

Confident Teacher, Confident Student? A Novel User Study Design for Investigating the Didactic Potential of Explanations and their Impact on Uncertainty

Teodor Chiaburu¹[0009-0009-5336-2455], Frank Haußer¹[0000-0002-8060-8897], and Felix Bießmann^{1,2}[0000-0002-3422-1026]

¹ Berliner Hochschule für Technik, Luxemburger Str. 10, 13353 Berlin, Germany
{chiaburu.teodor, frank.hausser, felix.biessmann}@bht-berlin.de
² Einstein Center Digital Future, Wilhelmstraße 67, 10117 Berlin, Germany

Abstract. Evaluating the quality of explanations in Explainable Artificial Intelligence (XAI) is to this day a challenging problem, with ongoing debate in the research community. While some advocate for establishing standardized offline metrics, others emphasize the importance of human-in-the-loop (HIL) evaluation. Here we propose an experimental design to evaluate the potential of XAI in human-AI collaborative settings as well as the potential of XAI for didactics. In a user study with 1200 participants we investigate the impact of explanations on human performance on a challenging visual task - annotation of biological species in complex taxonomies. Our results demonstrate the potential of XAI in complex visual annotation tasks: users become more accurate in their annotations and demonstrate less uncertainty with AI assistance. The increase in accuracy was, however, not significantly different when users were shown the mere prediction of the model compared to when also providing an explanation. We also find negative effects of explanations: users tend to replicate the model's predictions more often when shown explanations, even when those predictions are wrong. When evaluating the didactic effects of explanations in collaborative human-AI settings, we find that users' annotations are not significantly better after performing annotation with AI assistance. This suggests that explanations in visual human-AI collaboration do not appear to induce lasting learning effects. All code and experimental data can be found in our GitHub repository: <https://github.com/TeodorChiaburu/beexplainable>.

Keywords: XAI · uncertainty · human-in-the-loop

1 Introduction

XAI strives to bridge the gap between complex AI models and human users by providing explanations for the models' decisions. However, evaluating the effectiveness of XAI methods remains a challenging question[31,20,27,3,36,30].

Currently, two primary methodological approaches dominate the XAI evaluation landscape: **fully automated evaluation metrics** and **HIL approaches**. While the former strategy offers an objective and scalable solution [18,10,19], it often lacks the richness of human understanding.

HIL experiments are supposed to provide a powerful alternative by directly assessing user interaction with explanations. While human comprehension is the ultimate goal of explainability, evaluating model explanations through human subject studies presents a significant challenge. These studies necessitate rigorous experimental design and can be resource-intensive in terms of time, money or logistics [24]. Nonetheless, due to the inherently human-centric nature of explainability, a substantial body of literature advocates for HIL experiments as a means to assess explanation quality [12,22,9,23,2,28].

Several aspects of understanding the effect of explanations are of interest and worth exploring for a given XAI method, as discussed in the following section. While allowing for analyzing some of these points, the HIL framework we propose in this work enables practitioners to also measure the didactic potential of explanations, as well as correlations between the subjects' and the machine's uncertainty. To the best of our knowledge, these aspects have yet to be investigated in the current XAI literature.

2 Related Work

The definition of a "good" explanation remains an ongoing debate, leading to a diverse range of HIL experiments focusing on various aspects and metrics. Comprehensive reviews of this field can be found in [33,15,37].

Perhaps the most extensively explored facet of XAI research concerns the influence of explanations on **human performance**. Do explanations actually help users perform better? This can be investigated through two main lenses: *human-AI collaboration* and *knowledge transfer*.

In human-AI collaboration tasks, explanations can enhance user performance by effectively communicating the AI's reasoning. This allows users to better utilize the AI's suggestions, adjust their own decisions alongside the AI's input and potentially improve overall task efficiency through smoother collaboration. HIL experiments can shed light on these aspects by observing user behavior in such team-settings. The literature abounds in studies investigating scenarios where users work with AI assistants to solve various tasks. These studies often demonstrate that AI assistance improves human performance in tasks like sentiment analysis [34], poisonous mushroom classification [26], insulin dosage decisions for virtual patients [38] or prostate cancer classification in MRI scans [14].

In terms of long-term benefits of explanations, the question is whether explanations can empower users to learn from the AI and improve their independent performance on future tasks. Effective explanations might enhance user understanding of relevant rules and patterns used by the AI. This understanding could then translate to improved task execution without AI assistance, potentially promoting long-term learning that can be utilized for various tasks beyond

the specific context of the AI system. HIL experiments can be designed to assess whether explanations facilitate such knowledge transfer, basically crystallizing a pedagogic effect of XAI. As previously stated, to the best of our knowledge, there are no studies available at the time of writing this paper, that deal with the didactics of explanations. Our proposed HIL framework attempts to cover this gap.

Still related to performance, the aspect of **simulatability** is also frequently investigated in HIL studies. Simulatability refers to whether users can understand and replicate the model’s reasoning based on the explanations provided. This topic is extensively discussed in [16] and [12], where the authors distinguish between "forward simulation" (users predict the model’s output for a given input and explanation) and "counterfactual simulation" (given an input, a model’s output and an explanation, users predict the model’s output for a perturbation of that input).

