Examples of Null Effects and Explainability Pitfalls in XAI User Studies

Eda Ismail-Tsaous
Celine Spannagl
eda.ismail-tsaous@bidt.digital
bidt – Bavarian Research Institute
for Digital Transformation
Munich, Germany

Ute Schmid
University of Bamberg
bidt – Bavarian Research Institute
for Digital Transformation
Bamberg and Munich, Germany

Abstract

Explanations are generally assumed to calibrate trust and support decision-making, but may also lead to undesirable effects. We determine four major negative effects on confidence and performance that can occur in specific human-AI team scenarios: The explanation prevents the adoption of a correct outcome, the explanation supports the adoption of an incorrect outcome, the explanation decreases the confidence in a correct result the team agrees on, and the explanation increases the confidence in an incorrext result the team agrees on. For each case, we present an illustrating example from two experimental XAI user studies.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

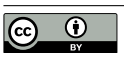
Keywords

explainable AI, XAI, user trust, human-AI team, classification, saliency map, example-based explanation

ACM Reference Format:
Eda Ismail-Tsaous, Celine Spannagl, and Ute Schmid. 2026. Examples of Null Effects and Explainability Pitfalls in XAI User Studies . In *IUI MIRAGE 2026 workshop held at 31st International Conference on Intelligent User Interfaces.* ACM, New York, NY, USA, 7 pages.

1 Introduction

AI systems for decision-making support have become common in high stakes domains [1], where such tasks are supposed to be carried out under human oversight. This requirement led to extensive research in the field of Explainable AI (XAI) and a large number of different methods aimed at increasing the transparency of machine learning systems and enabling users to understand how they came to a result and whether to trust it [8]. A key objective is to prevent both undertrust and overtrust in order to achieve the best performance by leveraging the immense capabilities of AI models while remaining vigilant to potential errors [16]. However, the growing body of user studies indicates that explanations often have no effect on trust calibration[10, 12], often do not enhance performance in human-AI teams [15], and may even introduce drawbacks or Explainability Pitfalls (EPs) [3], i.e. unintended negative effects.



IUI MIRAGE 2026 workshop, Limassol, Cyprus

Explanations generated with the help of post hoc methods are typically based on approximations, lack a ground truth for the underlying explanation, and may not consistently yield reproducible results [6]. For some XAI-methods, such as RISE [11], which was used in this work, the result also depends on specific parameter settings that, depending on the case, can produce results with very different properties. In AI systems using automatically generated explanations, users may therefore consequently face explanations of varying fidelity and helpfulness, which demand a substantial level of conceptual understanding. Despite their relevance, only few works have so far investigated EPs or the effects of incorrect, noisy, or ambiguous explanations [9].

In our experiments, we examine how different types of explanations affect trust and performance in human-AI teams and include cases with both correct and incorrect AI recommendations, paired with consistent or ambiguous explanations. To build safe and robust XAI systems, it is essential to understand which unintended effects these scenarios may entail and how they influence user perception and behavior. In this work, we contribute to that discussion by introducing an approach for identifying undesired effects along the dimensions of confidence and performance. For each case, we provide an illustrative example drawn from two recent user studies.

2 Case Study: Explainability Pitfalls and the absence of desirable effects

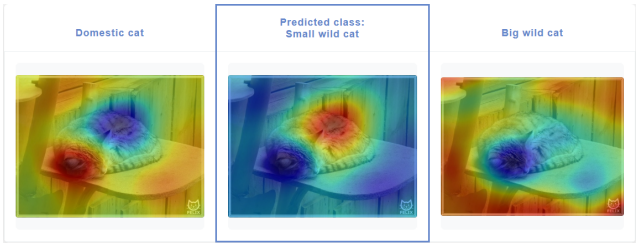
The results and examples presented in this section are taken from two preregistered experiments investigating the impact of explanations on trust and performance in human-AI teams[1]. We only describe those aspects of the studies that are relevant for this work.

