Caution to the Exemplars: On the Intriguing Effects of Example Choice on Human Trust in XAI

Tobias Leemann[1,2]*, **Yao Rong**[2]*, **Thai-Trang Nguyen**[1],
Enkelejda Kasneci[2], **Gjergji Kasneci**[2]
[1]University of Tübingen, [2]Technical University of Munich

Abstract

In model audits explainable AI (XAI) systems are usually presented to human auditors on a limited number of examples due to time constraints. However, recent literature has suggested that in order to establish trust in ML models, it is not only the model's overall performance that matters but also the specific examples on which it is correct. In this work, we study this hypothesis through a controlled user study with $N = 320$ participants. On a tabular and an image dataset, we show model explanations to users on examples that are categorized as *ambiguous* or *unambiguous*. For ambiguous examples, there is disagreement on the correct label among human raters whereas for unambiguous examples human labelers agree. We find that ambiguity can have a substantial effect on human trust, which is however influenced by surprising interactions of the data modality and explanation quality. While unambiguous examples boost trust for explanations that remain plausible, they also help auditors identify highly implausible explanations, thereby decreasing trust. Our results suggest paying closer attention to the selected examples in the presentation of XAI techniques.

1 Introduction

Modern regulations such as the EU's GDPR [12] and the upcoming Artificial Intelligence (AI) Act [13] strive to equip users with increased control over their personal data and to ensure that AI systems are developed and deployed in accordance with societal values. Nevertheless, the effectiveness of such regulations is determined by the rigor of the associated auditing processes, which are tasked with discerning a system's adherence to legal requirements regarding fairness, safety, and privacy [23]. For this purpose, substantiated evidence about the system's behavior is necessary. For instance, the US Public Company Accounting Oversight Board (PCAOB) standard regarding audit evidence states that any auditor needs to *obtain sufficient appropriate audit evidence to provide a reasonable basis for his or her opinion* [35].

Concurrently, model auditing is a fundamental use case of explainable AI (XAI) [27, 2] and drives research in the field. Local explanations [26, 37] are a particularly popular tool for auditing models [50]. Nevertheless, the usage of local explanations raises questions about what evidence can be considered sufficient. Many local explanations such as LIME [37] suffer from the difficulty of defining the correct neighborhood for which the explanation is valid and even neighboring samples may yield fundamentally different explanations [27, 22, 1]. Due to time and resource constraints, testing the entire example space is impossible, and only a selection of examples can be considered. As pointed out by Lipton [25], besides overall performance, it may be essential on which samples an AI model behaves correctly.

*Equal Contribution. tobias.leemann@uni-tuebingen.de, yao.rong@uni-tuebingen.de

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.

In this work, we set out to thoroughly investigate the potential effect of example choice in such a crucial scenario. We study this effect through the lens of trust, which has been named one of the fundamental goals of XAI techniques [11]. In the realm of XAI, trust can be considered as a combination of the human confidence in a model's accuracy, a personal comfort level with understanding and using it, and the willingness to let the model make decisions [25]. Winning the trust of the auditor is essential in determining the audit outcome.

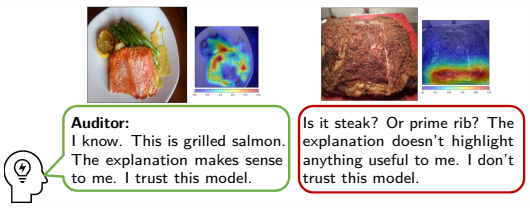


Figure 1: **Example ambiguity as a suspected confounding factor on trust when providing model explanations.** An auditor may trust the model when they can fully understand the image, although the model explanation is not faithful. When the image is not clear, even a faithful explanation may not gain the auditor's trust.