Another key factor is the effect of explanations on **trust**, namely how they calibrate user trust in the model’s predictions or how they can mitigate the issue of blind trust in situations where the model might be less confident. For instance, in [22] participants are shown the prediction of a computer vision model along with an explanation in the format "Class A because ..." and are asked how confident they are in the model’s prediction. On a different note, [5] investigate in a Turing test inspired approach whether subjects are able to distinguish between human-generated and AI-generated explanations and argue that such a quantitative metric, employed alongside trust calibration techniques, would offer valuable insights into how intuitive an explanation is. Clear, comprehensive and accurate explanations can help users assess the AI system’s competence and expertise. If explanations effectively reflect the model’s reasoning process, users are more likely to believe the AI is knowledgeable and capable of assisting with the task at hand. This fosters trust in the AI’s ability to provide accurate suggestions and recommendations. Explanations can also mitigate the risk of blind trust. By highlighting the AI’s limitations and uncertainties, explanations can encourage users to approach the AI’s suggestions with a healthy dose of skepticism. This allows them to maintain an appropriate level of critical thinking and intervene when necessary without entirely disregarding the AI’s input. Ultimately, explanations should strive to create a balance between trust and critical engagement with the AI system.

A natural question stems from the issue of trustworthiness: does **user uncertainty** align with **model uncertainty** in XAI contexts? This is particularly important for building trust in situations where the model might be less confident in its predictions [8]. Again, as far as we are aware, this aspect has yet to be investigated in HIL studies in the available literature so far.

The perceived **cognitive effort** required to understand explanations is another important consideration. Several studies highlight that explanations may not always have a solely positive impact and could even yield negative effects on human subjects during cognitive tasks [35,4,11,32]. These studies often attribute

performance declines in cognitive tasks to the increased cognitive load imposed by explanations [32].

Lastly, XAI research also considers the perceived **usefulness** and perceived **ease of use** of explanations, e.g., [29]. This focus acknowledges that explanations must not only be understandable, but also practically valuable for users. If explanations are deemed unhelpful or difficult to grasp, they are unlikely to enhance user experience or performance.

The following sections present our proposed HIL approach and the methods of investigating the didactic effect of explanations, the human-machine collaboration, the degree of trust users have in AI, as well as the relation between the users' and the model's uncertainty within this framework.

3 Dataset and Classification Problem

For the experiments described in this work, a subset of the iNaturalist dataset [1] was used. A dataset of wild bee images was constructed by scraping 30k images from the online database <https://www.inaturalist.org/>. These images depicted the top 25 most frequent wild bee species native to Germany within their natural habitats. Following preliminary experiments, the dataset was refined to focus on three particularly challenging and frequently confused species: *Andrena bicolor*/*flavipes*/*fulva* (see Fig. 1). This refinement resulted in a final subset containing 657 wild bee images. For more details on the scraping process, data split and annotation, as well as training of the model - a ResNet50 [17] - please consult [7] and our repository - <https://github.com/TeodorChiaburu/beexplainable>.



Fig. 1: Prototypical examples of the three wild bee species used for our HIL experiment. These examples were also shown to the participants in the instructions at the beginning of the trial. The difficulty in distinguishing the three species from one another consists in morphological features present on the bees' thorax and abdomen: *A. bicolor* has a fuzzy orange thorax and a shiny brown abdomen; *A. flavipes* has a fuzzy brown thorax and shiny brown abdomen; *A. fulva*'s thorax and abdomen are both fuzzy orange.

4 Experimental Design

Our experimental setup consists of three tasks - see Figure 2. The users are shown images of wild bees and are required to recognize the three species depicted in Fig. 1. In Task 1, subjects are left to assign on their own the correct label to the images they see. In Task 2, they are aided by a computer vision model trained to recognize wild bees. In the third and final task, the photos are again shown without any AI hints. Each task comprises 10 images. The reason for adding a second 'control task' (Task 3) after the AI-assisted Task 2 is to enable the investigation of the potential didactic effect that explanations may have. We hypothesize that, if explanations are able in a pedagogical sense to teach laypersons relevant classification rules, then the users' accuracy in Task 3 should be higher than in Task 1, when participants were just getting acquainted with the problem.

The 30 images that every participant sees throughout the trial are randomly drawn from a pool of 45 samples (selected from the test set). The pool is a mixture of 'easy' and 'hard' examples with respect to the model's confidence in classifying those samples. We label an image as 'hard' if the model assigned a true class probability below 80%. Inherently, some of the hard samples were misclassified by the network. When compiling the set of 30 images shown to a subject, we ensured that each task contains 5 easy samples and 5 hard ones.

The participants were informed that the data gathered would solely be used for research purposes. Their identities were anonymous and before starting they were given a detailed description of what they were required to do. For each possible class, a representative example was shown in the introduction, which they could refer back to any time during the trial. The experiment was approved by our institutional research board.