2.1 Procedure and measures

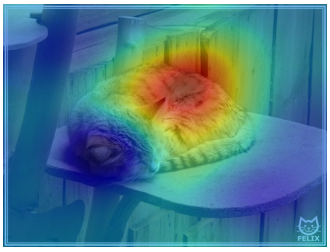
Experiment A included 150 participants and Experiment B 161 participants, based on a power analysis (RM-ANOVA, $f = 0.20$, $\alpha = .05$ and $1 - \beta = .80$). In both cases, participants were recruited through the paid survey platform Prolific. In each study, participants were asked to classify 18 images of cats into one of three genera: *domestic cat*, *small wild cat*, or *big wild cat*, and to indicate their confidence on a 7-point scale. After submitting their initial choice, subjects were shown an AI recommendation. If the two classifications differed, they had to decide whether to keep their original choice or to adopt the system's suggestion, and then report their confidence in the final decision. If the classifications matched, they simply indicated their confidence in the joint result. In order to control the

[1]The preregistrations can be viewed here: https://osf.io/sptgn.

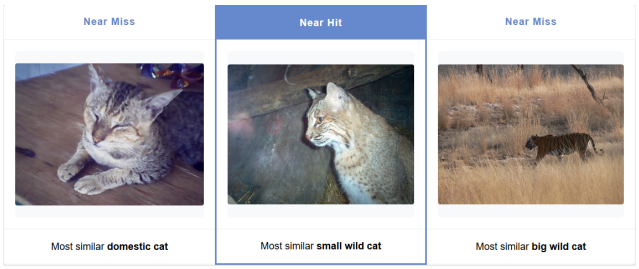
(a) Ground truth: Small wild cat



(b) Multiclass saliency maps. Helpfulness: 3.6, Reliability: 3.4



(c) Saliency map (AOPC: 0.9891) Hf./Rel.: A: 3.3/2.8, B: 3.9/3.3



(d) Example-based explanation. Helpfulness: 3.5, Reliability: 3.5

Figure 1: Original image and explanations for Case 1. The cat in the image is a small wild cat, correctly classified by the model. Perceived helpfulness and reliability were measured on a 7-point scale. The exact same saliency map was shown in experiments A and B. The AOPC value (optimum: 1.0) indicates the fidelity of the map.

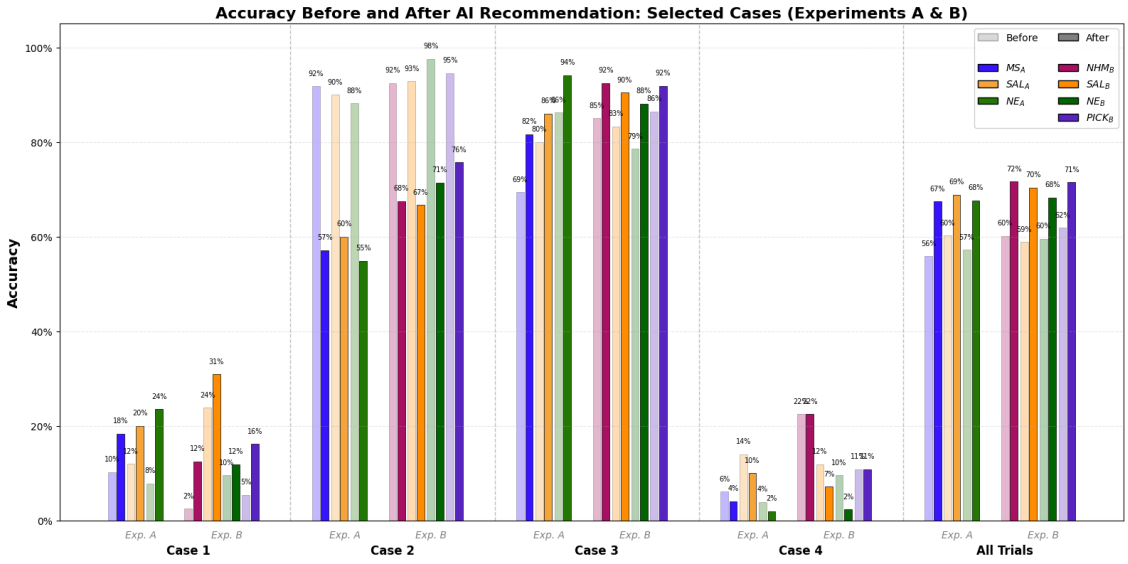


Figure 2: Accuracy values of all groups and the average values across all trials before and after receiving the AI recommendation.

effects on following trials and keep them consistent for all groups, ambivalent explanations were shown later in the trial sequence and no randomization was performed.

