Explainable Agents for Less Bias in Human-Agent Decision Making

Avleen Malhi[1](✉), Samanta Knapic[2], and Kary Främling[1,2]

[1] Department of Computer Science, Aalto University, Espoo, Finland
{avleen.malhi,kary.framling}@aalto.fi
[2] Department of Computing Science, Umeå University, Umeå, Sweden
kary.framling@cs.umu.se, sakn0011@student.umu.se

Abstract. As autonomous agents become more self-governing, ubiquitous and sophisticated, it is vital that humans should have effective interactions with them. Agents often use Machine Learning (ML) for acquiring expertise, but traditional ML methods produce opaque results which are difficult to interpret. Hence, these autonomous agents should be able to explain their behaviour and decisions before they can be trusted by humans. This paper focuses on analyzing the human understanding of the explainable agents behaviour. It conducts a preliminary human-agent interaction study to investigate the effect of explanations on the introduced bias in human-agent decision making for the human participants. We test the hypothesis where different explanation types are used to detect the bias introduced in the autonomous agents decisions. We present three user groups: Agents without explanation, and explainable agents using two different algorithms which automatically generate different explanations for agent actions. Quantitative analysis of three user groups (n = 20, 25, 20) in which users detect the bias in agents' decisions for each explanation type for 15 test data cases is conducted for three different explanations types. Although the interaction study does not give significant findings, but it shows the notable differences between the explanation based recommendations and non-XAI recommendations in human-agent decision making.

Keywords: Explainable agents · Explanation type · Human-agent interaction · Human-agent decision making

1 Introduction

For the machine learning experts to rely upon the model's recommendations, explainability is an issue. It is easy for the decision makers to rely on a statistical tool which is easy to understand and convince the analytical results which is not possible for the machine learning models. Hence, the requirement is to find the methods by which the computational system can be explained to the decision maker for the complete understanding of the whole system. Explainable Artificial

D. Calvaresi et al. (Eds.): EXTRAAMAS 2020, LNAI 12175, pp. 129–146, 2020.
https://doi.org/10.1007/978-3-030-51924-7_8

Intelligence (XAI) shows up as a new branch of AI to benefit any intelligent agent or machine to explain their predictions. For instance, it is vital for an intelligent agent to explain its behaviour to the end user to make them more trustworthy. These type of explanations builds the trust in the classifier decisions even if the class is predicted wrongly as it explains for its unexpected behavior.

Recently, Explainable Artificial Intelligence (XAI), and explainable machine learning in particular, has gained increased attention in the research community. The main facilitation behind XAI is that although the machine learning models have gained attention is last few years, they are not interpretable from the human perspective. To address this shortcoming, researchers have developed algorithms that facilitate post-hoc explainability of machine learning-based classifications. While a range of such algorithms exists, the line of research that evaluates these algorithms from a Human-Computer Interaction (HCI) perspective is still in its infancy. The research questions addressed in the article are: (i) If an AI system is presented to a user, how will the developer know that the explanation is working correctly and the user is able to understand the machine learning decisions completely? (ii) How good the explanations are? (iii) How can we measure the goodness of explanations. (iv) Are users satisfied with the explanations provided? (v) If the end-user's trust and dependence on AI is enough? (vi) how the human-agent system behaves?

In this work, we advance the state of art of the HCI perspective by evaluating how two different post-hoc explanation algorithms–SHAP and LIME–influence bias in human decision-making. For this, we generate a (synthetic) data set of loan application decisions. The loan applications largely follow a set of simple decision rules, but are biased against women. We then train a machine learning model (neural network) on this dataset. In a user study, we then assess how decision support provided by the model is affected in regards to bias when explanation with LIME and SHAP are added.

2 Background

Despite plenty of research on transparent and interpretable machine learning models, providing explanations to technical users is an imperative area of study. The comprehensive surveys on explainable artificial intelligence [2,4] provide an insight into the machine learning, data analytics and visualization, challenges and future research directions for explainable deep learning. The research [15] uses two approaches for image classification using explainable deep learning where first explains sensitivity with respect to changes in input and second decomposes decision for its important input variables. Further, an interesting study on XAI understanding in a comprehensive form [10] can be generally grouped into three classes for understanding, diagnosing and refining. It also presents applicable examples relating to the prevailing state-of-the-art with upcoming future possibilities. The DARPA project [8] provides literature related to motivation and state of the work related to the examples for basic concept and application in the areas of legal advices, finance military, transportation, medicine and security for instance. The machine learning explainable system has been studied for various applications, for example in plant stress phenotyping [7] and heat recycler fault detection in air handling unit [11]. The authors also applied the same technique for providing explanations in medical images collected for capsule Gastroenterology [12]. However there are unprecedented obstacles with the current efforts made by researchers since the traditional machine learning models are less interpretable and more complex with AI used for majority of the tasks. Further, AI is used more for making autonomous decisions than ever by introducing the agents. Hence there is no doubt that the agent autonomy will continue rising in eminence with more exciting work in the future [3].