Users were divided into 6 groups that differed from one another in the type of AI hint revealed in the second task (see Figure 2):

1. *Control Group*: the AI hint consists solely of the model's predicted class (which can be wrong)
2. *Control-Confidence Group*: the model's prediction is accompanied by the corresponding Softmax probability (model confidence)
3. *Concepts-CoProNN Group*: the model's prediction and confidence are shown together with an explanation computed by the concept-based XAI method CoProNN [7]. The explanation is visualized in a 'traffic-lights' format, where a representative patch of the concepts learned by the XAI method is marked as relevant (green) or not (red).
4. *Concepts-TCAV Group*: the model's prediction and confidence are shown together with an explanation computed by the concept-based XAI method TCAV [21]. The explanation modality is the same as for CoProNN.
5. *Examples-GradSim Group*: the model's prediction and confidence are shown together with an explanation computed by the example-based method Gradient Similarity [6]. The explanation is visualized in the form of the top 3 most similar samples from the training set that were classified similarly.

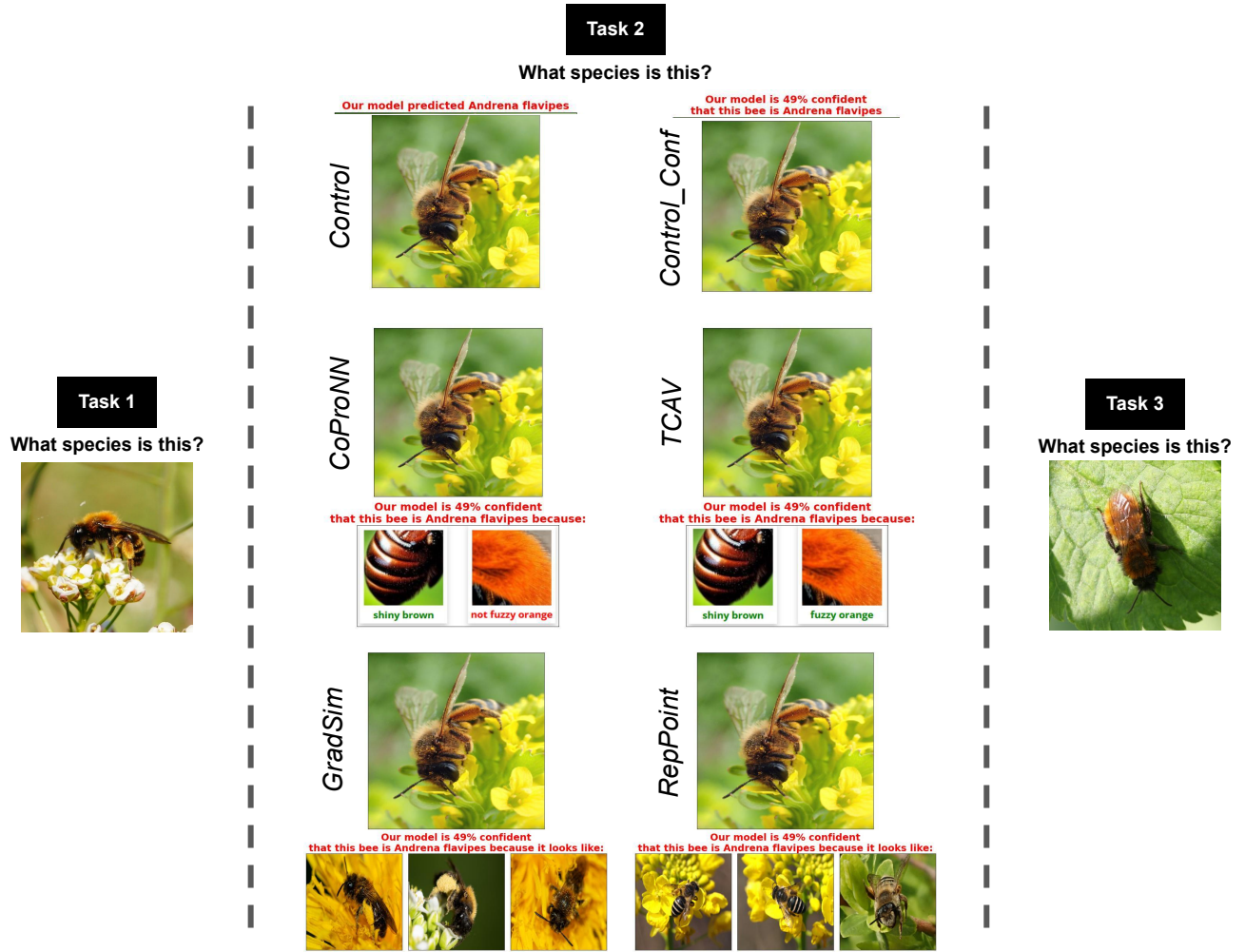


Fig. 2: Experimental Design. From left to right: Task 1 - users classify images on their own; Task 2 - further AI assistance is provided, presented differently depending on the assigned group; Task 3 - users again classify images alone.