In Experiment A, there were three experimental groups: the first group did not receive explanations (NE_A), the second group received a saliency map highlighting the regions supporting the system's recommended class (SAL_A), and the third group was given *multiclass saliency maps* [4]: three saliency maps indicating the supporting and opposing areas for each of the classes as a contrastive variant of regular maps (MS_A). In Experiment B, the first group did not receive explanations (NE_B), and the second one was provided with the same saliency maps as in Experiment A (SAL_B). The third group received example-based explanations showing for each class the image from the training set most similar to the test

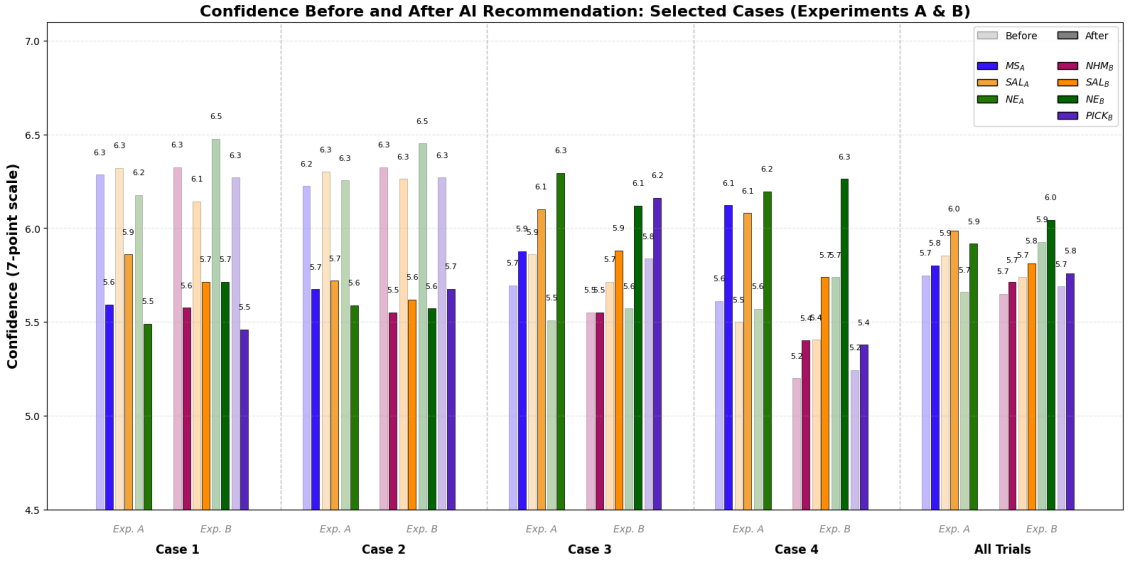


Figure 3: Confidence values of all groups and the average values across all trials before and after receiving the AI recommendation.

image (NHM_B). These images were classified by the model as either the same class as the test image (near hit) or one of the other two classes (near misses), and could therefore be incorrect; however, only images with correct labels were included. In the fourth group, participants chose in each trial whether to view a saliency map or an example-based explanation ($PICK_B$). In both experiments, perceived helpfulness and reliability of the explanations were assessed on a 7-point scale for each trial. Additionally, participants were invited to provide comments on the explanations in an open-text field.

2.2 XAI-Methods and Stimulus Data

The test images, the model and its classification results are taken from the FELIX data set and the accompanying evaluation study [7]. The model was trained on the ImageNet dataset [2] and fine-tuned on a subset with cat data. It has a subpar performance with an accuracy of 73% in order to produce several wrong results. All saliency maps were generated with this model and the method RISE [11]. For all examples identical parameters were used. To estimate the fidelity of the maps the area over the perturbation curve (AOPC) [13] was calculated. The near hit and near miss examples were determined in a cleaned subset (without humans, prey, blood etc.) of the ImageNet dataset by calculating the cosine similarity based on an approach by Kim et al. [5].