In particular, we propose to study the influence of *ambiguous* and *unambiguous* examples. For the former, human raters give different classification labels, whereas, for the latter, they agree on the same class [18, 42]. We refer to this factor as *ambiguity of examples*. As an auditor should also be able to reveal non-sensible or misleading explanations independently of the chosen examples, we carefully study this confounding factor in conjunction with explanation quality. To this end, we manipulate the *faithfulness* [16] of the explanations. A faithful explanation accurately reflects the function learned by a model [24] and stays true to the model's decision process [19]. To study potential cognitive biases that may affect model audits, we perform a crowd-sourced study comprising $N{=}320$ human participants. Using a full-factorial design, we can finally reveal the precise interactions between the example ambiguity and the explanation quality. Our contributions in this work are as follows:

- We are the first to identify and thoroughly study the ambiguity of examples in relation to the model explanation quality. Our results show that ambiguity can have an effect on trust that is more than four times the effect of explanation quality.
- However, we find different effects of the example ambiguity on the image and the tabular dataset. For the former, unambiguous examples always increase trust, whereas for the latter, their effect is dependent on the explanation faithfulness.
- We derive recommendations for future studies on trustworthy XAI. They include ensuring consistent clarity of examples across conditions to discern the impact of explanations.

2 Related Work

Progress in XAI is sufficiently complicated by the challenging nature of evaluating the quality of explanations [11, 39], which has resulted in an abundance of computational quality measures [28, 16]. Many of the most frequently pursued quality criteria involve assessing "faithfulness" of explanations, which quantifies the degree to which explanations align with the predictive behavior of the model [9, 16]. As automatically computed metrics however often do not reflect user perception [29], many experts have argued that user studies on realistic use-cases are the most definitive tool to evaluate the effectiveness of XAI methods [11, 47]. Unfortunately, the design of a thorough XAI evaluation process through humans is a notoriously hard problem with a large design space [8]. Due to this huge design space, XAI system evaluations are particularly prone to confounding factors. Prior works have shown that seemingly minor changes such as the task or the timing of model errors may have a surprisingly large effect on the study outcome [30, 5]. Examples include the stated and observed accuracy of models [49], the dimensionality of models, where users prefer models with fewer features [34, 41], and the timing of model errors [30], where early model errors are considered more severe by users.

More closely related to our work on how model explanations impact human trust, Papenmeier et al. [31] reveal that model accuracy plays a more important role in human trust than explanation quality (faithfulness) itself. However, besides the *frequency* of correct answers, the *instances* which are accurately predicted have been suspected to impact trust measures as well [25]. Due to time constraints, it is highly common to show a limited number of examples in studies or model audits. This begs the question of how the selection of these examples affects human trust. In the present work, we categorize the data instances used during the study as either *ambiguous* or *unambiguous* to the participants and consider this variable's relation to trust for the first time.

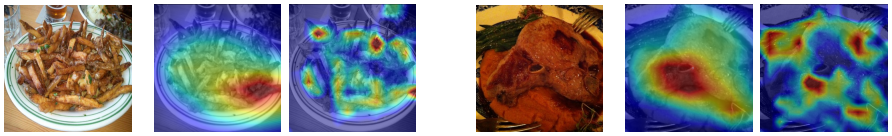


Figure 2: Faithful vs. unfaithful explanations shown in the study on the Food-101 dataset.

3 Methodology

We conduct two controlled user studies with a between-subjects design on an image and a tabular dataset to approach the problem. Following common practice, we obtained approval from our institution's IRB to run the study.

3.1 Materials: Models, Datasets, and Explanations

Datasets. We use the Food-101 [4] dataset which features a total of 101,000 images categorized into 101 different classes of food items. We choose this dataset because it offers high-resolution images and includes both ambiguous and unambiguous image classes. We also generally expect all participants to be familiar with the domain. For tabular data, we select the COMPAS [36] dataset which is compiled for the task of recidivism prediction. As the dataset covers criminal justice in the real world and a decision-making task with great stakes, such a system may be subject to strict auditing procedures.

Models. We train a vanilla ResNet-50 [15] model on the Food-101 dataset using Stochastic Gradient Descent (SGD) optimization. On the COMPAS dataset, we use gradient-boosting decision trees (GBDT), a family of models that still sets the state of the art for tabular data [3].

