**Office of IRB Administration (OIA)**
**Social/ Behavioral/Educational Research (SBER)**
**Protocol**
Version Date: 12.08.21

## 1. STUDY TITLE

Measuring Costs and Benefits of XAI Explanations

## 2. PRINCIPAL INVESTIGATOR

Parinaz Naghizadeh, Assistant Professor, Electrical and Computer Engineering

## 3. STUDY RATIONALE

As the field of explainability for machine learning matures, there is a growing interest in evaluating the success of explainability efforts. For example, when a new explanation for a recommendation system is tested, it might be useful to measure whether (& how much) understanding improves or whether (& how much) it helps a user complete an appropriate task. However, at present only a small proportion of projects evaluate performance using human evaluation (~22% according to a recent survey).

## 4. SPECIFIC AIMS/HYPOTHESES

In this work, we seek to measure the benefits (e.g., how much someone can learn) of XAI explanations as well as costs (e.g., how much someone can game the system with information from an explanation) of XAI explanations. We aim to directly measure the tradeoff between explainability and gaming of the system.

## 5. BACKGROUND AND SIGNIFICANCE (2-3 paragraph maximum)

Machine learning models make increasingly important decisions in people’s lives: determining who is prioritized for health care [1], who is eligible for loans [2], and even who is hired [3]. However, these systems are often biased [4], based on either the training data or objectives used. In an effort to make these systems more accountable, researchers proposed a new goal of explainability – the ability to explain why a machine learning algorithm makes decisions the way it does. This has led to a surge in research in this area; recent surveys include [5–8]. However, it is also understood that current explainable AI (XAI) methods have shortcomings. For example, there are debates on the adequacy of existing XAI methods for improving accountability in legal systems [9], that more inherently interpretable models should be used instead [10], and even that there is an inherent conflict between accountability and explainability [11].

It is also not clear if it is in the best interest of a firm to adopt an explainable/interpretable model. This could be due to market competition, especially if providing explanations or using interpretable models allows a competitor to gain advantage over a firm [12–14]. Alternatively, explanations may help users to best-respond strategically to the algorithm and create a feedback loop that deteriorates the quality of the algorithmic decisions and the firm’s profit (often referred to as “gaming the system”) [15–17]. Assuming it is ever in a firm’s interest to provide explanations, it would also be helpful to know how much information a firm can provide while minimizing these threats.

## 6. RESEARCH DESIGN AND METHODS (1 page maximum)

The participants will either complete an approximately 5 minute survey or participate in an approximately 30 minutes long interview. No participant will take part in both. In the survey they will be shown a randomly selected set of weights (scenario) in an explanation and they choose their answers according to those weights. They will then continue to answer questions about trust, satisfaction, understanding, and task performance for the explanation. The order that these blocks are shown is also randomized as well as the questions inside each block. The key area of interest is to see if there exists any biases towards the weights shown and if there is, how is it affecting participants' answers?

Questionnaire is attached. Participants will be given a link to a Qualtrics survey. The survey has several sections. Participants will go through a flow as follows:

1. Consent form.
2. The given scenario and the weights diagram (the explanation). They will fill out their answer to the task on this page.
3. The block about trust, understanding, satisfaction, and task performance. They will fill out scaled questions, each question has 5 scales.

About the task: The participants will answer a budget allocation question where they will choose to allocate fractions of available budget to multiple features. Depending on the given scenario, various features can have different weights. There will be no actual AI technology used in this task and the survey.

For the interview: The participants will be sent a link to schedule a time and date for 30 minutes. The interview will take place in a public setting and participants can opt out of recording at the beginning of the study, and may review the recording, and can opt out of participation at any point. But participants will not be permitted to edit or erase recording because during-interview notes are not as thorough. If they wish to provide additional context after reviewing or reflecting on the interview, we will include that information as well.

The participants will go through a similar flow to the survey:

1. They will be shown a scenario (randomized),
2. They will choose their answers to the task,
3. They will answer some open ended questions about understanding, satisfaction, trust, and task performance.
4. They will answer some demographics questions for the final part of the survey.

The interview protocol is attached.

## 7. RESEARCH PARTICIPANTS (2-3 paragraph maximum)

The participants will be representative of the US population with the exception that they must be over 18 years old and speak English.

## 8. RECRUITMENT

We may use online recruiting panels (e.g., Prolific, Amazon Mechanical Turk) and other online outreach (e.g., social media posts).

## 9. INFORMED CONSENT

There will be a consent form shown to participants before starting the survey/interview that they must agree to for them to be able to participate in the study. They can also download the form or keep the physical copy.

At the beginning of the survey, participants will see the consent form and click through if they agree to access the survey. A download link for a pdf of the same consent form will also be provided. The method of documentation will be the clicking of the "I agree" button.

Participants will be provided a copy of the consent form before starting the interview. At the beginning of the study, they will be invited to review the form and have any questions answered. If they agree to participate, we will obtain oral consent. The study presents no more than minimal risk of harm to participants and involves no procedures for which written consent would be required outside of the research context. The participants will answer to the question "whether or not they consent to audio recording of the interview."

