2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW) | 979-8-3503-8403-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDEW61823.2024.00037

Exploring Fairness Interpretability with *FairnessFriend*: A Chatbot Solution

Chiara Criscuolo
DEIB – Politecnico di Milano
Milan, Italy
chiara.criscuolo@polimi.it

Tommaso Dolci
DEIB – Politecnico di Milano
Milan, Italy
tommaso.dolci@polimi.it

***Abstract*—In the contemporary world, artificial intelligence and machine learning algorithms are an important driver for decision-making, by leveraging real-world data for future predictions. Despite clearly improving efficiency, the lack of transparency in their predictions raises concerns about the degree of fairness of machine learning models, well highlighted by recent instances of algorithmic unfairness, from automated decisions on criminal recidivism to disease prediction. Increased user awareness of algorithmic fairness is met with a deficiency in systems guiding data analysts and practitioners in comprehending the implications of their outputs. To tackle the challenge of fairness interpretability, we propose *FairnessFriend*, a chatbot solution that combines data science with a human-computer interaction perspective. Given a dataset and a trained machine learning model with established fairness metrics, our system facilitates users in understanding these metrics and their significance in the context of the training data. *FairnessFriend* provides meanings for various statistical fairness metrics, and presents the resulting metrics values with detailed explanations, offering specific insights into their implications.**

***Index Terms*—fairness, interpretability, machine learning, human-computer interaction, chatbot**

I. Introduction

In today's digitally interconnected world, decision-making systems increasingly rely on machine learning algorithms. These algorithms learn from real-world datasets to predict an output from unseen input data, helping humans in making critical decisions in a variety of domains. While these systems improve efficiency and reduce working time, the rationale behind their decisions often remains obscure. This lack of transparency is one of the main reasons for the increasing concern over fairness in systems based on artificial intelligence (AI) and machine learning (ML). This concern is strengthened by many recent examples of algorithmic unfairness in AI systems, from racism in crime recidivism prediction [1] to gender bias in automated tools for recruitment [2].

One of the most famous examples of lack of fairness in AI and ML applications was exposed by ProPublica's analysis of the COMPAS recidivism prediction algorithm [1]. This tool, used by judges to predict the likelihood of criminal defendants committing another crime, was shown to exhibit racial bias. Specifically, all things being equal, black defendants are more likely to be incorrectly assessed with a higher risk of re-offending compared to white defendants. As demonstrated in the literature [3], [4], the COMPAS system evaluates and satisfies a specific fairness metric, but it does not satisfy other fairness metrics, ultimately leading to unequal treatment. How could have COMPAS engineers avoided this issue and created a fair system according to all fairness metrics? Unfortunately, further studied showed that this was not possible, as demonstrated by the so-called impossibility theorem [4], asserting that, for binary classification, equalising a specific set of fairness metrics between protected classes is impossible, except in two special cases. The first case is when an algorithm is a perfect predictor, and the second is when the prevalence of the outcome being predicted is equal across groups. Unfortunately, training a perfect predictor is not possible, and often the true distributions between groups are not equal, such as in the case of COMPAS [5]. This example highlights the fundamental need to improve the understanding of the different fairness metrics, in order to select the most appropriate for the specific use case, or at least make informed decisions. Unfortunately, algorithmic fairness is a recent topic, and despite growing awareness by the public, there is a lack of tools to adequately inform and support data analysts and ML practitioners. This includes understanding the degree of fairness in decision-making systems, as well as the consequences and significance of their predictions. To the best of our knowledge, there is a lack of practical tools to help users in understanding fairness metrics, correctly interpreting their results, and guiding them towards informed decisions.

This work proposes *FairnessFriend*, a chatbot solution to address the **fairness interpretability problem** from a human-computer interaction (HCI) perspective. *FairnessFriend* provides a chatbot interface to guide users in exploring the outcomes of ML models in terms of fairness. By providing a dataset and the corresponding model, users can query to system for insights, including the display of the resulting fairness metrics, detailed explanations of each metric meaning, and implications in the specific context. Through progressive interactions with *FairnessFriend*, the user is guided in the comprehension of the most correct and appropriate fairness metric and can subsequently understand whether the model meets it or not. The overall goal is to assist data analysts in correctly assessing the algorithmic fairness of the ML model created, providing a tool for practitioners without expertise in fairness.

The paper is divided into the following sections: after

979-8-3503-8403-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDEW61823.2024.00037

the introduction, Section II introduces preliminary notions on fairness and HCI for AI. Section III presents the scenario that we considered, describes the interface of our solution, and discusses the interaction between the user and *FairnessFriend*. Section IV provides technical details on *FairnessFriend*. Section V describes the related work. Finally, Section VI concludes the paper and illustrates future work.