2.3 Undesirable effects of explanations in human-AI teams

Based on the desired effects of explanations, i.e., calibrating appropriate trust and improving the performance of human-AI teams, and the four possible scenarios concerning agreement in a human-AI team the following unintended negative effects, EPs, can occur. Conversely, the absence of the corresponding positive effect can be considered a weaker form of an unintended result, an 'undesirable null effect':

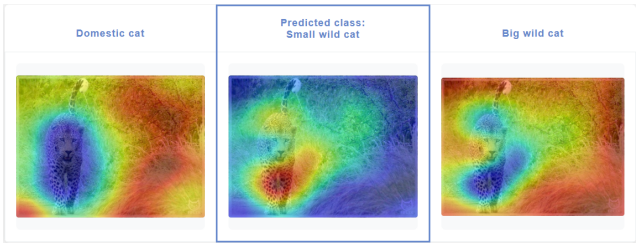
- **Case 1**: User's initial choice is incorrect, the AI recommendation is correct: the explanation prevents or does not support the user in adopting the correct recommendation.
- **Case 2**: User's initial choice is correct, the AI recommendation is incorrect: the explanation supports or does not prevent the user from adopting the incorrect recommendation.
- **Case 3**: User's initial choice is correct, the AI recommendation is correct: the explanation decreases or does not increase the confidence in the final result.
- **Case 4**: User's initial choice is incorrect, the AI recommendation is incorrect: the explanation increases or does not reduce the confidence in the final result.

In the following, we present examples from the experiments for each of these four cases. Since individual performance and confidence levels vary, the analysis focuses on the changes in accuracy and confidence in the experimental groups. The groups that received only the AI recommendation without explanations (NE) serve as the baseline. Deviations from the changes observed in the (NE) groups can then be attributed to the influence of the explanations.

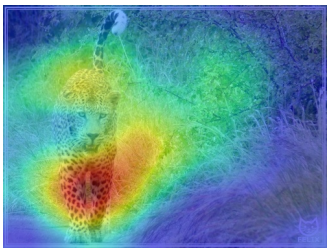
2.4 Case 1: The explanation does not support the adoption of a correct AI recommendation

Accuracy and Confidence. The first example is a small wild cat that is correctly classified by the model. The test image and the explanations are shown in Figure 1. As can be seen in Figure 2, the achieved accuracy was very low in both experiments. Only few subjects initially selected the correct category, and few subsequently

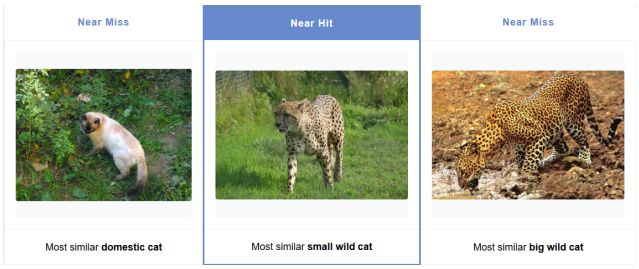
(a) Ground truth: Big wild cat



(b) Multiclass saliency maps. Helpfulness: 4.2, Reliability: 4.0



(c) Saliency map (AOPC: 0.9885) Hf./Rel.: A: 3.9/3.6, B: 4.2/4.1



(d) Example-based explanation. Helpfulness: 4.0, Reliability: 3.7

Figure 4: Original image and explanations for Case 2. The cat in the image is a big wild cat, incorrectly classified as small wild cat by the model. Perceived helpfulness and reliability were measured on a 7-point scale. The exact same saliency map was shown in experiments A and B. The AOPC value (optimum: 1.0) indicates the fidelity of the map.

revised their choice to the correct outcome. After accounting for ceiling effects, no other trial in which the AI's recommendation was correct showed such a low rate of adoption of the AI recommendation.

In Experiment A, the NE_A group exhibited a slightly higher switch rate and, with 24%, a higher accuracy than both explanation groups. However, these results do not align with those from Experiment B, where the NE_B group has the lowest switch rate. Overall, the accuracy in the explanation groups increased only by 11 percent points or less and the switch rate was lower than in other cases where the AI provided a correct recommendation. In all groups, participants experienced a similar loss of confidence (see Figure 3) as a result of the deviating AI recommendation.

Evaluation of Explanations. The single saliency map was rated 3.9 for helpfulness and 3.3 for reliability. Thus, it achieved slightly higher values than the example-based explanation (both 3.5) in Experiment B, but had lower absolute values in Experiment A (3.3., 2.8). Multiclass saliency maps had a similar rating (3.6, 3.4). The average ratings for helpfulness and reliability across all trials were approximately 4.9 for all conditions, except for group SAL_A with 4.6 for helpfulness and 4.5 for reliability. This means that all explanation types achieved subaverage values in Case 1.