The ability of an agent to plan and act effectively on its own towards a goal is determined by the agent's actions, when it can perform actions and the outcomes of these actions. The progress is defined by an explainable agent which is able to learn the preconditions related to action and then perform preparatory planning process. Hence an agent performs both exploratory as well as goal-directed actions which opens up the research questions related to controlling actions of exploratory and goal planning and the explanation of agent's behaviour to any technical user [5]. The virtual agents' impact in the area of XAI is examined based on the trust in the autonomous intelligent systems [16].

For assessing the practicality of the trust in autonomous agents, a user study is conducted based on simple bank loan application. As a consequence of this study, we came to a significant evidence indicating that an interactive design of application by integrating the virtual agents with XAI, the trust of the user in the autonomous intelligent agents increase. The objectives of explanation comprises of investigating the questions such as, "How does the system work?"; how easy is it to understand?; What does it do?; Is the user able to trust the system?; and "Is the system able to justify user for its decisions?". The proposed work tries to address the following question: Suppose an AI system which explains its working is presented to a user, what are the ways to measure if it works or not, how accurately it performs, is the user able to have the practical understanding about the system. The aim of the paper is to measure the end-user confidence in understanding the machine learning recommendations with and without explanations, and how well the bias can be reduced with the help of human-agent decision making.

3 Explainable Artificial Intelligence

The machine learning black box models excel in their task of decision making but precisely do not permit to make human understandable decisions. An organization at the forefront of XAI research is the United States' *Defense Advanced Research Projects Agency (DARPA)*. DARPA report defines XAI as: XAI allows an "end user who depends on decisions, recommendations, or actions produced by an AI system [...] to understand the rationale for the system's decisions" [1]. According to a survey conducted by Miller [13], the major findings regarding the properties of explanations in human-like interactions are:

- **Explanations are contrastive:** People have the tendency of not asking why something happened but instead why something happened instead of something else. They try to create the reference between their expectations and the reality.
- **Explanations are Selected:** People rarely expect any explanation covering all aspects of reasoning.
- **Probabilities don't matter:** People consider casual explanations more relevant that pure correlations.
- **Explanations are social:** Explanations are considered as transfer of knowledge as part of interaction which also involves queries as well as preferences.

The behavioural and social challenges should also be taken into consideration for better design decisions while developing the explainable agents. A black box model decisions are sometimes too complex for a human to understand, or it is a model that is challenging to troubleshoot. The explanations need to be considered as a separate tool for replicating the black box behaviour. Most of the recent works on transparent and interpretable machine learning decisions only focus on the technical users. End user explanations are overlooked in many useful and practical applications. Unless humans understand the model's reason of assessment, they can not trust them [9].

SHAP value is a united approach for explaining any machine learning model's output. It has the following characteristics: (i) global interpretability – how much each predictor contributes to the target variable, either positively or negatively. (ii) local interpretability – SHAP values are calculated for each instance which greatly increases its transparency. It helps in explaining the prediction of a case and the major contributors in decision. (iii) the SHAP values can be computed for any model which is tree based.

Local Interpretable Model-agnostic Explanations (LIME) is another explanation tool for providing the explanations for the predictions using the most important contributors. It helps the decision makers in justifying the model behaviour with respect to the important input parameters. The overall purpose of LIME is to identify an interpretable model over the interpretable representation which can fit the classifier locally. The underlying model's approximation by interpretable model is used to generate the explanations by learning the original instance's disruptions. LIME is a simple tool which approximates the black box locally compared to approximation on a global scale. The original instance is weighted by similarity to the instance which we wish to explain. LIME provides the model agnostic explanations which makes it easier to use LIME to explain innumerable classifiers (such as Random Forests, Support Vector Machines and Neural Networks) Because our goal should be to have model-agnostic model, using textual or image data [14].

3.1 Challenges of Explainable Machine Learning

There are significant misconceptions related to the current work on explainability which can effect negatively on its wider social acceptance.

