Reassuring, Misleading, Debunking: Comparing Effects of XAI Methods on Human Decisions

CHRISTINA HUMER and ANDREAS HINTERREITER, Johannes Kepler University Linz, Linz, Austria
BENEDIKT LEICHTMANN, Johannes Kepler University Linz, Linz, Austria and Ludwig-Maximilians-Universität München, Munich, Germany
MARTINA MARA and MARC STREIT, Johannes Kepler University Linz, Linz, Austria

Trust calibration is essential in AI-assisted decision-making. If human users understand the rationale on which an AI model has made a prediction, they can decide whether they consider this prediction reasonable. Especially in high-risk tasks such as mushroom hunting (where a wrong decision may be fatal), it is important that users make correct choices to trust or overrule the AI. Various explainable AI (XAI) methods are currently being discussed as potentially useful for facilitating understanding and subsequently calibrating user trust. So far, however, it remains unclear which approaches are most effective. In this article, the effects of XAI methods on human AI-assisted decision-making in the high-risk task of mushroom picking were tested. For that endeavor, the effects of (i) Grad-CAM attributions, (ii) nearest-neighbor examples, and (iii) network-dissection concepts were compared in a between-subjects experiment with $N = 501$ participants representing end-users of the system. In general, nearest-neighbor examples improved decision correctness the most. However, varying effects for different task items became apparent. All explanations seemed to be particularly effective when they revealed reasons to (i) doubt a specific AI classification when the AI was wrong and (ii) trust a specific AI classification when the AI was correct. Our results suggest that well-established methods, such as Grad-CAM attribution maps, might not be as beneficial to end users as expected and that XAI techniques for use in real-world scenarios must be chosen carefully.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence*;

Additional Key Words and Phrases: Explainable artificial intelligence, trust calibration, AI-assisted decision-making, mushroom identification

This work was funded by Johannes Kepler University Linz, Linz Institute of Technology (LIT), the State of Upper Austria, and the Federal Ministry of Education, Science and Research under grant number LIT-2019-7-SEE-117, awarded to MM and MS, the Austrian Science Fund under grant number FWF DFH 23–N, and under the Human-Interpretable Machine Learning project (funded by the State of Upper Austria).
Authors' Contact Information: Christina Humer (Corrsponding author), Johannes Kepler University Linz, Linz, Austria; e-mail: christina.humer@jku.at; Andreas Hinterreiter, Johannes Kepler University Linz, Linz, Austria; e-mail: andreas.hinterreiter@jku.at; Benedikt Leichtmann, Johannes Kepler University Linz, Linz, Austria and Ludwig-Maximilians-Universität München, Munich, Germany; e-mail: benedikt.leichtmann@psy.lmu.de; Martina Mara, Johannes Kepler University Linz, Linz, Austria; e-mail: martina.mara@jku.at; Marc Streit (Corresponding author), Johannes Kepler University Linz, Linz, Austria; e-mail: marc.streit@jku.at.

ACM 2160-6463/2024/8-ART16
https://doi.org/10.1145/3665647

ACM Reference format:
Christina Humer, Andreas Hinterreiter, Benedikt Leichtmann, Martina Mara, and Marc Streit. 2024. Reassuring, Misleading, Debunking: Comparing Effects of XAI Methods on Human Decisions. *ACM Trans. Interact. Intell. Syst.* 14, 3, Article 16 (August 2024), 36 pages.
https://doi.org/10.1145/3665647

1 Introduction

Using **artificial intelligence (AI)** systems as the sole decision makers is undesirable in many contexts. While AI models outperform humans at specific tasks [36], these systems can also be fallible, for example, if the model receives unprecedented input [71]. In high-stakes domains, where wrong decisions potentially lead to harmful outcomes the synergy between a highly accurate AI model and a human providing logical insights is ideal for decision-making tasks [52, 80, 84]. For example, an incorrect medical diagnosis could lead to wrong medication treatment [43, 60], and incorrect identification of mushrooms could lead to poisoning symptoms [10]. Maximizing the performance of human–AI teams in decision-making requires proper calibration of user trust in and appropriate reliance on the AI system. However, to improve trust calibration, end-users need to understand a model's decision such that they can reject an AI's prediction if it is wrong (avoiding overtrust), or accept it if it is correct (avoiding distrust). Being able to comprehend system decisions better is particularly important for users with only limited domain knowledge (e.g., about poisonous mushrooms), as they may otherwise have difficulty recognizing a wrong decision. This "appropriate reliance" [see 75] can be achieved by (i) explaining the model's decision using **explainable AI (XAI)** methods [61, 65], or (ii) explaining how a system works in general such that users understand its abilities and limitations [61, 68]. While empirical research on the effect of explanations on user behavior has been growing rapidly, there remain gaps that have not yet been studied in depth.

First, the literature about AI-assisted decision-making considers only some of the available explanatory methods and design options [74]. In particular, little to no research has been conducted on image data, concept-based explanation techniques, or combinations of complementary explanation methods [74].

Second, existing studies usually focus on evaluating explanation techniques based on performance measures that aggregate data over larger sets of task items. While such aggregate measures lead to general insights, they often conceal the fact that the effects of explanations can vary strongly between individual items. For many existing studies, it is thus difficult to draw conclusions about how well methods can help avoid overtrust or distrust. Explanations of an AI or its predictions are not always helpful; they can also mislead users into trusting an untrustworthy model or model prediction [53] if they appear to be plausible [44]. A generally useful explanation method may sometimes produce *misleading* explanations for individual inputs [53], which can cause users to trust predictions that are actually incorrect (or distrust predictions that are actually correct). To better understand when individual explanations are actually *helpful* [28], various explanation methods need to be studied and compared on an item level.

In this article, we present new empirical results to fill these research gaps. We studied the decision behavior of a diverse group of 501 end-users. The core contributions of this study are two-fold:

— We compare the effects of three different explanation methods on decision correctness and user assessment in a collaborative human–AI decision-making scenario. We use an image-based context and include explanations from the previously under-explored class of concept-based explanations.

—We investigate why certain explanations lead to particularly good or bad decision-making performance by studying individual task items. We perform detailed analyses for items with high variance and for subsets of misleading and helpful explanations. Furthermore, we assess which explanation methods are useful for avoiding overtrust and distrust.

For the study of such effects of explanations, the context of application is crucial. Results from literature are mixed and vary strongly across the different contexts used, such as number recognition [47], nutrition [12], bargaining situations [24, 25], and many more [74, 75]. The context of the application must therefore be carefully selected and clearly described. In our previous research [57, 58], we introduced the use case of mushroom identification in the context of human AI-assisted decision-making. Mushroom identification is a suitable use case for trust research because it involves a potentially risky decision while being relatable to many participants. Since user knowledge of specific mushrooms is often limited, an AI can assist with decision-making. However, a wrong decision by the AI or the user could lead to harm.

We used mushroom hunting again as the context of this study. Results from our previous studies showed that XAI can indeed aid humans in AI-guided decision-making, but it remained unclear, which explanations were beneficial and why. The results of this study show that not all explanation methods have a clear positive influence on human–AI collaborative decision-making. The methods we studied differ in their ability to help users avoid overtrust and distrust. In particular, the results of our study suggest that nearest-neighbor examples helped participants to avoid overtrust, while network-dissection explanations seemed to be better at avoiding distrust. We also found that methods differ strongly in terms of when explanations were misleading or helpful, with helpful explanations having a higher positive influence on decision performance. By analyzing the strengths and weaknesses of various explanation methods we hope to help researchers make informed decisions about selecting methods or complementary combinations of methods for future studies and applications.

2 Related Work

2.1 Appropriate Reliance and Trust Calibration

Appropriate user reliance in the AI system is one of the key factors that influence joint performance in an AI-assisted decision-making task [39, 75]. Such a reliance is closely related to the concept of user trust. While reliance is defined as the actual behavior of following an AI's advice [75], trust has been defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [56, p. 51]. An individual's expression of this construct changes dynamically during the interaction with the automated system [38, 50, 56, 59]. According to psychological models, appropriate reliance and trust levels are affected by how well a user *understands* the automated systems [38, 51, 56].

One strategy for improving user understanding of AI systems is XAI. Machine learning developers have created a large number of explanation techniques for various types of models [2, 4, 14, 66], and the effects of XAI on user understanding have been the subject of several user studies in the AI literature [3, 9, 13, 46, 48, 70, 81, 82]. Similarly, the effects of XAI on perceived trust were studied previously [42, 55, 57, 58]. However, despite efforts to create benchmarks for objectively evaluating XAI techniques [1, 18, 40, 83], understanding how exactly XAI affects end-user trust and behavior in human–AI interaction has remained a challenge [14, 22, 23, 65]. This results in contradicting results for whether explanations are beneficial [31, 57, 58, 84] for trust calibration or not [15, 81, 82, 85].

Factors that contribute to this challenging situation are the many possible mediators (e.g., self-confidence [75]), confounding variables (e.g., cognitive errors [35]), and moderators

(e.g., error rates [47]), or the wide variety of contexts in which XAI can be studied and the variety of XAI techniques that can be evaluated. For example, a recent meta-analysis by Schemmer et al. [74] found that most studies so far used text or tabular data in their experiments. Image data had been less commonly used in studies despite being the subject of many AI tasks and being easily understandable by humans, making the task particularly interesting to explore. Explanation methods previously studied mainly belong to the categories of attribution-based [3, 48, 57, 58, 82, 85] and/or example-based [12, 13, 48, 57, 58, 81, 82] methods, but little research was done on concept-based methods [48] or the influence of misleading explanations [53]. Furthermore, the tasks vary strongly with regard to the level of vulnerability that is involved. As previously mentioned, user *uncertainty* and *vulnerability* are prerequisites for trust formation [33]. The risk of negative consequences due to wrong AI recommendations, for example, is relatively low in age prediction from images [15] or in the prediction of forest cover types [82]—contexts in which the effects of XAI have recently been studied. Likewise, for income prediction, Zhang et al. found no influence of local explanations on trust calibration and stated that future studies should use "scenarios with more significant real-world impact" [85, p. 304]. [52] studied the impact of explanations on the precision in human–AI collaboration for identifying social media posts as "toxic"; this might be a more relatable task. A recent study that tested the effects of explanations for assessing damage from satellite imagery entails more vulnerability [67]. In our previous studies [57, 58], we found that mushroom hunting is a promising use case for studying XAI because it is relatable and involves a high degree of vulnerability. Studies that implement high-risk use cases are difficult to execute because participants must never be placed in a potentially dangerous situation. This limits such studies to being set in virtual environments leading to concerns about real-world validity and data quality. Authors of such studies must therefore carefully design their studies in a way that conveys the importance of the task.

We chose the use case of mushroom hunting as a relatable and risky task and implemented it as an image classification task. This means that (i) participant decisions depended on real-world images rather than on abstract numbers, and (ii) we were able to make use of established explanation techniques for convolutional neural networks (CNNs) [20, 66]. In contrast to other studies, we used predictions and explanations that come from an AI that was trained on the mushroom classification task instead of using a fictional AI with handcrafted explanations [48, 55, 67, 81], making the task more faithful to real-world usage of XAI.

2.2 Explainability Methods

There are many ways to categorize XAI methods, for example, into global and local methods or model-agnostic and model-specific methods [20, 30, 66]. In this study, we distinguish XAI methods by the type of explanation that the technique produces. We identified three categories that apply to the image classification task used in this study: (i) *attribution-based* methods; (ii) *example-based* methods; and (iii) *concept-based* methods.

Attribution-based methods highlight regions in the input that were important for a classification. Gradient-based attributions, like saliency maps [77] or **Gradient-weighted Class Activation Mapping (Grad-CAM)** [76], compute the gradient of an output with regard to an input or intermediate layer. The calculated values are then used as an attribution map for the input image. Techniques like Local Interpretable Model-agnostic Explanations [71] or SHapley Additive exPlanations [62], on the other hand, extract super-pixels (i.e., areas of similar pixels) from an input image and compute the importance of each super-pixel to a prediction. In any case, these importance (or attribution) mappings are then visualized as heatmaps and overlaid with input images to highlight important areas in the input. *Example-based* methods extract or generate exemplary data items. For example, prototypes [8] are representative instances of a set of data, and are often used to

describe a cluster or sub-classes of data items. Nearest-neighbor examples [13, 42] are instances of a dataset that are closest to a target instance. *Concept-based* methods extract human-understandable concepts, often in textual form. Concept bottleneck models [49], for example, are neural networks that first map the input to concepts and then the concepts to the final output. The concepts from the intermediate layer can be used to explain the model decisions. Network dissection [5, 6]) is a method that assigns concepts to units of an already trained CNN. The activation maps of each unit are correlated with segmentation maps of pre-defined concepts. Based on these correlation statistics, certain units can be identified as "detectors" for specific scenes, colors, and/or textures.

2.3 Human Decision-Making in the Realm of Mushroom Hunting

The study presented in this article is closely related to two previous studies in which we investigated the effects of XAI on human decision-making and trust in the domain of mushroom hunting [57, 58].

Our pilot study [58] had an exploratory nature. During a virtual mushroom hunt participants had to pick edible and avoid poisonous mushrooms using an AI-based app for guidance. The study assessed the impact of XAI methods (with vs. without additional explanations of the AI's predictions) and an educational intervention (with and without extra information about how the AI functioned) on participants' decision-making behavior. The results showed that receiving explanations led to better performance and more appropriately calibrated trust levels. Surprisingly, the educational intervention, domain-specific knowledge (about mushrooms), and AI knowledge had no significant effect on performance. Finally, participants who received explanations trusted the AI classification less and reported a lower understanding of it.

In our second study [57], we tested the robustness of the results from the exploratory study in a more immersive environment. Due to a priori information from our pilot study, we were able to test specific hypotheses with stricter statistical error control. We conducted an in-person study at an Austrian media and art festival, where we set up an artificial forest. We developed a game for tablet computers, in which we implemented the mushroom-hunting scenario as an interactive task. In this conceptual replication of our previous online study, we confirmed that participants who received explanations made more correct decisions. This confirms that results from online studies can, at least partly, be generalized to setups that are closer to real-life scenarios. The effects of XAI on human self-reported trust could not be replicated in this setup.