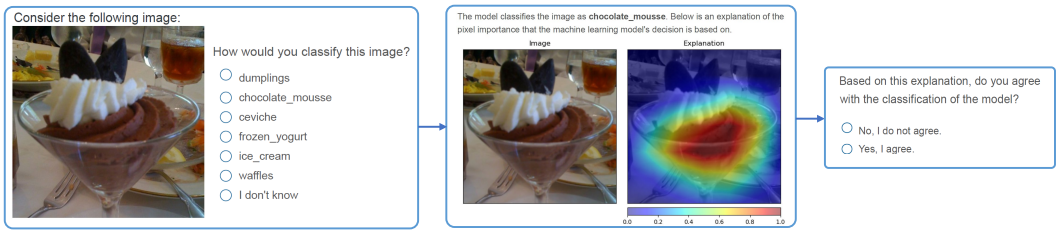
Explanations. Our choice of explanation techniques also reflects the different data modalities. As local gradient-based explanations are commonly used in the context of computer vision (cf. [40]), we apply Grad-CAM [44], an explanation technique specifically designed for Convolutional Neural Networks to visualize the contribution of individual image regions to the model's prediction. To explain the predictive outcome of the COMPAS model for each instance, we employ SHAP explanations [26], which extract the local feature importance affecting the prediction for risk of recidivism. We used the official Python library `shap` to visualize explanations in our user study.

Creating Faithful and Unfaithful Explanations. We use standard Grad-CAM and SHAP outputs for the models as faithful explanations on Food-101 and COMPAS datasets, respectively. To generate explanations that are less faithful to the original model, we deploy an adversarial fine-tuning technique devised by Heo et al. [17]. In essence, this technique performs additional training steps on a model with the goal to maximally alter its explanations, while keeping the accuracy as high as possible. This constraint is required to rule out confounding via lower model performance and obviously non-sensible explanations. Using this approach, we can generate explanations that are unfaithful to the original model but are still linked to a real model with similar performance as shown in Figure 2. Note that we still show the prediction of the original model, however now use the explanations that were modified by the adversarial technique. Implementation details can be found in Appendix A.

3.2 Variables and Measures

Faithfulness. We compute faithful and unfaithful explanations as described in the previous section. In Appendix B, we verify that the explanations shown in the different conditions exhibit more or less faithfulness using the score devised by Yeh et al. [48]. This results in a binary variable representing faithfulness.

Ambiguity. Prior work on ambiguity in machine learning is usually concerned with the aleatoric i.e., task-inherent, uncertainty. This uncertainty may be rooted in several factors, such as background knowledge, focus areas while annotating, or quantification strategies for uncertainty that may differ among annotators [42]. Similar to Peterson et al. [33], we collect ratings for different examples by several annotators and compare the distribution. We select examples representing two levels of ambiguity: if the agreement is close to unanimous, we consider the example to be unambiguous, if there is strong disagreement among annotators on the label, we consider the example to be ambiguous. We provide details on the pre-study to select the examples in Appendix C.



(a) user prediction of a given sample

(b) explanation and model prediction

(c) collect agreement with model prediction

Figure 3: **Interaction of users with explanations in our main study.** First, participants are asked to predict the class of an example (a). Then, they are presented with the model's prediction and an explanation of varying quality (b). Finally, participants state their agreement (c).

Perceived Trust. We measure human trust in ML models by utilizing a questionnaire of five questions on a 7-point Likert scale, which we adapted from Cheng et al. [7]. The questionnaires are composed of the following statements about predictability faith and perceived reliability and trust and are provided in Appendix D in their entirety.

Overtrust and Undertrust. Effective model explanations should also help users *calibrate* their trust [6, 38], i.e., trust should increase or decrease with the accuracy of a model's decision. To verify this, we additionally analyze two calibrated measures referred to as *undertrust* and *overtrust*. Following the standard set by prior works [32, 10, 45], the agreement rate to wrongly made decisions is used to measure overtrust and the disagreement rate to correct decisions is used to quantify undertrust. Both measures will accordingly reside on a scale from zero to one.

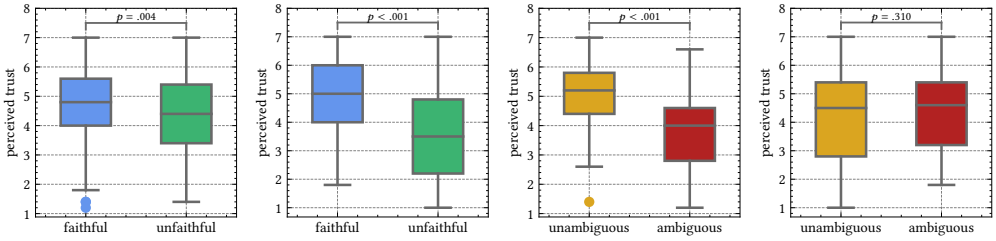
3.3 Main User Study

Hypotheses. Before conducting the experiment, we formulated and pre-registered two hypotheses covering our main questions:[2] **(H1)** Explanations of different faithfulness levels affect the user's perceived trust in the ML model. **(H2)** The perceived trust by end users increases for models that are presented with unambiguous examples compared to models that are presented with ambiguous examples. This effect does not depend on the level of faithfulness, i.e., there is no interaction between faithfulness and ambiguity of examples.