1. *Trade-off between interpretability and accuracy:* It is believed that the complex models present more accurate results which implies that best predictive results can only be obtained by a complex black-box model which is not interpretable at all. The fact is that interpretibility can be imbibed perfectly with the deep learning applications without affecting the performance of the system.
2. *Explainable Machine learning methods provides unfaithful explanations:* It is common belief that if the explanation is exactly what the original model computed, then we do not need the original model at first place. It leads to the situation where it is considered that explanations are the original model's inaccurate representation in parts of the feature space. The explanations methods actually compute the summary of the prediction results of the model instead of exact explanations.
3. *Incomplete explanations:* Sometimes, the explanation may not give complete information that the meaning becomes unclear. It might impart a false confidence in the black box explanation method.
4. *Non-compatibility of the black box models to assess risk:* Some machine learning decisions can increase of decrease the estimated risk. The additional information provided by black box model may increase or decrease the level of risk assessment.

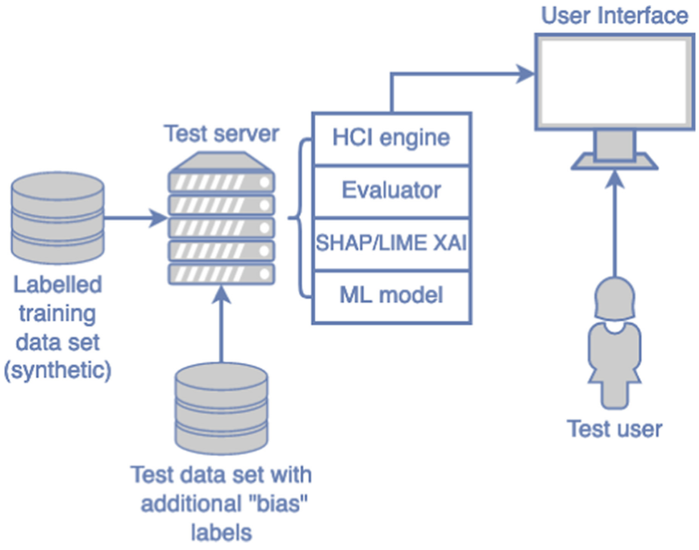
User Interface
Test server
HCI engine
Evaluator
SHAP/LIME XAI
ML model
Labelled training data set (synthetic)
Test data set with additional "bias" labels
Test user

Fig. 1. Test setup and architecture

4 Human-Agent Interaction Method

This section provides an overview of the methodology used to evaluate the impact of explanations on the bias introduced in the models in the human decision making (Fig. 1). We want to emphasize that by agents, we refer to the AI systems which are capable of decision making by themselves. We design an application which is responsible for recommending the decisions to the users. We generate synthetic data, a machine learning model predictions, and post-hoc explanations that allow for the evaluation of the ability of the post-hoc explanation techniques Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to avoid biased decision making in humans. For this, firstly we generate a (synthetic) data set of loan application decisions. The loan applications largely follow a set of simple decision rules, but are biased against women. The potential biases which can be introduced in the dataset are due to *gender* and *age*. The potential features which can be used in the dataset are listed below:

- Gender
- Age
- Income
- Unpaid debt
- Wealth
- Educational background/profile: ties to the place/country
- Other liabilities
- Credit history
- Job stability

The general architecture for human agent interaction[1] has been explained in the Fig. 1 where labelled training data generated synthetically is used for training a machine learning model. The explainable recommender agent gives the decision of the model and also explains the decision using various XAI based methods such as LIME, SHAP etc. The recommendations and explanations provided by the machine are evaluated by designing an appropriate user interface for a test user.

4.1 Generate Test Data

We generate a data set of loan applications and their decisions. Each loan application has the following parameters:

1. Age (age) of the applicant in years;
2. Income (income) of the applicant in €;
3. Debt/assets (assets) of the applicant in €;
4. Employment type (employment) of the applicant (fixed-term, permanent);
5. Gender (gender) of the applicant (female, other, male);
6. Loan size (loan) in €.

[1] https://colab.research.google.com/drive/1-iq1xZhYuKZgH5NgYUBETYmun9yoKMdC.

Table 1. Decision rules for test data

If age < 18: reject
Else if income < 20000 and loan > 10000: reject
Else if assets < 20000 and loan > 10000: reject
Else if assets < 0: reject
Else if employment != permanent and loan > 400000: reject
Else if loan > 500000: reject
Else: accept

Table 2. Bias added

If Gender == 'female' or 'other': reject with probability of 80%
Else: call the first set of rules

The basis for classifications in the test data set are the decision rules given in Table 1. In addition, we induce the rule given in Table 2 that adds bias to the classifications. For loan application data, we create a data set with a size of 10000 entries with the properties given in Table 3.

With the exception of gender, all parameters are approximately uniformly distributed, i.e. we use Python's random.uniform() function to assign any of the possible values. Note that for the scope of the study, it is not necessary to create a representative data set; instead, it is important to have a data set that contains a large amount of entries that will be affected by the gender-biased decision rules. Gender is distributed approximately as follows: 10% other, 50% female, 40% male.

