Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance

HONG SHEN, Carnegie Mellon University, USA
HAOJIAN JIN, Carnegie Mellon University, USA
ÁNGEL ALEXANDER CABRERA, Carnegie Mellon University, USA
ADAM PERER, Carnegie Mellon University, USA
HAIYI ZHU, Carnegie Mellon University, USA
JASON I. HONG, Carnegie Mellon University, USA

Ensuring effective public understanding of algorithmic decisions that are powered by machine learning techniques has become an urgent task with the increasing deployment of AI systems into our society. In this work, we present a concrete step toward this goal by redesigning confusion matrices for binary classification to support non-experts in understanding the performance of machine learning models. Through interviews (n=7) and a survey (n=102), we mapped out two major sets of challenges lay people have in understanding standard confusion matrices: the general terminologies and the matrix design. We further identified three sub-challenges regarding the matrix design, namely, confusion about the direction of reading the data, layered relations and quantities involved. We then conducted an online experiment with 483 participants to evaluate how effective a series of alternative representations target each of those challenges in the context of an algorithm for making recidivism predictions. We developed three levels of questions to evaluate users' objective understanding. We assessed the effectiveness of our alternatives for accuracy in answering those questions, completion time, and subjective understanding. Our results suggest that (1) only by contextualizing terminologies can we significantly improve users' understanding and (2) flow charts, which help point out the direction of reading the data, were most useful in improving objective understanding. Our findings set the stage for developing more intuitive and generally understandable representations of the performance of machine learning models.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Algorithmic decision-making; Machine Learning; Explainable AI; Human-centered AI; Confusion Matrix

ACM Reference Format:
Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 153 (October 2020), 22 pages. https://doi.org/10.1145/3415224

Authors' addresses: Hong Shen, hongs@cs.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Haojian Jin, haojian@cs.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Ángel Alexander Cabrera, cabrera@cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Adam Perer, adamperer@cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Haiyi Zhu, haiyi@cs.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA; Jason I. Hong, jasonh@cs.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA.

2573-0142/2020/10-ART153 $15.00
https://doi.org/10.1145/3415224

1 INTRODUCTION

Imagine you are a journalist who needs to write a story covering issues of differential treatment of two demographic groups by a recidivism prediction algorithm [4]. This task involves explaining the performance of a machine learning classifier on two demographic groups to facilitate informed public debate. Effectively communicating the machine performance to your readers, the vast majority of whom have little or no knowledge of computer science or machine learning, is not easy, as most representations of these results are designed for machine learning experts. Lay people often lack the technical background required to understand relatively complex concepts and terms (e.g., true positive, false negative) as well as the performance of the underlying machine learning models.

The need for explaining and presenting the performance of algorithmic decision-making systems powered by machine learning techniques to the non-expert public is a widely shared challenge across many domains. With the increasing deployment of algorithmic decision-making systems in many critical parts of our society, such as college admissions [42], loan decisions [55], and child maltreatment prediction [12], many social and ethical concerns are being raised (e.g., [4, 15, 17, 43]) and urgently need public evaluation and debates. Effective public understanding of algorithmic decisions – in particular, the performance of the underlying machine learning models – serves as one way to support such public debates.

To date, the majority of past work on explaining and visualizing AI systems has primarily focused on supporting *expert* decision-making ([13, 16, 30, 31, 46]). With the growing impacts of algorithmic decision-making in our society, a growing body of work in HCI has called for more efforts to explore how to better explain and present algorithmic decisions to multiple stakeholders, especially to lay people without technical knowledge [58]. On the other hand, although researchers in Explainable AI (XAI) have made progress in simplifying and explaining complicated machine learning models, this line of work was critiqued for lacking a deep understanding of or evaluation with actual users or stakeholders [39]. As a result, recent studies in HCI (e.g., [13, 56]) have started to adopt a human-centered approach to both create and evaluate interactive user interfaces in this specific domain.

Our work contributes to this emerging line of research. In this work, we explore how to create **non-expert-oriented representations of confusion matrices for binary classifications**. Confusion matrices are a widely used tool in machine learning to communicate the performance of a classifier on a set of training or test data where ground truth is known [2]. They help explain the fundamental performance metrics in binary classifications, including true positives, false positives, true negatives, and false negatives, which are often used as the basis of other evaluation metrics.

More recently, with the surging interest in explaining and visualizing AI, confusion matrices have also been used as one of the basic explanation components and core visual elements in many existing XAI toolkits and tutorials, including the ones explaining fairness metrics that have attracted increasing attention among ML and HCI researchers (e.g., [40, 48, 56]). For example, in a widely circulated tutorial, Narayanan used binary confusion matrices to explain a series of fairness metrics developed by the ML community [40]. Despite such wide applications, however, there exists very limited understanding in terms of whether confusion matrices are a useful and effective way to communicate ML performance to the non-expert public. Previous studies have also shown that lay people in general face great challenges in probabilistic reasoning (e.g., [22]), which might impact their ability in effectively interpreting confusion matrices. However, there lacks a systematic investigation in terms of whether lay people are able to interpret information contained in confusion matrices without additional help. In this paper, we set out to tackle this increasingly critical yet under-explored problem.

Our research questions are: **R1:** What are the main challenges non-expert lay people have in understanding the information contained in standard confusion matrices? **R2:** How effective different alternative representations perform in addressing those challenges, in terms of objective understanding, time costs and subjective understanding?

Our work explores how to create non-expert-oriented alternative representations of standard confusion matrices for binary classification, as a step towards facilitating public understanding of the performance of machine-learning-powered algorithmic decision-making systems. Towards this end, we first conducted interviews (n=7) and a survey (n=102) to identify the major challenges lay people have in understanding standard confusion matrices. Based on these initial studies, we conducted an online experiment where 483 participants used different alternative information representations of confusion matrices to understand the performance of an algorithm for making recidivism predictions. We developed three levels of questions to assess participants' understanding of the performance of the algorithm, and assessed these representations for accuracy in answering those questions, completion time, and understanding, which refers to self-reported perceived understandability. Our contributions are three-fold:

- First, our work identified major challenges lay people have in understanding standard confusion matrices, a widely used method in machine learning to present the performance of a classifier on a set of test data where ground truth is known.
- Second, we designed a few alternative representations to tackle those challenges: Contextualized confusion matrices that map all the terminologies (e.g., false positive and false negative) back in specific problem domains, Tree diagrams, Bar charts, and Flow charts.
- Third, we evaluated our designs for their effectiveness. Our evaluations suggested several design implications that might serve the foundation for further exploration in this space: (1) contextualizing the terminologies can significantly improve users' objective and subjective understanding; (2) flow charts, which help point out the direction of reading the data, were most useful in improving objective understanding.

2 RELATED WORK

In this section, we outline relevant past work in two areas. First, we survey the related work in explainable AI and visualizing AI, and describe how our work is positioned in this space. Next, we present an overview of existing work on human perceptions of algorithmic decisions, a quickly growing field that has received increasing attention in CSCW and describe how our work contributes to this emerging line of research.

2.1 Explaining and Visualizing AI

While interpreting intelligent systems has a long history in AI and HCI, the recent widespread deployment of machine learning enabled algorithmic systems in many critical and complex social environments has lead to a renewed attention to interpretability. Generally labeled as "explainable AI (XAI)," this growing body of work aims at providing a human-understandable explanation for decisions made by "black box" machine learning models in order to support decision-making, improve transparency, and increase trust (see review articles [1, 10, 57]).