6. *Examples-RepPoint Group*: the model’s prediction and confidence are shown together with an explanation computed by the example-based method Representer Point Selection [39]. The explanation is visualized the same way as for Gradient Similarity.

We carried out our experiment on the crowdsourcing platform Toloka (<https://toloka.ai/>). For each of the above mentioned groups, 200 users were accepted (not counting automatically discarded suspicious bot accounts that complete the whole experiment in a matter of seconds). At the end of the experiment, a separate outlier detection would be applied to every group, filtering out submissions with less than 3 correct answers in any of the three tasks. The number 3 corresponds to the 20% quantile and was determined based on a pilot study. The subjects were selected from the top 50% of English-speaking Tolokers and remunerated for their participation with 0.06\$ per task suite.

Before conducting this experiment at larger scale, we designed and performed a smaller pilot study with five wild bee species and only 80 participants divided into two groups: *Control* and *CoProNN*. The experiment was deployed as a jsPsych app [25]. A demo is still freely available online for the CoProNN group at <https://hgyl4wmb21.cognition.run>. For more details about the pilot study, please refer to [7] and our repository.

The following section discusses the results of our experiment. As a forenote regarding the quantification of the user-related and model-related metrics: we report the entropy³ of the Softmax-normalized probability vector output by our model for every image as the *model’s uncertainty* and the entropy over the Softmax-normalized vector of the users’ submitted answers for each image as the *users’ uncertainty*. We acknowledge that more sophisticated metrics exist for quantifying epistemic uncertainty within model predictions, see e.g. [13]. However, for the purposes of this study, a proxy measure as defined above was deemed sufficient.

5 Results and Discussion

We summarize below the insights we gained from our user study. Figure 3 offers a broad overview of the subjects’ performance throughout the three tasks and the six groups, while the correlation plots in Figures 4 and 5 display the relation between the model’s and the users’ performance.

³ <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.entropy.html>

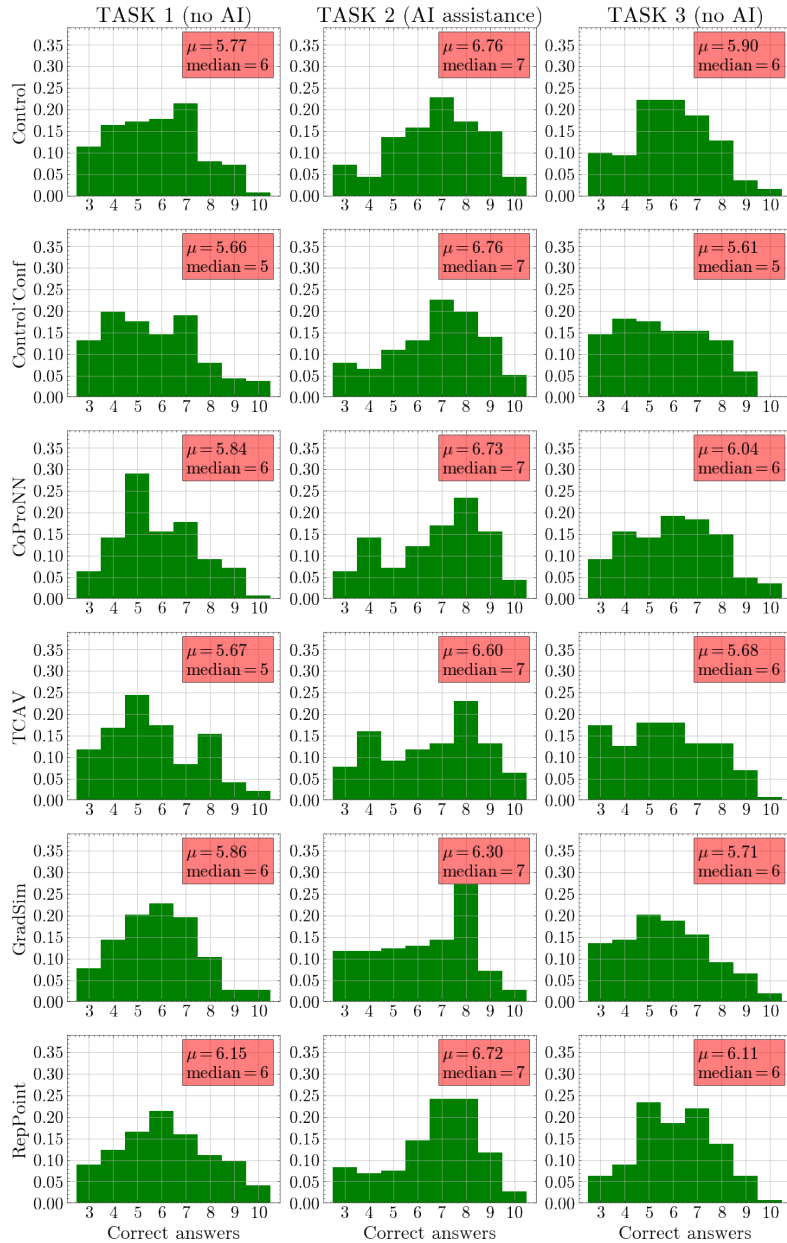


Fig. 3: AI assistance improves users’ performance, but showing only AI predictions helps as much as showing explanations. Throughout all the 6 human-AI collaboration conditions we notice higher user accuracies in Task 2 (where AI hints were provided) in the mean and median, as opposed to Tasks 1 and 3 (both without AI assistance). In Task 3 participants are not performing any better than in Task 1, suggesting that there is no substantial didactic effect derived from the assistive AI in Task 2. Moreover, there is also no noticeable difference between the two control groups and the four explanation groups. This suggests that explanations were not more effective than simply reporting the model’s prediction.