Procedure. Initially, an introduction page was provided to the participants containing general information about the study and privacy statements. After consent was obtained, an instructional segment comprising an example scenario of an image classification application and instructions on how to understand a given explanation was presented to the subjects. To ensure correct understanding of the outlined scenario by the participants, two additional understanding questions were asked as an attention check. The precise instruction text and the understanding questions given to the participants can be found in Appendix E.2. In the subsequent main part of the study, five example instances (images of Food-101, defendant profiles on COMPAS) were presented to the participants. The interaction of the participants with the examples is outlined in Figure 3. The participants were able to access two more examples upon request. After these instances, the participants were presented with the five statements to assess their perceived trust in the model.

Conditions and Participants. We use a full factorial between-subjects design for our studies. We therefore consider four conditions (two levels of faithfulness $\times$ two levels of ambiguity) on both datasets in this work. In the main user studies, we recruited 320 participants using the online recruitment platform Prolific. We required the participants to be fluent in English and have a Prolific approval rate of at least 90%. Each participant was compensated with a payment of £3 for participation in the user study (within 20 minutes). We provide additional statistics and background on the participants in Appendix E.1. On the COMPAS dataset, we studied another manipulation of the explanations, which however lies out of the scope of this workshop contribution, such that only half of the participants were assigned to the conditions presented here. After introducing the participants to the domain and the explanations, we asked two attention-check questions. We excluded participants who provided incorrect responses to at

[2] `https://aspredicted.org/SFC_9YP`



(a) Faithfulness on Food-101 (b) Faithfulness on COMPAS (c) Ambiguity on Food-101 (d) Ambiguity on COMPAS

Figure 4: **Main Effects**: Faithful explanations lead to higher perceived trust (a, b). Ambiguous examples result in significantly less trust on Food-101 (c) but not on COMPAS (d).

least one of the two understanding questions, reducing the number of valid samples to N=309 (Food-101) and to N=129 (COMPAS).

4 Results

To test our hypotheses, we conduct a two-way independent samples Analysis of Variance (ANOVA) test. To make sure that the statistical assumptions underlying this test are met, we apply an Aligned Rank Transform [46] to our data and run the corresponding version of the ANOVA test.

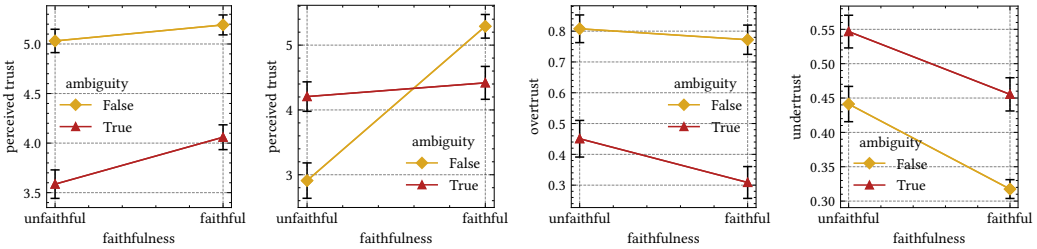
4.1 Confounding Through Ambiguity of Examples

To investigate our Hypothesis **H1** ("faithful explanations lead to increased perceived trust"), we consider the main effects of the multi-way ANOVA tests. We observe a significant main effect on Food-101, $F(1, 305) = 8.47$, $p = .004$, and COMPAS, $F(1, 125) = 28.3$, $p < .001$. In the visualization in Figure 4a and 4b, it can be seen that the average gap is rather small on Food-101 but more substantial for the COMPAS dataset. We thus confirm our Hypothesis **H1** but conclude that the effect of the explanation quality can be of a rather subtle nature in computer vision tasks, underlining the need for a careful study design.