With our two prior studies [57, 58], we introduced the task of mushroom hunting and its usefulness for testing the effects of XAI on human–AI collaboration and showed that XAI, in general, helped participants in making more correct decisions. However, these studies did not investigate which XAI methods were particularly useful for participants to verify the quality of an AI prediction. In the previous studies, we only compared the effects of a combined XAI interface (Grad-CAM explanations plus nearest-neighbor examples) to an interface with no explanation at all. This means that we could not investigate *which* of the two methods might have been more useful for end-users to make their decision and whether the information the explanations provided was complementary or redundant. Furthermore, although we found that participant performance varied for different subsets of items, we were unable to investigate *why* this was the case, and which characteristics of an explanation might have caused a larger or smaller effect. Finally, the influence of explanations on participants' self-reported trust—which was not replicated in the second study—remained uncertain after these two studies.

3 Research Questions

With the insights gained from the XAI literature, human factors, and psychology, as well as our previous studies in particular [57, 58] we formed new research questions that we aim to answer (see the preregistration of this study [41]).

- **(RQ1) Which explanation methods improve human AI-assisted decision-making?** We investigate the influence of each method from our previous studies individually (i.e., Grad-CAM and nearest-neighbor examples) on participants' decision-making. Additionally, we investigate the effects of another method from the previously untouched category of concept-based explanations. Based on our previous results, we assume that participants who receive any of the three explanations will identify more mushrooms correctly compared to participants who do not receive any explanations.
- **(RQ2) Why do certain explanations lead to higher variances of human decision-making?** We analyze individual task items where explanations lead to large variances in human decision-making to better understand the characteristics of the explanations that lead to these variances.
- **(RQ3) Which explanation methods help participants to avoid overtrust or distrust?** We again compare the decision-making performance of each explanation method. However, we use the subset of task items that were correctly classified by the AI as a measure of distrust and the subset of task items that were wrongly classified by the AI as a measure of overtrust.
- **(RQ4) How do misleading or helpful explanations influence decision performance?** We investigate the effects of misleading explanations on decision performance. In previous studies, we speculated that some explanations are more misleading and some are more helpful to users, but we have not tested this in detail. Misleading explanations should lead to a worse decision performance because participants are persuaded to follow a prediction that is not trustworthy. Helpful explanations, on the other hand, should result in improved decision-making performance because they aid participants in seeing the faithfulness of a prediction.
- **(RQ5) How large are the effects of explanation methods on participants' assessment of the system?** In our previous study [58], participants who received explanations also showed more appropriate self-reported trust levels with regards to the AI system (i.e., less trust in a poorly accurate system). However, in our follow-up study [57] this effect was not replicated. We therefore test the robustness of this effect for the three explanation methods used in this study. Based on the results of our first study we assume that participants without explanations will (i) report higher trust in the system, (i) evaluate the system more positively, and (iii) report higher intentions to use the system in future compared to participants with one of the explanation methods. Additionally, we analyze the influence of explanation methods on the correlation between participants' propensity to trust in automated systems in general and their self-reported trust in the AI-based system shown in the mushroom-picking task. This will shed light on how explanations moderate the effects of propensity to trust on trust ratings.

4 Methods

In a between-subjects online experiment, we recruited 501 participants who were randomly assigned to one of four groups. Participants were asked to solve a number of task items, where they had to identify mushrooms as poisonous or edible[1] with the aid of an AI-based mushroom identification app. Depending on the group, participants received the AI classification result only, or one of three

[1]Note that we do not distinguish between poisonous and inedible mushrooms in this study.

explanations. The following sections give details about the materials we used in the study, as well as the study setup and how we analyzed the data we gathered from this study.

Our study complied with the tenets of the Declaration of Helsinki, the ethical guidelines of the APA Code of Conduct, European data protection regulations, local legislation, and institutional requirements. Before participation, informed consent was obtained from every participant. Participation in our study was voluntary and could have been terminated at any point without consequences. The obtained study data were completely anonymous. To adhere to values of Open Science and transparency, we conducted a preregistration of our experiment [41], available at https://doi.org/10.17605/OSF.IO/SD953 (including a priori hypotheses, design plan, sampling plan, variables, and analysis plan). We also provide the materials to reproduce this study at https://osf.io/bq85c/.

We did not put participants in a real-world mushroom hunting scenario due to ethical reasons, which might have lowered the quality of the collected data. However, the task of mushroom hunting is, at its core, more risky than tasks proposed in related work. Furthermore, we described the task in a way that participants would be able to imagine themselves to be on an actual mushroom hunt. The relatability of the task and the gamified setting in which the task was implemented made the importance of the decision clear to participants.

4.1 Materials

4.1.1 Dataset and AI. We reused the mushroom classification model from our previous study [58]. The model is a pre-trained ResNet50 [37] CNN that was fine-tuned on a mushroom dataset [78] to assign an image of a mushroom to one of 18 species with a test accuracy of 71 %. To measure the model confidence in a prediction, we calculate certainty estimates for the prediction with Monte Carlo dropout [29]. This measure can be interpreted as the likeliness that the classification is correct.

We also created a ground truth dataset that maps each mushroom species to the binary value of edible or poisonous as this is needed for the user tasks.

4.1.2 Explanation Methods. In this study, we compared three XAI methods: (E1) Grad-CAM [76] as an example for attribution-based methods that is widely used and proven to pass so-called "sanity checks" [1]; (E2) nearest-neighbor as an example-based technique that was found to be a preferred explanation method for image classification tasks [42]; and (E3) explanations based on network-dissection [5, 6] as a hybrid concept- and attribution-based method that highlights important parts and indicates why they were important.

We reused the Grad-CAM and nearest-neighbor explanations from our previous study [58] so that the results for the individual explanation methods are comparable to the results from the combined interface of the previous studies. Grad-CAM attributions were computed by calculating the gradient of the highest predicted class with regard to the last convolutional layer [76]. The resulting attribution map was then upscaled to match the resolution of the input image. The nearest-neighbor examples are those images from the training dataset with the smallest Euclidean distance from the input image, measured in the latent space of the model [42].

To compute network-dissection explanations, we determined for each image the three units in the final convolutional layer that had the largest integrated gradient maps with respect to the top-ranked species. To this end, we calculated all gradient maps of the final convolutional layer with respect to the output of the highest predicted class. We averaged over the values of each gradient map and picked the three units with the highest value (i.e., the most influence on the predicted output class). We then extracted the top five representative images from the training dataset for each of these units (i.e., the five images for which this unit had the largest integrated

activation map). In each of the five images per unit, we extracted the regions for which the unit activation was in the 95th percentile. These regions revealed that the three units typically detected similar concepts, such as the gills of a specific mushroom species. In contrast to Bau et al. [6], we did not have a segmentation network for the concepts needed in our task (i.e., identify different parts of mushrooms) to automatically assign concepts to each unit. For the sake of this study, we therefore manually created a textual label for the shared concept of the three units (see an example for this in Figure 11 in the Appendix). In the future, segmentation networks can be used to automate this process and make it scaleable to larger tasks[2]. As an explanation for our study, we extracted a textual label previously created for the shared unit concept. We also computed an attribution map by averaging the areas with the strongest activation of the three most active units (again using the 95th percentile as a cutoff). The code to produce network-dissection explanations can be downloaded at https://osf.io/bq85c/.

We would like to note that the representation and conceptualization of the explanations (that is the concrete design of abstract concepts) of course affect study results. In our study, we aimed to ensure that participants are able to determine whether or not to trust a model prediction using the corresponding explanation. The focus is thus a user perspective. We conceptualized the explanations in a way that we deemed to be most comprehensible to study participants. In previous studies, we learned that some explanation methods are easier to understand than others for laypersons.

The core concept of the nearest-neighbor (E2) approach is to show users different training examples of what the model perceives as most similar to the target image. In other words, users can determine whether or not to trust a model prediction by comparing whether the example images align with the prediction and target image.

In contrast to the nearest-neighbor examples, we only provide one image for the Grad-CAM and network-dissection approaches instead of a Grad-CAM or network-dissection image for each class. This is because the explanation is situated within the target image and there is only one target image—not one per predicted class. An alternative conceptualization would be to calculate Grad-CAM and network-dissection explanations with regard to each of the top predicted classes. However, we decided not to do this (i) because we aimed to test the core concepts of these approaches and (ii) due to the higher complexity for the participants this would involve. Regarding the latter argument, participants would need to focus on three different explanations, where each explanation itself contains a lot of information (including the core concept that is conveyed with the explanation) and is hard to understand, leading to a cognitive overload.

The core concept of Grad-CAM is to show users whether a model "looks" at regions in the input image that make sense to be important for classification. In other words, users can determine whether or not to trust a model prediction by looking at an attribution map and deciding whether or not this makes sense with regard to the prediction (e.g., in most cases, users should not trust a prediction if the attribution map focuses on the background). Similarly, the core concept of network-dissection is to show users the important regions for a prediction, but also give them a textual description of these regions. In other words, users can determine whether or not to trust a model prediction by looking at the important region and by checking whether the textual description aligns with the important region and the prediction.

4.1.3 Forestly App. Similarly to our previous studies [57, 58], we used a mock-up mushroom identification app—*Forestly*—to present the AI classifications and explanations. While the mock-up app screen was a static image of what such an app could look like, the model classifications and explanations were actual results of a trained neural network. In its base layout, the app showed (i)

[2]Note that training such a segmentation model would have been too much overhead for this study. An additional set of pixel-level masked training data would have been necessary to train such a model.

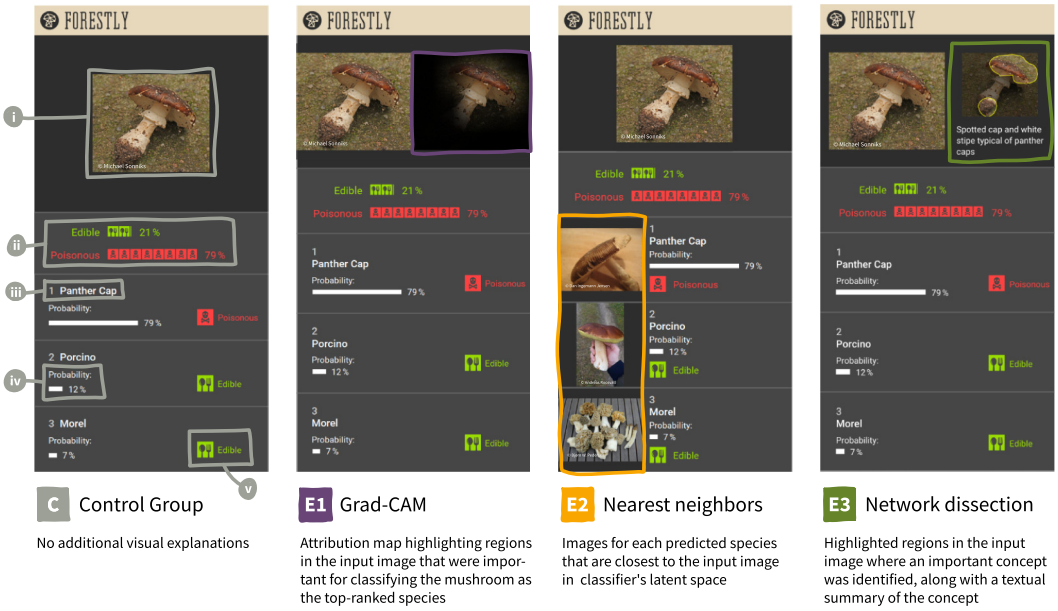


Fig. 1. The *Forestly* mock-up app showing the (i) mushroom image, (ii) total edibility prediction based on species, (iii) rank and name of predicted species, (iv) probability of predicted species, and (v) edibility of predicted species. Each variant of the *Forestly* app shows additional information corresponding to the explanation method. Note: Translated from the German original version.

the image of a mushroom (i.e., the input image), (ii) the cumulative estimated certainty that the mushroom was edible or poisonous based on the species predicted, (iii) the rank and name of the predicted species, (iv) the certainty estimation of the top three classes, and (v) the edibility for each of the top three predicted classes (i.e., whether the species is edible or poisonous) (see Figure 1 (C)).

We created three modifications of the base *Forestly* layout, one for each explanation method we want to test. For E1 [76] attributions, we show a second image, where the original image was overlaid with a heatmap that represents the attribution scores for each region of the image (see Figure 1 (E1)). In the nearest-neighbor version, we show the nearest-neighbor image for each of the top three predicted classes as an example (see Figure 1 (E2)). For the network-dissection (E3) explanation, we showed the textual label of the concept computed for this image. Additionally, we showed a version of the input image, where we highlighted the region to which the concept applies (see Figure 1 (E3)).

4.2 Study Setup

4.2.1 User Task. Participants of our study were asked to complete a set of 10 task items in random order. For each item, an image of a mushroom was shown together with the *Forestly* app screen. Participants then had to answer two binary questions: (i) would they classify the mushroom as poisonous or edible (i.e., a measure of a participant's best guess on the correct decision; see Figure 9 in the Appendix); and (ii) would they pick the mushroom for later consumption or leave it (i.e., a participant's actual decision that also entails the risk of the task; see Figure 2) [57]. The answers to the two questions may vary because even if participants believe that a mushroom is edible they might not want to take the risk of consuming it. While the decision if a mushroom is edible or poisonous involves knowledge and trust, the item whether to consume or leave it might

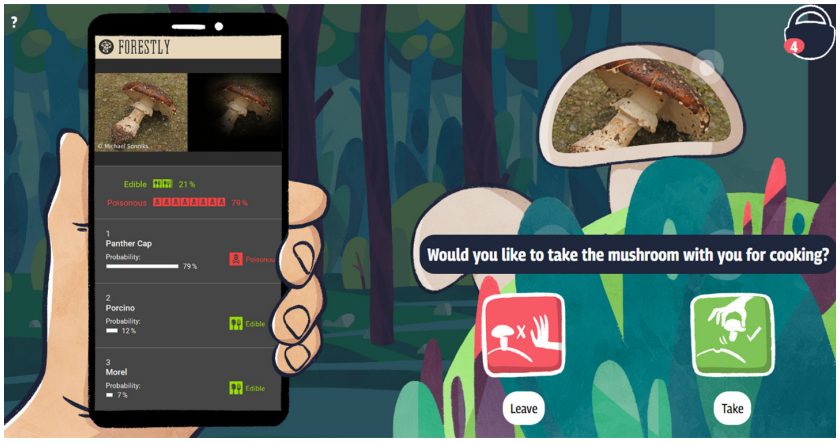


Fig. 2. Example of *AI Forest—The Schwammerl Hunting Game* and the *Forestly* app with a Grad-CAM (E1) explanation. Participants must decide whether they want to pick the mushroom or leave it. Note: Translated from the German original version.

also reflect inter-individual personality differences such as risk aversion. On the other hand, it would not make sense if participants decide to pick a mushroom although they think it is poisonous. Our previous studies showed different effect sizes for these variables [57, 58], making it important to differentiate between edibility assessments and take/leave decisions. For take/leave decisions, in particular, participants might be influenced by personality traits like their propensity to avoid risk [58].

