-making: Healthcare experts encountering digital automation. In *Everyday Automation*, pages 140–153. Routledge, 2022.

[4] W. Blanchard. Evaluating social equity: What does fairness mean and can we measure it? *Policy Studies Journal*, 15(1), 1986. URL https://www.proquest.com/scholarly-journals/evaluating-social-equity-what-does-fairness-mean/docview/1300125838/se-2.

[5] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 2022. ISBN 9781003278290.

[6] W. De La Rosa and C. J. Bechler. Unveiling the adverse effects of artificial intelligence on financial decisions via the ai-impact model. *Current Opinion in Psychology*, page 101843, 2024.

[7] S. Greco, K. Zhou, L. Capra, T. Cerquitelli, and D. Quercia. Nlpguard: A framework for mitigating the use of protected attributes by NLP classifiers. *Proc. ACM Hum. Comput. Interact.*, 8(CSCW2):1–25, 2024. doi: 10.1145/3686924. URL https://doi.org/10.1145/3686924.

[8] L. Grozdanovski. In search of effectiveness and fairness in proving algorithmic discrimination in eu law. *Common Market Law Review*, 58 (1), 2021.

[9] H. Hofmann. Statlog (German Credit Data). https://doi.org/10.24432/C5NC77, 1994.

[10] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. 1(2), June 2024. doi: 10.1145/3631326. URL https://doi.org/10.1145/3631326.

[11] M. Jorgensen, M. Waller, O. Cocarascu, N. Criado, O. Rodrigues, J. Such, and E. Black. Investigating the legality of bias mitigation methods in the united kingdom. *IEEE Technology and Society Magazine*, 42 (4):87–94, 2023. doi: 10.1109/MTS.2023.3341465.

[12] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[13] Northpointe. Practitioner's Guide to COMPAS Core, 2019. URL https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf.

[14] D. Pessach and E. Shmueli. *Algorithmic Fairness*, pages 867–886. 2023. ISBN 978-3-031-24628-9. doi: 10.1007/978-3-031-24628-9_37.

[15] ProPublica. COMPAS Recidivism Racial Bias. https://www.kaggle.com/datasets/danofer/compass, 2024.

[16] S. Verma and J. Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357463. doi: 10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

[17] S. Wachter, B. D. Mittelstadt, and C. Russell. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law Security Review*, 41:105567, 2021. doi: 10.1016/j.clsr.2021.105567.

[18] M. Waller, O. Rodrigues, and O. Cocarascu. Beyond consistency: Nuanced metrics for individual fairness. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. ACM, 2025 (Forthcoming). doi: 10.1145/3715275.3732141.

[19] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, page 1–1, 2019. ISSN 2160-9306. doi: 10.1109/tvcg.2019.2934619. URL http://dx.doi.org/10.1109/TVCG.2019.2934619.

[20] A. Xiang and I. D. Raji. On the legal compatibility of fairness definitions, 2019.

[21] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML*, volume 28, pages 325–333, 2013. URL http://proceedings.mlr.press/v28/zemel13.html.