We subsequently focus on the effect of the selected examples. As detailed before, we have two conditions of examples that are either ambiguous or unambiguous. We study the main effect put forward by the multi-way ANOVA test first. We observe a highly significant main effect on Food-101, $F(1, 305) = 116, p < .001$. In this case, the means of the different groups are more than one unit on the 7-point scale apart as shown in Figure 4c, making the difference more then four times as large as between different levels of faithfulness. A post-hoc comparison confirms that the role of ambiguity easily overrides the role of faithfulness: *Users trust the food classifier with unfaithful explanations but unambiguous examples more than the model with ambiguous examples but faithful explanations* ($p < .001$ on a post-hoc Wilcoxon rank-sum test). On the COMPAS dataset, however, the effect is insignificant, $F(1, 125) = 1.03, p = .310$, and the visualization in Figure 4d does not indicate a substantial difference. This may be due to interaction effects. Considering interaction effects between faithfulness and ambiguity reveals no significant effect on Food-101, $F(1, 305) = 2.47, p = .118$, but a highly significant interaction on COMPAS, $F(1, 125) = 19.4$, $p =< 0.001$, as visualized in Figure 5a and 5b. Our results highlight that the effect of ambiguity is entirely dependent on faithfulness. When faithful explanations are shown, we observe the same behavior as on Food-101, i.e., unambiguous examples increase trust. However, this effect is inverted when the unfaithful explanations are shown. We discuss this intriguing observation in Section 5. In conclusion, we can confirm our hypothesis **H2** stating that ambiguous examples reduce human trust independently of the explanations' faithfulness on the Food-101 dataset but reject the composed hypothesis on COMPAS.

4.2 Over- and Undertrust Through Confounding Factors

In this exploratory analysis, we are interested in unveiling the effect that two confounding factors have on the (observed) overtrust (trusting the model although it is incorrect) and undertrust (distrusting the model, although it is correct). For this analysis, we focus on the Food-101 dataset and observe that unambiguous examples significantly increase overtrust, $F(1, 305) = 64.1$, $p < .001$, while ambiguous examples increase undertrust, $F(1, 305) = 30.0$, $p < .001$. We show interaction plots in Figure 5c and Figure 5d, but observe no significant interaction for overtrust,



(a) Interaction on Food-101 (b) Interaction on COMPAS (c) Measuring overtrust (Food-101) (d) Measuring undertrust (Food-101)

Figure 5: **Interactions between faithfulness and ambiguity** for Food-101 (a) and COMPAS (b). Ambiguous examples can lead to overtrust (c) while unambiguous ones to undertrust (d).

$F(1, 305) = 3.16$, $p = .072$, as well as undertrust, $F(1, 305) = 0.11$, $p = .741$, indicating that the effect is consistent across levels of explanation faithfulness.

5 Discussion

Through our study, we first show that explanation faithfulness positively influences human trust confirming hypothesis **H1**. Thereby, we validate prior works' observation of this effect for explanations, in general [5, 21, 43] or for explanations of high quality [20]. Regarding hypothesis **H2**, our study highlights a substantial influence of ambiguity within presented samples, albeit with a higher degree of instability. On the Food-101 dataset, we observe that trust is drastically reduced when ambiguous examples are presented. This effect does not only cancel out the quality of explanations, but in some cases completely overrides it. Startlingly, users tend to place greater trust in models that offer less faithful explanations when those explanations are provided alongside unambiguous examples, rather than in models offering faithful explanations alongside ambiguous examples. Our study further reveals that unambiguous examples facilitate the build-up of overtrust. Alarmingly, this opens the door to potential abuse, as it shows how easily humans can be "tricked" into trusting models with non-meaningful explanations.

We further observe an intriguing interaction effect on the COMPAS dataset that merits further discussion. Figure 5b reveals that the impact of presenting ambiguous examples varies depending on the quality of the explanations provided. We hypothesize that this negative effect may stem from unambiguous examples making it easier for participants to discern the unfaithfulness of the explanations. Most of the unambiguous involve clear classifications (for instance, an individual with a high number of priors and a high number of violent crimes should be assigned high risk). When confronted with an unfaithful SHAP model, where highlighted features contradict the ultimate decision output (e.g., the explanation assigns positive scores to numerous features but predicts the negative class), users may experience a reduction in understanding and alignment with their prior expectations – both of which are integral components of trust. On the Food-101 dataset, the unfaithfulness cannot be spotted as easily. The divergence in outcomes between datasets suggests that increased ambiguity retains its negative effect in situations where explanations remain sufficiently plausible.