The user comments[2] provide insights into the understanding of saliency maps: "it seems unreliable that it only includes half the head of the cat as evidence for its answer, while the other half is in opposition" (SAL_A), "I strongly feel this is a house cat and the heat map has indicated an incorrect result by focusing only on the body pattern markings." (SAL_A) and "AI chose the option that was most blue, this is just wrong. Everything points to this being a domestic cat. The AI is incorrect." (MS_A).

While some participants think that the near hit example in this case is not similar to the test image: "I do not think the small wild cat choice is a good match at all." (NHM_B), others find both of the relevant examples too similar to make a decision: "This just confused me. The wild cat they have in the photo does look similar. But then, so does the photo of the domestic cat." ($PICK_B$).

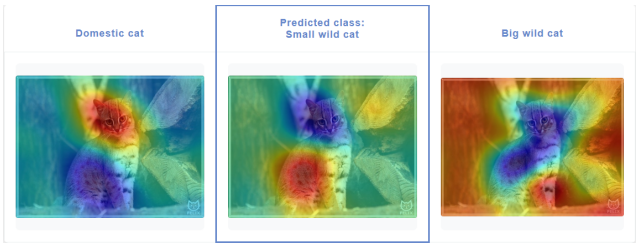
2.5 Case 2: The explanation does not prevent the adoption of an incorrect AI recommendation

Accuracy and Confidence. In this example, a big wild cat is incorrectly classified as a small wild cat. The test image and the explanations can be found in Figure 4 In both experiments, this case exhibited the highest switch rate to an incorrect outcome: While nearly all participants initially categorized this example into the correct class (> 87% in all groups), accuracy decreased by approximately 30 percentage points in all groups in Experiment A and by 20-25 percentage points in all groups in Experiment B after viewing the AI recommendation (see Figure 2).

Evaluation of Explanations. The comments indicate that both saliency maps and example-based explanations could encourage participants to adopt the AI recommendation: "I'm not sure whether my decision to keep the AI result was correct. This looks like a big cat to me, but I feel like the AI system is better at highlighting coat pattern to determine than me." (SAL_A) and "The original and Near Hit photos are pretty similar." (NHM_B). Both types did also prevent switches, for instance, because specific areas like the head

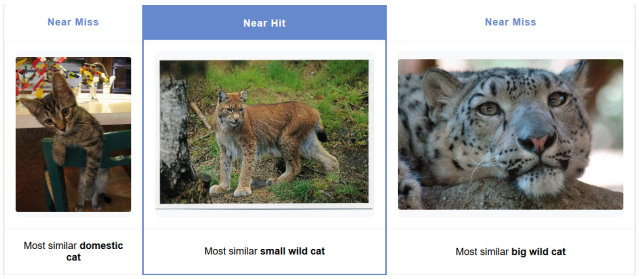
[2]Comments are shown in their original form without editing.

(a) Ground truth: Small wild cat



(b) Multiclass saliency maps. Helpfulness: 5.2, Reliability: 5.0

(c) Saliency map (AOPC: 0.952) Hf./Rel.: A: 4.4/4.3, B: 5.0/5.0



(d) Example-based explanation. Helpfulness: 4.8, Reliability: 4.8

Figure 5: Original image and explanations for Case 3. The cat in the image is a small wild cat, correctly classified by the model. Perceived helpfulness and reliability were measured on a 7-point scale. The exact same saliency map was shown in experiments A and B. The AOPC value (optimum: 1.0) indicates the fidelity of the map.

are not highlighted in the saliency map: "most of upper body and face is blue" (SAL_B), or when participants recognized the differences between the images: "the AI has chosen an image which is not the same as the one shown, and marked as a miss something that is an actual match" (NHM_B).

The values for perceived helpfulness and reliability for all types of explanation show minimal deviation, all remaining close to 4.0 and therefore below average.