II. Preliminaries

Fairness is defined as *the absence of prejudice or favouritism towards an individual or group on the basis of their inherent or acquired characteristics* [6, p.100]. Unfairness results from the presence of bias, which can be of various types; the two main ones being historical and algorithmic biases. *Historical bias* is an existing bias caused by socio-technical issues in the world [7]. On the other hand, *algorithmic bias* occurs when the bias is not present in the input data and is added purely by the algorithm [8]. *Algorithmic fairness* aims at understanding and correcting algorithmic bias.

Various metrics for measuring the fairness of a model can be found in the literature, but they are frequently difficult to understand by non-experts. The creation of a catalogue of fairness definitions and guidelines for the use of each definition is an open challenge. In fact, there are dozens of existing definitions, and each of them depends on and is appropriate only in specific contexts [9]. The first work to describe a catalogue of fairness metrics was presented in [10]. Later, this catalogue has been enriched in [11] by including metrics from [12]. In this paper, we adopt the methodology suggested by [11]: given a dataset and a ML binary classification algorithm trained on a dataset, the algorithm is evaluated on a test set to measure the fairness of the resulting model predictions.

Fairness metrics are based on the idea of a protected attribute. A **protected attribute** is a characteristic for which non-discrimination should be established, such as religion, ethnicity, gender, etc. [10]. Metrics analyse only one protected attribute at a time in a binary setting: for instance, Caucasian and Afro-American for ethnicity, or young and adult for age. These metrics are multifaceted and often their meaning is not easy to understand in practice; in fact, they are defined using mathematical formulas that involve statistical concepts and notation, which require a solid foundation in these disciplines. Furthermore, they depend on the context in which a machine learning model is applied. Different applications may require different fairness considerations, making it challenging to have a one-size-fits-all understanding. Finally, studying algorithmic fairness is an interdisciplinary field that involves concepts from computer science, ethics, and social sciences, thus understanding these metrics requires knowledge in multiple domains.

The metrics from the aforementioned works [10]–[12] are based on the performance values extracted from the confusion matrix: true positive, true negative, false positive, and false negative. These values summarise the prediction results from a statistical classification problem: for this reason, the considered metrics are called **statistical fairness metrics**. The *ideal fairness value for each metric is 1*, meaning that both groups are treated equally. If the metric is between 0 and $1 - t$, the expected discriminated group is treated unfairly, whereas if the metric is greater than $1 + t$, the expected privileged group is treated unfairly. The parameter t is a threshold value to be set at the beginning of the experiment. The list in Table I presents the taxonomy of statistical fairness metrics collected and presented in [11].

Since understanding the significance and implications of these metrics is not an easy task for the user, this work focuses on their interpretability following the paradigm of human-computer interaction. **Human-computer interaction** (HCI) is concerned with the design, evaluation, and implementation of interactive computing systems for human use, and with the study of major phenomena surrounding them [13]. This discipline is connected with the notion of *user-centred design*, which is focused on understanding stakeholders and their needs, analysing their behaviour and attitude with technology, and finally evaluating how design affects users' performance and satisfaction. In this process three aspects are fundamental:

- Target users: the process needs to focus on all stakeholders who will use or will interact with the system.
- Needs: the process focuses on the users' needs to design and later prototype the system properly.
- Knowledge and attitudes: the process focuses also on users' knowledge and attitudes toward the product to evaluate the system's features directly perceived by users.

This process is particularly relevant when the output system uses AI or ML algorithms. This is because the stochastic behaviour of these algorithms can confuse users, erode their confidence, be offensive, or – in this specific case – it can discriminate against people, leading to the abandonment of the technology. For this reason, a new field of study named **interactive human-centred AI** [14] has emerged, highlighting the following needs:

- Clearly stating how humans can benefit from AI (the goal it is designed for).
- Explaining the risks of AI on individuals and society.
- Making visible who has control on the AI, and who has the power over data, models, and algorithms.
- Making visible what data, knowledge base, and information are used to create the AI.

People + AI Guidebook [15] presents a set of methods, best practices, and examples for designing with AI and ML systems. This guideline is based on five points:

- User needs, to reason on how the system can support user's goal.
- Mental models, to prepare users for change helping them understand how to train the system.
- Explainability+Trust, to focus on the explanation for AI and ML systems based on probability and uncertainty.
- Feedback+Control, to let users give feedback to products.
- Error+Graceful Failure, to allow users to move forward after a failure.