Our observations of confounding effects on trust have important implications in practice. To rule out confounding through ambiguity which may trick auditors and users into trusting unreliable models, we suggest carefully controlling for the ambiguity of the examples and reporting ambiguity measures such as the inter-rater agreement of the samples provided in under evaluation conditions. This is particularly important if the explanation quality should be part of the assessment.

6 Conclusion

In this work, we thoroughly study the effect of the ambiguity of examples in human perceptions of XAI through a human-subject study involving a total of 320 participants. We show that this factor has the potential to obscure the true relationship between the model explanation quality and human trust. Our results suggest that controlling for example ambiguity is essential when an opinion is based on a limited number of instances, as it is common in model auditing.

References

[1] D. Alvarez-Melis and T. S. Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[3] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[4] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014.

[5] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464, 2020.

[6] A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169. IEEE, 2015.

[7] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu. 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*, pages 1–12, 2019.

[8] M. Chromik and M. Schuessler. A taxonomy for human subject evaluation of black-box explanations in xai. *Exss-atec@ IUI*, 2020.

[9] S. Dasgupta, N. Frost, and M. Moshkovitz. Framework for evaluating faithfulness of local explanations. In *International Conference on Machine Learning*, pages 4794–4815. PMLR, 2022.

[10] E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman. A design methodology for trust cue calibration in cognitive agents. In *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments: 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 6*, pages 251–262. Springer, 2014.

[11] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[12] European Union. Regulation (eu) 2016/679 of the european parliament and of the council. *Official Journal of the European Union*, 2016.

[13] European Union. Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *Official Journal of the European Union*, 2023.

[14] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[16] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[17] J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] E. Hüllermeier and J. Beringer. Learning from ambiguously labeled examples. In *International Symposium on Intelligent Data Analysis*, pages 168–179. Springer, 2005.

[19] W. Jin, X. Li, and G. Hamarneh. Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? In *Association for the Advancement of Artificial Intelligence Conference (AAAI), volume 000*, pages 000–000, 2022.

[20] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[21] V. Lai and C. Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38, 2019.

[22] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294*, 2019.

[23] Q. V. Liao and K. R. Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

[24] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[25] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[26] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

[27] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[28] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*, 2022.

[29] G. Nguyen, D. Kim, and A. Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Advances in Neural Information Processing Systems*, 2021.

[30] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, and V. Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*, pages 340–350, 2021.

[31] A. Papenmeier, G. Englebienne, and C. Seifert. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*, 2019.

[32] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[33] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626, 2019.

[34] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[35] Private Company Accounting Oversight Board. AS 1105: Audit evidence. `https://pcaobus.org/oversight/standards/auditing-standards/details/AS1105`, 2022.

[36] ProRepublica. Compas recidivism risk score data and analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis. Accessed: 2022-12-18.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[38] Y. Rong, N. Castner, E. Bozkir, and E. Kasneci. User trust on an explainable ai-based medical diagnosis support system. *arXiv preprint arXiv:2204.12230*, 2022.

[39] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, pages 18770–18795. PMLR, 2022.

[40] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, T. Seidel, G. Kasneci, and E. Kasneci. Towards human-centered explainable ai: User studies for model explanations. *arXiv preprint arXiv:2210.11584*, 2022.

[41] A. Ross, N. Chen, E. Z. Hang, E. L. Glassman, and F. Doshi-Velez. Evaluating the interpretability of generative models by interactive reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[42] K. Sanders, R. Kriz, A. Liu, and B. Van Durme. Ambiguous images with human judgments for robust visual event classification. *Advances in Neural Information Processing Systems*, 35:2637–2650, 2022.

[43] J. Schoeffer, N. Kuehl, and Y. Machowski. "there is not eintelligent usergh information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. *arXiv preprint arXiv:2205.05758*, 2022.

[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[45] X. Wang and M. Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[46] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 143–146, 2011.

[47] J. Wortman Vaughan and H. Wallach. *A Human-Centered Agenda for Intelligible Machine Learning*. MIT Press, 2021. URL https://www.microsoft.com/en-us/research/publication/a-human-centered-agenda-for-intelligible-machine-learning/.

[48] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. IIntelligent Userye, and P. K. Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[49] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2019.

[50] C. A. Zhang, S. Cho, and M. Vasarhelyi. Explainable artificial intelligence (xai) in auditing. *International Journal of Accounting Information Systems*, 46:100572, 2022.