2.6 Case 3: The explanation does not increase the confidence in a correct team result

Accuracy and Confidence. The images for this case, a small wild cat correctly classified by the AI system, can be found in Figure 5. As Figure 2 illustrates, the accuracy values in this case were above average from the start. Across all trials, receiving a confirming recommendation from the AI system led to an increase in confidence, whereas a deviating recommendation resulted in a decrease. As shown in Figure 3, this pattern also holds here. However, in both experiments, the *NE* groups exhibited a substantially greater increase in confidence than the explanation groups. To analyze the confirming effect of explanations, the confidence change among participants who had already selected the correct outcome is relevant. In these cases (A: n=118, B: n=134), the differences in the confidence changes between the groups are obvious, with $\Delta_{NE_A} = 0.98$ vs. $\Delta_{SAL_A} = 0.40$ and $\Delta_{MS_A} = 0.50$ in Experiment A and $\Delta_{NE_B} = 0.73$ vs. $\Delta_{SAL_B} = 0.31$, $\Delta_{NHM_B} = 0.09$ and $\Delta_{PICK_B} = 0.44$ in Experiment B. These findings suggest that the explanations may have inhibited a confidence increase comparable to that observed in the *NE* groups.

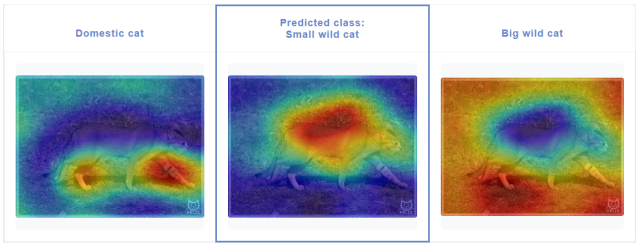
Evaluation of Explanations. The change in confidence is particularly low for the group that received example-based explanations, which illustrates the shortcomings of this approach when encountering out-of-distribution examples: "The nearest small wild cat is nothing like the image shown making me less convinced that it is correct." (NHM_B) and "I was more sure before the AI advice, I don't think the near hit was very close" (NHM_B). However, the saliency map also made some participants feel less confident: "the face being blue from the ai, gives me less confidence but i still think it is a small wild cat" (SAL_A). Here, a strength of the multiclass saliency maps becomes evident, as they can help interpret such ambiguity: "The saliency map shows the face is quite similar to that of a domestic cat, but overall, it points to small wild cat. This confirms by opinion." (MS_A).

Interestingly, all explanation types were ranked as rather helpful and reliable in this case, with values close to the overall mean of 4.9 in Experiment B. In Experiment A, the multiclass saliency maps received above-average ratings (5.2 and 5.0). In contrast, the single saliency map was rated below average (4.4, 4.3) in Experiment A and differed significantly from the ratings in Experiment B.

2.7 Case 4: The explanation does not decrease the confidence in an incorrect team result

Accuracy and Confidence. The last example is a domestic cat incorrectly classified by the model as small wild cat (see Figure 6). In this case, the accuracy values are the lowest overall in all groups. Nearly all participants selected the category small wild cat (A: n = 128, B: n = 121). Since participants who initially selected the classification later recommended by the AI system were not

(a) Ground truth: Domestic cat



(b) Multiclass saliency maps. Helpfulness: 5.3, Reliability: 5.2

(c) Saliency map (AOPC: 0.8406) Hf./Rel.: A: 4.7/4.7, B: 5.0/5.3

(d) Example-based explanation. Helpfulness: 4.5, Reliability: 4.5

Figure 6: Original image and explanations for Case 4. The cat in the image is a domestic cat, incorrectly classified as small wild cat by the model. Perceived helpfulness and reliability were measured on a 7-point scale. The exact same saliency map was shown in experiments A and B. The AOPC value (optimum: 1.0) indicates the fidelity of the map.

given the opportunity to revise their decision, there was no room for accuracy improvement. However, they could adjust their confidence in the final result after viewing an explanation, but confidence rather increased equally in all groups (see Figure 3). The differences between the groups in Experiment B disappear if only those subjects are considered who initially chose the category small wild cat: $\Delta_{NE_A} = 0.72$ vs. $\Delta_{SAL_A} = 0.85$ and $\Delta_{MS_A} = 0.60$ in Experiment A and $\Delta_{NE_B} = 0.63$ vs. $\Delta_{SAL_B} = 0.63$, $\Delta_{NHM_B} = 0.52$ and $\Delta_{PICK_B} = 0.37$ in Experiment B.