TABLE I
LIST OF STATISTICAL FAIRNESS METRICS WITH THE CORRESPONDING DESCRIPTION.

#	Metric	Description
1	Group Fairness	Requires that discriminated and privileged individuals have the same probability of a positive outcome.
2	Conditional Statistical Parity	Requires that discriminated and privileged individuals, with the same values for the legitimate attributes, have the same probability of a positive outcome.
3	Predictive Equality	Requires that discriminated and privileged individuals with a negative decision have the same probability of having a positive prediction.
4	Predictive Parity	Requires that discriminated and privileged individuals who are predicted to have a positive outcome should have the same probability of actually having a negative decision.
5	FP/GS Parity	Requires that discriminated and privileged individuals with a negative decision have the same probability of a positive prediction.
6	Equal Opportunity	Requires that discriminated and privileged individuals who have a positive decision should have an equal probability of a negative prediction.
7	FOR Parity	Requires that discriminated and privileged individuals who are predicted to have a negative outcome should have an equal probability of a positive decision in the dataset.
8	FN/GS Parity	Requires that discriminated and privileged individuals with a positive decision have equal probability to be predicted to have a negative outcome.
9	Treatment Equality	Requires that the ratio between wrong and correct predictions is the same for privileged and discriminated individuals.
10	Conditional Use Accuracy Equality	Requires that discriminated and privileged individuals with a positive decision have an equal chance of being predicted to have a positive outcome. Similarly, discriminated and privileged individuals with a negative decision should have an equal chance of being predicted to have a negative outcome.
11	Overall Accuracy Equality	Requires that discriminated and privileged individuals have an equal probability of a positive decision and of a positive prediction; they should also have an equal probability of a negative decision and of a negative prediction.
12	Equalized Odds	Requires that discriminated and privileged individuals who have a positive decision should have an equal probability of a positive prediction; similarly, discriminated and privileged groups who have a negative decision should have an equal probability of a positive prediction.

III. FAIRNESSFRIEND

In this section, we first focus on the aspects related to HCI in the considered scenario, giving more details about the target users, their needs, knowledge and attitudes in relation to our system. After that, we present the web user interface and discuss the interaction between the user and our system. Through interaction with *FairnessFriend*, the user is guided towards:

1) **understanding the general meaning** of fairness metrics, presenting the respective formula and its significance with respect to the context of the data;
2) **understanding the results** of the various metrics, presenting and explaining them when they reveal unfairness;
3) **describing the motivation for choosing the most suitable metric**, given the context and the problem at hand.

A. HCI Scenario

In the rest of the paper, we discuss *FairnessFriend* considering the following scenario. The target user is a *data analyst* who wants to create a ML model for predicting credit risk for a bank company. She is aware of the possible fairness issues related to the task, and she wants to understand whether the trained algorithmic model is fair with respect to the results of the risk prediction. Specifically, given the group of women, the aim is to understand the following: *(a)* whether the algorithm discriminates female individuals, *(b)* which metrics show unfairness results towards them, and *(c)* what is the meaning and of these metrics.

The goal is to concretely assess which metric is the most suitable for the context and the task under analysis, and therefore to make an informed decision on the deployment of the system. The user has good experience with ML systems and has limited experience in fairness issues, thus needing guidance and explanations to understand the results.

The dataset considered in this scenario is the **German Credit Risk**[1]. In this dataset, each tuple represents a person who has applied for credit from a bank. Each person is associated with a set of attributes (e.g., sex, occupation, savings) and with a target class describing the risk prediction: either low or high. To measure fairness metrics, a dataset with at least one protected attribute (e.g., sex, age, religion) is required. In this case, the protected attribute is "sex". Fairness metrics are traditionally calculated considering the protected attributes, which is why it is necessary to retain this information in the dataset. In fact, the moment this attribute is removed, it is no

[1]https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk

TABLE II
VALUES OF FAIRNESS METRICS AND DISCRIMINATED GROUP (WITH $t = 0.15$) ON THE GERMAN CREDIT RISK DATASET.