4.2.2 Mushroom Hunting Game. In our first study [58], we showed the task items as static images within an online survey tool. Participants only saw the static *Forestly* app for each task item, selected their decisions with radio buttons, and stepped to the next task item. For the second study [57], we decided to make the task more immersive and embedded the task items in an interactive game that we named *AI Forest—The Schwammerl Hunting Game*. The goal of the game was to virtually pick mushrooms for a stew. Classifications of an AI system were given to support the participants' decision-making. In the introduction of the game, participants were encouraged to pick as many edible mushrooms as possible (to improve the flavor of the stew) but cautioned not to add any poisonous mushrooms (which would ruin the dish or could be harmful) [57]. The game also showed the number of mushrooms that participants already picked up, to remind them to collect as many mushrooms as possible. While the previous study was conducted in person, we adapted the game such that it could be used in an online setting, but still provides the benefits of the gamification aspects [73] (e.g., the interactive task setup and the counter showing the amount of mushrooms already picked). In the end, participants were shown the outcome of the game. The code for the game can be downloaded at https://osf.io/bq85c/.

4.2.3 Study Procedure. We conducted a between-subjects online experiment with four independent groups. A control group that received no further explanations (C) was compared to three experimental groups that received different explanations. The explanations were based on Grad-CAM (E1), nearest-neighbor (E2), and network-dissection (E3) explanations. Participants were randomly assigned to one of the four groups. The experiment started with tests about participants' AI and mushroom knowledge using tests from our previous study [58]. Following the knowledge tests, participants answered questions from the *Propensity to Trust* scale of the *Trust in Automation*

Questionnaire [51] that measures participants' general willingness to trust automated systems. Participants then received a step-wise onboarding tour that explained each part of the *Forestly* app screen including the corresponding explanation methods (for an example, see Figure 10 in the Appendix). Then, participants completed *AI Forest—The Schwammerl Hunting Game* with the 10 decision task items in random order. After the game, participants completed a post-survey. They answered questions for the *Trust in Automation* scale of the *Trust in Automation Questionnaire* [51] that measures participants' self-reported trust in the system. Similarly to our previous study [57], participants reported their intention to use *Forestly* in the future and evaluation of the *Forestly* app using five-point Likert scales. Finally, participants provided demographic information (i.e., gender, age, level of education, and mushroom-picking experience), and optionally further feedback in an open text field. At the end of the experiment, participants were informed of their success in their mushroom hunt and were provided with information about the risks of mushroom picking.

4.3 Sample Description and Sample-Size Justification

As preregistered [41], the minimum sample size of $N > 452$ was calculated in an a priori power analysis. To be on the safe side, we initially collected data from $N = 623$ participants. After excluding subjects who dropped out of the study prematurely or who answered attention-check items incorrectly, we were left with a total number of $N = 501$ complete datasets for analysis. See Section A.3 in the Appendix for calculation details.

Participants were recruited by a market research company, who managed the invitation of participants and initial screening. The market research company invited a diverse set of participants which were drawn from an online panel under consideration of demographic information. A diversified sample reflects the diversity within the population and can contribute to improving the generalizability of study results, thus aiding in mitigating biases. Participants received €2 for finishing the study. With the initial screening, only people who would—in general—eat mushrooms are forwarded to our survey. In doing so we aimed to ensure that participants can imagine themselves well in the task scenario. Participants were then forwarded to our website with a unique identifier that we later used to cross-check the validity of the collected data. In the pre-and post-survey, we included attention check items [63] to screen out inattentive participants and increase data quality as done in our previous studies [57, 58]. An example attention-check item is the following: "Several edible mushrooms can be mistaken for similar-looking poisonous ones. This question checks your attention. Please simply select 'Yellow foot'. (A) Common ink cap, (B) Pinecone cap, (C) Snowy waxcap, and (D) Yellow foot" [58]. Here, an incorrect response (all responses other than Yellow foot) would lead to a screen out of the individual due to inattentive responding.

Of the 501 participants, 240 identified as female, 258 as male, 2 as non-binary, and 1 person did not indicate their gender. The mean age was $M = 45.72$ years (SD = 15.81), and 145 participants had a university degree (≈29 %). Most participants reported that they would generally eat mushrooms (≈97 %) and had been mushroom picking before (≈84 %), but most had not used an app for mushroom picking (≈91 %). Descriptive results of a six-item mushroom knowledge test with $M = 2.88$ (SD = 1.42) correct answers and of a five-item AI knowledge test with $M = 2.83$ (SD = 1.34) correct answers indicate that the sample consisted of non-experts with varying degrees of prior knowledge. The result distributions over each group for the mushroom knowledge test (C: $M = 2.73$, SD = 1.39; E1: $M = 2.92$, SD = 1.29; E2: $M = 2.96$, SD = 1.6; E3: $M = 2.97$, SD = 1.43) and AI knowledge test (C: $M = 2.9$, SD = 1.27; E1: $M = 3.04$, SD = 1.37; E2: $M = 2.73$, SD = 1.43; E3: $M = 2.64$, SD = 1.27) are similar, which means that within each group participants had varying degrees of prior knowledge, but the groups do not differ considerably. Participants were randomly assigned to one of the four groups. The numbers of participants in each group were $n(\text{C}) = 155$, $n(\text{E1}) = 117$, $n(\text{E2}) = 112$,

and n(E3) = 117. For a detailed breakdown of the demographic information of each participant group, see Table 6 in the Appendix.

4.4 Data Analysis

We computed the performance of the human–AI teams by counting their correct decisions using the ground-truth edibility of the mushrooms for each task item. Note that the AI model is a multi-class classification (i.e., the model predicts the mushroom species), but participants were asked to make binary decisions (i.e., edible/poisonous and take/leave). The performance score is calculated using the binary participant decisions. With the performance score, we can compare the four user groups and evaluate how well each explanation method supports participants' decision-making.

The evaluation of a diverse set of task items gives insights into the overall performance of each explanation method. We analyze subsets of task items with particular characteristics to understand different aspects of user trust and reliance. In particular, we want to investigate how explanations can help against overtrust or distrust (which are concepts of appropriate trust), and how misleading or helpful explanations can influence decision-making.

4.4.1 Appropriate Trust. For measuring overtrust, we can use the subset of task items with only wrong AI classifications. High performance in this subset of task items indicates that participants were able to identify the fallibility of the model correctly, which means that they were able to avoid overtrusting the system. A low performance, on the other hand, shows that participants were unable to identify that the AI was wrong with the given explanation.

Conversely, distrust can be measured with the subset of task items of correct AI classifications. In that case, a high performance shows that participants were able to determine the good quality of the AI classification with the given explanation. Table 9 in the Appendix shows the items each subset comprised.

4.4.2 Helpful and Misleading Explanations. We also investigate the influence of helpful and misleading explanations on human AI-assisted decision-making. To select subsets that aid us in our investigation, we quantified this phenomenon with the following definition: An explanation is helpful if (i) correct (or trustworthy) AI classifications are accompanied by according explanations and (ii) incorrect AI classifications by discording explanations. On the other hand, an explanation is misleading if (i) correct (or trustworthy) AI classifications are accompanied by discording explanations and (ii) incorrect AI classifications by according explanations. An explanation accords with the classification, if the explanation makes the model seem trustworthy for the given classification. On the other hand, an explanation discords with the classification makes the model seem untrustworthy for the given classification. As an example, consider that an AI suggests a certain type of mushroom, but the nearest-neighbor examples for this suggestion do not look similar to the input image. This incoherence could indicate an error in the system (i.e., the explanation discords with the classification).

We also mapped this definition of *according* and *discording* to the explanation methods we use in this study: An explanation accords with the model's decision if (E1) the Grad-CAM attribution map focuses on (parts of) the mushroom; (E2) the mushroom depicted in the top-ranked nearest-neighbor example looks similar to the one in the input image; and (E3) the textual network-dissection description matches the highlighted regions in the attribution map and top-ranked species.

With these definitions, we asked four independent raters to categorize each explanation as *according* or *discording*. We calculated the inter-rater reliability using Fleiss' kappa [27] to measure the categorization agreement. For Grad-CAM (E1) and network-dissection (E3) explanations, agreement was "substantial" [54] ($\kappa = 0.739$ and $\kappa = 0.78$, respectively). However, for the nearest-neighbor (E2) examples agreement was only "fair" [54] ($\kappa = 0.4$). Table 7 in the Appendix lists all agreement

statistics calculated. For the final categorization of the ambiguous items (i.e., items for which the categorizations differed between raters), the raters discussed and finally decided on a categorization. We show the four categorizations and the final decisions in Table 8 in the Appendix. With this categorization, we are able to select sets of helpful and misleading explanations. Table 9 in the Appendix shows the items of each category.

4.4.3 Software and Statistical Methods. For data analysis, we used the open-source software R (version 4.2.1) [72]. The analysis code and data can be downloaded at https://osf.io/bq85c/. We performed multiple statistical tests for each hypothesis. Multiple testing increases the family-wise error rate (FWER), which results in a higher global α level than the typically used level of 0.05. We used Dunnett's tests [21] to control the FWER in many-to-one comparisons (i.e., a control group is compared to many experimental groups). We also calculated local Brunner–Munzel tests and report the results for completeness. We want to emphasize that the p-values reported for the Brunner–Munzel tests were not corrected for multiple testing and therefore not used for interpretation. For single tests, we report Brunner–Munzel tests [11] for which the p-values were adjusted using the Bonferroni–Holm procedure [7] to control the FWER.

As preregistered [41], effect-size estimates and confidence intervals are reported and visualized for non-binary analyses [17, 19, 32, 82]. We calculated effect sizes using Cohen's d [16] between each group and the control group and plotted the point estimation along with the 95 % confidence interval.

5 Results

5.1 Overall Effects on Human Decision Performance

We tested the effects of each explanation method on decision-making performance aggregated over all 10 items (*RQ1*). The results of the statistical tests are listed in Table 1. We first tested whether participants who received one of the explanations assessed edibility more correctly (●) and made better take/leave decisions (▲). For both tasks, we found that nearest-neighbor (E2) examples significantly improved the performance over that of the control group (see Dunnett's test results in Table 1). This aligns with the findings of our previous studies [57, 58], which reported that the combination of nearest-neighbor (E2) and Grad-CAM (E1) explanations improved participant performance.

In contrast, for Grad-CAM (E1) explanations alone (i.e., without nearest-neighbor (E2) examples) effects did not reach statistical significance. It seems like Grad-CAM explanations were less effective in our previous studies [57, 58] than we originally anticipated and that the main effects probably resulted from the nearest-neighbor examples. For network-dissection (E3) explanations, effects also did not reach statistical significance.

Figure 3 shows the effect-size point estimates and confidence intervals for each method. The direction of the estimates for the group that received Grad-CAM (E1) explanations is unclear, and there seems to be a discrepancy between the edibility assessment (●) and take/leave decision (▲). Effect-size estimates of the groups with nearest-neighbor (E2) and network-dissection (E3) explanations are similar; for both groups, the estimates are positive (i.e., they indicate improved performance). Note that confidence intervals are large, and they must be interpreted with caution.

5.2 Analysis of User Behavior for Individual Items

To investigate, which characteristics of an explanation lead to varied performances (*RQ2*), we looked into the results for individual items. We noticed a general trend among participant decisions: The number of participants who classified a mushroom as edible was typically similar across all groups and close to the AI's prediction for the mushroom's edibility. However, for some items, it seemed that certain explanation groups deviated more markedly from the control group. Based on

Table 1. Effects of Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations on Participant Performance in Assessing the Edibility (•) of Mushrooms and the Take/Leave Decision (▲)

		Control G.		Exp. G.		Dunnett			BM			Effect Size		
	XAI	*M*	*SD*	*M*	*SD*	*SE*	*t*	*p*	*t*	*df*	*p*	*d*	CI_{Lo}	CI_{Hi}
	E1	5.03	1.11	5.19	1.15	0.142	1.1	0.571	0.537	236	0.296	0.138	−0.104	0.379
•	E2	5.03	1.11	5.39	1.32	0.143	2.51	**0.0341**	2.11	193	0.018	0.299	0.0534	0.544
	E3	5.03	1.11	5.32	1.04	0.142	2.01	0.119	1.86	250	0.0318	0.262	0.0203	0.505
	E1	4.67	1.16	4.7	1.09	0.138	0.216	0.994	−0.331	253	0.630	0.0264	−0.215	0.268
▲	E2	4.67	1.16	5.03	1.28	0.14	2.55	**0.0315**	2.06	217	0.0201	0.294	0.0489	0.54
	E3	4.67	1.16	4.97	0.955	0.138	2.14	0.0885	1.81	269	0.0356	0.274	0.032	0.516

Mean (*M*) and standard deviation (*SD*) of user performance for control group (Control G.) and experimental group (Exp. G.) for each explanation method (XAI). Dunnett's standard errors (*SE*), test statistics (*t*), and *p*-values (*p*). Brunner–Munzel's (BM) test statistics (*t*), degrees of freedom (*df*), and *p*-values (*p*). Effect-size point estimations (*d*), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals. $P < 0.05$ (adjusted for multiple testing) are indicated in bold.