5.1 No Observable Didactic Effects Detected

Our experiment did not reveal any considerable evidence of long-term learning or knowledge transfer from explanations to independent task performance. This is indicated by the comparable user accuracy observed in Task 3 (where users classified images on their own again) compared to Task 1 (initial independent classification). Across all six groups, both average and median accuracy scores in Task 3 remained similar to those in Task 1 (Figure 3). This suggests that while explanations may improve performance during collaboration with the AI (as shown in Subsection 5.3), they may not necessarily equip users with the ability to retain that knowledge and apply it to solve similar tasks independently over time. Tiredness and cognitive load may also play a role once users arrive at the final Task 3.

5.2 Users' Uncertainty Decreases when Collaborating with an AI Assistant

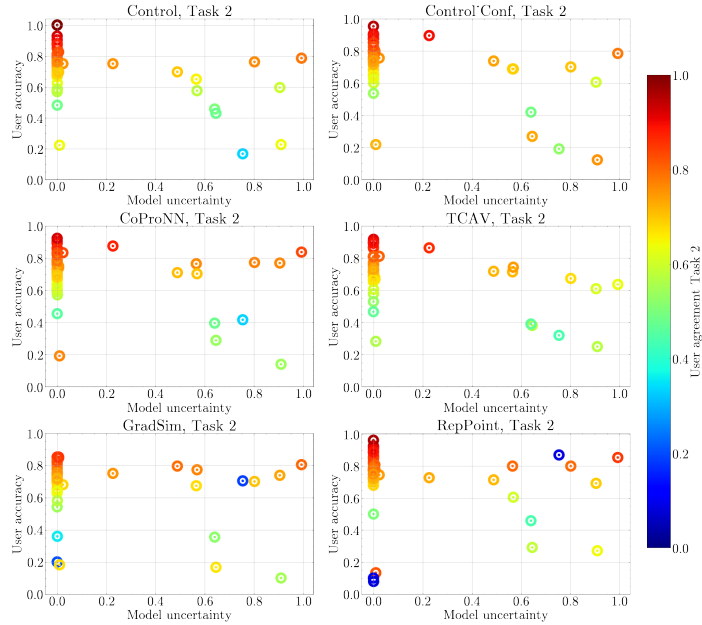
Our study also revealed a positive impact on user confidence when collaborating with an AI assistant. This is reflected in the user uncertainty levels observed across the different tasks and computed as described at the end of Section 4. User uncertainty in Task 2, where participants received help from our model, was considerably lower compared to the uncertainty levels observed in Tasks 1 and 3 (Table 1 and Figure 5). This suggests that hints provided by the AI assistant helped users feel more certain about their classifications in Task 2 and allowed them to approach the task with greater confidence.

5.3 Human Performance Improves when Collaborating with an AI Assistant

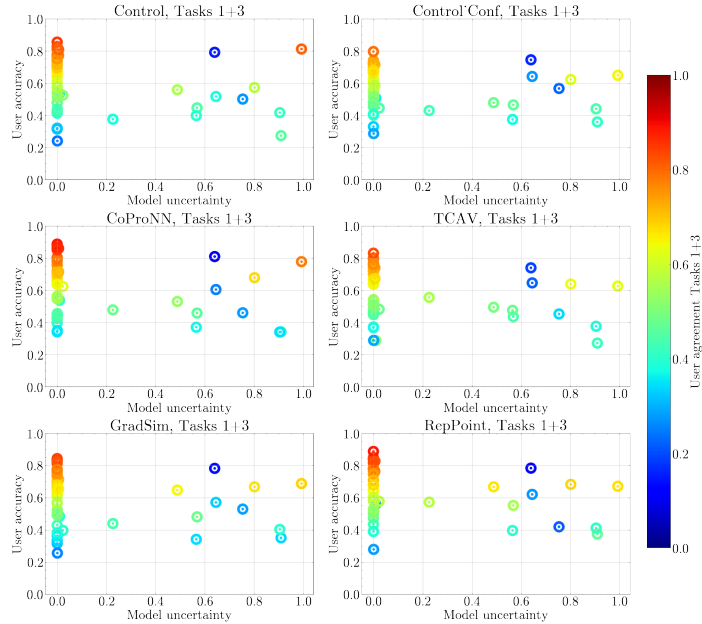
Our findings demonstrate that human performance improves when collaborating with an AI assistant. Across all user groups participating in Task 2, both average and median user accuracy scores are consistently higher compared to Tasks 1 and 3 (as illustrated in Figure 3). Overall, user performance in Task 2 is generally superior to that observed in Tasks 1 and 3, particularly for samples where the model exhibited a high degree of certainty (Figure 4). This indicates that explanations and AI assistance were most beneficial for tasks where the model was most certain, potentially aiding users in making more accurate classifications.