Metric	**Value**	**Discrimination**
Group Fairness	0.866	-
Conditional Statistical Parity	0.969	-
Predictive Equality	0.757	woman
Predictive Parity	0.993	-
FP/GS Parity	0.867	-
Equal Opportunity	0.534	woman
FOR Parity	1.150	-
FN/GS Parity	0.577	woman
Treatment Equality	0.867	-
Conditional Use Accuracy Equality	1.109	-
Overall Accuracy Equality	0.623	woman
Equalized Odds	0.694	woman

longer possible to measure the overall fairness of the model. To compute these metrics, the dataset is split into two parts using hold-out evaluation (2/3 training set and 1/3 test set). Then, a binary ML classifier (i.e., *random forest*) is trained on the training set. Fairness metrics are calculated on the model's predictions of the test set data. Table I details the fairness metrics considered.

To illustrate the importance and impact of the different options, Table II presents the results of the statistical fairness metrics, obtained for the German Credit problem. We adopted a threshold t equal to 0.15, a value small enough to understand whether there is discrimination between men and women. The different formulas gave different results: the same system can be considered fair or unfair according to the metric adopted: seven definitions have values between 0.85 and 1.15, i.e., fairness is achieved; five definitions have values between 0 and 0.85, i.e., men are privileged. As evident from these results, different metrics measure different dimensions of fairness. This makes it challenging to interpret the results and understand whether the classifier is actually fair or not. For this reason, it is important to understand the meaning of each metric and its implications in the context under analysis.

To choose the most appropriate fairness metric in this specific case, we applied a decision model from the literature, i.e., the *Fairness Decision Tree* [11]. Selecting the appropriate metric depends on the context and the problem at hand, factors that the decision tree takes into account to assist the user in determining the optimal definition. In this specific context, the most appropriate metric is the *Conditional Statistical Parity*, because we are only concerned with the results from the decision system and "savings" should be at the same level to have the same predictions. *FairnessFriend* is designed to clearly explain the motivation of this selection, through a user-friendly interface.

B. Interaction

In this section, we describe the design of the interaction between *FairnessFriend* and the user. In the next section, we provide details on the implementation of the chatbot.

To give an overview of the expected dialogue between the user and *FairnessFriend*, Figure 1 represents the expected

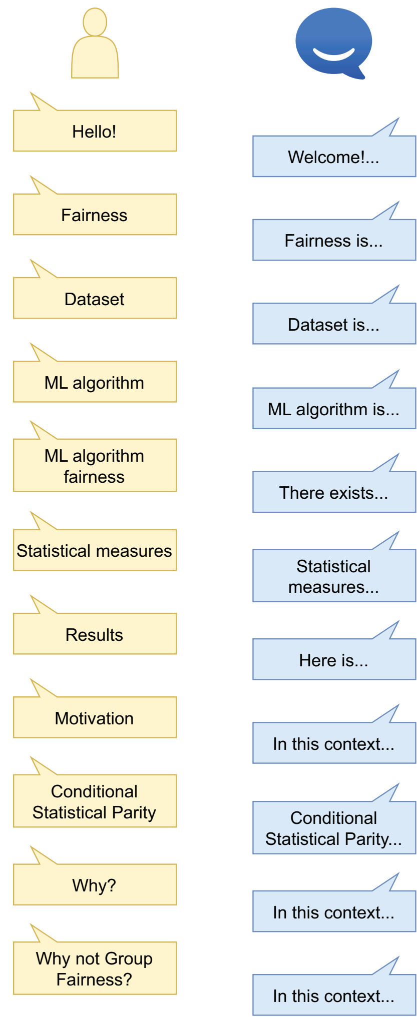

Fig. 1. Expected interaction between user and chatbot.

interaction flow of the system. On the left, the user's messages are displayed; on the right, the chatbot responses. For brevity's sake we only provide shortcuts; the full dialogue related to the scenario under analysis is available on the authors' Google Drive[2]. A demonstration video showing the user-chatbot interaction is also available[3].

The user starts the interaction with a word such as "Hello!" or "Hi!", then the chatbot explains what its task is and suggests some commands to facilitate the interaction. After that, the user can ask different questions to better understand the context in which *FairnessFriend* operates, through commands regarding the following concepts: "Fairness", "Dataset", "ML algorithm", "ML algorithm fairness", "Statistical measures". The chatbot responds by giving information on, respectively,

[2]https://drive.google.com/file/d/1E7JxzJdb0_TTeVlMzmbKiCWPTiyCe2_6/view?usp=sharing
[3]https://drive.google.com/file/d/1JbcpCveOA0CUiOa2w4vsFHXJbZPkAOmv/view?usp=sharing