4.2 Training of Model

We then train a machine learning model (Random Forest with gradient boosted trees) on this dataset. The model is trained with 80% of the data and 20% is used for testing the data. The trained data is biased with gender as explained in an earlier section. The model is then tested with rest of the 20% of the data which provides the recommendation for the loan application in the form of approve or reject. In a user study, we then assess how decision support provided by the model is affected in regards to bias when explanation with LIME and SHAP are added.

4.3 Explanation Types

Out of the XAI approaches previously discussed in Sect. 3, we used LIME and SHAP to explain the decisions of our Test use case: bank loan approval. We used the following methods of providing explanations for the explanation agents:

Table 3. Dataset variables

Age	17–70 years;
Income	0–200000€;
Assets	100000–1000000€;
Employment type	Fixed-term or permanent;
Gender	Female, male or other;
Loan amount	5000–520000€

Table 4. Sample data for generating explanations

Income	68100
Gender	Male
Employment	Fixed
Loan	479000
Assets	271900
Age	54

No Explanation. The agent does not provide any form of explanation for recommendations made. The black-box XAI acts as a baseline for our empirical assessment.

Explanation I: LIME. The agent explicitly states the explanation of the decision providing post-hoc explainability of the model decision. The model recommendation provided is complemented with the explanations to justify the machine recommendations. The explanations are used to test the bias-preventing effects of XAI. We use Local Interpretable Model-agnostic Explanations (LIME) as our first post-hoc explainability algorithm and generate the explanations that will be used in the human-computer interaction study. The explanations provided for a particular test case (Table 4) are depicted in Fig. 2 with *Reject* recommendation.

Explanation II: SHAP. We use SHapley Additive exPlanations (SHAP) as our second post-hoc explainability algorithm for generating the explanations to be used in the human-computer interaction study. Figure 3 where the recommendation provided by machine learning model is *Reject*.

5 Empirical Assessment

To investigate the effect of explainable agents, we conducted a human-computer interaction study as a foremost step for providing an empirical assessment of the proposed concept.

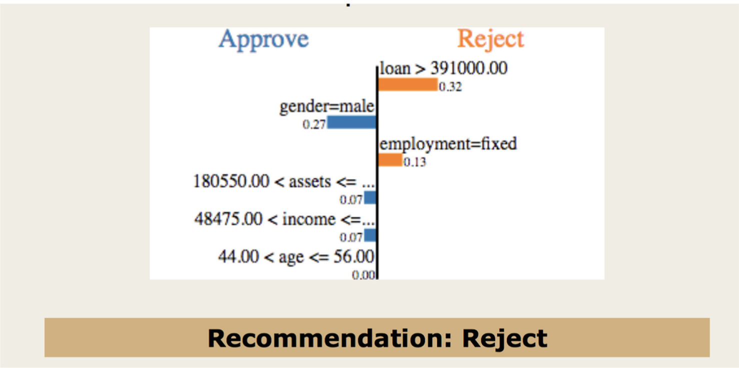

Fig. 2. Lime explanations with recommendation

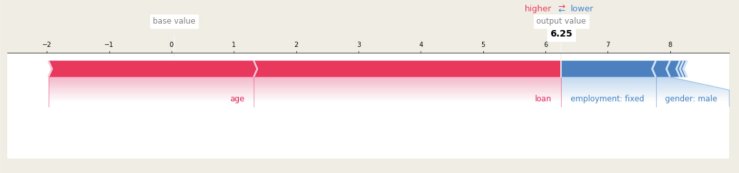

Fig. 3. Shap explanations

5.1 Study Description

The aim of this study is to gather preliminary facts for the explanation of the bias-preventing effects of XAI methods: LIME, SHAP with respect to the black-box as baseline with no XAI. The study is conducted with 65 participants with 20 for black-box XAI, 25 for LIME and 20 for SHAP. Fifteen different case data for loan application are generated and the recommendations are provided to approve or reject the loan application. The user has to select if he approves or rejects the loan application based on the recommendation provided. Three different interactive applications were generated as:

- Black-box based recommendation for loan application without XAI.
- XAI based recommendation for the loan application with visual explanations using XAI tool LIME.
- XAI based recommendation for the loan application with visual explanations using XAI tool SHAP.

Hypotheses. The aim of the study is to evaluate the following hypothesis:

1. H_a**:** Number of "overridden" recommendations that are biased is higher for SHAP then without explanations (true positive).
2. H_b**:** Number of "overridden" recommendations that are not biased is lower for SHAP then without explanations (false positive).