A large body of past work has explored different explanation styles to probe the inner working of such "black box" models. For example, Datta et al. [16] offered a solution to help users measure the influence of a series of features in inputs on system outputs. Nugent et al. [41] selected similar cases from the training data to offer explanation to the decision in question. Lakkaraju et al. [32] aimed at generating short rule-based explanation. Despite making greater progress, this line of work was critiqued for lacking a deep understanding of or evaluation with actual users or stakeholders [39].

Meanwhile, although previous non-expert-oriented visualizations in (e.g., [28, 38]) used simple diagrams to target specific probability problems such as Bayesian Reasoning, the recent surge of XAI has led to more complex visual techniques, systems, and toolkits to facilitate *expert* users' evaluation of model performance (e.g. see [3, 21]). For example, Google's People + AI Research group (PAIR) released the open-source "What-if" tool [52] to help people who are not formally trained in machine learning visualize the effects of bias metrics. Prospector [30] enabled data scientists to understand how a given feature influences algorithmic prediction and support for manipulating feature values for response. Similarly, studies have also been conducted to create alternative visualizations to confusion matrices, to help machine learning practitioners as well as non-ML-experts adjust their models. For example, Square [46] introduced a novel visual system to help machine learning practitioners evaluate the performance of a multi-class classifier. Classee [7] offered a simplified barchart to help end-users with no expertise in machine learning choose and tune parameters and understand classification errors. While these systems provided valuable methods to explore machine learning data and model, they are designed primarily for *expert* users with relatively advanced technical literacy – including both technical experts like machine learning experts and data scientists as well as domain experts like doctors.

Recently, researchers in HCI seek to expand this line of research by taking a human-centered design approach, with an emphasis on both improving usability of these explainable tools and performing empirical tests for evaluation. For example, Cheng et al. [13] conducted non-expert-oriented experiments to test the performance and trade-offs among interactive vs. static explanation as well as white-box vs. black-block explanations. Their findings suggested that while both interactive and white-box explanations can improve non-expert users' comprehension, interactive approach is more time consuming.

Our work contributes to this emerging line of research in HCI by focusing on confusion matrices, a basic and widely used representation in the evaluation of algorithm performance. It helps set the stage for understanding and developing laypeople-oriented representations of machine learning model performance.

2.2 Human Perceptions towards Algorithmic Decisions

With the increasing penetration of algorithmic decision-making systems in our society, there is a widely shared concern that ML-enabled algorithmic systems may produce socially and ethically questionable decisions that are not aligned with human values (e.g. see [4, 6, 18, 19]).

In response to these concerns, an emerging line of work in the HCI community has started to look at those decisions from a human-centered perspective. This body of research has used interviews, surveys, and experiments to empirically probe people's perceptions towards algorithmic decisions. Examples include studies on how humans perceive algorithmic decisions versus human decisions in managerial contexts [34], whether people perceive certain features (such as criminal history or neighborhood safety) as fair to be used to predict criminal risk [25, 50], how explanation styles might matter in shaping people's justice perceptions [9], how members of traditionally marginalized communities feel about algorithm (un)fairness [54], how affected communities feel about algorithmic decisions in the context of a child welfare system [12], how the general public perceives online behavioral advertising that used demographic factors (e.g., race) as targeting variables [44], which statistical definitions of fairness people perceive to be the fairest in the context of loan decisions [47], as well as how humans use AI systems to make decisions [23, 24]. This past body of work mostly used storyboards or text to present several algorithmic scenarios to their study participants, often without tackling the results and performance of the underlying machine learning models. This is understandable, as lay people often lack the required technical knowledge to fully understand and evaluate highly specialized algorithmic results.

As a result, researchers in HCI are calling for more efforts to explore how to better explain and present algorithmic decisions to multiple stakeholders, especially to non-experts with limited technical literacy [58]. Recent studies (e.g., [13, 35, 50, 56]) have started to use visualizations or create user interfaces to communicate algorithmic decisions to their study participants.

Our research contributes to this line of work by offering a new and complementary angle, looking at making the performance of machine learning model more comprehensible. By creating alternative representations of standard confusion matrices, we aim at developing more easily understandable methods to represent the performance of machine learning models for the *non-expert* public. Our work has the potential to be used to better solicit public perception of algorithmic decisions (e.g., help lay users evaluate fairness) for future research in this domain.

3 OVERVIEW OF THE STUDIES

We conducted a series of studies, both qualitative and quantitative, to explore how to better help non-expert users interpret a very common output of machine learning classifiers: confusion matrices. We chose binary confusion matrices as a starting point to improve model literacy for the non-expert public for several reasons. First, they are commonly used representations of the performance of machine learning models. Second, they work for all kinds of classifiers regardless of the underlying algorithm and so have general applicability. Third, they help convey some important aspects of how well a machine learning model is (or is not) working. And fourth, they help explain the fundamental performance metrics in binary classifications (i.e., true positives, false positives, true negatives, and false negatives), which can be further used as the basis of other evaluation metrics.

Our team for this project was composed of researchers with diverse backgrounds, including machine learning, software engineering, visualization, and social science. We began with a session for rapid ideation to identify our task domain.

3.1 Choice of Task Domain and Dataset

3.1.1 Task Domain. Recidivism prediction is an important public issue where algorithms have been increasingly deployed to help judges assess the risk of re-offending among defendants. This topic has also generated one of the most controversial cases so far in the debates around fairness and public algorithmic decision-making [4]. Instead of looking at issues around fairness, in this work, we used it as our task domain to test and facilitate public *understanding* of algorithmic outputs and performance. Here, we chose to focus on a specific domain, instead of abstract algorithmic results, because past studies in psychology (e.g., Wason's selection task [51]) suggest people tend to perform better with their reasoning powers on realistic cases than with abstract formulations.

3.1.2 Dataset. We used the dataset gathered by ProPublica in its 2016 report, which examined the performance of COMPAS – a widely used recidivism algorithm in the US. This dataset contains two years worth of COMPAS scores from Broward County, Florida in 2013 and 2014 [33]. COMPAS assigns risk scores to criminals to determine their likelihood of re-offending, with 1-4 constituting "low" risk, 5-7 "medium" risk, and 8-10 "high" risk. Following the study by ProPublica [33], which used 4 as the threshold, we similarly present this as a binary classification task with score at and below 4 as negative prediction ("Labeled as low risk"), and above 4 as positive prediction ("Labeled as high risk"). Ground-truth labels – "Re-offended" and "Didn't re-offend" – correspond to whether a defendant released on bail was arrested for another crime within 2 years of their release.

3.2 Study 1

To understand what challenges non-expert lay people have in understanding standard confusion matrices, we conducted a qualitative study comprised of interviews (n=7) and an Amazon MTurk

survey (n=102). In both studies, we asked our study participants to answer a series of factual questions about information contained in confusion matrices as well as what were the challenges they have in reading the table and answering those questions. We identified two sets of main challenges in Study 1.

3.3 Study 2

Building on the challenges identified in Study 1, in Study 2 we developed and evaluated alternative representations of confusion matrices. We conducted a between-subjects online experiment (n=483) on Amazon Mturk to assess the relative effectiveness of our alternatives in addressing those challenges,in terms of comprehension, time costs and subjective preference.