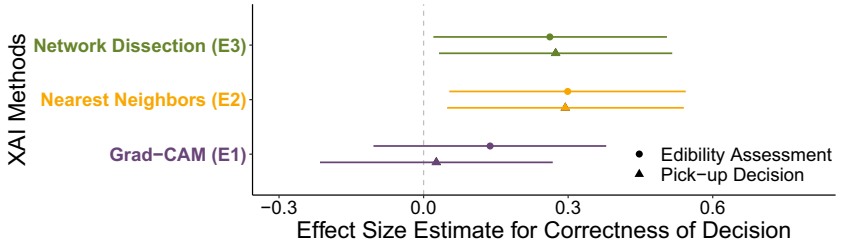
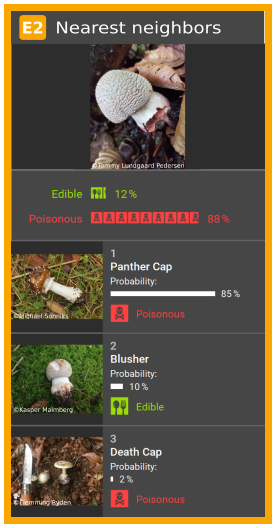


Fig. 3. Effect-size point estimates and 95 % confidence interval for participant performance of the three experimental groups. Estimates for groups with nearest-neighbor (E2) and network-dissection (E3) explanations suggest improved participant performance. The edibility assessment is marked with a dot (•) and the take/leave decision with a triangle (▲). The dashed line indicates 0 effect (i.e., the further away from 0, the larger the effect).

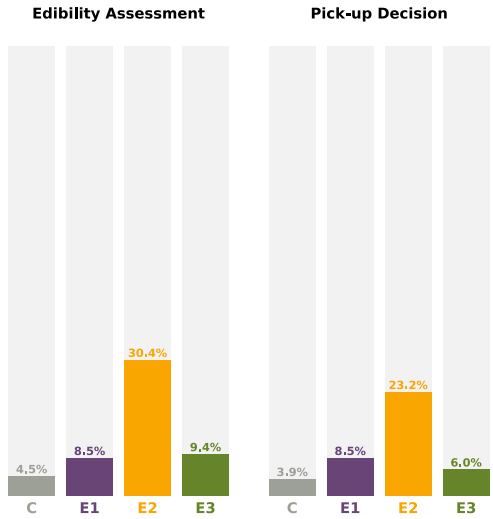
performance statistics we identified three outlier items for which the answers for one experimental group deviated markedly from all others (for details see Section B.1 in the Appendix).

We found that there were two outlier items for the nearest-neighbor (E2) method and one outlier item for the network-dissection (E3) method. One of the outliers was a blusher (i.e., an edible mushroom) with an AI classification of 12 % for the total probability of the mushroom being edible (see Figure 4(a)). Figure 4(c) shows plots of the edibility assessments (•) and the take/leave decisions (▲) of participants. Most participants in the control group classified the mushroom as poisonous, likely due to the low edibility score predicted by the AI. However, participants who received nearest-neighbor (E2) examples were much more likely to correctly classify the mushroom as edible. A possible explanation for this decision behavior is that the example of the mushroom with the highest probability differed strongly from the target mushroom. However, the example image of the second-ranked species (which was the correct mushroom species in this case) looked very similar to the target image. The explanation for this item discorded with the classification and thus hinted at an incorrect AI decision. In contrast, the Grad-CAM (E1) and network-dissection (E3) explanations (see Figure 4(b)) do not show the incorrectness of the model as clearly. This item is an example of how a nearest-neighbor (E2) explanation can help to avoid overtrust in an AI system (i.e., the participant correctly overrode the AI's incorrect suggestion).



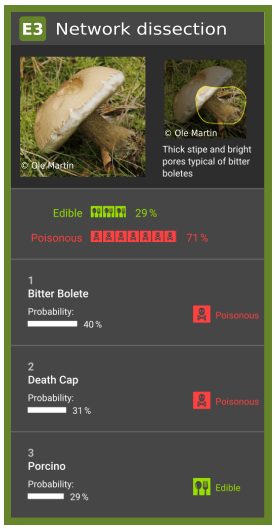
(a) *Forestly* app with nearest-neighbor (E2) examples.

(b) Grad-CAM (E1) and network-dissection (E3) explanations.

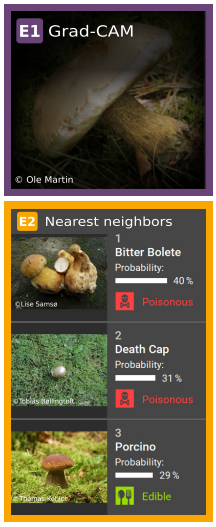


(c) Proportions of users assessing the mushroom as edible (left) and deciding to take the mushroom (right) per group.

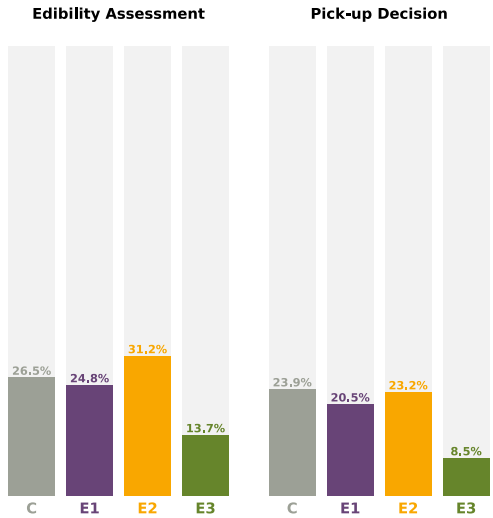
Fig. 4. Example of the *Forestly* app (a, b) and user responses for this task item (c). Participants of the nearest-neighbor (E2) group are more likely to correctly believe that the mushroom is edible. Note: Translated from the German original version.



(a) *Forestly* app with a network-dissection (E3) explanation.



(b) Grad-CAM (E1) and nearest-neighbor (E2) explanations.



(c) Proportions of users assessing the mushroom as edible (left) and deciding to take the mushroom (right) per group.

Fig. 5. Example of the *Forestly* app (a, b) and user responses for this task item (c). Participants of the network-dissection (E3) group are more likely to believe that the mushroom is poisonous. Note: Translated from the German original version.

Table 2. Effects of Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations on Participant Performance in Assessing the Edibility (●) of Mushrooms, and the Take/Leave Decision (▲) for *Correctly Classified* Items

		Control G.		Exp. G.		Dunnett			BM			Effect Size		
	XAI	*M*	*SD*	*M*	*SD*	*SE*	*t*	*p*	*t*	*df*	*p*	*d*	CI_{Lo}	CI_{Hi}
	E1	4.59	1.33	4.55	1.21	0.152	−0.307	0.982	−0.62	251	0.732	−0.0364	−0.277	0.205
●	E2	4.59	1.33	4.51	1.22	0.154	−0.55	0.911	−0.852	247	0.802	−0.066	−0.31	0.178
	E3	4.59	1.33	4.89	1.17	0.152	1.95	0.135	1.81	258	0.0359	0.234	−0.00761	0.476
	E1	4.21	1.29	4.11	1.06	0.144	−0.664	0.856	−1.04	266	0.849	−0.0796	−0.321	0.162
▲	E2	4.21	1.29	4.21	1.21	0.145	0.0539	1	−0.25	235	0.598	0.00624	−0.238	0.25
	E3	4.21	1.29	4.48	1.07	0.144	1.9	0.15	1.68	266	0.0473	0.227	−0.015	0.469

Mean (*M*) and standard deviation (*SD*) of user performance for control group (Control G.) and experimental group (Exp. G.) for each explanation method (XAI). Dunnett's standard errors (*SE*), test statistics (*t*), and *p*-values (*p*). Brunner–Munzel's (BM) test statistics (*t*), degrees of freedom (*df*), and *p*-values (*p*). Effect-size point estimations (*d*), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals.

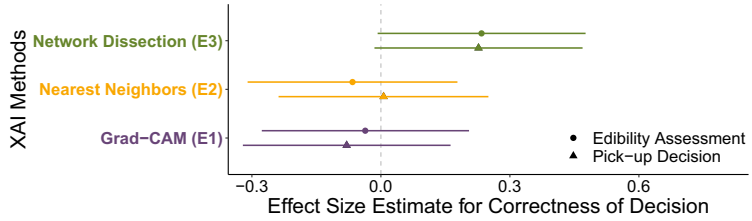
Another outlier is a bitter bolete (i.e., an inedible mushroom) with a predicted edibility of 29 % (see Figure 5(a)). Most participants classified the mushroom as poisonous/inedible, but on average, 24 % of all participants distrusted the system and wrongly classified the mushroom as edible (see Figure 5(c)). In contrast, only 14 % of participants who received network-dissection (E3) explanations misclassified the mushroom. Upon closer inspection of this item, the network-dissection (E3) explanation seemed to confirm the validity of the model's classification because the highlighted region focuses on parts of the mushroom that humans would also deem relevant (i.e., cap and stipe) and the semantic text matches the highlighted region (see Figure 5(b) for comparison to the remaining explanations)—the explanation thus *accorded* with the classification. This is an example of how network-dissection (E3) explanations can help to overcome distrust in an AI system (i.e., participants followed a correct AI suggestion). On the other hand, participants who received nearest-neighbor (E2) examples seem to perform slightly worse on average. The example images seem to mislead participants into not trusting the AI system.

5.3 Effects on Human Decision Performance for Item Subsets

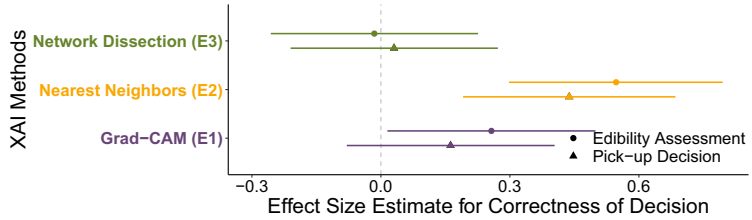
As shown in Section 5.2, we did not observe the same trends for all items in the classification tasks. Group differences may have been larger for some items due to specific explanation characteristics that moderated the effects. In this section, we report the effects of each explanation method on the decision-making performance for subsets of task items. In particular, we analyze subsets that can be used to measure overtrust or distrust (*RQ3*), and subsets of helpful or misleading (*RQ4*) explanations.

5.3.1 Appropriate Trust. We split the items into the sets of (i) correct, and (ii) wrong model classifications. We chose these two sets to measure how explanation methods can help mitigate overtrust and distrust. If participants decide to follow a model's suggestion although it is wrong, this means that participants are not able to determine that the classification is wrong and trust the system too much (i.e., overtrust it). Conversely, if the model gives a correct classification and participants reject it, they exhibit too little trust (i.e., distrust) in the system. We used the set of correctly classified items to measure distrust, and the set of wrongly classified items to measure overtrust.

For items that were correctly classified by the model, effects did not reach statistical significance (see Table 2). Figure 6(a) shows the effect-size point estimates and confidence intervals for each method. The effect-size estimates for the groups that received Grad-CAM (E1) and nearest-neighbor



(a) The group with network-dissection (E3) explanations seems to exhibit less *distrust*.



(b) The group with nearest-neighbor (E2) examples seems to exhibit less *overtrust*.

Fig. 6. Effect-size point estimates and 95 % confidence interval for participant performance for the subset of items (a) *correctly* and (b) *wrongly* classified by the AI.

Table 3. Effects of Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations on Participant Performance in Assessing the Edibility (•) of Mushrooms and the Take/Leave Decision (▲) for *Wrongly Classified* Items

		Control G.		Exp. G.		Dunnett			BM			Effect Size		
	XAI	*M*	*SD*	*M*	*SD*	*SE*	*t*	*p*	*t*	*df*	*p*	*d*	CI_{Lo}	CI_{Hi}
	E1	0.439	0.765	0.641	0.814	0.0969	2.09	0.0989	2.45	242	0.00749	0.257	0.0151	0.499
•	E2	0.439	0.765	0.884	0.878	0.0981	4.54	**0.00002**	5.03	237	5.00×10^{-7}	0.547	0.298	0.795
	E3	0.439	0.765	0.427	0.711	0.0969	−0.117	0.999	0.195	256	0.423	−0.0153	−0.256	0.226
	E1	0.465	0.75	0.59	0.8	0.0956	1.31	0.431	1.4	239	0.082	0.162	−0.0793	0.404
▲	E2	0.465	0.75	0.812	0.855	0.0969	3.59	**0.00107**	3.84	228	7.85×10^{-5}	0.438	0.191	0.685
	E3	0.465	0.75	0.487	0.727	0.0956	0.237	0.992	0.483	251	0.315	0.0306	−0.21	0.272

Mean (*M*) and standard deviation (*SD*) of user performance for control group (Control G.) and experimental group (Exp. G.) for each explanation method (XAI). Dunnett's standard errors (*SE*), test statistics (*t*), and *p*-values (*p*). Brunner–Munzel's (BM) test statistics (*t*), degrees of freedom (*df*), and *p*-values (*p*). Effect-size point estimations (*d*), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals. *P*-values < 0.05 (adjusted for multiple testing) are indicated in bold.

(E2) explanations are close to zero. However, effect-size estimates of the group with network-dissection (E3) explanations seem to point in a positive direction. Note that the confidence intervals are large and must be interpreted with caution. To confirm whether network-dissection (E3) explanations are actually beneficial to countering participant distrust in systems, we need to conduct further studies.

For the set of items that were wrongly classified by the model, we found statistically significant effects for nearest-neighbor (E2) examples, but not for Grad-CAM (E1) and network-dissection (E3) explanations (see Table 3). The group with nearest-neighbor (E2) examples performs better within the set of wrongly classified items, which leads to the assumption that nearest-neighbor (E2) examples can help participants avoid overtrusting the system.

Table 4. Effects of Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations on Participant Performance in Assessing the Edibility (•) of Mushrooms and the Take/Leave Decision (▲) of Items with a Helpful Explanation

		Control G.		Exp. G.		BM			Effect Size			
	XAI	M	SD	M	SD	t	df	p	d	CI_{Lo}	CI_{Hi}	p_{adj}
	E1	4.83	4.93	1.16	1.17	0.287	228	0.387	0.0855	−0.156	0.327	0.774
•	E2	3.46	3.95	0.855	0.948	4.51	196	5.50×10^{-6}	0.546	0.297	0.794	$\mathbf{3.31 \times 10^{-5}}$
	E3	4.3	4.55	1	0.856	1.81	263	0.0357	0.259	0.0166	0.501	0.1428
	E1	4.47	4.44	1.17	1.12	−0.754	247	0.774	−0.0306	−0.272	0.211	0.774
▲	E2	3.2	3.56	0.929	1.03	2.64	214	0.00443	0.373	0.127	0.619	**0.02215**
	E3	4.11	4.38	1.05	0.828	1.7	269	0.0451	0.278	0.0354	0.52	0.1428

The adjusted p-values (p_{adj}) of the tests are marked in bold when < 0.05. Mean (M) and standard deviation (SD) of user performance for control group (Control G.) and experimental group (Exp. G.) for each explanation method (XAI). Test statistics (t), and p-values (p). Brunner–Munzel's (BM) test statistics (t), degrees of freedom (df), and p-values (p). Effect-size point estimations (d), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals. Note that the mean (M) and standard deviation (SD) for the control group vary because in this analysis we compare a different subset for each explanation method.