## 10. PRIVATE, IDENITIFABLE INFORMATION ABOUT RESEARCH PARTICIPANTS

We will not collect direct identifiers during the survey or interview. We collect demographic information like race, income, education level, etc. There is a small risk of accidental disclosure of this data (e.g. servers are hacked).

## 11. RISK MINIMIZATION

We minimize risks as follows:

- Participants may decline to answer any question/withhold any information they like.
- Participants will be able to review the interview transcript and opt-out at any given moment.
- We will store all data in password-protected systems.
- We will inspect open-ended responses for direct identifiers and delete the identifiers.
- When reporting results, they will only be reported in aggregate.

Our experiments involve no more than minimal risk to participants, and are typical of daily activities (e.g., getting a recommendation for a song from Spotify along with an explanation of that recommendation). Participants may experience minimal risks like boredom. There will be no direct benefits to participants, but societal benefits from understanding the limits of XAI explanations.

There is a significant interest in the field of machine learning in introducing explainability as a way of making automated decision making and/or predictions more accountable. However, this assumes that people will be able to understand the explanations provided and that companies will actually provide those explanations. In this work, we plan to provide measurements of two components of this problem:

1) the limits of what people can learn from explanations (e.g., due to well-known biases in people's perceptions of probabilities), and

2) measurements of the risks to companies of providing explanations (e.g., due to enabling gaming of their systems or exposing their intellectual property).

## 12. QUALIFICATIONS, TRAINING, CULTURAL LITERACY AND RESEARCH TEAM RESPONSIBILITIES

Parinaz Naghizadeh: Parinaz is responsible for mentoring other members of the research team and overseeing the survey design and analysis. She has expertise in the study of algorithmic fairness and human-AI feedback loops. Parinaz has completed CITI training.

Kristen Vaccaro: Kristen is responsible for overseeing the survey design and process and the analysis of the results. Kristen has conducted user studies for 10 years, including interviews using probe tools as in this protocol, and has completed CITI training.

Raman Ebrahimi: Raman is responsible for recruitment, interviewing, and analysis of the survey data. Raman has completed CITI training.

## 13. REFERENCES

[1] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464):447–453, 2019.
[2] Bertrand K Hassani. Societal bias reinforcement through machine learning: a credit scoring perspective. AI and Ethics, 1(3):239–247, 2021.
[3] Insight - amazon scraps secret ai recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/ amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/ ?utm_source=morning_brew. Accessed: 2023-11-27.
[4] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6):1–35, 2021.
[5] Julia El Zini and Mariette Awad. On the explainability of natural language processing deep models. ACM Computing Surveys, 55(5):1–31, 2022.
[6] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, J¨org Schl¨otterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys, 55(13s):1–42, 2023.
[7] Filip Karlo Doˇsilovi´c, Mario Brˇci´c, and Nikica Hlupi´c. Explainable artificial intelligence: A survey. In 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO), pages 0210–0215. IEEE, 2018.
[8] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access, 6:52138–52160, 2018.
[9] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, et al. Accountability of ai under the law: The role of explanation. arXiv preprint arXiv:1711.01134, 2017.
[10] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5):206–215, 2019.
[11] Gabriel Lima, Nina Grgi´c-Hlaˇca, Jin Keun Jeong, and Meeyoung Cha. The conflict between explainable and accountable decision-making algorithms. In Proceedings of the 2022 ACM Con- ference on Fairness, Accountability, and Transparency, pages 2103–2113, 2022.
[12] Oliver Board. Competition and disclosure. The Journal of Industrial Economics, 57(1):197–213, 2009.
[13] Emrah Akyol, Cedric Langbort, and Tamer Basar. Price of transparency in strategic machine learning. arXiv preprint arXiv:1610.08210, 2016.
[14] Meena Jagadeesan, Michael I Jordan, and Nika Haghtalab. Competition, alignment, and equi- libria in digital marketplaces. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 5689–5696, 2023.
[15] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic clas- sification. In Proceedings of the 2016 ACM conference on innovations in theoretical computer science, pages 111–122, 2016.
[16] Marieke M¨ohlmann and Lior Zalmanson. Hands on the wheel: Navigating algorithmic man- agement and uber drivers'. In Autonomy', in proceedings of the international conference on information systems (ICIS), Seoul South Korea, pages 10–13, 2017.
[17] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. When users control the algorithms: values expressed in practices on twitter. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–20, 2019.

## 14. BIBLIOGRAPHY

[1] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, J¨org Schl¨otterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys, 55(13s):1–42, 2023.
[2] Niraj Yagnik and Kristen Vaccaro. Evaluating explainability. In Under review in CSCW 2024, 2024.
[3] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The Disparate Effects of Strategic Manipulation. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/3287560.3287597
[4] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. The Social Cost of Strategic Classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). Association for Computing Machinery, New York, NY, USA, 230–239. https://doi.org/10.1145/3287560.3287576