3.4 Evaluation Framework of Objective Understanding

Over the years, the machine learning community has developed different error metrics to assess the performance of classification algorithms (e.g., see [27]). However, these metrics are designed primarily for expert users.

In this work, we developed three levels of questions – comprehension, comparison and simulation – to assess and facilitate non-experts' understanding of the statistical results and performance of machine learning classifiers. Our evaluation questions were developed and adapted from previous frameworks in learning science [20], which identified three levels of graph understanding of statistical results for lay people: Level 1 focuses on extracting or locating information; Level 2 focuses on finding relationships; and Level 3 focuses on analyzing implicit relations contained in the graph. In total, we created 11 questions, including 4 comprehension questions, 4 comparison questions and 3 simulation questions.

Note that the framework developed in this study offers a general guideline and all the questions can be adapted to different domains and problem statements in future research (see Table 1).

3.4.1 Question Level 1: Comprehension. In this level, we created questions to measure how accurately users can identify and locate certain information in the information representation.

Questions we created and used in this level include:

- "In Group A, how many defendants were labeled high risk and re-offended?"
- "In Group A, how many defendants were labeled high risk but did not re-offend?"
- "In Group B, how many defendants were labeled low risk but re-offended?"
- "In Group B, how many defendants were labeled low risk and did not re-offend?"

3.4.2 Question Level 2: Comparison. In this level, we measured how accurately users can compare information and performance of the machine learning classifier on two demographic groups, as well as how well they can perform simple analysis. Questions in this level also served as a bridge to lead users to the next level.

Questions we created and used in this level include:

- "Between Group A and B, which one has more defendants who were labeled as high risk and re-offended?"
- "Between Group A and B, which one has more defendants who were labeled as low risk but re-offended?"
- "In Group A, among people who did not re-offend, what percentage were labeled as high risk?"
- "In Group A, among people who were labeled as high risk, what percentage did not re-offend?"

Level	Purpose	Question type (X may equal to Y)	Examples
1	Comprehension	In Group A, how many people were classified as X but/and belonged to Y?	In Group A, how many defendants were labeled high risk but did not re-offend?
2	Comparison	Between Group A and B, which one has more people who were classified as X but/and belonged to Y? In Group A/B, among people who belonged to X, what percentage were classified as Y? In Group A/B, among people who were classified as Y, what percentage belonged to X?	Between Group A and B, which one has more defendants who were labeled as high risk but did not re-offend? In Group A, among people who did not re-offend, what percentage were labeled as high risk? In Group A, among people who were labeled as high risk, what percentage didn't re-offend?
3	Simulation	For a new defendant in Group A/B, who we already know belongs to X, how likely will he be labeled as Y? For a new defendant, who we already know belongs to X, in which group he/she is more likely to be classified as Y?	For a new defendant in Group A, who we already know won't re-offend, how likely will he be labeled as high risk? For a new defendant, who we already know won't re-offend, in which group he/she is more likely to be labeled as high risk?

Table 1. Our evaluation framework used three levels of questions to assess non-experts' objective understanding of the statistical outputs of binary classification. Note that the framework offers a general guideline and all the questions can be adapted for different domains and problem statements in future research.

3.4.3 Question Level 3: Simulation. In this level, we wanted to measure how well users can perform simulations based on the presented information, e.g., can identify implicit information contained in the information representation.

Questions we used in this level include:

- "For a new defendant in Group A, who we already know will not re-offend, how likely will he/she be labeled as high risk?"
- "For a new defendant in Group B, who we already know will not re-offend, how likely will he/she be labeled as high risk?"
- "For a new defendant, who we already know will not re-offend, in which group he/she is more likely to be labeled as high risk?"

To ensure that our participants could interpret and understand the questions above, we tested all questions using cognitive interviews [45, 53]. We asked our participants to think aloud as they answered the questions, and we identified and addressed any concerns and questions that emerged from the process. We conducted cognitive interviews with four volunteer participants who had not seen or heard of confusion matrices before. Two main changes were made to modify the questions

based on their feedback: (1) we adjusted the wording of the level 1 and 2 questions by adding conjunctions like "and" and "but"; and (2) we also adjusted the order of the clauses (e.g., "who we already know won't re-offend") for level 3 questions to better facilitate understanding. We iteratively refined our questions based on participants' feedback until no new concerns or questions were identified.

4 STUDY 1: AN EXPLORATORY INVESTIGATION OF THE CHALLENGES LAY PEOPLE HAVING IN UNDERSTANDING STANDARD CONFUSION MATRICES

In Study 1, we began by interviewing non-expert lay people to understand what challenges they might have in reading standard confusion matrices with original terminologies, which was followed by a survey. In both interviews and the survey, we used the three levels of questions developed for measuring objective understanding (see section 3.4) to test their comprehension and then asked open-ended questions to probe in greater detail about the challenges.

4.1 Method

4.1.1 Interview (n=7). We first conducted semi-structured interviews with seven participants (three males and four females) to learn about the challenges they are having. We recruited our participants in a university located around a northeast US metropolitan area from non-technical majors. None of them had seen or heard of a confusion matrix before. During our interviews, we showed them a standard confusion matrix of the ProPublica data, along with factual questions regarding the information contained in the confusion matrix. We asked them to think aloud as they worked through the process. We then probed in detail what made them struggle in answering the questions. Each interview lasted for 20-30 minutes and participants received $10 for time compensation. All interviews were audio-recorded and transcribed.

4.1.2 Survey (n=102). In addition to qualitative interviews, we also conducted a quantitative survey (n=102) on Amazon Mechanical Turk in August, 2019. The survey consisted of two sets of questions. The first part included fill-in-the-blank and single choice questions. We asked factual questions about information contained in confusion matrices, using the three levels of questions developed for measuring objective understanding (see section 3.4). The second part included two open-ended questions and probed in greater detail about users' challenges in understanding. The average time for completing the survey was 7 minutes and each participant received $1.

4.1.3 Data analysis. We applied thematic analysis [11] to all interview data as well as all the open-ended questions in our survey. Two researchers read through the data set individually and held weekly discussions. The aim of the coding is to find out comprehension challenges that the study participants had. With this idea in mind, the two researchers coded the data set independently, looking for as many basic codes as possible. After this step, they met again, discussing and refining their initial codes and sorted the different codes into themes. Through iterative discussion and deductive and inductive thinking, they further reviewed, defined and named themes until a satisfactory thematic map has been reached.

4.2 Results

Below we discuss our findings from our interviews and survey. Interviewees are identified with a "P" and survey participants are identified with a "S". Survey responses are also accompanied with percentages.

In general, users were frustrated with the information presented using standard confusion matrices in original terminologies. In some cases, this led to negative feelings toward the decision-making system as well as the people working behind it. For example, S35 commented: *"I think*

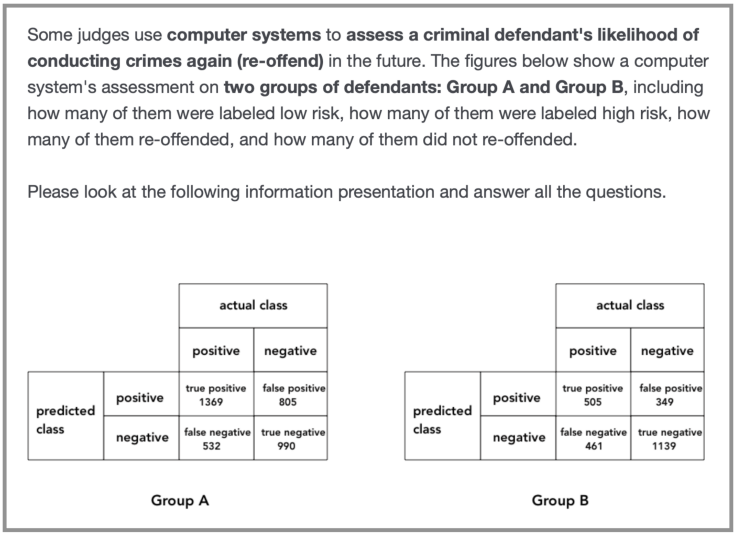

Fig. 1. The standard confusion matrix shown to participants in Study 1.

that the individual who chose to present the data this way should be fired. If someone working for me presented this in a meeting, I would let them go."