As shown in Figure 6(b), the effect-size estimates for the group that received nearest-neighbor (E2) examples indicate a clear positive direction. The effect-size estimates of Grad-CAM (E1) explanations also indicate a positive direction, but due to the non-significance of this difference and the large confidence intervals, this must be interpreted with caution. For the group that received network-dissection (E3) explanations, the effect-size estimates are close to zero.

5.3.2 Helpful and Misleading Explanations. We performed three independent one-to-one comparisons, to measure the effect of helpful explanations. This means that each experimental group was compared to the control group using the corresponding subset of items with helpful explanations. Table 9 in the Appendix lists the subsets for each explanation method.

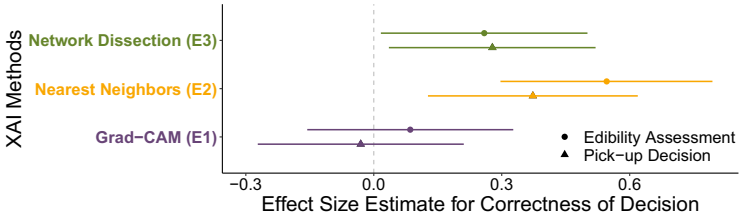
Based on the adjusted p-values for the edibility assessment (•) and the take/leave decision (▲) (see Table 4) the effects for nearest-neighbor (E2) examples are statistically significant. The results are similar to those reported in Section 5.1 but with a larger effect size. This makes sense because we only test the subset of items that are deemed helpful with respect to such decision-making tasks. For network-dissection (E3) and Grad-CAM (E1) explanations the effect did not reach statistical significance based on the adjusted p-values.

The effect-size estimates (see Figure 7(a)) for the group that received network-dissection (E3) explanations tend toward positive values, while the sign of effect-size estimates for Grad-CAM (E1) is uncertain. To obtain better estimates, replications with larger sample sizes should be conducted in the future.

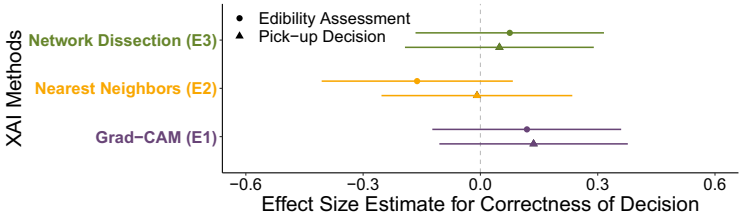
We also performed three independent one-to-one comparisons, to measure the effect of misleading explanations (i.e., each experimental group was compared to the control group using the corresponding subset of items with misleading explanations).

The adjusted p-values did not reach statistical significance (see Table 5), and the directions of the effect-size estimates (see Figure 7(b)) are uncertain. Although we expected that performances would be worse with misleading explanations, this assertion can neither be confirmed nor rejected with the current analysis. However, it is interesting to see that the results of subsets of task items with misleading explanations differ from the results of all task items.

Note that the subsets tested here consisted only of two task items for Grad-CAM (E1) and network-dissection (E3), as shown in Table 9. This is a very low number of items to consider for



(a) Helpful nearest-neighbor (E2) and network-dissection (E3) explanations seem to improve participant performance.



(b) Misleading explanations seem to lower participant performance.

Fig. 7. Effect-size point estimates and 95 % confidence interval for user performance for the subset of items with (a) *helpful* and (b) *misleading* explanations.

Table 5. Effects of Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations on Participant Performance in Assessing the Edibility (●) of Mushrooms and the Take/Leave Decision (▲) for the Subset of Items with Misleading Explanations

		Control G.		Exp. G.		BM			Effect Size			
	XAI	*M*	*SD*	*M*	*SD*	*t*	*df*	*p*	*d*	CI_{Lo}	CI_{Hi}	p_{adj}
	E1	0.2	0.462	0.256	0.494	1.11	237	0.865	0.119	−0.123	0.36	1
●	E2	1.57	0.755	1.45	0.837	−1.36	223	0.0873	−0.162	−0.406	0.0829	0.5238
	E3	0.729	0.573	0.769	0.48	0.826	268	0.795	0.0751	−0.166	0.316	1
	E1	0.2	0.462	0.265	0.498	1.26	236	0.896	0.136	−0.105	0.377	1
▲	E2	1.47	0.759	1.46	0.782	−0.0259	231	0.49	−0.0087	−0.253	0.235	1
	E3	0.561	0.615	0.59	0.544	0.7	263	0.758	0.0486	−0.193	0.29	1

Mean (*M*) and standard deviation (*SD*) of user performance for control group (Control G.) and experimental group (Exp. G.) for each explanation method (XAI). Dunnett's standard errors (*SE*), test statistics (*t*), and *p*-values (*p*). Test statistics (*t*), degrees of freedom (*df*), and *p*-values (*p*). Effect-size point estimations (*d*), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals. Note that the mean (*M*) and standard deviation (*SD*) for the control group vary because in this analysis we compare a different subset for each explanation method.

statistical testing or interpretation and might yield artifacts. Future studies should consider this, and a larger set of task items should be included.

5.4 Effects on Self-Reported Trust, System Evaluation, and Intention to Use

We tested the effects of each explanation method on self-reported user trust, evaluation of the *Forestly* app, and intention to use a mushroom-identification app in the future (*RQ5*). The effects did not reach statistical significance for any explanation method or variable (detailed results of Dunnett's tests are included in Table 12 in the Appendix).

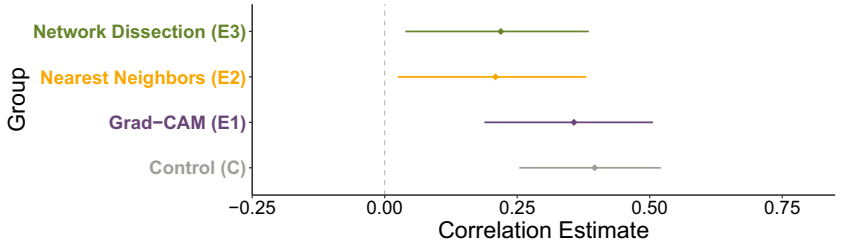


Fig. 8. Point estimates and 95 % confidence intervals of correlation coefficients for participants' propensity to trust and their self-reported trust. Trust in the *Forestly* app seems to be less influenced by participants' propensity to trust for groups with nearest-neighbor (E2) and network-dissection (E3) explanations.

Effect-size estimates for nearest-neighbor (E2) examples ($d = -0.29$, $CI_{95} = [-0.535, -0.0444]$) and Grad-CAM (E1) explanations ($d = -0.209$, $CI_{95} = [-0.450, 0.0333]$) on self-reported trust tend toward negative values, as visualized in Figure 16 in the Appendix. This aligns with results from our previous study [58], but confidence intervals are large and the results must be interpreted with care. Similarly, for the evaluation of the *Forestly* app, the effect-size estimates of nearest-neighbor (E2) examples tend toward negative values, but again with a large confidence interval (see Figure 16 in the Appendix).

Finally, we also analyzed the correlation between user propensity to trust automated systems in general (measured before the mushroom-picking task) and self-reported user trust in the *Forestly* app (measured after the mushroom-picking task) as propensity to trust is an important predictor in current trust models [e.g., 38, 51]. We calculated the correlation for each experimental group using Pearson's product-moment correlation. Based on the correlation coefficients and confidence intervals shown in Figure 8, correlations between the general propensity to trust automated systems and the self-reported trust in the *Forestly* app seem to differ between groups. On a descriptive basis, explanations seem to moderate this correlation.

In our interpretation of this moderation effect, a strong correlation between the general propensity to trust and the self-reported trust in the app would mean that participants base their trust in the *Forestly* app substantially on their general tendency to trust automated systems (perhaps, as in the control conditions without explanation, the app itself gave too little information relevant to trust formation). In contrast, a smaller correlation, as observed in two of the experimental conditions with explanations, would indicate that the explanations *had* an effect. Participants may have adapted their trust levels because the explanations provided valuable information relevant to trust formation. Thus they based their trust ratings for the *Forestly* app less on their overall tendency to trust automated systems. The results of our correlation analysis are listed in detail in Table 13 in the Appendix. For participants who received nearest-neighbor (E2) or network-dissection (E3) explanations, trust in the *Forestly* app seems to have been influenced less by their propensity to trust automated systems, as correlation coefficients are lower in these groups than in the control group (C) and in the group of participants who received Grad-CAM (E1) explanations. This would mean that nearest-neighbor (E2) and network-dissection (E3) provided more useful information for trust formation than the other groups.

6 Discussion

In this work, we investigated the effects of Grad-CAM (E1), nearest-neighbor (E2), and network-dissection (E3) explanations on human decision-making behavior in the high-risk task of mushroom hunting, building on two previous studies [57, 58].

6.1 Summary of Results

We performed statistical tests and calculated effect-size estimates and confidence intervals. The method yielding the most positive impact on participants' mushroom-identification performance was nearest-neighbor (E2) examples, as indicated by statistically significant effects and the highest effect-size estimates. Especially in situations in which a model made an incorrect classification, this method helped participants to realize that the classification was wrong (i.e., it helped to avoid overtrust)—as previously speculated [57, 58]. An advantage of nearest-neighbor (E2) compared to Grad-CAM (E1) and network-dissection (E3) explanations is that users see examples for each of the top three predicted classes, which might enable them to make a more informed decision.

Most other effects (for network-dissection (E3) or Grad-CAM (E1) explanations) did not reach statistical significance. Although effect-size estimates for network-dissection (E3) explanations seem to show a similar positive trend as nearest-neighbor (E2) when testing all task items, confidence intervals are large. According to the effect-size estimates, network-dissection (E3) explanations seem to help participants realize when a model's classification is indeed correct (i.e., it helps to avoid distrust).

Interestingly, effects for Grad-CAM (E1) explanations did not reach statistical significance in any performance tests, and effect directions are inconclusive. This was surprising to us because we chose to include attribution-based techniques due to their widespread use for explaining image classification tasks [1]. It is therefore possible that the significant effects from our previous studies [57, 58] with a combination of nearest-neighbor (E2) examples and Grad-CAM (E1) are due more to nearest-neighbor (E2) examples and less to Grad-CAM (E1). Possible reasons are that this technique is less understandable to users, who have little experience with AI systems, than originally anticipated or that these explanations are less effective for identifying a models' incorrect decisions [28].

Note that effects might be hidden due to various factors. First, true effect sizes might be lower than expected (i.e., lower than estimated in the preregistration), and therefore the statistical power in the many-to-one comparisons might have been too low to detect them. Second, only a specific realization of each explanation method was used. Different explanation designs might yield better results (e.g., more elaborate textual explanations for network-dissection (E3) which give more detailed information on features could lead to larger effects). Third, effects might be moderated by other factors and only become significant under specific circumstances. One such moderator could be particular characteristics of the stimulus material (i.e., mushroom images). This can be seen in the exploratory analyses of specific item subsets of the mushroom-picking task of this study.

In our exploratory analysis of item subsets, we found that explanations differed in their effects depending on whether they were helpful or misleading. We defined helpful explanations as those where (i) the explanation accords with a correct AI classification and (ii) the explanation discords with an incorrect AI classification, and vice versa for misleading explanations. We found increased effects for item subsets that contained only helpful explanations. In the case of misleading explanations, we were unable to reach a conclusion on whether or not decision performance gets worse for the groups with explanations compared to the control group. However, participants' performance is lower than for the subset of helpful explanations. Such insights must be confirmed in future studies, where these conditions are tested with more task items. This analysis of item subsets allowed us to understand *why* some XAI methods work better for participants than others. We, therefore, recommend that future studies also test the effects of subsets of items with specific characteristics to explore *why* specific XAI methods instead of just *which* methods work better in terms of user performance and trust.

Tests regarding the effects of XAI on self-reported participant trust in the system, their evaluation of the *Forestly* app, and their intention to use such an app in the future did not reach statistical significance similar to our festival-based study [57] but unlike our previous online study [58]. To better understand trust calibration, we tested how much participants' trust in a specific AI-based system within a given task is influenced by their propensity to trust such AI-based systems in general, and how this relationship is affected by different explanation designs. Analysis regarding the correlation between participant propensity to trust automated systems and self-reported participant trust in the *Forestly* app showed differences between groups. This differential influence of personality traits (e.g., propensity to trust in automation) and machine characteristics (e.g., the understandability of AI-based explanations) in human–technology interaction has also been shown in previous human factors research [64] and illustrates the dynamic nature of trust formation. Explanations allow participants to rely less on their general propensity when deciding how much to trust a specific system with a task. Participants adjust their trust specifically to the system as it provides them with trust-relevant cues (e.g., according and discording elements between classification and explanation).

6.2 Limitations and Future Work

To keep the scope and duration of our study within reasonable limits, we had to make a number of decisions that give rise to certain limitations and leave potential for future work. First, we selected three different XAI techniques. While we chose one technique for each of the three explanation types, our study can only provide a limited view of all possible XAI methods. We also had to consider how to represent the explanations (i.e., make certain design choices). We discussed the core concept of each explanation method and how participants can use an explanation to determine whether or not to trust a prediction. We discussed various conceptualizations but had to decide on one in the end. The conceptualization and representation of the explanations can moderate participant trust levels of a prediction. For example, we decided on showing images of each of the top three predicted classes for nearest-neighbor examples, while Grad-CAM and network-dissection show their explanations on the input image only, which might have moderated the participant's trust. Therefore, different representations of the explanations should be considered in the future to understand whether the effects of this study are a function of certain design choices or indeed effects of the concepts. However, our fine-grained analysis of different item subsets allowed us to filter out various characteristics of the respective explanations that led to a more appropriate trust. These are corresponding cues that stand out more clearly with some explanation strategies than with others. We therefore believe that these effects are also attributable to the core concepts and can be abstracted.