5.4 Limited Impact of Explanation Type on Task Performance

While Subsections 5.3 and 5.2 highlighted the overall benefits of collaboration with an AI assistant in Task 2, user performance within this task did not exhibit noticeable differences across the six groups (as depicted in Figure 3). This suggests that, in the context of our experiment, the specific format or level of detail provided in the explanations (concept-based or example-based) did not have a substantial impact on user accuracy when compared to the two Control groups that received only the model's prediction (with and without model confidence).

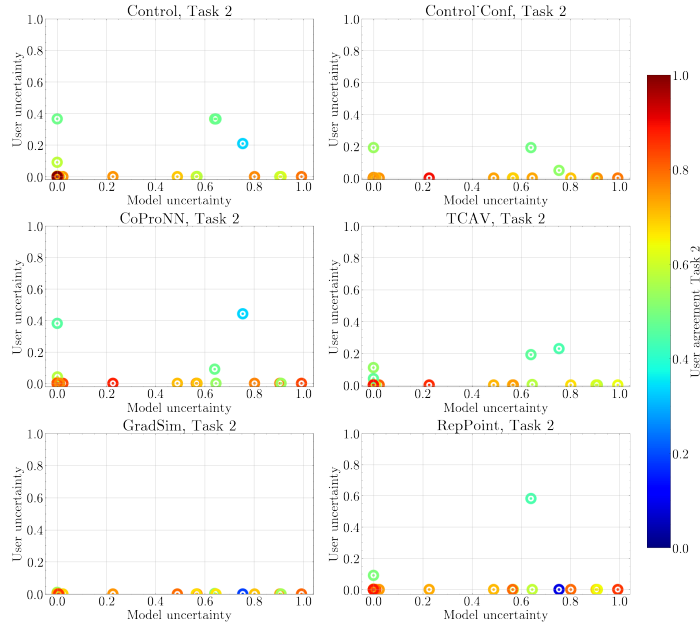


(a) Task 2 (AI Assistance)

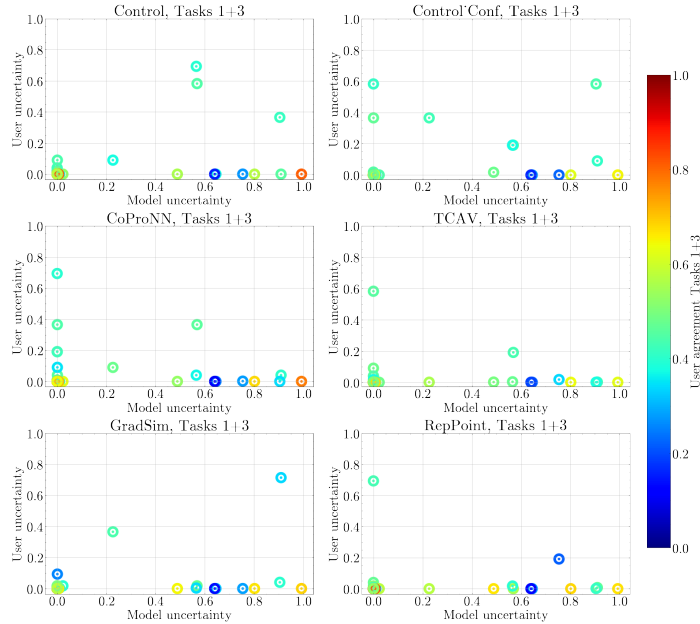


(b) Tasks 1 and 3 (no AI Assistance)

Fig. 4: Explanations improve user accuracy in human-AI collaboration (Task 2). Every dot represents one of the 45 images with corresponding user metrics averaged over all users' responses. **(a) Task 2 (with AI assistance):** Throughout all groups, samples classified with high certainty by the model were also classified with high accuracy by users. These samples are also associated with high acceptance rate for the AI suggestion. **(b) Tasks 1+3 (without AI assistance):** User accuracy is lower than in Task 2, regardless of whether the model was certain or not. The user-AI agreement rate is also notably lower.



(a) Task 2 (AI Assistance)



(b) Tasks 1 and 3 (no AI Assistance)

Fig. 5: Uncertainty of users’ responses decreases when explanations are shown (Task 2) compared to no AI assistance (Task 1+3). Every dot represents one of the 45 images with corresponding user metrics averaged over all users’ responses. **(a) Task 2 (with AI assistance):** Independent of model uncertainty, the users’ uncertainty is near 0 for most samples. User-AI agreement is also high with these samples. **(b) Tasks 1+3 (without AI assistance):** Users are less certain than without AI assistance (Task 2), regardless of whether the model was certain or not. Also the acceptance rate of the AI’s suggestions is lower.