They also felt incompetent and "dumb." S36 commented: *"Can you dumb it down for us regular folks? Otherwise, those of us who are not so bright will not be able to pair the information appropriately."*

4.2.1 Challenge 1: Confusion about terminology. The first set of challenges that emerged from our data is that users in general have difficulty interpreting the terminology used in confusion matrices (true positive, true negative, false positive, and false negative). They did not know what positive and negative meant and had a hard time mapping it to the original problem. This was a widely shared concern in both our interviews (7/7) and survey (98%).

P3 explained : *"What does positive mean? If this is from my doctor's office, I know there might be something wrong with my test. But in this (recidivism predication) case, does this mean 'high risk' or 're-offended'?"*

S2 felt the repeated use of positive and negative in the matrix increased the difficulty of the task, i.e., that the term "positive" and "negative" can refer to both predicted and actual classes: *"Positive and negative make it too sciency. Also, what does true mean? Does it mean positive positive?"*

S14 suggested replacing the terminologies and labeling the matrix appropriately according to the context: *"Label the tables better with terms match what is asked in the questions. I had to guess to start which axis X and Y was represented in the question about the tables."*

4.2.2 Challenge 2: Confusion about the underlying relationships between the four types of predictions. Apart from terminology issues, users also reported that they need help in clarifying the underlying structure of the confusion matrix. In general, they felt that the matrix design does not convey the underlying relationships between the four types of predictions. Three sub-challenges were further identified under challenge 2.

Sub-challenge 2.1: Confusion about the direction of reading the data. Users were confused about how the data flows from one category into other categories, i.e., direction of reading the data in confusion matrices, in particular, about "False Positive" and "False Negative". They made mistakes

in answering questions like "In Group A, how many people were classified as X but belonged to Y?" Four of seven interview participants and 54% of survey participants reported this challenge.

For example, P1 said: *"You are asking me how many re-offending defendants were labeled as high risk, there is a 'directional relationship' here – like, re-offending defendants were thrown into another group. But from this table, it is really hard to tell. There is no clear clue in terms of which direction the data go."*

S55 asked, *"Can you better explain how the different groups went into the other groups? Like 'in Group A, how many defendants were labeled high risk but did not re-offend?' I think understand that 'positive' means 'labeled high risk,' but I'm confused after that."*

P2 suggested to add direction metaphor to the representation: *"Maybe you can identify which one (class) is the 'starting point' and which one (class) is the 'end point'? So it can be read as a map – I will have a better idea of how the data travels ..."*

Sub-challenge 2.2: Confusion about layered relations. Some users are confused about the layered relations embedded in confusion matrices. In particular, when answering questions like "In Group A, how many people were classified as X and belonged to X," they tend to mistake "Positive" with "True Positive" (or "Negative" with "True Negative"). Four out of seven interview participants and 45% of our survey participants reported this challenge.

When asked why this is difficult to answer, P4 explained: *"There are too many layers in the table! You see, the table has a top and bottom row and when you ask 'how many people were labeled as high risk and re-offended?' I just added everything up here."*

S13 suggested: *"Can you break down the graph into subsections so it can only have one set of headings – right now there are too many things messed up here."*

S27 said:*"While I understand the terminologies, I found the true vs. false in addition to positive and negative to be difficult to understand."*

S29 reported: *"I just needed better organization. This setup with triple layers on each side wasn't easy to read."*

Sub-challenge 2.3: Confusion about quantities involved. Finally, a small group of users (one out of seven of interviewees and 9.8% of survey participants) reported that they need support for comparing the quantities involved in the confusion matrix, especially when they needed to compare the performance of two classifiers. Even if they could locate the information correctly, it was hard for them to compare.

P7 said: *"I think everything is OK until you asked me to compare in which group a new defendant who won't re-offend is more likely to be labeled as high risk. You see, this involved two steps – I need to figure out, first, how likely a new defendant who won't re-offend is going to be labeled as high risk in Group A and Group B, and second, do comparison. It would be much easier if you can make everything proportionally so I don't have do the rough calculation – I can just tell directly from the graph."*

S6 suggested: *"It was difficult to tell percentages at a glance. If the different percentages could be represented with visual size differences, it might be easier to compare groups at a glance."*

4.3 Summary

We identified two major sets of challenges non-expert lay people have in understanding standard confusion matrices. First, we found that the majority of our participants had difficulties understanding the terminology and thus had a hard time mapping them back to the scenario. Second, they also reported structural issues, i.e., that the matrix design does not convey the underlying relationships between the four types of predictions. We then further identified three sub-challenges under this structural issue: (1) some of our participants reported problems in understanding how the data flows into other categories, i.e., the direction of reading the data; (2) others were confused

about the multiple layers contained in the table and suggested us to unpack these layers; and (3) a few of them suggested to represent the data proportionally to facilitate easy comparisons.

5 STUDY 2: AN EXPERIMENT OF USERS' PERFORMANCE OF ALTERNATIVE REPRESENTATIONS

Building on the findings of our first study, we conducted a design workshop to develop alternative representations of confusion matrices that could address the challenges raised by our participants. We then conducted a crowd-sourced experiment on Mturk to test the effectiveness of our alternatives in helping them understand the performance of a machine learning algorithm for making recidivism predictions.

5.1 Design Workshop

We invited one machine learning expert, one software engineer, one former journalist, two visualization experts, two UX designers (current graduate students at our university), and one psychologist to join the workshop. We introduced the goal of facilitating public understanding of the performance of machine learning classifiers using alternative representations of confusion matrices. We shared the challenges we identified from Study 1 and asked our participants to come up with multiple alternatives that could address those challenges. To avoid potential interaction effects among different designs, in the workshop we asked participants to think about making each representation targeting one specific challenge.

During our initial ideation and prototyping phase, we drew on previous studies that used simple non-expert-oriented diagrams to target probability problems like Bayesian reasoning (e.g., [28, 38]) and developed 20 initial alternatives. Participants were also encouraged to develop visual representations to tackle those challenges, minimizing the use of text descriptions.

We performed three steps to narrow down these initial prototypes. First, we discussed and assessed the initial alternatives for their feasibility in targeting the specific structural challenges identified in Study 1, usability for our targeted audience, and generalizability. Second, we clustered similar ideas, identified common themes, and combined different options. Third, over the course of our design process, we also iteratively conducted informal user testing with 12 participants. In these tests, we showed our mock-ups and assessed them for understandability (e.g., we asked questions like "Tell us what this table shows?"), how well they helped the participants in answering a set of questions (e.g., "How many people were correctly labeled as 'high risk?"'), as well as subjective preference.