Second, we chose a fixed set of 10 items for the identification task. We aimed for a balance between the different item characteristics (e.g., poisonous/edible, correct/incorrect AI decision, and helpfulness of explanations) that we identified as possible moderators. Still, our specific choice is likely to have influenced the results and we were not able to consider all factors (e.g., balancing according vs. discording explanations, and helpful/misleading explanations) because some of the factors also vary between items and explanations, which makes it hard to balance all possible subsets.

Third, we kept the AI model itself fixed (choosing a model with relatively low accuracy) and did not disclose the accuracy to the participants because we wanted to focus on the explanations as variables. As shown in other work [34], accuracy can influence user reliance on an AI system and should be investigated further. Future studies may cover more explanation techniques, draw from a bigger set of items, and/or experiment with different models.

Finally, we decided to implement a random order of task items for each participant to filter out temporal moderators. However, with this implementation, we were unable to measure how first impressions of our system affected trust formation over time [69, 79]. Another confounding factor that can mediate effects is the risk that individual participants are willing to take. Future studies could also first test human decisions without the aid of the AI-based system and afterward with the aid of an AI-based system. Measuring whether humans change their decision can give further insights into their cognition during decision-making [see 75]. As experiments are limited by time and resources, we were unable to take all these mediating factors into account and focused on a subset of variables.

6.3 Ethical Considerations

In addition to the limitations discussed above, one difficulty with studying XAI in high-risk tasks relates to ethical challenges in the study design. Ideally, the risks associated with decisions (which form the prerequisite for studying trust) can be simulated in a realistic way. However, for the mushroom-hunting task, a field study in a real forest environment, where participants could be endangered by wrong decisions, is infeasible. To mitigate the disadvantages of an online study, we used gamification to properly motivate participants like in our previous festival-based study [57]. We explained the task such that participants could imagine being in the given situation and highlighted the importance of the decision for participants. We additionally differentiated between participants' edibility assessment and the actual pick/leave decision that was relevant to the game score. Instead of a hypothetical award (i.e., collecting the mushrooms for a stew), we could have opted for real rewards to create a more realistic incentive for participants. For example, a monetary reward [52, 85] could be a motivator for participants to increase their performance. However, it is questionable if a monetary reward would be able to properly convey the risk of the task and how much hypothetical rewards affect the results of user studies compared to real rewards [45]. The implementation of such a reward must be considered carefully and can be another moderating factor. Finally, from an ethical point of view, giving unequal payment to participants for the same effort is considered unfair.

We also recognized that the study might prompt participants to go mushroom picking for real and that they might want to use a similar app to *Forestly*. Therefore, we informed participants after the study about the dangers of mushroom hunting. We highlighted that mushroom identification apps similar to *Forestly* can be fallible and dangerous to use when hunting for edible mushrooms.

Finally, we want to discuss ethical considerations from an app developer's point of view. Results from our previous studies hint that the presence of certain explanations—especially those that uncover potential errors in the system—might reduce user trust and intention initially [58]. This may prompt developers to refrain from showing explanations in their apps. In our study, results regarding trust were inconclusive. However, our results do show that properly chosen explanations can increase decision performance and may lead to higher user acceptance in the long run. We thus advocate the use of explanations and transparency about the AI model in general.

6.4 Practical Implications

We found that nearest-neighbor (E2) explanations (which show an example image of each suggested class of mushroom species) led to reliable improvements in human decision performance, but the effectiveness of other methods remains to be clarified. While this holds for mushroom picking, future studies are needed to see whether the results can be generalized or whether effects change in different settings and domains. We found that a single explanation method that works for some task items within one context is not necessarily beneficial in another (i.e., some methods might aid users

in avoiding overtrust, others in avoiding distrust). This indicates that there is no one-size-fits-all solution, and that explanation methods must be carefully selected, combined, and tested with target users before implementation in a production environment. By analyzing *why* particular explanations improved user performance more than others, future researchers can make more informed decisions on which explanations to test. In particular, explanation methods should be able to highlight possible mistakes of an AI prediction and not mislead users into following wrong classifications.

In future research, different effects for different target users can also be expected, since users may come from different domains or have varying degrees of prior knowledge related to the task. In real life, there are various domains in which AI and XAI are already applied. Furthermore, an AI's classification and explanation might be shown on various devices, which might favor some XAI techniques over others. Our study relied on a set of visual explanations to foster user understanding; future work should address situations in which users are not able to perceive visual explanations (e.g., people with visual impairments).

7 Conclusion

We have reported the results of an online experiment about the effects of explanations on human behavior in an AI-assisted, gamified mushroom-picking task, building on the results of two previous studies in this context [57, 58]. In particular, we compared the effects of three types of methods that explain AI suggestions. Surprisingly, XAI had little effect on participants' self-reported trust in the AI, their evaluation of the system, and their intention to use a mushroom identification app in the future. Regarding user performance, we found consistent positive effects for participants who received nearest-neighbor examples. We found that the hybrid attribution- and concept-based network-dissection method is a promising explanation technique that should be investigated further in future studies. Furthermore, we found that Grad-CAM explanations were less useful to participants than originally anticipated. We also tested the effects for various subsets of task items and found that effects were more distinct for helpful explanations that allow users to verify a model prediction. Misleading explanations, on the other hand, resulted in a performance drop. We also found differing effects when testing for overtrust and distrust. It seems that some explanations are better suited to avoid overtrust, while others are better suited for avoiding distrust in AI. We believe that trust calibration with the help of explanations is a promising direction for future studies.

Appendices

A Additional Information on Methods

This section contains additional information for Section 4.

A.1 Study Workflow

Figure 10 shows an example of the onboarding screens shown to participants prior to the identification task for the group. The image shown is the final view of a slideshow in which bullet points were added one by one. Other groups received similar onboarding.

A.2 AI Assistance

An example of the creation of textual labels for network-dissection (E3) concepts and attribution is depicted in Figure 11.

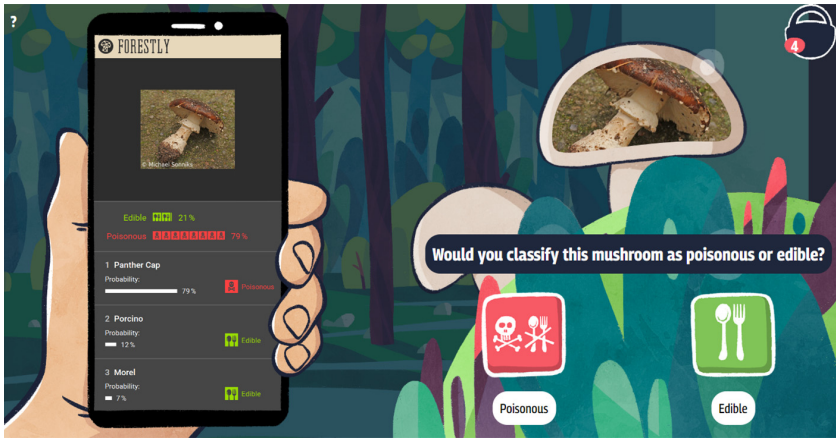


Fig. 9. Example of the *AI Forest—The Schwammerl Hunting Game* and the *Forestly* app without explanations (C). The participant is asked to assess the edibility of the mushroom. Note: Translated from the German original version.

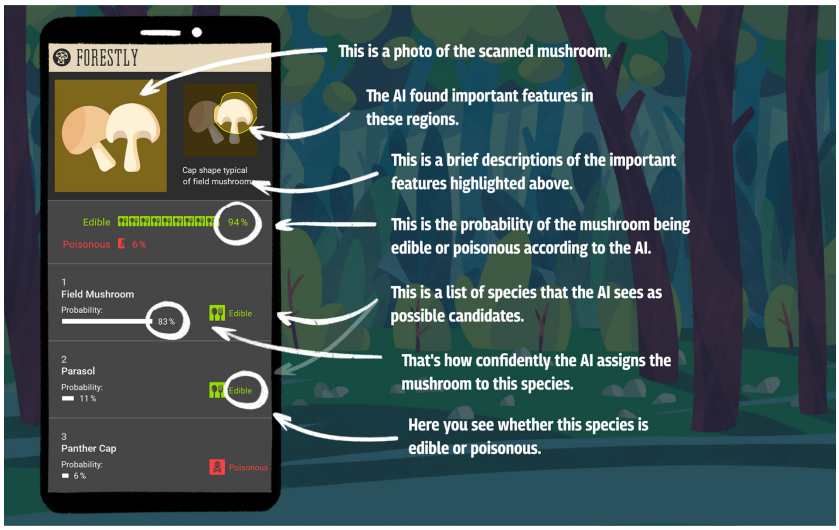


Fig. 10. Screenshot of the final onboarding view for *Forestly* with network-dissection (E3) explanations, translated from the original German version.

A.3 Sample Description and Sample-Size Justification

As preregistered [41], the minimum sample size was calculated in an a priori power analysis using G*Power [26] based on the many-to-one comparison of the control group with the three experimental groups. The local α-level for each of the three comparisons should be smaller than 0.01667 to control for multiple testing conservatively. For the power analysis, we expected an effect size of $d = 0.44$ based on the results of our previous study [58]. Thus, to achieve a minimum of 80 % statistical power with the study characteristics mentioned, a total sample size of $N > 452$ participants was needed for this study.

See Table 6 for information about demographic distributions per group.

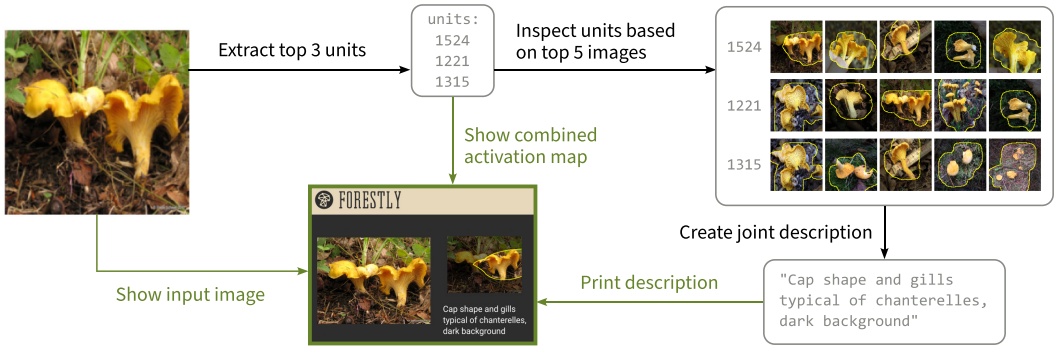


Fig. 11. The heatmap shows the differences in participant assessment of the edibility of the mushroom (left) and the decision to take or leave the mushroom (right) between each of the experimental groups and the control group.

Table 6. Demographic Information About the Participants of Our Study for All Participants (Overall) and Participants with No (C), Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations

	Overall	C	E1	E2	E3
Gender (Female)	48%	45%	42%	50%	56%
Gender (Male)	51%	54%	58%	49%	43%
Gender (Non-Binary)	0.4%	0.6%	0%	1%	0%
Gender (N/A)	0.2%	0%	0%	0%	1%
Age (*M*)	45.72	45.36	44.65	46.94	46.1
Age (*SD*)	15.81	15.2	14.89	16.3	17.07
University Degree	29%	32%	26%	29%	28%
Eat Mushrooms	97%	99%	95%	97%	96%
Mushroom Picking	84%	82%	85%	86%	82%
Used App	9%	12%	11%	5%	6%
Mushroom Knowledge (*M*)	2.88	2.73	2.92	2.96	2.97
Mushroom Knowledge (*SD*)	1.42	1.39	1.29	1.60	1.43
AI Knowledge (*M*)	2.83	2.90	3.04	2.73	2.64
AI Knowledge (*SD*)	1.34	1.27	1.37	1.43	1.27

Categorical values are stated as percent (%) of samples in a group for each category. Numerical values are stated with mean (*M*) and standard deviation (*SD*).

A.4 Categorization of Explanations

Table 7 shows the statistics for calculating the inter-rater reliability of the four categorizations for each explanation method. Table 8 shows the four independent categorizations and the final decision for each explanation method and item.

B Additional Information, Tables, and Figures on Results

This section contains additional information for Section 5.

Table 7. Fleiss' Kappa for Measuring Inter-Rater Reliability for the Explanation Categorization with 10 Subjects and Four Raters

Method	κ	z	p-value
Grad-CAM (E1)	0.739	5.72	1.06×10^{-8}
Nearest Neighbor (E2)	0.4	3.1	1.95×10^{-3}
Network Dissection (E3)	0.78	6.04	1.51×10^{-9}

Table 8. Categorization of Each Explanation and Item by Four Independent Raters (R1–R4), and the Final Decision

	Grad-CAM (E1)					Nearest Neighbor (E2)					Network Dissection (E3)				
ID	R1	R2	R3	R4	**Final**	R1	R2	R3	R4	**Final**	R1	R2	R3	R4	**Final**
1	+	+	−	−	**−**	+	+	+	−	**+**	−	−	−	−	**−**
2	−	−	−	−	**−**	−	+	−	−	**−**	−	−	−	−	**−**
3	+	+	+	+	**+**	−	−	−	−	**−**	−	−	+	−	**−**
4	+	+	+	+	**+**	+	+	+	+	**+**	+	+	+	+	**+**
5	+	+	+	+	**+**	+	+	−	+	**+**	+	+	+	+	**+**
6	+	+	+	+	**+**	+	+	+	−	**+**	+	+	+	+	**+**
7	+	+	+	+	**+**	−	−	−	−	**−**	+	+	+	+	**+**
8	+	+	+	+	**+**	−	+	+	+	**+**	+	+	+	+	**+**
9	+	+	+	+	**+**	−	−	−	−	**−**	+	−	−	−	**−**
10	+	+	+	+	**+**	−	+	+	+	**+**	+	+	+	+	**+**

According and discording explanations are indicated by + and − respectively.

B.1 Analysis of User Behavior for Individual Items

Figure 12 shows a heatmap of the difference in participants' decision behavior between each experimental group with explanations and the control group without explanations for each task item (i.e., we calculated the percentage of how many participants would classify a mushroom as edible and how many participants would pick a mushroom and calculate the difference between the experimental groups and the control group).