5.5 Potential for Blind Trust with Explanations

Our study also identified a potential concern regarding the use of explanations, particularly in relation to fostering blind trust. Figures 4 and 5 use color coding to represent user agreement with the model’s suggestion (low agreement in blue, high agreement in red). These figures reveal a notable presence of "hot spots" where user agreement is high (green-yellow-orange-red) despite high model uncertainty. Table 1 also shows that, on average, users’ responses matched more often the model’s prediction in Task 2 than in the other tasks. On the one hand, following the AI’s recommendation lead to higher user accuracy scores, as discussed above. On the other hand, when zooming in only on the misclassified samples (accompanied by a matching wrong explanation in the four XAI groups), we report the following agreement rates: Control - 47.29%, Control-Confidence - 52.31%, CoProNN - 47.67%, TCAV - 45.25%, GradSim - 62.95%, RepPoint - 69.1%. This suggests that users may, in some cases, predominantly in the two example-based XAI groups, exhibit a tendency to blindly trust the model’s suggestions, even when presented with explanations for demonstrably incorrect predictions.

Table 1: Aggregated user uncertainty scores and user-AI agreement rates for control tasks (1 and 3) and Task 2. The users’ uncertainty was computed as described at the end of Section 4.

	<i> User Uncertainty </i>	<i>User-AI Agreement</i>
<i>Tasks 1+3 (no AI help)</i>	0.0359	0.5732
<i>Task 2 (with AI assistance)</i>	0.0151	0.7187

6 Conclusion

In this work, we proposed a novel HIL experiment design that allows to analyze the didactic effect of explanations, as well as correlations between the users’ and the model’s uncertainty. Apart from these new considerations, more traditional investigative points such as human-machine performance as a team or blind trust were also taken into account. We examined human-AI collaboration with explanations in image classification; nonetheless, our framework can be readily applied to any other machine learning task.

We found that explanations considerably improved user performance during collaboration, especially when the AI was certain of its prediction. User uncertainty also decreased with explanations. However, our study also identified certain limitations. Explanations did not show notable benefits for long-term knowledge transfer and the specific explanation format had little to no impact

on user accuracy. In line with previous work, our results also show that explanations tend to bias users responses to replicate the AI predictions, even when they are wrong; this finding highlights the need for well calibrated trust relationships in human-AI interactions in order to counteract blind trust in AI systems. We hope that the experimental paradigms quantifying the trust relationship in XAI developed in this study contribute to a better understanding of the trust relationship.

Overall, our findings support the potential of human-AI collaboration with explanations to enhance performance and trust. Nevertheless, further research is needed to optimize the design of explanations for knowledge transfer and mitigate blind trust. Future work should explore how XAI can empower users not just to collaborate effectively, but also develop their own problem-solving skills. By striking a balance between trust and critical thinking, we believe that explanations can pave the path for a future of successful human-machine collaboration.

References

1. inaturalist challenge dataset. https://github.com/visipedia/inat_comp/tree/master/2021, accessed: 2023-03-1
2. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity Checks for Saliency Maps. arXiv:1810.03292 [cs, stat] (Nov 2020), <http://arxiv.org/abs/1810.03292>, arXiv: 1810.03292
3. Alangari, N., El Bachir Menai, M., Mathkour, H., Almosallam, I.: Exploring evaluation methods for interpretable machine learning: A survey. *Information* **14**(8) (2023). <https://doi.org/10.3390/info14080469>, <https://www.mdpi.com/2078-2489/14/8/469>
4. Alufaisan, Y., Marusich, L.R., Bakdash, J.Z., Zhou, Y., Kantarcioglu, M.: Does explainable artificial intelligence improve human decision-making? (2020)
5. Biessmann, F., Treu, V.: A Turing Test for Transparency. arXiv:2106.11394 [cs] (Jun 2021), <http://arxiv.org/abs/2106.11394>, arXiv: 2106.11394
6. Charpiat, G., Girard, N., Felardos, L., Tarabalka, Y.: Input similarity from the neural network perspective (2021)
7. Chiaburu, T., Haußer, F., Bießmann, F.: Copronn: Concept-based prototypical nearest neighbors for explaining vision models (2024)
8. Chiaburu, T., Haußer, F., Bießmann, F.: Uncertainty in xai: Human perception and modeling approaches. *Machine Learning and Knowledge Extraction* **6**(2), 1170–1192 (2024). <https://doi.org/10.3390/make6020055>, <https://www.mdpi.com/2504-4990/6/2/55>
9. Colin, J., Fel, T., Cadène, R., Serre, T.: What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods (2022), http://papers.nips.cc/paper_files/paper/2022/hash/13113e938f2957891c0c5e8df811dd01-Abstract-Conference.html
10. Cugny, R., Aligon, J., Chevalier, M., Roman Jimenez, G., Teste, O.: Autoxai: A framework to automatically select the most adapted xai solution. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. p. 315–324. CIKM '22, Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3511808.3557247>, <https://doi.org/10.1145/3511808.3557247>