3. H_c**:** Number of "overridden" recommendations that are biased is higher for LIME then without explanations (true positive).
4. H_d**:** Number of "overridden" recommendations that are not biased is lower for LIME then without explanations (false positive).
5. H_e**:** Number of "overridden" recommendations that are biased is higher for LIME then for SHAP (true positive).
6. H_f**:** Number of "overridden" recommendations that are not biased is lower for LIME then for SHAP (false positive).

The study introduces the bias in *four out of fifteen* test cases and our hypothesis aims at evaluating whether the humans are able to detect the bias in agent supported recommendations or not. We are testing whether we can reject the null hypothesis (H_{a0}, H_{b0}, H_{c0}, H_{d0}, H_{e0}, H_{f0}) being the negations of our six hypotheses.

5.2 Data Collection

Study Protocol. For this user-centric study, we got the participants from the University's environment which means most of the participants have a technical university degree.

1. Initially, the study participant is introduced to the user study. The study instructions are given to participant by a facilitator in the form of written instructions. The bias introduced was not disclosed until the end of the study.
2. After providing the instructions, the study is carried out under the supervision of one researcher who helps in controlling the experiments as planned.
3. The study participant is asked to give the recommendation if he accepts or rejects the loan application based on the case data provided for any of the above 3 applications discussed.
4. The process is iterated for 15 rounds of different case data.
5. After all the fifteen rounds of the application are completed, the participant is guided through the questionnaire. Since these questions could potentially affect the respondent's assessment about the process, so these questions are asked at last after the application assessment has been carried out and could not be accessed by the participant beforehand.

Questionnaire Design. Initially the users are asked to interact with the application and provide the data, **Q0:** Received the user responses in the form of approve or reject. We asked the users to provide the following demographic data **Q1:** Age (number); **Q2:** Gender (Selection: male, female, other); **Q3:** Highest educational degree (Selection: Pre-high school, High school, Bachelor, Master, Ph.D. or higher); **Q4:** Background in science, technology, engineering, or mathematics (STEM) [Boolean]; To evaluate the interactions between study participants, the following data was taken regarding their performance:

- Were they able to understand the (explanations of the) recommendations provided by application [Boolean];

- To rate their satisfaction level of (explanations of the) recommendations on a scale of 0-5;
- Which parameters they consider important in deciding if to approve or reject the loan application (multiple selections from income, gender, employment, loan, assets, age);
- To rate the user interface of the application on a scale of 0–5;

There were few different questions for the questionnaire generated for the study without explanations[2] and with explanations[3]. The following additional questions were designed for the study without explanations apart from the above defined questions:

- Do the users see themselves trusting the recommendation without an appropriate reason for its decision [boolean];
- If the decisions would be more satisfying with explanations along with recommendations;
- What kind of explanations the users expect to support the recommendations;

The following additional questions were designed for the both studies with explanations:

- If they heard about explainable machine learning before [boolean];
- If the user answers yes to the above question, then describe their knowledge about explainable machine learning in few words;
- Do they consider the parameters analysed by application as important [boolean];
- Do they think provided explanation is good enough to let them trust or not the recommendations provided;
- Describe possible improvements of the explanations to improve understandability;
- Describe interaction experience with application.
- Describe possible improvements of the user interface in terms of design;
- Can the users see themselves using the decision making application with given explanation;

Analysis Methods. In order to investigate the effect of explanations provided by the autonomous agents on the human participants we performed a comparison between three different user groups: Agents without explanation, and explainable agents using two different algorithms (LIME and SHAP) which automatically generate different explanations for agent actions. We analysed the results using Excel XLMiner Data analysis ToolPak to run hypothesis tests as well as exploratory statistics. Firstly, we determined the differences between means and medians of human decision making in different settings. For each hypothesis, we

[2] https://docs.google.com/forms/d/1nxJpzdo8y5QiCFeHo6LY8M86u6istdAmau67pUdDZ1g/viewform.

[3] https://docs.google.com/forms/d/1CTatrqSgjX_PUYxxRGjktOdHaPepaKA1clgL41io3t4/viewform.

Table 5. Demographics of study participants for noEXP, LIME, SHAP

Methods	Total	Gender			Highest degree				STEM background		Age (years)
		Male	Female	OTH	Ph.D. (or higher)	Master	Bachelor	High school	Yes	No	
noEXP	20	10	9	0	1	5	7	6	13	7	21 (2), 23, 24 (2), 26(2), 27(2), 28(3), 30(3), 31(2), 34, 50, 57
LIME	25	18	5	1	4	12	6	2	24	1	20, 24(3), 25(2), 28, 29(4), 30(4), 32(3), 33(2), 37(2), 38, 51, 53
SHAP	20	11	7	1	7	9	3	1	18	2	21, 23, 24, 25(2), 26, 27(3), 28, 29, 32, 33(2), 34, 35, 36(2), 38, 41

tested the difference between distribution of decisions using two-Sample t-Test assuming equal variances with significance level of α set to 0.05. The correlation is calculated using Pearson correlation coefficient between demographic values and count of the right decisions (the decision overriding biased recommendation and approving non biased recommendations) by study participants.