At the end of the process, we settled on two sets of solutions. The first set addresses the terminology issue by replacing general terminology with a **(a) contextualized confusion matrix** that uses terminology that maps back to the specific case in question, e.g., "Reoffended" and "Labeled high risk" instead of "Positive". We opted for this design since confusion over terminology was a common challenge identified in Study 1.

In addition to changing terminology, we also developed a second set of representations, each of which addresses a specific sub-challenge identified in Study 1. These representations include a **(b) a tree diagram** to unpack layered relations, a **(c) a flow chart** to point out the direction of reading the data, and a **(d) a bar chart** to clarify the quantities involved.

We made two additional choices in terms of visual encoding. First, in the design of the tree diagram and flow chat, we explored how the matrix can be read from prediction to true label. We followed the best practice in data visualization design to use lines (and arrows) to connect disparate regions to help readers follow the visualization, based on the Gestalt principle of connectedness [36]. Second, when choosing the color used in the design, we followed best practices in information visualization to use different color hues to represent different categorizations [26]. In particular, we

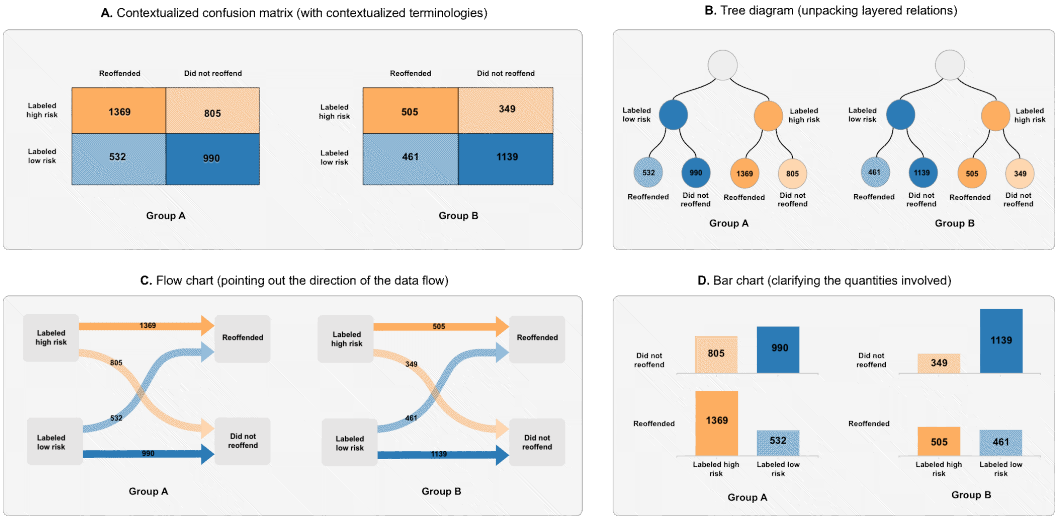

Fig. 2. The design workshop led to the creation of the following alternative representations to address each challenge identified in Study 1, including (a) a contextualized confusion matrix with all terminologies mapping back to the case in question but without structural change; (b) a tree diagram that unpacks layered relations; (c) a flow chart that points out the direction of the data flow; and (d) a bar chart that clarifies the quantities involved.

used blue, a more neutral color to represent "low risk," and orange, a more risky color to represent "high risk". These mappings of risk to semantically resonant colors may aid understanding through semantic facilitation and require less conscious thought [5] as well as improve memory during the tasks [8]. In addition, blue and orange are common colors for qualitative/categorical palettes, because they are colorblind friendly [49]. We used these colors with decreased saturation for incorrect predictions (i.e., False Negative and False Positive), based on feedback from the aforementioned iterative informal users tests. Where possible, we kept size, color, and fonts consistent across all four alternatives (see Figure 2).

5.2 Experiment

We conducted a between-subjects experiment to assess users' performance across our four alternative representations (see Figure 2). We used two baselines for our experiment: (1) the standard confusion matrix with the original terminologies as Baseline 1; and (2) the contextualized confusion matrix without structural change as Baseline 2. This results in total 5 conditions for our experiment.

5.2.1 Participants (n=483). The experiment was active on Amazon MTurk in August 2019 (see experiment interface in Figure 1). In total, 574 participants completed the experiment. We removed 91 respondents who did not finish the trial, were not attending to the experiment (i.e., failed the attention check question or provided gibberish answers), or have seen or heard about confusion matrices before. Our final dataset contains 483 completed responses. Our sample was diverse in terms of demographics. It included 182 female, 296 male, and 3 prefer not to say, ranged in age from 18 to 75+ (6% from 18-24, 45% from 25-34, 25% from 35-44, 14% from 45-54, 7% from 55-64, 1% from 64-74, and 0.2% above 75).

To ensure the quality of survey responses, we only recruited participants with a HIT Approval Rate greater or equal to 95% and Number of HITs Approved greater than or equal to 50, who

reside in the US, and are aged 18 or above. We randomly assigned each participant to one of the 5 conditions. Each participant completed a single trial.

The average time for completing the trial was 7.05 (std=5.6) minutes. We assigned a bonus payment to participants based on the number of correct answers they gave for understanding questions. Each participant received a base payment of $2 and an additional bonus (up to $2) based on the number of correct answers they gave for the understanding questions.

5.2.2 Study Procedure. In our experiment, participants first filled out a consent form that explained the purpose, procedure, and compensation of our study. They were then introduced to the recidivism prediction algorithm. We told our participants that a computer system had been developed to assess the possibility of re-offending of two groups of defendants. Each participant was then randomly assigned to one of the five conditions. Participants explored the information representation presented to them and answered all of the aforementioned 11 questions we created for understanding measurement (see details of the evaluation questions in 3.4). They were allowed to go back to questions during their interpretation of the representations. They were then presented with questions measuring their subjective understanding. A final page captured demographic information. We also included an instructed-response question for an attention check [37].

5.2.3 Evaluation Metrics. We recorded three metrics: objective understanding, time cost, and subjective understanding. We used the evaluation questions we developed in section 3.4 for assessing understanding, looking at number of questions correctly answered. Time cost was measured as the number of seconds from when the question was displayed until the participant hit the submit button on the interface. Subjective understanding, which refers to self-reported perceived understandability of the confusion matrix design, was measured by asking the participants to subjectively rate their agreement with the following statement on a scale from one (strongly disagree) to seven (strongly agree): "I found this information representation easy to understand."

5.2.4 Data analysis. We used linear regression models for data analysis, examining whether the different alternative representations led to different levels of objective understanding, time costs, as well as subjective understanding (See Table 2). We performed two sets of analyses: (1) Analysis 1 compared the contextualized confusion matrix against the standard confusion matrix; (2) Analysis 2 compared bar, tree and flow charts against the contextualized confusion matrix to explore whether the structural change of the representations led to different level of understanding, while other variables remaining the same.

5.3 Results

5.3.1 Analysis 1: Contextualized confusion matrices versus standard confusion matrices. In Analysis 1, we compared contextualized confusion matrices against standard confusion matrices to test whether replacing general terminologies with contextualized ones will significantly improve users' performance, given terminology challenge is a widely shared issue among our participants in Study 1 (see Figure 3).

In general, our results suggested that compared to the standard confusion matrix, participants in contextualized condition performed significantly better in terms of both objective understanding and subjective understanding.