11. David, D.B., Resheff, Y.S., Tron, T.: Explainable ai and adoption of financial algorithmic advisors: an experimental study (2021)
12. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning. Tech. Rep. arXiv:1702.08608, arXiv (Mar 2017). <https://doi.org/10.48550/arXiv.1702.08608>, <http://arxiv.org/abs/1702.08608>, arXiv:1702.08608 [cs, stat] type: article
13. Fellaji, M., Pennerath, F., Conan-Guez, B., Couceiro, M.: On the calibration of epistemic uncertainty: Principles, paradoxes and conflictual loss (2024), <https://arxiv.org/abs/2407.12211>
14. Hamm, C.A., Baumgärtner, G.L., Biessmann, F., Beetz, N.L., Hartenstein, A., Savic, L.J., Froböse, K., Dräger, F., Schallenberg, S., Rudolph, M., Baur, A.D.J., Hamm, B., Haas, M., Hofbauer, S., Cash, H., Penzkofer, T.: Interactive explainable deep learning model informs prostate cancer diagnosis at mri. *Radiology* **307**(4), e222276 (2023). <https://doi.org/10.1148/radiol.222276>, <https://doi.org/10.1148/radiol.222276>, PMID: 37039688
15. Hamm, P., Klesel, M., Coberger, P., Wittmann, H.F.: Explanation matters: An experimental study on explainable AI. *Electronic Markets* **33**(1), 1–21 (December 2023). <https://doi.org/10.1007/s12525-023-00640->, https://ideas.repec.org/a/spr/elmark/v33y2023i1d10.1007_s12525-023-00640-9.html
16. Hase, P., Bansal, M.: Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? (2020)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). <https://doi.org/10.48550/ARXIV.1512.03385>, <https://arxiv.org/abs/1512.03385>
18. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.M.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research* **24**(34), 1–11 (2023), <http://jmlr.org/papers/v24/22-0142.html>
19. Herman, B.: The promise and peril of human evaluation for model interpretability (2019)
20. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable ai: Challenges and prospects (2019)
21. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav) (2018)
22. Kim, S.S.Y., Meister, N., Ramaswamy, V.V., Fong, R., Russakovsky, O.: HIVE: Evaluating the human interpretability of visual explanations. In: *European Conference on Computer Vision (ECCV)* (2022)
23. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (Un)reliability of saliency methods. arXiv:1711.00867 [cs, stat] (Nov 2017), <http://arxiv.org/abs/1711.00867>, arXiv: 1711.00867
24. Leavitt, M.L., Morcos, A.: Towards falsifiable interpretability research (2020)
25. de Leeuw, J.: jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods* (2015). <https://doi.org/10.3758/s13428-014-0458-y>
26. Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., Mara, M.: Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* **139**, 107539 (2023). <https://doi.org/https://doi.org/10.1016/j.chb.2022.107539>, <https://www.sciencedirect.com/science/article/pii/S0747563222003594>

27. Ma, J., Lai, V., Zhang, Y., Chen, C., Hamilton, P., Ljubenkov, D., Lakkaraju, H., Tan, C.: Openhexai: An open-source framework for human-centered evaluation of explainable machine learning (2024)
28. Mei, A., Saxon, M., Chang, S., Lipton, Z.C., Wang, W.Y.: Users are the north star for ai transparency (2023)
29. Meske, C., Bunde, E.: Design principles for user interfaces in ai-based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers* **25**(2), 743–773 (mar 2022). <https://doi.org/10.1007/s10796-021-10234-5>, <https://doi.org/10.1007/s10796-021-10234-5>
30. Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems (2020)
31. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlöterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.* **55**(13s) (jul 2023). <https://doi.org/10.1145/3583558>, <https://doi.org/10.1145/3583558>
32. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.: Manipulating and measuring model interpretability (2018)
33. Rong, Y., Leemann, T., Nguyen, T.T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., Kasneci, E.: Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(4), 2104–2122 (2024). <https://doi.org/10.1109/TPAMI.2023.3331846>
34. Schmidt, P., Biessmann, F.: Quantifying interpretability and trust in machine learning systems. In: *AAAI-19 workshop on network interpretability for deep learning* (2019)
35. Schmidt, P., Biessmann, F., Teubner, T.: Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* **29**(4), 260–278 (2020), publisher: Taylor & Francis
36. Schuff, H., Adel, H., Qi, P., Vu, N.T.: Challenges in explanation quality evaluation (2022), <https://api.semanticscholar.org/CorpusID:257427664>
37. Schwalbe, G., Finzel, B.: A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* (Jan 2023). <https://doi.org/10.1007/s10618-022-00867-8>, <http://dx.doi.org/10.1007/s10618-022-00867-8>
38. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence* **291**, 103404 (2021). <https://doi.org/https://doi.org/10.1016/j.artint.2020.103404>, <https://www.sciencedirect.com/science/article/pii/S0004370220301533>
39. Yeh, C.K., Kim, J.S., Yen, I.E.H., Ravikumar, P.: Representer point selection for explaining deep neural networks (2018)