65 people participated in the study (n = 65) out of which 20 were given no explanation during application interaction, 25 were given the explanations given by LIME and 20 by SHAP. When performing the correlation analysis we excluded two of the participants identified as other in terms of gender. The demographics of the participants are shown in Table 5. The study participants were predominantly male and predominantly had high education and background in science and technology. Majority is in their twenties or thirties with few outlining cases.

5.3 Result Analysis

Quantitative Analysis. The analysis starts with calculation of true positives, false positives, true negatives and false negatives which signify the following with respect to our evaluation criteria:

True Positive = *Overrides biased recommendation*
False Positive = *Overrides non biased recommendation*
True Negative = *Supports not biased recommendation*
False Negative = *Supports biased recommendation*

Table 6. Mean and median measures

	Measures	noEXP	LIME	SHAP
False negative	Count	32	37	23
	Mean	4.05	3.48	3.05
	Median	2	1	1
False positive	Count	81	87	61
	Mean	1.6	1.48	1.15
	Median	3	3	3
True negative	Count	139	188	159
	Mean	6.95	7.52	7.95
	Median	8	8	8
True positive	Count	48	63	57
	Mean	2.4	2.52	2.85
	Median	2	3	3

Table 7. Hypothesis analysis

	t-test	Hypothesis	p-value (two-tailed)	p-value (one-tailed)
1	true positive (SHAP, noEXP)	Ha0	0.18	0.09
2	false positive (SHAP, noEXP)	Hb0	0.13	0.06
3	true positive (LIME, noEXP)	Hc0	0.71	0.35
4	false positive (LIME, noEXP)	Hd0	0.35	0.17
5	true positive (LIME, SHAP)	He0	0.36	0.18
6	false positive (LIME, SHAP)	Hf0	0.39	0.19

Table 8. t-Test: two-sample assuming equal variances (true positives)

	noEXP	LIME	noEXP	SHAP	LIME	SHAP
Mean	2.4	2.52	2.4	2.85	2.52	2.85
Variance	0.7789473684	1.426666667	0.7789473684	1.397368421	1.426666667	1.397368421
Observations	20	25	20	20	25	20
P(T<=t) one-tail	0.3549150958		0.09027080274		0.180026242	
P(T<=t) two-tail	0.7098301915		0.1805416055		0.3600524839	

Table 9. t-Test: two-sample assuming equal variances (true negatives)

	noEXP	LIME	noEXP	SHAP	LIME	SHAP
Mean	6.95	7.52	6.95	7.95	7.52	7.95
Variance	5.628947368	2.76	5.628947368	2.681578947	2.76	2.681578947
Observations	20	25	20	20	25	20
P(T<=t) one-tail	0.1745331086		0.06455759198		0.1950434934	
P(T<=t) two-tail	0.3490662172		0.129115184		0.3900869869	

The count, mean and median for each of the above type of recommendation for each of the three user study groups is calculated as shown in Table 6. It depicts that there are notable differences in means aligned with the assumption

Table 10. t-Test: two-sample assuming equal variances (false positives)

	noEXP	LIME	noEXP	SHAP	LIME	SHAP
Mean	4.05	3.48	4.05	3.05	3.48	3.05
Variance	5.628947368	2.76	5.628947368	2.681578947	2.76	2.681578947
Observations	20	25	20	20	25	20
P(T<=t) one-tail	0.1745331086		0.06455759198		0.1950434934	
P(T<=t) two-tail	0.3490662172		0.129115184		0.3900869869	

Table 11. t-Test: two-sample assuming equal variances (false negatives)

	noEXP	LIME	noEXP	SHAP	LIME	SHAP
Mean	1.6	1.48	1.6	1.15	1.48	1.15
Variance	0.7789473684	1.426666667	0.7789473684	1.397368421	1.426666667	1.397368421
Observations	20	25	20	20	25	20
P(T<=t) one-tail	0.3549150958		0.09027080274		0.180026242	
P(T<=t) two-tail	0.7098301915		0.1805416055		0.3600524839	

Table 12. Correlation between demographics and decision making

Demographics	no EXP (correlation)	no EXP (p-value)	LIME (correlation)	LIME (p-value)	SHAP (correlation)	SHAP (p-value)
Age	−0.3943338142	0.08534479016	−0.2122591802	0.3083793126	−0.2735918895	0.2431317331
Gender	−0.01899685628	0.9366405456	−0.2145147339	0.314135297	0.08552499375	0.7277463724
Education	−0.1699636098	0.4737466067	0.01314368012	0.9502790615	−0.1209607494	0.6114580972
STEM background	−0.05775093751	0.80890297	0.09386465089	0.6553926784	−0.1433608882	0.5465249821

that motivate our first five hypothesis regarding the differences between modes with explanation versus modes without explanation. However the differences are statistically not significant.