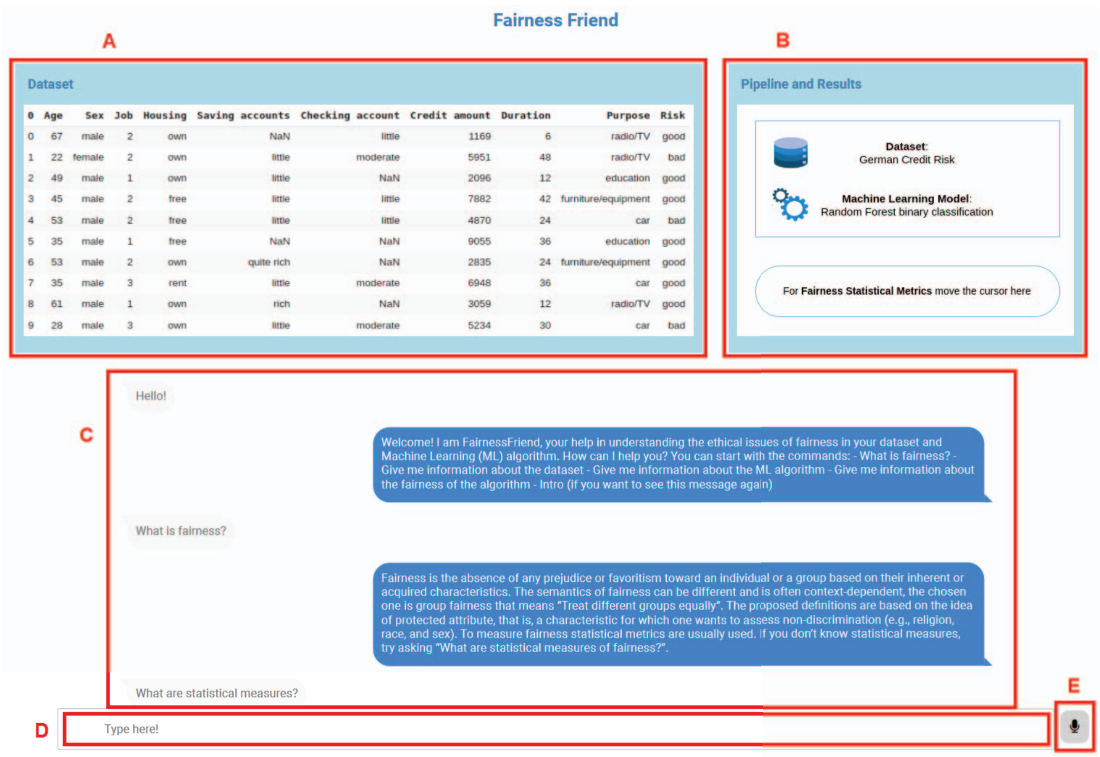

Fig. 2. *FairnessFriend* user interface.

the general concept of fairness, a description of the dataset and the machine learning algorithm used, the concept of fairness with respect to the specific machine learning algorithm used, and the statistical metrics used to measure fairness.

Secondly, *FairnessFriend* presents and explains the results of the computation of the statistical fairness metrics from Table I, responding to the the following two commands: "Results" and "Motivation". The last three commands in the interaction flow answer any further doubts the user might have with respect to the specific fairness metrics and their computation. This is the crucial moment when *FairnessFriend* allows the user to fully understand the fairness metric results, and interpret them accordingly to the chosen dataset, context and algorithm. In our scenario, the chatbot responds to the command "Conditional Statistical Parity" presenting the definition of the metric, the meaning of the metric considering the specific dataset, and the explanation of the computed values. The user can further query the chatbot with the command "Why?", to obtain the motivation and an explanation on the choice of the best fairness metric; in this case, it is Conditional Statistical Parity. The user can also ask "Why not X?" where X is a fairness metric other than Conditional Statistical Parity: this is a counterfactual question that further improves the interpretability of the result and clarifies the motivation behind the suggested metric. This allows the user to better explore the world of statistical fairness metrics, and get important insights for training other machine learning models in the same context, or in different ones. The following extract from a conversation with *FairnessFriend* shows the final steps of the interaction described before:

User: Give me information about the results of the statistical measures

FairnessFriend: The fairness metric that is more suitable for you is Conditional Statistical Parity, if you want to more about this, try asking 'What is the fairness metric most suitable in this context?'. This metric has a value near to 1, thus the model is fair.

User: What is the fairness metric most suitable in this context?

FairnessFriend: In this context, the protected attribute is 'sex', thus there is an interest in assessing fairness between women and men. We are only concerned with the results obtained from the ML-based decision system and the savings should be at the same level to have the same predictions, thus the proper definition is Conditional Statistical Parity. To know more about this metric, ask 'What is Conditional Statistical Parity?'. If you are not convinced, ask 'Why not Group Fairness?' or 'Why?' or 'Why not X?' where X is a statical measure.