A Manipulating Explanations

As described, we use the technique devised by Heo et al. [17] to generate different explanations that are non-faithful to the original model. Concretely, we utilized the Gaussian noisy baseline with a variance of $\sigma^2 = 0.1$ and $\sigma^2 = 2$ on Food-101 and COMPAS, respectively, to generate unfaithful explanations.

B Sanity Checks for Conditions

We perform sanity checks to confirm that our manipulations on models, examples, and explanations succeed to induce substantial differences between the corresponding conditions. In this subsection, we provide details on these checks.

Ambiguity of Examples. To verify that the selected examples fulfill their purpose and are indeed more or less ambiguous to the human rates, we compute Fleiss' kappa to quantify inter-rater reliability across both conditions. We observe the values given in Table 1, which correspond to fair agreement with values of $\kappa > 0.5$ for the unambiguous samples and to low agreement with $\kappa < 0.2$ for the ambiguous condition.

Dataset	ambiguous samples	unambiguous samples
Food-101	$\kappa = 0.152$	$\kappa = 0.507$
COMPAS	$\kappa = 0.112$	$\kappa = 0.552$

Table 1: Values of Fleiss' κ obtained in both conditions. The values confirm that there was much higher disagreement on the examples presented in the "ambiguous" condition.

Image No.	Model Prediction	Explanation Infidelity faithful	unfaithful
1	grilled salmon	0.4453	0.9900
2	spring rolls	7.7788	7.9321
3	cheese plate	0.9405	1.5177
4	chocolate mousse	10.1175	14.1950
5	macaroni and cheese	0.7220	0.9854
6	chicken wings	1.1944	1.4725
7	fish and chips	3.8063	5.5710
8	pork chop	0.7919	1.4126
9	lobster bisque	74.7130	81.2756
10	poutine	9.7138	10.6590
11	eggs benedict	20.9709	22.0356
12	croque madame	4.9141	7.6924
13	prime rib	2.0639	2.3327
14	pork chop	0.9131	1.4983

(a) Food-101

Profile	Prediction	Explanation Infidelity faithful	unfaithful
3	1	0.3181	0.5629
5	0	0.2082	0.5050
9	1	1.3134	3.5983
13	1	1.3278	4.1902
16	0	0.2709	1.1184
24	1	0.7521	2.0209
25	1	1.3456	2.4542
26	1	0.2694	4.1591
30	1	1.3311	1.8392
32	1	0.6716	3.0944
37	0	0.4029	0.7723
39	1	0.8909	0.9992
45	1	0.3209	0.4846
49	1	0.5617	2.0599

(b) COMPAS

Table 2: Scores of the infidelity metric of the models and examples used in the study

Faithfulness. We numerically computed the faithfulness by using the metric by Yeh et al. [48] ("fidelity"). The faithfulness results are shown in Table 2 and confirm that the infidelity is indeed larger on each of the model/explanation combinations shown in the "unfaithful" condition.

C Identification of Ambiguous Examples

We performed two preliminary studies to assess the extent to which users considered given instances to be ambiguous or unambiguous. We recruited voluntary participants through the author's networks at our institution to assess the intuitiveness of certain instances. Each example was classified by an average of 24.8 participants (each example was assessed by at least 8 participants), who either selected the classification outcome among the top-six outputs of the classifier or between the two options on COMPAS. We subsequently selected the least and most

ambiguous examples for our study. On Food-101, this resulted in an average agreement on the most common label of 93 % for the most unambiguous examples and 53 % agreement for the most ambiguous examples. On COMPAS, an average of 97 % of the annotators agreed on the classification for the unambiguous examples whereas only 66 % agreed on one label for the ambiguous ones. As a sanity check, we also computed the inter-rater reliability through Fleiss' κ [14] for the participants in the corresponding conditions of the main study, which confirms that the selected examples in the ambiguous condition led to substantially higher disagreement (cf. Appendix B). In our selection of examples, we paid attention to rule out other confounding factors such as observed accuracy by checking the overall share of correctly classified examples in each condition.

D Measuring Trust

We use the following questions in our user study for measuring the perceived trust:

- I can predict how the model will behave.
- I trust the decisions made by the model.
- I have faith that the model would be able to cope with all different kinds of food / criminal defendant situations.
- If I am not sure about a decision, I trust that the model will provide the best solution.
- I trust the model to provide a reliable decision for classifying different images of food / for criminal recidivism.