Objective Understanding. Overall, our results suggested that people's level of understanding decreased as the question level increased. In Table 2, when compared to the standard confusion matrix, the contextualized condition led to significant increases in participants' understanding of the algorithmic performance. In total, they answered 1.54 more questions (out of 11) correctly,

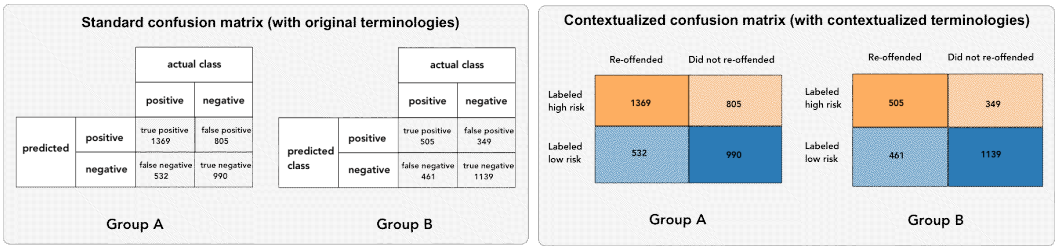

Fig. 3. In analysis 1, we compared the contextualized confusion matrix with the standard confusion matrix. Our results suggested that compared to the standard confusion matrix, participants in contextualized condition performed significantly better in terms of accuracy in answering understanding questions and in subjective understanding.

compared to participants using the standard confusion matrix with original terminologies (Coef. = 1.54, $p < 0.001$). Figure 4 shows the accuracy rate across standard and contextualized conditions.

Completion Time. The "time" column in Table 2 shows the time participants spent on the understanding task using different alternative representations. Our results suggest that compared to the standard confusion matrix, there was no statistical significance in terms of the time our study participants spent on the understanding task in contextualized condition. We also broke down the time to complete at each task-level and calculated the mean. The differences between the means were not significant.

Subjective Understanding. The "subjective understanding" column in Table 2 shows the level of self-reported perceived understanding our study participants reported with the alternative representations, in the form of "I found this information representation easy to understand" and on a 7-point scale, where 7 means "strongly agree". Compared against the standard confusion matrix, participants in contextualized conditions reported significant improvement in this measurement (Coef. = 2.20, $p < 0.001$).

5.3.2 Analysis 2: Comparing three alternatives against confusion matrices, all using contextualized terminologies. (see Figure 2).

In the second set of analysis, we compared tree, flow, and bar charts against the confusion matrix, all using contextualized terminologies. Our results showed that there was no significant difference for bar charts and tree diagrams in terms of overall objective understanding, time, and subjective understanding. Bar charts outperformed the matrix condition only in answering level 3 questions. In contrast, participants performed significantly better using the flow chart in understanding the algorithmic performance compared to using confusion matrix. Flow chart in particular helped participants answer level 2 and level 3 questions correctly. However, it didn't improve users' subjective understanding.

Objective Understanding. When we compared tree, flow, and bar against the matrix condition, all using contextualized terminologies, participants in the flow condition performed significantly better in overall objective understanding tests (Coef. = 0.73, $p < 0.05$), as we saw in Table 2. Bar charts performed significantly better than the matrix condition only in answering level 3 questions (Coef. = 0.28, $p < 0.05$). There was no significant difference in understanding task when we comparing tree against the contextualized confusion matrix. Figure 4 shows the accuracy level among all four experimental conditions in our second set of analysis.

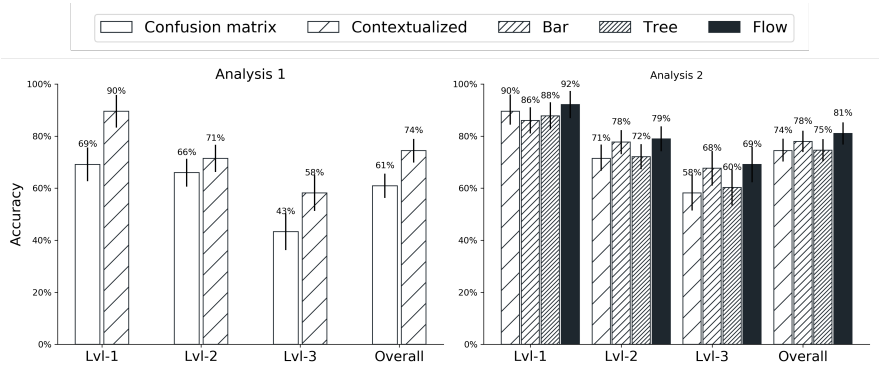

Fig. 4. The accuracy rate in our two sets of analyses across different conditions (error bars show the 95% confidence interval). The left one shows the comparison between the standard confusion matrix versus the contextualized one. The right one shows the comparison of bar, tree and flow against the contextualized condition, all using contextualized terminologies. The three levels of questions refer to Level 1: Comprehension, Level 2: Comparison, and Level 3: Simulation, see Evaluation Framework in section 3.4.

We further preformed Welch's t-test to compare Bar vs. Tree vs. Flow. We only found a significant effect ($p < 0.05$, Cohen's d = 0.30) with flow condition outperforming tree. There was no statistical significance between Bar vs. Tree or Flow vs. Bar.

Completion Time. Our results also suggested that there was no significant difference in time costs when we compared all the three representations (tree, flow and bar) against the confusion matrix, when they all used contextualized terminologies. Similarly, there was no significant difference when we compared the means of the time participants spent on each level with these representations.

Subjective Understanding. Similarly, there was no significant difference in subjective understanding when we compared all the three representations against the contextualized confusion matrix.

5.4 Interpretation of the Results

Study 2 investigated the relative effectiveness of four alternative representations in addressing the challenges identified in Study 1, with an aim to better facilitate non-experts in understanding the information contained in standard confusion matrices.

There are three major findings from our data. First, the contextualized confusion matrix (with no structural change) led to significant improvement on users' objective and subjective understanding of the algorithm. This suggests that there are straightforward opportunities for creating more comprehensive alternatives of confusion matrices. That is, by mapping the terminologies back to the specific scenario, we can significantly improve users' performance on both objective and self-reported understanding.

Second, when comparing tree, bar, and flow conditions against the contextualized confusion matrix, all using contextualized terminologies, only flow charts performed significantly better than the matrix condition in terms of improving *overall* objective understanding. One possible explanation is that flow charts articulate the directional relationships between categories, which is effective to help users compare information across categories and predict the algorithm's outputs. Bar charts outperformed the matrix condition only in accuracy of answering level 3 (simulation) questions. One potential reason is that bar charts help clarify the quantities involved, making it easy

	Lvl-1[1] Accu. (# = 4)	Lvl-2 Accu. (# = 4)	Lvl-3 Accu. (# = 3)	Overall Accu. (# = 11)	Time (s)	Subjective Understanding
	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)	Coef. (S.E.)
Analysis 1: Comparing contextualized terminologies and standard terminologies						
Baseline: Standard confusion matrix[2]	2.75*** (0.12)	2.63*** (0.10)	1.27*** (0.10)	6.65*** (0.25)	489*** (35)	2.91*** (0.17)
Contextualized vs Standard (Δ)[3]	**0.83***** (0.16)	0.23 (0.14)	**0.47**** (0.15)	**1.54***** (0.35)	−79 (48)	**2.20***** (0.24)
Analysis 2: Comparing bar, tree and flow with confusion matrix, all using contextualized terminologies						
Baseline: Contextualized confusion matrix	3.58*** (0.10)	2.86*** (0.10)	1.74*** (0.10)	8.18*** (0.23)	410*** (32)	5.12*** (0.17)
Bar vs Contextualized (Δ)	−0.14 (0.15)	0.25 (0.13)	**0.28*** (0.15)	0.39 (0.33)	+18 (45)	−0.18 (0.23)
Tree vs Contextualized (Δ)	−0.07 (0.15)	0.03 (0.14)	0.06 (0.15)	0.02 (0.34)	−9 (46)	−0.07 (0.24)
Flow vs Contextualized (Δ)	0.10 (0.15)	**0.30*** (0.14)	**0.33*** (0.15)	**0.73*** (0.33)	-22 (46)	−0.36 (0.24)

p-value significance: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.
Bold font indicates statistical significance.