User: Why not Group Fairness?

FairnessFriend: In this context we are only concerned with the results obtained from the ML-based decision system, but Group Fairness metric

```
does not consider other aspects, such as '
savings' that are important for our context. To
know more about this metric, ask 'What is Group
Fairness?'.
```

C. Interface

Figure 2 shows the web user interface of *FairnessFriend*, highlighting in red the different areas. On the left side (Panel A), the original input dataset is described, showing the first ten tuples. In this case, the German Credit Risk dataset. On the right side (Panel B), an image illustrates the machine learning pipeline under analysis, and – if the user moves the cursor over the panel – *FairnessFriend* lists all fairness metrics and their respective results, in a similar fashion as Table II.

On the bottom side of the interface, there is an input bar to interact with the chatbot. The user can interact directly with their voice by pressing the button with the microphone icon, located in the lower right corner (Area E). Otherwise, they can type directly on the text bar (Area D). The conversation with the chatbot is displayed in the centre of the page (Panel C).

IV. Technical Details

In this section, we present the technical details of our chatbot solution. *FairnessFriend* is a conversational agent based on text interaction through a web application. The architecture is shown in Figure 3. The conversational platform adopted is *DialogFlow*,[4] a standalone service offered by Google. DialogFlow is a platform with a series of facilities to create a wide range of conversational interfaces, typically integrated into "suites" of cloud services. It supports multiple languages and has REST API to interact with different programming languages. This framework is based on *intents* and *entities*. An intent is a mapping between what a user says and what action should be taken by the conversational agent. Typically, an intent is composed by:

- what a user says, e.g., a question or a statement;
- actions and parameters related to what the user says;
- a response, i.e., text, images or other form of media that the agent delivers to the user.

The full set of intents that we implemented are represented in Figure 4. On the other hand, entities represent actionable concepts for extracting parameter values from natural language. However, as our application does not take actions, we did not implement any.

DialogFlow exposes two services: *text-to-speech* and *speech-to-text* to interact with the system without the need for the user to write, thus improving accessibility to the application. For implementing the web user interface we used *Flask*, a flexible and lightweight Python framework to develop dynamic and interactive web applications.

At the moment, the platform embeds the results of the statistical fairness metrics: these metrics are implemented and computed inside a Python notebook under a *Colab*[5] environment. In this paper, we considered a single scenario,

[4]https://dialogflow.com
[5]https://colab.research.google.com

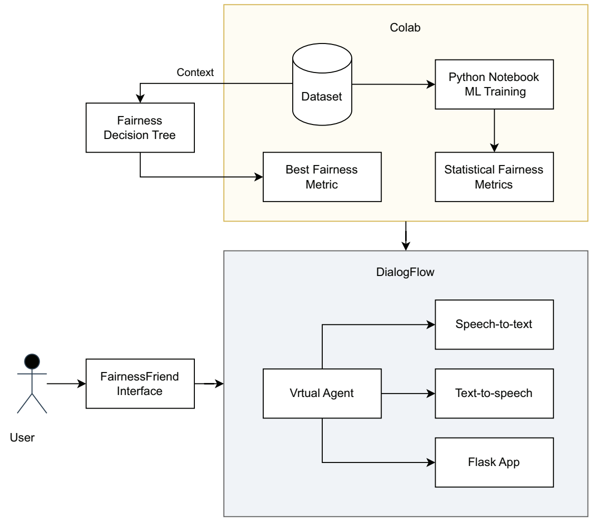

Fig. 3. *FairnessFriend* architecture.

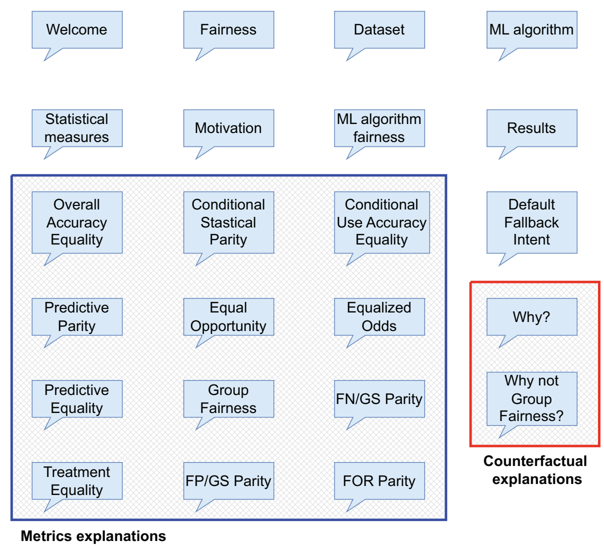

Fig. 4. Intents implemented in DialogFlow.

i.e., the German Credit Risk dataset on which a random forest binary classification model is trained. In the future, we plan to integrate an API to dynamically load any dataset, choose different ML models and embed the logic of the *Fairness Decision Tree* [11] to further customise the solution.