Our hypotheses are applicable for only true positives and false positives and the Table 7 gives the calculated p-values for two-tailed as well as one-tailed tests to test our null hypothesis (negations of our hypotheses). There are not observed significant differences but the notable differences can be seen for overriding the bias based recommendations higher for LIME than no explanation. There are also notable differences for overriding non-biased recommendations lower for SHAP than no explanation. Hence, the results supports our hypotheses H_b, H_c to a little extent. However the results are not in favour of our last sixth hypothesis since number of overridden recommendations that are not biased is not lower for LIME than for SHAP. Considering the small sample size, the results can not be generalized. Tables 8, 9, 10 and 11 are showing the means, variances, number of observations and p-values (both one and two tailed) which indicate whether there were observed differences between different groups of participants.

Further we performed the correlation analyses using Pearson coefficient between the complete count of the right decisions (true negative and true positive decisions) made by participants and the demographic variables. Table 12 is showing values of Pearson correlation coefficient for all conditions - no explanation, LIME and SHAP - and the p-values which depict if the computed coefficients are showing statistical significant correlations between demographic variables and the right decisions (true negative and true positive) of the participants. We did not find a significant correlation, although the correlation between age and count of the right decision in group without explanation seems plausible which may be indicating that the lower age was somehow predicting the higher number of right decisions.

Qualitative Analysis Interaction Experience. In a group without provided explanation, participants generally considered income as the most important parameter when deciding for approving or rejecting the loan whereas the gender was noted as the least important and majority was satisfied with the user interface of the application. Participants in group with given explanations (SHAP and LIME) similarly rated income as the most important parameter when deciding for approving or rejecting the loan whereas the gender was noted as the least important and majority perceived the user interface of application as good.

Explanation Evaluation. Most of the participants in group without provided explanation, answered that they did not understand recommendations provided by application and that they can not see themselves trusting the recommendations provided by the given application without the provided explanation. They were mostly satisfied with the given recommendations but also noted that they would like to have an explanation added in the application. In groups with provided explanations (SHAP and LIME) participants mainly answered that they understood the explanations of the recommendations provided by application and were satisfied with the explanations provided. About half of the participant answered that the given explanation was good enough in order to let them judge when they should trust or not trust provided explanations and they could also see themselves using the application with the given explanation. By analyzing the participants' free-form feedback, we additionally found that:

- End-users want additional linguistic explanations along with visual based explanations.
- End users want explanations to be suitable for intuitive comparisons.
- End users want to interact with agent for more information.

6 Discussion and Perspectives

From observed results comparing human decisions from different groups of participants (two with explanations versus no explanation group), we observed notable differences between groups in mean/median of their decisions. Those differences may reflect the initial assumptions stating that both SHAP and LIME explanation will cause less overriding of non biased recommendations and more overriding of biased recommendations than having no explanation at all. However, as our hypothesis testing showed these differences are statistically not significant, and we therefore can not draw empirically valid inferences. Our initial assumption that participants with LIME will perform better in overriding more biased recommendations than participants with SHAP explanation, was also to some extent supported by our results, however participants having LIME explanation did not perform better in overriding less non-biased recommendations than those with SHAP. This could indicate that participants with SHAP explanation to some extent engaged in more understanding of the explanation since SHAP explanation has higher complexity compared to LIME which can also be concluded from the participants comments from their interaction with application.

Our results also give insight into the initial research questions. The user study and interaction with no explanation and explanation study depicts that users are able to understand the explainable AI systems more profoundly and are comfortable in the recommendations provided by these systems compared to the noEXP systems. This imbibes more confidence in the developers to have more explainable systems which will instill more confidence in users to trust such systems. The two explanation types have been used to understand the goodness of the explanations and the parameters which users consider as important in regard with explanations provided. Thus, the study provided a deep insight into the details of these systems from the perspective of users which can be used as positive feedback for the design of such AI systems.