[1] The three levels of questions refer to Level 1: Comprehension, Level 2: Comparison, and Level 3: Simulation, see Evaluation Framework in section 3.4.

[2] The baseline row describes the performance of the baseline. For example, participants answered 2.75 level-1 questions (out of four) correctly and spent 489 seconds to complete three-level questions using the original confusion matrix.

[3] The comparison row describes the performance delta between an alternative representation and the baseline. For example, participants using contextualized confusion matrix can answer 0.83 more questions correctly and spend 79 seconds less time, compared to participants using the original confusion matrix.

Table 2. Results of our two sets of analyses. In Analysis 1, compared to the standard confusion matrix, participants in contextualized condition performed significantly better in terms of accuracy and subjective understanding. In Analysis 2, when comparing bar, tree and flow conditions against the confusion matrix, all using contextualized terminologies, there was no significant difference for bar chart and tree diagram in terms of overall understanding, time, and subjective understanding, while participants performed significantly better using flow chart in understanding the algorithmic decision.

for people to simulate the future performance of ML models. But the actual mechanism remains unclear. There was no statistical significance between tree and matrix.

Additionally, although some alternative designs (i.e., flow and bar charts) had some positive effects on improving users' objective understanding of the model compared to the standard confusion matrix, none of them improved users' subjective understanding of the model. To better understand this finding, we looked into participants' responses to our open-ended questions. The most likely explanation from our data is that our participants were less familiar with the new designs. Such unfamiliarity may lower the level of subjective understanding, even if the representation itself might facilitate the understanding of the algorithm.

For example, S96 highlighted this unfamiliarity: *"It was OK, slightly confusing until I got used to it."*. S65 proposed using more conventional information representations: *"I felt if the information was presented on a table it would've been a lot easier to read."*.

Therefore, the data suggests that there might be possible trade-offs between addressing identified structural issue of standard confusion matrices and improving understandability perception. Introducing new structures can facilitate objective understanding but might reduce subjective understanding due to unfamiliarity. However, more studies are needed to explicitly test if such a relationship exists and identify innovative approaches that might satisfy both goals.

6 DISCUSSION

In this paper, we present an exploratory study on how to create non-expert-oriented alternative representations of confusion matrices, as a step towards facilitating public understanding of the performance of machine-learning-powered algorithmic decision-making systems. We conducted a series of studies, both qualitative and quantitative, investigating better ways of presenting the performance of machine learning models through redesigning standard confusion matrices with original terminologies. We have identified major challenges lay people face in front of standard confusion matrices and evaluated a couple of designs that were chosen specifically for targeting those challenges. Our results contribute to the emerging line of research on building human-centered algorithm design.

6.1 Summary of the Research Questions and Results

Below we summarize our results in response to the research questions raised in the introduction.

6.1.1 What are the major challenges lay people have in understanding standard confusion matrices?

- Our results in Study 1 suggest that lay people have two sets of major challenges in understanding the information contained in standard confusion matrices.
- First is the terminology issue. We found that the majority of our participants had difficulties understanding the terminology and thus had a hard time mapping them back to the scenario.
- Second is the structural issue, i.e., that the matrix design does not convey the underlying relationships between the four types of predictions. We then further identified three sub-challenges under this structural issue: (1) some of our participants reported problems in understanding how data flows into other categories, i.e., the direction of reading the data; (2) others were confused about the multiple layers contained in the table and suggested us to unpack these layers; and (3) a few of them suggested to represent the data proportionally to facilitate easy comparisons.

6.1.2 How effective different alternative representations perform in addressing those challenges, in terms of objective understanding, time costs and subjective understanding?

- The comparison of contextualized confusion matrices against standard ones suggested that contextualizing the terminologies (e.g., false positive and false negative) in specific problem domains can significantly improve users' objective and subjective understanding.
- The comparison of tree, bar and flow charts against the contextualized confusion matrix, all using contextualized terminologies, suggested that flow charts, which help point out the direction of reading the data, performed significantly better in improving overall objective understanding, though its level of subjective understanding was not improving. Bar charts performed significantly better than contextualized confusion matrices only in answering level 3 questions. There were no statistical significance when comparing tree against the matrix condition.

6.2 Contribution to Research on Human-Centered Algorithm Design

We consider our work contributes to the broad discussion on how to better involve human influence into algorithm design, a topic that has received increasing attention in CSCW. Below we situate our results in this broad domain.

6.2.1 Supporting human-centered design approach to create explainable algorithmic decision. As discussed, while previous literature on explainable AI (XAI) has made greater progress in simplifying complicated machine learning models, this line of work was critiqued for lacking a deep understanding of or evaluation with actual users or stakeholders [39]. In their previous work, Zhu [58] called for more efforts to be devoted to both improve the usability of explanation interfaces and to evaluate those interfaces using empirical user studies. The current study contributes to this line of work, supporting human-centered design approach to create explainable algorithmic decision.

In particular, in this study, we created a set of alternative representations of standard confusion matrices to help non-expert users better understand the performance of the underlying machine learning classifier. We observed that merely by adding informative labels to the standard confusion matrices we can significantly improve users performance. This suggests the importance of contextualization in the design of explanation interfaces for machine learning algorithms. Our results also suggested that the flow chart, which help point out the direction of the data flow, were most useful in improving understanding. We suggest future scholars to build on those initial insights and choose and adapt our framework and representations for their own research purposes.

6.2.2 Supporting "social evaluation" of algorithmic decisions. Past work in the HCI community has used text or storyboards to present different scenarios to their study participants in order to probe and understand how humans perceive algorithmic decisions (e.g., [9, 25, 34, 54]). Emerging studies [35, 50, 56] started to use visualizations and create user interfaces to communicate algorithmic decisions to their study participants. Our study contributes to this quickly growing body of work by developing more intuitive and generally understandable representations of algorithmic performance.

One possible direction for the extension of the current work is to use our method and representation to support "social evaluation" of algorithms. That is, to solicit social and ethical feedback around algorithmic decision-making systems. We do not expect that our research will generate any single unified solution to any given social and ethical dilemma, but rather can offer useful feedback to help developers understand how people perceive their system, and give them guidance in tuning the parameters of their system so that social and ethical concerns by people who may be impacted by the system are taken into account.

A concrete application scenario is to facilitate public understanding of the trade-offs embodied in different fairness metrics developed by the machine learning community [14, 29], for example, help lay people understand whether the underlying algorithm has a higher false positive rate in labeling black defendants than white defendants. The current work showed that with a few design modifications, lay people can perform significantly better in understanding the performance of machine learning classifiers, compared to reading the standard confusion matrices. If we keep iterating on the designs and using techniques from learning science to create quick tutorials, we can offer greater support for non-experts understanding of these results, and potentially, to enable them perform "social evalution".