V. Related Work

The concept of fairness in machine learning systems has been studied extensively in recent years, particularly focusing on statistical metrics. New techniques to solve unfairness in ML systems are designed every year and have been compared from different points of view in recent surveys, focusing on bias and fairness in data [16], algorithms [17], or ML systems in general [18]. Fairness in AI and ML systems is significantly influenced by the interplay between the opportunities presented by these technologies and how human end-users adapt and collaborate with them over time, shaping their perceptions of fairness. This perspective studied in the HCI area, outlined in [19], underscores the importance of organisations and societies in approaching the adoption of AI to ensure fair perceptions among end-users.

Explanation to the user is a central topic in HCI: receiving explanations is one of the first end-users' expectations in order to trust these technologies. Explanation has a central role for trustworthiness of AI- and ML-based systems, and it is related to two different concepts: interpretability and explainability. *Interpretability* is the ability to provide explanations regarding the system's reasoning process and outcomes in terms that humans will understand, while *explainability* is the ability to create explanations that will be used as an interface between humans and the decision makers [20]. An in-depth survey has been formulated on the explainability problem in AI and ML algorithm-based systems [21]. Furthermore, [22] shows that fair decision-making requires extensive contextual understanding, and explanations help identify potential variables that are driving the unfair outcomes. For this reason, as expressed in [23], conversational agents are the perfect setting, since they are intuitive for humans and computationally processable. This is also indicated by a recent user study on virtual agents for explainable AI interaction systems [24]. Moreover, in [25]–[27] HCI researchers have studied, through extensive interviews, the human perception of fairness metrics, focusing on a limited set of definitions. All these works underline the importance of interpreting and explaining fairness notions, suggesting the creation of *ad hoc* solutions and frameworks to help users in understanding, reasoning on, and choosing the right definition of fairness.

Currently, most of the solutions that try to solve the fairness problem through explainability are not based on statistical metrics, but on causal metrics instead, that require causal graphs to analyse relationships between attributes and output [28]. Causality, through the visual form of a causal graph, has the benefit of immediately conveying the relationship between protected attributes (gender, ethnicity, religion, etc.) and output. Unfortunately, these techniques do not analyse, due to their settings, correlation relationships captured by statistical metrics, which are the ones most used in fairness assessment.

VI. Conclusions and Future Work

A. Conclusions

Machine learning algorithms drive decision-making systems, leveraging real-world data to predict outcomes. While enhancing efficiency, the lack of transparency in their predictions raises concerns about algorithmic fairness, exacerbated by notable recent examples. Researchers, data analysts and AI practitioners are increasingly aware of the risks and dangers of algorithmic fairness, but there is a lack of tools and supporting systems to adequately guide them in understanding the consequences and significance of machine learning outputs.

To address the problem of fairness interpretability, by combining data science with an HCI perspective, we propose a chatbot solution named *FairnessFriend*. After selecting a dataset, and training a machine learning model on it, our solution compute a series of statistical fairness metrics well established in the literature, providing insights on the significance of each metric, both in general and in the specific context of the task. In this paper, we focused on a real-case scenario involving the task of predicting the risk for banks of granting a loan, learning from the German Credit Risk dataset. We described our solution and showed the interaction between a user and our system. *FairnessFriend* illustrates the meaning of the different statistical fairness metrics, it shows the resulting fairness metrics with detailed explanations for each of them, and describes their implications, guiding the user towards better understanding of the problem of algorithmic fairness, and more informed decisions for the deployment of AI- and ML-based decision-making systems. Moreover, our solution is the first of such kind resulting from an interdisciplinary approach at the intersection of data science and HCI.