We performed a visual outlier analysis using boxplots (see Figure 13) and identified outliers using 1.5 times the interquartile range. We found three outliers for the edibility assessment (●) (mushrooms 2 and 3 for the group with nearest-neighbor (E2) examples, and mushroom 7 for the group with network-dissection (E3) explanations) and one outlier for the take/leave decision (▲) (mushroom 3 for the group with nearest-neighbor (E2) examples).

B.2 Effects on Human Decision Performance for Item Subsets with Specific Characteristics

Table 9 shows how task items were mapped to subgroups used in the analysis.

B.2.1 Poisonous Mushrooms. In this section, we investigate the subset of items that contains only poisonous mushrooms. This subset of items involves decisions of the highest risk. Misclassifying poisonous mushrooms as edible may have serious consequences (i.e., possible poisoning), while

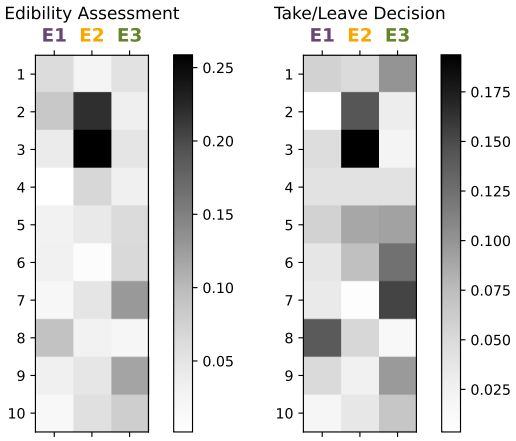


Fig. 12. The heatmap shows the differences in participant assessment of the edibility of the mushroom (left) and the decision to take or leave the mushroom (right) between each of the experimental groups and the control group.

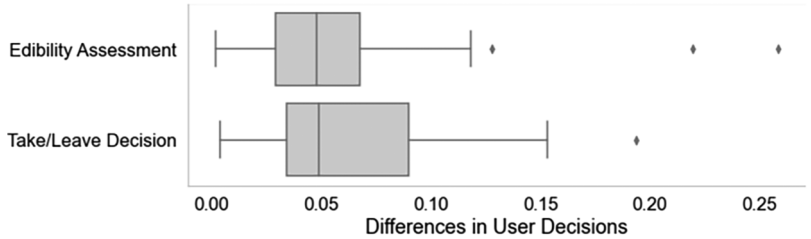


Fig. 13. Boxplots showing the distribution and outliers of differences in participant decisions between the control group and experimental groups (the differences for the three groups are shown in one boxplot). The upper boxplot shows the distribution for the edibility assessment. The lower boxplot shows the distribution for the take/leave decision.

wrongly classifying edible mushrooms as poisonous only results in missing out on a possibly delicious mushroom.

Testing this subset, we found a statistically significant effect for the take/leave decision performance of the group with network-dissection explanations. All other tests did not reach statistical significance (see Table 10).

Based on the effect-size estimates in Figure 14, explanations seem to have had a greater influence on participant decisions to take a mushroom home for consumption than on edibility assessment. For all three experimental groups, the point estimates of the effect sizes for the participants' pick-up decisions (▲) are consistently higher than for the edibility assessments (●). For the subset of poisonous mushrooms, this increased effect of explanations on the more important decision may be desirable to avoid poisoning, although further experiments are needed.

B.2.2 Edible Mushrooms. In this section, we investigate the subset of items that contain only edible mushrooms analogously to the analysis we did for poisonous mushrooms in Section B.2.1. This subset is less critical than the subset of poisonous mushrooms, which involves decisions of the highest risk, but we add it here for the sake of completeness.

Table 9. Lists All Task Items and Their Subgroup Assignments

ID	AI correct	Edibility	Helpful E1	Helpful E2	Helpful E3
1	⛔	☠	✅	⛔	✅
2	⛔	🍴	✅	✅	✅
3	⛔	🍴	⛔	✅	✅
4	✅	☠	✅	✅	✅
5	✅	🍴	✅	✅	✅
6	✅	🍴	✅	✅	✅
7	✅	☠	✅	⛔	✅
8	✅	🍴	✅	✅	✅
9	✅	🍴	✅	⛔	⛔
10	⛔	☠	⛔	⛔	⛔

AI correct indicates whether the AI system correctly classified this mushroom. *Edibility* states the ground-truth edibility of a mushroom. *Helpful* shows for each explanation method (i.e., Grad-CAM (E1), nearest-neighbor (E2), and network-dissection (E3) explanations) whether the combination of explanation and AI classification was helpful (i.e., if (i) the explanation accorded with a correct AI classification or (ii) discorded with an incorrect AI classification).

Table 10. Effects of Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations on Participant Performance in Assessing the Edibility (●) of Mushrooms and the Take/Leave Decision (▲) for the Subset of *Poisonous* Mushrooms

		Control G.		Exp. G.		Dunnett			BM			Effect Size		
	XAI	*M*	*SD*	*M*	*SD*	*SE*	*t*	*p*	*t*	*df*	*p*	*d*	CI_{Lo}	CI_{Hi}
	E1	1.92	0.644	2.01	0.676	0.0831	1.11	0.562	1.2	230	0.116	0.14	−0.101	0.382
●	E2	1.92	0.644	1.77	0.805	0.0841	−1.76	0.197	−2.01	207	0.977	−0.207	−0.452	0.0378
	E3	1.92	0.644	2.05	0.585	0.0831	1.63	0.255	2.06	255	0.02	0.218	−0.0236	0.46
	E1	1.99	0.655	2.14	0.668	0.0835	1.79	0.186	1.69	239	0.0465	0.227	−0.0153	0.468
▲	E2	1.99	0.655	1.96	0.782	0.0846	−0.27	0.988	−0.557	215	0.711	−0.0321	−0.276	0.212
	E3	1.99	0.655	2.21	0.627	0.0835	2.71	**0.0195**	3.03	252	0.00134	0.352	0.109	0.595

Mean (*M*) and standard deviation (*SD*) of user performance for control group (Control G.) and experimental group (Exp. G.) for each explanation method (XAI). Dunnett's standard errors (*SE*), test statistics (*t*), and *p*-values (*p*). Brunner–Munzel's (BM) test statistics (*t*), degrees of freedom (*df*), and *p*-values (*p*). Effect-size point estimations (*d*), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals. *P*-values < 0.05 (adjusted for multiple testing) are indicated in bold.

We found a statistically significant effect for nearest-neighbor (E2) examples, as shown in Table 11. Other tests did not reach statistical significance.

According to the effect-size estimates in Figure 15, as for the subset of poisonous mushrooms, it seems that participants' take/leave decisions (▲) differed considerably from participants' edibility assessments (●). Unlike for the set of poisonous mushrooms, it seems that the effects of XAI methods

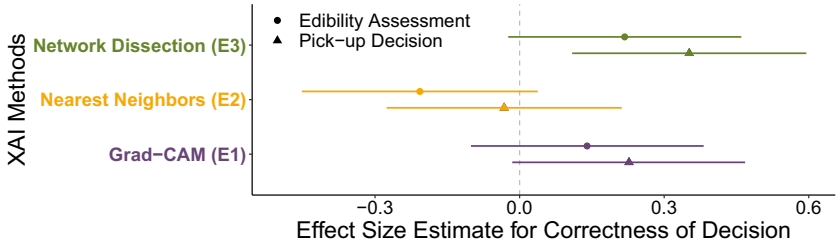


Fig. 14. Effect size point estimates and 95 % confidence interval for participant performance for the subset of *poisonous* mushrooms. Note the discrepancy between the edibility assessment (●) and take/leave decision (▲).

Table 11. Effects of Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations on Participant Performance in Assessing the Edibility (●) of Mushrooms and the Take/Leave Decision (▲) for the Subset of *Edible* Mushrooms

		Control G.		Exp. G.		Dunnett			BM			Effect Size		
	XAI	*M*	*SD*	*M*	*SD*	*SE*	*t*	*p*	*t*	*df*	*p*	*d*	CI_{Lo}	CI_{Hi}
	E1	3.12	0.926	3.18	1.03	0.12	0.529	0.92	0.271	223	0.393	0.0652	−0.176	0.306
●	E2	3.12	0.926	3.62	1.01	0.121	4.19	**0.000102**	4.08	197	3.22×10^{-5}	0.53	0.282	0.778
	E3	3.12	0.926	3.26	0.968	0.12	1.24	0.474	1.44	230	0.0751	0.158	−0.0839	0.399
	E1	2.68	1.04	2.56	0.995	0.131	−0.913	0.701	−1.17	259	0.879	−0.117	−0.358	0.124
▲	E2	2.68	1.04	3.06	1.2	0.133	2.85	**0.0131**	2.43	204	0.00806	0.341	0.0954	0.587
	E3	2.68	1.04	2.75	1.06	0.131	0.52	0.923	0.487	249	0.313	0.0651	−0.176	0.306

Mean (*M*) and standard deviation (*SD*) of user performance for control group (Control G.) and experimental group (Exp. G.) for each explanation method (XAI). Dunnett's standard errors (*SE*), test statistics (*t*), and *p*-values (*p*). Brunner–Munzel's (BM) test statistics (*t*), degrees of freedom (*df*), and *p*-values (*p*). Effect-size point estimations (*d*), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals. *P*-values < 0.05 (adjusted for multiple testing) are indicated in bold.

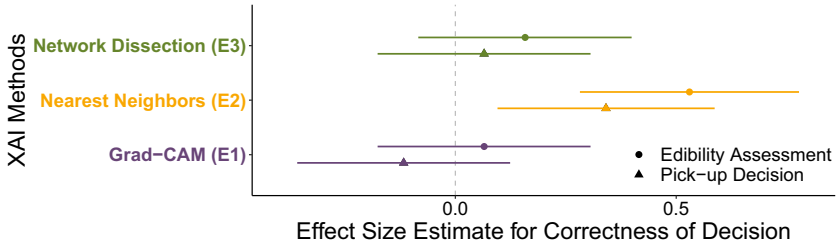


Fig. 15. Effect-size point estimation and 95% confidence interval of participant performance for the subset of *edible* mushrooms.

were more beneficial to the edibility assessment (●) than to the take/leave decision (▲). Since the confidence intervals are large, interpretations should be treated with care.

B.3. Effects on Trust, System Evaluation, and Intention to Use

Table 12 shows the results of the statistical tests for self-reported user trust, evaluation of the *Forestly* app, and intention to use such an app in the future. The effect size intervals for those hypotheses are shown in Figure 16.

Table 12. Effects of Grad-CAM (E1), Nearest-Neighbor (E2), and Network-Dissection (E3) Explanations on Self-Reported User Trust (*Trust*), Evaluation of the *Forestly* App (*Eval*), and Intention to Use Such an App in the Future (*Use*)

		Control G.		Exp. G.		Dunnett			BM			Effect Size		
	XAI	*M*	*SD*	*M*	*SD*	*SE*	*t*	*p*	*t*	*df*	*p*	*d*	CI_{Lo}	CI_{Hi}
trust	E1	3.59	0.75	3.42	0.91	0.107	−1.6	0.269	−1.56	232	0.0602	−0.209	−0.45	0.0333
	E2	3.59	0.75	3.34	1	0.109	−2.31	0.0582	−1.69	198	0.0466	−0.29	−0.535	−0.0444
	E3	3.59	0.75	3.56	0.867	0.107	−0.244	0.991	−0.0154	230	0.494	−0.0327	−0.274	0.208
eval	E1	4.11	0.85	4.04	0.941	0.113	−0.592	0.892	−0.353	244	0.362	−0.0752	−0.316	0.166
	E2	4.11	0.85	3.87	0.991	0.115	−2.13	0.0903	−1.93	212	0.0273	−0.267	−0.513	−0.022
	E3	4.11	0.85	4.09	0.934	0.113	−0.214	0.994	0.0395	234	0.516	−0.0273	−0.268	0.214
use	E1	3.5	1.06	3.33	1.17	0.135	−1.21	0.497	−1.11	234	0.135	−0.148	−0.389	0.0938
	E2	3.5	1.06	3.32	1.18	0.137	−1.28	0.449	−1.07	219	0.143	−0.158	−0.402	0.0867
	E3	3.5	1.06	3.48	1.02	0.135	−0.134	0.998	−0.293	253	0.385	−0.0174	−0.259	0.224

Mean (*M*) and standard deviation (*SD*) of user performance for control group (Control G.) and experimental group (Exp. G.) for each explanation method (XAI). Dunnett's standard errors (*SE*), test statistics (*t*), and *p*-values (*p*). Brunner–Munzel's (BM) test statistics (*t*), degrees of freedom (*df*), and *p*-values (*p*). Effect-size point estimations (*d*), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals.

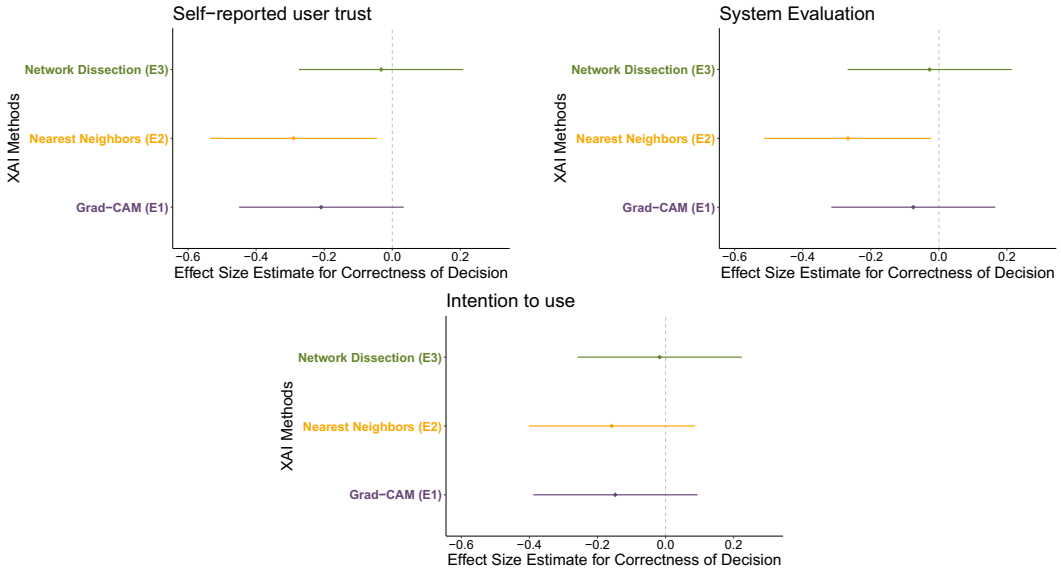


Fig. 16. Effect-size point estimation and 95% confidence interval of self-reported user trust, evaluation of the *Forestly* app, and intention to use such an app in the future.