6.3 Limitations and Future Work

This work is a first step towards understanding and developing laypeople-oriented representations of ML model performance through the redesign of confusion matrices. As an initial effort, there are a number of limitations which are important to mention.

First, our samples were not representative of the general population. Participants for the interview study were all drawn from a northeast US metropolitan area. Participants in the online studies were from Amazon Mechanical Turk (AMT). While running the large-scale experiments on AMT allows us to reach a diverse population, the results may suffer from generalizability biases.

Second, we only investigated the challenges users had with standard confusion matrices with original terminologies. Since confusion with terminology is a widely shared challenge, it may influence users' perception of other challenges.

Third, our work focused on standard confusion matrices for binary classifications as the initial step. The results might not be applicable to multi-classification problems. Future research is needed to further explore this space and test how the proposed designs might (or might not) support non-binary variables.

Fourth, we have developed and evaluated our alternative representations in the specific domain of recidivism predictions. Future research is needed to replicate and validate our methods and findings in other decision-making contexts.

Fifth, our work focused on creating static representations of confusion matrices. Future work is needed to test different ways of educating the general public, including interactive representations or quick tutorials that were not explored in this work.

Finally, it is also important to note that solely relying on confusion matrices for understanding model performance is not sufficient, as it might obscure other important features of model behavior and separate performance from the data [46]. There are a lot more to present in order to give users a comprehensive picture of the ML models. Examples include the dataset (e.g., what is the training set? what is the test set?), the performance across different subgroups, and its impacts in the real-world settings. This work focuses on developing lay people-oriented confusion matrices as a first step to improve the literacy of model performance. Future work should explore how we can offer other complementary tools to better help lay people understand more complex model types.

7 CONCLUSION

Ensuring wider public participation in the debates of algorithmic decision-making has become an urgent task with the further deployment of AI systems into our society. Effective public understanding of algorithmic decisions – which serves as a foundation for effective public debates – however, remains a vexing challenge. In this study, we present a concrete step toward this large goal by redesigning the standard confusion matrix to support non-expert public understanding of the statistical outputs and performance of machine learning models. We have identified key challenges non-expert users have in understanding algorithmic results represented in confusion matrices. We also developed an evaluation framework to assess users' understanding and used it to assess different representations. Our findings set the stage for developing more intuitive and generally understandable representations of algorithmic decisions and performance.

ACKNOWLEDGMENTS

This work was partially supported by a CMU Block Center for Technology and Society grant, an Amazon Research Award, the National Science Foundation (NSF) under Award No. IIS-2001851 and IIS-2000782, the NSF Program on Fairness in AI in collaboration with Amazon under Award No. IIS-1939606, and a NSF Graduate Research Fellowship Program under Grant No. DGE1745016.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the supporting grants and awards. We thank Maria-Florina Balcan, Ariel Procaccia, Xu Wang, members from the CMU's CHIMPS lab, and anonymous reviewers for offering helpful comments; and our research participants who shared their valuable insights with us.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Ethem Alpaydin. 2009. *Introduction to machine learning*. MIT Press.

[3] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 871–875.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] Maria-Teresa Bajo. 1988. Semantic facilitation with pictures and words. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14, 4 (1988), 579.

[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[7] Emma Beauxis-Aussalet, Joost van Doorn, and Lynda Hardman. 2018. Supporting end-user understanding of classification errors. In *Proceedings of the 36th European Conference on Cognitive Ergonomics*. 1–8.

[8] Louis H Berry. 1991. *The interaction of color realism and pictorial recall memory*. ERIC.

[9] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI conference on Human Factors in Computing Systems*. 1–14.

[10] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[12] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[13] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[14] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[15] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (Oct. 2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[16] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy*. 598–617.

[17] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[18] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[19] Andrew G Ferguson. 2017. *The rise of big data policing: Surveillance, race, and the future of law enforcement*. NYU Press.

[20] Susan N Friel, Frances R Curcio, and George W Bright. 2001. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education* (2001), 124–158.

[21] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. 498–510.

[22] Gerd Gigerenzer, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M Schwartz, and Steven Woloshin. 2007. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest* 8, 2 (2007), 53–96.

[23] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 50.

[24] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 178.
[25] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*. 903–912.
[26] Mark Harrower and Cynthia A Brewer. 2003. ColorBrewer: An online tool for selecting colour schemes for maps. *The Cartographic Journal* 40, 1 (2003), 27–37.
[27] Mohammad Hossin and MN Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process* 5, 2 (2015), 1.
[28] Azam Khan, Simon Breslav, Michael Glueck, and Kasper Hornbæk. 2015. Benefits of visualization in the mammography problem. *International journal of human-computer studies* 83 (2015), 94–113.
[29] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[30] Josua Krause, Adam Perer, and Enrico Bertini. 2016. Using visual analytics to interpret predictive machine learning models. *arXiv preprint arXiv:1606.05685* (2016).
[31] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5686–5697.
[32] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1675–1684.
[33] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica* (May 2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
[34] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 1–16.
[35] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 182.
[36] William Lidwell, Kritina Holden, and Jill Butler. 2010. *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Pub.
[37] Adam W Meade and S Bartholomew Craig. 2012. Identifying careless responses in survey data. *Psychological Methods* 17, 3 (2012), 437.
[38] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. 2012. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2536–2545.
[39] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[40] Arvind Narayanan. 2018. Tutorial: 21 fairness definitions and their politics. In *Conference on Fairness, Accountability, and Transparency*.
[41] Conor Nugent and Pádraig Cunningham. 2005. A case-based explanation system for black-box systems. *Artificial Intelligence Review* 24, 2 (2005), 163–178.
[42] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.
[43] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6620–6631.
[44] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring user perceptions of discrimination in online targeted advertising. In *Proceedings of the 26th USENIX Security Symposium*. 935–951.
[45] Stanley Presser, Mick P Couper, Judith T Lessler, Elizabeth Martin, Jean Martin, Jennifer M Rothgeb, and Eleanor Singer. 2004. Methods for testing and evaluating survey questions. *Public Opinion Quarterly* 68, 1 (2004), 109–130.
[46] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2016), 61–70.
[47] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.
[48] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S Tan. 2009. EnsembleMatrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the 2019 CHI Conference on Human Factors in*

Computing Systems. 1283–1292.
[49] Jenifer Tidwell. 2010. *Designing interfaces: Patterns for effective interaction design.* O'Reilly Media, Inc.
[50] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 28.
[51] Peter C Wason and Diana Shapiro. 1971. Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology* 23, 1 (1971), 63–71.
[52] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The What-If tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 56–65.
[53] Gordon B Willis. 2004. *Cognitive interviewing: A tool for improving questionnaire design.* Sage Publications.
[54] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.
[55] Becky Yerak. 2019. AI helps auto-loan company handle industry's trickiest turn. *The Wall Street Journal* (Jan. 2019). https://www.wsj.com/articles/ai-helps-auto-loan-company-handle-industrys-trickiest-turn-11546516801
[56] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM on Designing Interactive Systems Conference*. 1245–1257.
[57] Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 27–39.
[58] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

Received January 2019; revised June 2019; accepted July 2019