B. Future Work

We plan several future works to improve *FairnessFriend*. Currently, our system deals with a single scenario (i.e., credit risk prediction using the German Credit dataset), one machine learning model, and all the statistical metrics from the literature illustrated in Table I. We plan to create an API to easily integrate more datasets, different models, the possibility to include additional custom fairness metrics, and embed the *Fairness Decision Tree* [11] to build an expansible solution. Additionally, *FairnessFriend* is based on the DialogFlow paradigm offered by Google. In recent years, large language models (LLMs) are becoming pervasive, offering a more advanced option for developing conversational agents. We plan to test LLMs in our solution, to improve usability and interaction. Moreover, some answers provided in the current version of *FairnessFriend* are dense and very detailed, hence the conversation flow may be difficult to follow: we plan to conduct evaluation sessions to gather feedback from users and improve the interaction. Finally, *FairnessFriend* offers precise definitions taken from the literature. We plan to improve the initial explorative phase of the conversation, to allow users with little knowledge on statistics and machine learning to progressively discover the concepts underlying fairness metrics. In the future, our solution will assist users in implementing fairness by suggesting parameters and solutions for mitigating any unfairness detected.

Acknowledgements

We want to thank professors Maristella Matera, Massimo Zancanaro, Luigi de Russis, and Federico Cerutti for introducing us to HCI for AI and for giving precious comments and feedback to improve *FairnessFriend*. Additionally, we want to also thank professor Letizia Tanca for her contribution.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," ProPublica, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed: Feb. 2024.

[2] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," in *Ethics of data and analytics*. Auerbach Publications, 2022, pp. 296–299.

[3] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[4] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), vol. 67, 2017, pp. 43:1–43:23.

[5] M. Raghavan, "What should we do when our ideas of fairness conflict?" *Communications of the ACM*, vol. 67, no. 1, pp. 88–97, 2024.

[6] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, "How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations," *Artif. Intell.*, vol. 283, p. 103238, 2020.

[7] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *CoRR*, vol. abs/1901.10002, 2019. [Online]. Available: http://arxiv.org/abs/1901.10002

[8] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2125–2126.

[9] Y. Brun and A. Meliou, "Software fairness," in *Proceedings of the 2018 ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2018*. ACM, 2018, pp. 754–759.

[10] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018*. ACM, 2018, pp. 1–7.

[11] L. Baresi, C. Criscuolo, and C. Ghezzi, "Understanding fairness requirements for ml-based software," in *2023 IEEE 31st International Requirements Engineering Conference (RE)*. IEEE, 2023, pp. 341–346.

[12] P. Saleiro, K. T. Rodolfa, and R. Ghani, "Dealing with bias and fairness in data science systems: A practical hands-on tutorial," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2020, pp. 3513–3514.

[13] T. T. Hewett, R. Baecker, S. Card, T. Carey, J. Gasen, M. Mantei, G. Perlman, G. Strong, and W. Verplank, *ACM SIGCHI curricula for human-computer interaction*. ACM, 1992.

[14] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 4, oct 2020. [Online]. Available: https://doi.org/10.1145/3419764

[15] G. PAIR, "People + ai guidebook," 2019, https://pair.withgoogle.com/guidebook Accessed: Feb. 2024.

[16] A. Balayn, C. Lofi, and G.-J. Houben, "Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems," *The VLDB Journal*, vol. 30, no. 5, pp. 739–768, 2021.

[17] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, feb 2022. [Online]. Available: https://doi.org/10.1145/3494672

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, jul 2021. [Online]. Available: https://doi.org/10.1145/3457607

[19] D. De Cremer and G. Kasparov, "The ethical ai—paradox: why better technology needs more and not less human responsibility," *AI and Ethics*, vol. 2, no. 1, pp. 1–4, 2022.

[20] A. Shulner-Tal, T. Kuflik, and D. Kliger, "Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system," *Ethics and Information Technology*, vol. 24, no. 1, p. 2, 2022.

[21] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[22] J. Zhou, F. Chen, and A. Holzinger, "Towards explainability for ai fairness," in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 2020, pp. 375–386.

[23] S. F. Jentzsch, S. Höhn, and N. Hochgeschwender, "Conversational interfaces for explainable ai: a human-centred approach," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers 1*. Springer, 2019, pp. 77–92.

[24] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, ""let me explain!": exploring the potential of virtual agents in explainable ai interaction design," *Journal on Multimodal User Interfaces*, vol. 15, no. 2, pp. 87–98, 2021.

[25] M. K. Lee and S. Baykal, "Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017, pp. 1035–1048.

[26] M. Srivastava, H. Heidari, and A. Krause, "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2459–2468.

[27] A. Woodruff, S. E. Fox, S. Rousso-Schindler, and J. Warshaw, "A qualitative exploration of perceptions of algorithmic fairness," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–14.

[28] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *Advances in neural information processing systems*, vol. 30, 2017.