Table 13 shows the results of the correlation tests between user propensity to trust and self-reported user trust for each group. The correlation point estimates and confidence intervals are shown in Figure 8.

Table 13. Correlation Estimates between User Propensity to Trust Automated Systems in General and Self-Reported User Trust in the *Forestly* App Presented in This Study by User Group

Group	t	df	p	corr	CI_{Lo}	CI_{Hi}
C	5.34	153	3.00×10^{-7}	0.396	0.254	0.521
E1	4.1	115	7.65×10^{-5}	0.357	0.188	0.506
E2	2.25	110	0.0267	0.209	0.0248	0.38
E3	2.41	115	0.0177	0.219	0.039	0.385

Pearson's product-moment correlation test statistics (t), degrees of freedom (df), and p-values (p) for each user group. Correlation point estimations (cor), lower (CI_{Lo}), and upper (CI_{Hi}) confidence intervals.

Acknowledgements

We thank Moritz Heckmann for helping with the implementation of the *AI Forest—The Schwammerl Hunting Game* and Stefan Eibelwimmer for the graphic design of the game. We thank Dr. Otto Stoik, the members of the Mycological Working Group (MYAG) at the Biology Center Linz, Austria, and the German Mycological Society (DGfM) for providing mushroom images for this study. Finally, we thank Alfio Ventura for helping with the study setup.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 31. Retrieved from https://papers.nips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[2] Gulsum Alicioglu and Bo Sun. 2022. A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics* 102 (Feb. 2022), 502–520. DOI: https://doi.org/10.1016/j.cag.2021.09.002

[3] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: A user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. 275–285. DOI: https://doi.org/10.1145/3377325.3377519

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (Jun. 2020), 82–115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. 3319–3327. DOI: https://doi.org/10.1109/CVPR.2017.354

[6] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences* 117, 48 (Dec. 2020), 30071–30078. DOI: https://doi.org/10.1073/pnas.1907375117

[7] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. DOI: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[8] Jacob Bien and Robert Tibshirani. 2011. Prototype selection for interpretable classification. *The Annals of Applied Statistics* 5, 4 (Dec. 2011), 2403–2424. DOI: https://doi.org/10.1214/11-AOAS495

[9] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 807–819. DOI: https://doi.org/10.1145/3490099.3511139

[10] William E. Brandenburg and Karlee J. Ward. 2018. Mushroom poisoning epidemiology in the United States. *Mycologia* 110, 4 (2018), 637–641. DOI: https://doi.org/10.1080/00275514.2018.1479561

[11] Edgar Brunner and Ullrich Munzel. 2000. The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal* 42, 1 (2000), 17–25. DOI: https://doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U

[12] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464. DOI: https://doi.org/10.1145/3377325.3377498

[13] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. 258–262. DOI: https://doi.org/10.1145/3301275.3302289

[14] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. 2020. The state of the art in enhancing trust in machine learning models with the use of visualizations. *Computer Graphics Forum* 39, 3 (Jun. 2020), 713–756. DOI: https://doi.org/10.1111/cgf.14034

[15] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? A case study in model-in-the-loop prediction. arXiv:2007.12248. Retrieved from http://arxiv.org/abs/2007.12248

[16] Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences. Routledge*.

[17] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological Science* 25, 1 (Jan. 2014), 7–29. DOI: https://doi.org/10.1177/0956797613504966

[18] Piotr Dabkowski and Yarin Gal. 2017. Real time image saliency for black box classifiers. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 30, 6970–6979. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/0060ef47b12160b9198302ebdb144dcf-Abstract.html

[19] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern Statistical Methods for HCI*. Judy Robertson and Maurits Kaptein (Eds.), 291–330. DOI: https://doi.org/10.1007/978-3-319-26633-6_13

[20] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Communications of the ACM* 63, 1 (Dec. 2019), 68–77. DOI: https://doi.org/10.1145/3359786

[21] Charles W. Dunnett. 1955. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50, 272 (Dec. 1955), 1096–1121. DOI: https://doi.org/10.1080/01621459.1955.10501294

[22] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6. DOI: https://doi.org/10.1145/3411763.3441342

[23] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O. Riedl. 2022. Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–7. DOI: https://doi.org/10.1145/3491101.3503727

[24] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with AI systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18. DOI: https://doi.org/10.1145/3491102.3517734

[25] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8 (Oct. 2020), 43–52. DOI: https://doi.org/10.1609/hcomp.v8i1.7462

[26] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (2007), 175–191. DOI: https://doi.org/10.3758/BF03193146

[27] Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33, 3 (Oct. 1973), 613–619. DOI: https://doi.org/10.1177/001316447303300309

[28] Raymond Fok and Daniel S. Weld. 2023. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. arXiv:2305.07722. Retrieved from https://doi.org/10.48550/ARXIV.2305.07722 arXiv.

[29] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*. 1050–1059. Retrieved from https://proceedings.mlr.press/v48/gal16.html

[30] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM Computing Surveys* 51, 5 (Sep. 2019), 1–42. DOI: https://doi.org/10.1145/3236009

[31] Francisco Gutiérrez, Nyi Nyi Htun, Vero Vanden Abeele, Robin De Croon, and Katrien Verbert. 2022. Explaining call recommendations in nursing homes: A user-centered design approach for interacting with knowledge-based health decision support systems. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*. 162–172. DOI: https://doi.org/10.1145/3490099.3511158

[32] Lewis G. Halsey. 2019. The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? *Biology Letters* 15, 5 (May 2019), 20190174. DOI: https://doi.org/10.1098/rsbl.2019.0174

[33] Glenda Hannibal, Astrid Weiss, and Vicky Charisi. 2021. "The robot may not notice my discomfort" – Examining the experience of vulnerability for trust in human-robot interaction. In *Proceedings of the 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN '21)*. 704–711. DOI: https://doi.org/10.1109/RO-MAN50785.2021.9515513

[34] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sep. 2023), 1–29. DOI: https://doi.org/10.1145/3610067

[35] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18. DOI: https://doi.org/10.1145/3544548.3581025

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV '15)*. 1026–1034. DOI: https://doi.org/10.1109/ICCV.2015.123

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

[38] Kevin A. Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434. DOI: https://doi.org/10.1177/0018720814547570

[39] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for explainable AI: challenges and prospects. arXiv:1812.04608. Retrieved from https://doi.org/10.48550/arXiv.1812.04608 [cs] version: 2.

[40] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 32. Retrieved from https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html

[41] Christina Humer, Andreas Hinterreiter, Benedikt Leichtmann, Martina Mara, and Marc Streit. 2022. Effects of Explainable Artificial Intelligence Methods on Human Trust and Behavior: A Comparison of Nearest Neighbor, Grad-CAM, and Network Dissection. DOI: https://doi.org/10.17605/OSF.IO/SD953

[42] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. 2020. How can i explain this to you? An empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems* 33 (2020), 4211–4222.

[43] Zhicheng Jiao, Ji W. Choi, Kasey Halsey, Thi M. L. Tran, Ben Hsieh, Dongcui Wang, Feyisope Eweje, Robin Wang, Ken Chang, Jing Wu, Scott A. Collins, Thomas Y. Yi, Andrew T. Delworth, Tao Liu, Terrance T. Healey, Shaolei Lu, Jianxin Wang, Xue Feng, Michael K. Atalay, Li Yang, Michael Feldman, Paul J. L. Zhang, Wei-Hua Liao, Yong Fan, and Harrison X. Bai. 2021. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: A retrospective study. *The Lancet Digital Health* 3, 5 (2021), e286–e294. DOI: https://doi.org/10.1016/S2589-7500(21)00039-X

[44] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2023. The XAI alignment problem: Rethinking how should we evaluate human-centered AI explainability techniques. arXiv:2303.17707. Retrieved from http://arxiv.org/abs/2303.17707 [cs].

[45] Matthew W. Johnson and Warren K. Bickel. 2002. Within-subject comparison of real and hypothetical money rewards in delay discounting. *Journal of the Experimental Analysis of Behavior* 77, 2 (2002), 129–146. DOI: https://doi.org/10.1901/jeab.2002.77-129

[46] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14. DOI: https://doi.org/10.1145/3313831.3376219

[47] Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence* 294 (May 2021), 103459. DOI: https://doi.org/10.1016/j.artint.2021.103459

[48] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help me help the AI": Understanding how explainability can support human-AI interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17. DOI: https://doi.org/10.1145/3544548.3581001

[49] Pang W. Koh, Thao Nguyen, Yew S. Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*. 5338–5348.

[50] Johannes Kraus, David Scholz, Dina Stiegemeier, and Martin Baumann. 2020. The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Journal of the Human Factors and Ergonomics Society* 62, 5 (2020), 718–736.

[51] Moritz Körber. 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA '18)*. Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.), Vol. 823. 13–30. DOI: https://doi.org/10.1007/978-3-319-96074-6_2

[52] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. 1–18. DOI: https://doi.org/10.1145/3491102.3501999

[53] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85. DOI: https://doi.org/10.1145/3375627.3375833

[54] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. DOI: https://doi.org/10.2307/2529310

[55] Retno Larasati, Anna De Liddo, and Enrico Motta. 2020. The effect of explanation styles on user's trust. In *Proceedings of the 2020 Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies, 2582*.

[56] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Journal of the Human Factors and Ergonomics Society* 46, 1 (Mar. 2004), 50–80. DOI: https://doi.org/10.1518/hfes.46.1.50_30392

[57] Benedikt Leichtmann, Andreas Hinterreiter, Christina Humer, Marc Streit, and Martina Mara. 2023. Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human–Computer Interaction* 0, 0 (Jun. 2023), 1–18. DOI: https://doi.org/10.1080/10447318.2023.2221605

[58] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (Feb. 2023), 107539. DOI: https://doi.org/10.1016/j.chb.2022.107539

[59] Benedikt Leichtmann, Thomas Meneweger, Christine Busch, Bernhard Reiterer, Kathrin Meyer, Daniel Rammer, Roland Haring, and Martina Mara. 2024. Teaming with a robot in mixed reality: Dynamics of trust, self-efficacy, and mental models affected by information richness. *International Journal of Human–Computer Interaction* (Apr. 2024), 1–18. DOI: https://doi.org/10.1080/10447318.2024.2331878

[60] Geert Litjens, Thijs Kooi, Babak E. Bejnordi, Arnaud A. A. Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42 (Dec. 2017), 60–88. DOI: https://doi.org/10.1016/j.media.2017.07.005

[61] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (ACM Digital Library)*, Regina Bernhaupt (Ed.), 1–16. DOI: https://doi.org/10.1145/3313831.3376727

[62] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 4768–4777.

[63] Michael R. Maniaci and Ronald D. Rogge. 2014. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality* 48 (2014), 61–83. DOI: https://doi.org/10.1016/j.jrp.2013.09.008

[64] Stephanie M. Merritt and Daniel R. Ilgen. 2008. Not all trust Is created equal: Dispositional and history-based trust in human-automation interactions. *Journal of the Human Factors and Ergonomics Society* 50, 2 (Apr. 2008), 194–210. DOI: https://doi.org/10.1518/001872008X288574

[65] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. DOI: https://doi.org/10.1016/j.artint.2018.07.007

[66] Christoph Molnar. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). Retrieved from christophm.github.io/interpretable-ml-book/

[67] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the impact of human explanation strategies on human-AI visual decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (Apr. 2023), 1–37. DOI: https://doi.org/10.1145/3579481

[68] Davy T. K. Ng, J. K. L. Leung, K. W. S. Chu, and Maggie S. Qiao. 2021. AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology* 58, 1 (2021), 504–509. DOI: https://doi.org/10.1002/pra2.487

[69] Mahsan Nourani, Donald R. Honeycutt, Jeremy E. Block, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan, and Vibhav Gogate. 2020. Investigating the importance of first impressions and explainable AI with interactive video analysis. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8. DOI: https://doi.org/10.1145/3334480.3382967

[70] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer W. Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52. DOI: https://doi.org/10.1145/3411764.3445315

[71] Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*. 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

[72] RStudio Team. 2020. RStudio: Integrated Development Environment for R. Retrieved March 29, 2022 from http://www.rstudio.com/.

[73] Michael Sailer, Jan Hense, Heinz Mandl, and Markus Klevers. 2013. Psychological perspectives on motivation through gamification. *Interaction Design & Architecture Journal*, 19, 28–37.

[74] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 617–626. DOI: https://doi.org/10.1145/3514094.3534128

[75] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422. DOI: https://doi.org/10.1145/3581641.3584066

[76] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV '17)*. 618–626. DOI: https://doi.org/10.1109/ICCV.2017.74

[77] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR) Workshop Track Proceedings*.

[78] Statens Naturhistoriske Museum et al. 2022. Danmarks svampeatlas. https://svampe.databasen.org/ Accessed March 29, 2022.

[79] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 77–87. DOI: https://doi.org/10.1145/3450613.3456817

[80] Aybike Ulusan, Uttkarsh Narayan, Sam Snodgrass, Ozlem Ergun, and Casper Harteveld. 2022. "rather solve the problem from scratch": Gamesploring human-machine collaboration for optimizing the debris collection problem. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*. 604–619. DOI: https://doi.org/10.1145/3490099.3511163

[81] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (Feb. 2021), 103404. DOI: https://doi.org/10.1016/j.artint.2020.103404

[82] Xinru Wang and Ming Yin. 2021. Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. 318–328. DOI: https://doi.org/10.1145/3397481.3450650

[83] Mengjiao Yang and Been Kim. 2019. Benchmarking attribution methods with relative feature importance. arXiv:1907.09701 [cs, stat]. Retrieved from http://arxiv.org/abs/1907.09701

[84] Qiaoning Zhang, Matthew L. Lee, and Scott Carter. 2022. You complete me: Human-AI teams and complementary expertise. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI '22)*. 1–28. DOI: https://doi.org/10.1145/3491102.3517791

[85] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20)*. 295–305. DOI: https://doi.org/10.1145/3351095.3372852

Received 3 May 2023; revised 15 February 2024; accepted 17 April 2024