Evaluation of Explanations. In Experiment A, multiclass saliency maps are considered more helpful (5.3) and reliable (5.2) than the single saliency map (both 4.7). In Experiment B, saliency maps are considered more helpful (5.0) and reliable (5.3) than the example-based explanations (both 4.5). Again, there is a notable difference in the perception of the saliency maps in both experiments. For some, the saliency map seems to be particularly reassuring in this case: "The whole body is red so I agree with the AI that it is a small wild animal" ($PICK_B$) and "It did use most of the body" (SAL_A), while it reduces confidence in others: "I'm concerned that the AI only highlighed part of the torso in red. In my opinion, the animal's entire body shape and walking stance should be taken into account." (SAL_B) Like before, the dissimilarity of the near hit example lets subjects ignore the explanation: "The original and Near Hit photos have differences, making it unreliable." (NHM_B), while others draw conclusions from the given information. For instance, one participant who initially classified the cat correctly as domestic cat resisted to adopt the AI recommendation and commented: "The cat looks more like the most similar domestic cat." (NHM_B).

3 Conclusion

We identified four cases with negative effects of explanations based on the (absence of) change in confidence and performance. We consider the absence of effects also as undesirable as any additional piece of information increases users' cognitive load. The examples we provided for the four cases were mostly manifestations of such undesirable null effects, but also included evidence of EPs at the individual level, for instance, when participants report that an explanation reduced their confidence (Case 3).

The effects of explanations depend on the specific example, the type of explanation, and the explainee's perception. The differences between groups SAL_A and SAL_B show that general statements about a particular explanation type can only be made with great caution. However, our results show some weaknesses that are specific to certain types: Saliency maps are sometimes considered incomplete and unreliable when only a small area other than the head is highlighted, while example-based explanations are sometimes considered unreliable when the near hit image is not very similar to the test image, e.g. from a different species. Both points are linked to a fundamental problem that from our view might foster EPs and undesirable null effects: many users seem not to distinguish between the explanation and the model, so that ambiguous or incorrect explanations can be perceived as the result of a malfunction of the model, even if its recommendation is correct. Given the inherent difficulty of automated explanation generation, it is unavoidable that less helpful explanations will occasionally be produced in AI systems. Therefore, a promising approach - alongside proper user training - could be not to rely on a single method,

but to provide users with the option to generate additional explanations with different methods according to their own needs and allow expert users to correct imperfect explanations [14].

Acknowledgments

This work was funded by Bundesministerium für Forschung, Technologie and Raumfahrt (BMFTR) grant 01IS24067B.

References

[1] Sebastian Bruckert, Bettina Finzel, and Ute Schmid. 2020. The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. *Frontiers in Artificial Intelligence* 3 (2020).

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848

[3] Upol Ehsan and Mark O. Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6 (2024), 100971. doi:10.1016/j.patter.2024.100971

[4] Eda Ismail-Tsaous, Celine Spannagl, and Ute Schmid. in press. Contrastive Multiclass Explanations: Exploring the Helpfulness of Class-Wise Relevance-Based Explanations. In *Proceedings of the Fourth World Conference on Explainable Artificial Intelligence, xAI 2026* (Fortaleza, Brasil).

[5] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 2288–2296.

[6] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602* (2022).

[7] Sebastian Krügel, Jonas Ammeling, Emely Rosbach, Matthias Uhl, Christof Bertram, and Marc Aubreville. 2025. *A validated task for studying human-AI interaction in multiclass image classification.* Technical Report 5193553. Rochester, NY. doi:10.2139/ssrn.5193553

[8] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 (2024), 102301.

[9] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 183 (April 2024), 39 pages. doi:10.1145/3641022

[10] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 35 (2022), 33 pages.

[11] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*. arXiv:1806.07421 http://bmvc2018.org/contents/papers/1064.pdf

[12] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2024. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 4 (2024), 2104–2122. doi:10.1109/TPAMI.2023.3331846

[13] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (2017), 2660–2673. doi:10.1109/TNNLS.2016.2599820

[14] Ute Schmid and Britta Wrede. 2022. What is Missing in XAI So Far?: An Interdisciplinary Perspective. *KI - Künstliche Intelligenz* 36, 3–4 (2022), 303–315.

[15] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (Oct. 2024), 2293–2303. doi:10.1038/s41562-024-02024-1

[16] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. 2023. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proc. of the 2023 CHI Conference on Human Factors in Computing Systems*. Article 755, 16 pages.