6.1 Limitations

The paper provides human-agent interaction study to reduce the bias in human decision making with the help of explanations provided with recommendations. The interaction study has a set of limitations, the most important of which are listed below:

- **The scope is limited to only two explanation tools:** The current approach focuses only on two explanation tools, LIME and SHAP which can further be validated with other more sophisticated tools such as CIU [6].
- **The time for a decision is not taken into account:** The user time for making a decision in the study is not considered which may effect the empirical validation of the study. It would be, for example, interesting to know if people presented with SHAP explanation took longer time in deciding to approve or not to approve loan in each case. Because the provided results are not showing significant differences in results between groups without explanation and groups with explanations for decision making which could indicate the presence of human bias.
- **Human bias in decision making:** The paper provides simplistic scenario of the loan application data which neglects the human related biases. It means that human participants possibly ignored the given recommendations, or did not pay that much attention to it as expected and/or that they ignored the explanations of recommendations by focusing more on their own assumptions.
- **The scope is limited to a synthetically generated data:** For facilitation of real life applicability, it is necessary to use the real application data in the context of real world situations in the actual real-life settings. The sample should be more diversified and it would be good to control the demographics as well.

6.2 Future Work

The following future research directions can be considered to address the limitations discussed in previous subsection:

- **To scale the study to other XAI tools:** We present the study with LIME and SHAP as explanation tools, it will be good to test the concept on a more wider perspective with other explainable artificial intelligence tools as well such as CIU, ELI5 etc.
- **Evaluate the study applicability with domain experts:** While we have provided a prototype that shows the applicability in the generated dataset, its exact usability can be validated with the domain experts.
- **Extend the scope to real-life case study:** It will be interesting to explore the actual case study with real life complexities to show that the agents act more rationally in real life applications.

7 Conclusion

We try to explain the behaviour of the autonomous agents to humans by conducting a preliminary human-agent interaction study to investigate the effect of explanations provided by agents to lower the biased based recommendations. In this paper, we explored the potential of the bias based recommendations in human decisions for three different groups of participants, 2 groups with explanations provided and 1 group without explanations. The results of our study show the improved trend of user's perceived trust in explanation based recommendations compared to the ones with no explanation for the less bias in agent supported human decision making. The results of our study are inline with our initial assumption that end-users experience could benefit from explanation based recommendations to reduce the bias in human decision making. The presented agent-supported interaction study for enabling human-agent decision making pave the way for exhaustive evaluations for the effectiveness of the agent supported decisions.

Current user study for supporting agent-human decision making concerns the integration of a system able to detect user's approval or rejection for the machine recommendations; further development of interaction strategies such as management of socio-emotional factors and the decision time in human agent interaction. We expect that such integrations will contribute in providing realistic interactions and improved results.

References

1. Defense advanced research projects agency (DARPA): Broad agency announcement- explainable artificial intelligence (XAI) (2016)
2. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

3. Biran, O., Cotton, C.: Explanation and justification in machine learning: a survey. In: IJCAI 2017 Workshop on Explainable AI (XAI), vol. 8, p. 1 (2017)
4. Choo, J., Liu, S.: Visual analytics for explainable deep learning. IEEE Comput. Graph. Appl. **38**(4), 84–92 (2018)
5. Dannenhauer, D., Floyd, M.W., Molineaux, M., Aha, D.W.: Learning from exploration: towards an explainable goal reasoning agent (2018)
6. Främling, K.: Explaining results of neural networks by contextual importance and utility. In: Proceedings of the AISB 1996 Conference. Citeseer (1996)
7. Ghosal, S., Blystone, D., Singh, A.K., Ganapathysubramanian, B., Singh, A., Sarkar, S.: An explainable deep machine vision framework for plant stress phenotyping. Proc. Natl. Acad. Sci. **115**(18), 4613–4618 (2018)
8. Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2017)
9. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923 (2017)
10. Liu, S., Wang, X., Liu, M., Zhu, J.: Towards better analysis of machine learning models: a visual analytics perspective. Vis. Inf. **1**(1), 48–56 (2017)
11. Madhikermi, M., Malhi, A.K., Främling, K.: Explainable artificial intelligence based heat recycler fault detection in air handling unit. In: Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.) EXTRAAMAS 2019. LNCS (LNAI), vol. 11763, pp. 110–125. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30391-4_7
12. Malhi, A., Kampik, T., Pannu, H., Madhikermi, M., Främling, K.: Explaining machine learning-based classifications of in-vivo gastral images. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7. IEEE (2019)
13. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
14. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
15. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)
16. Weitz, K., Schiller, D., Schlagowski, R., Huber, T., André, E.: Do you trust me?: increasing user-trust by integrating virtual agents in explainable AI interaction design. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, pp. 7–9. ACM (2019)