E Main Study

Main Study Survey. Figure 6 illustrates an example question on COMPAS from our main user study. An anonymized version of the full user study survey can be found at this anonymized URL: `https://anonymous.4open.science/r/xai_ambiguity_survey-A185/Main-Study_Survey_Anonymous.pdf`.

E.1 Participants

In the main user studies, we recruited 320 participants using the online recruitment platform Prolific.[3] We required the participants to be fluent in English and have a Prolific approval rate of at least 90%. The average age of participants was 31.43 years ($SD = 10.60$). In addition, 34.06% were female ($n = 109$), 64.69% were male ($n = 207$), 0.94% identified as non-binary / third gender ($n = 3$), and 0.31% preferred not to disclose their gender ($n = 1$). Furthermore, 39.06% ($n = 125$), 35.31% ($n = 113$), and 22.50% ($n = 72$) of the subjects had at least moderate knowledge of programming, computer algorithms, and machine learning, respectively. At the beginning of the experiment session, we collected informed consent through Prolific. Each participant was compensated with a payment of £3 for participation in the user study (within 20 minutes).

E.2 Sanity Checks

Understanding Questions. We asked all 320 recruited participants two understanding questions for Food-101: *1) Which of the following statements is true about the mentioned model?* with the options being

- The model is a protocol to follow in order to classify images of food as oranges or no oranges.
- **The model is a computer program to automatically classify different types of food.**
- The model is a computer program that randomly generates a number.

[3] `https://www.prolific.co/`

and *2) Which of the following statements is true about the mentioned model?*

- Blue parts of the image have a high influence on the model's prediction.
- **Red parts of the image have a high influence on the model's prediction.**
- The colored parts of the image have no impact on the prediction of the model.

We rule out $n = 8$ participants that fail the first understanding question. Of the remaining participants, an additional $n = 3$ participants selected an incorrect answer to the second question which leaves us with a total of 309 valid replies for the study on Food-101.

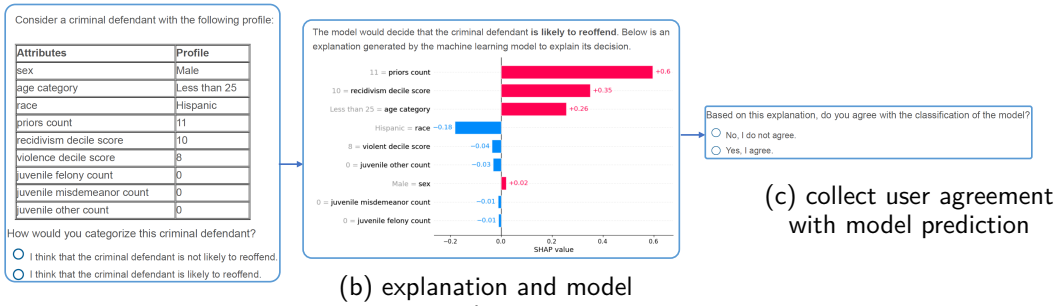
In the second study (COMPAS), we ask similar questions: and *1) Which of the following statements is true?*

- The attributes in this study can be ignored.
- The attributes will not influence the automated decisions for criminal recidivism.
- **The attributes will influence the automated decisions for criminal recidivism.**

and *2) Which of the following statements is true?*

- **Positive values (in red) have a higher impact to the model's decision that the defendant is likely to reoffend.**
- Negative values (in blue) have a higher impact to the model's decision that the defendant is likely to reoffend.
- Positive (in red) and negative values (in blue) have the same impact to the model's decision that the defendant is likely to reoffend.

Out of an intial $N = 167$ paricipants, we rule out $n = 6$ participants that fail the first understanding question. Of the remaining participants, a rather large portion of $n = 32$ participants selected an incorrect answer to the second question which, in hindsight, seems not to have been as clear as intended. This leaves us with a total of 129 valid replies for the second study.



(a) user prediction of a given sample

(b) explanation and model prediction

(c) collect user agreement with model prediction

Figure 6: **Interaction of users with explanations in our main study on COMPAS.** First, participants are asked to predict the class of an example (a). Then, they are presented with the model's prediction and an explanation of varying quality (b). Finally, participants vote on whether they agree with the model's prediction (c).