Computers in Human Behavior 139 (2023) 107539

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task

Benedikt Leichtmann [a,*], Christina Humer [b], Andreas Hinterreiter [b], Marc Streit [b], Martina Mara [a]

[a] LIT Robopsychology Lab, Johannes Kepler University Linz, Altenberger Straße 69, Linz, 4040, Austria
[b] Visual Data Science Lab, Institute of Computer Graphics, Johannes Kepler University Linz, Altenberger Straße 69, Linz, 4040, Austria

ARTICLE INFO

Keywords:
XAI
AI literacy
Domain-specific knowledge
Mushroom identification
Trust calibration
Visual explanation

ABSTRACT

Understanding the recommendations of an artificial intelligence (AI) based assistant for decision-making is especially important in high-risk tasks, such as deciding whether a mushroom is edible or poisonous. To foster user understanding and appropriate trust in such systems, we assessed the effects of explainable artificial intelligence (XAI) methods and an educational intervention on AI-assisted decision-making behavior in a 2×2 between subjects online experiment with $N = 410$ participants. We developed a novel use case in which users go on a virtual mushroom hunt and are tasked with picking edible and leaving poisonous mushrooms. Users were provided with an AI-based app that showed classification results of mushroom images. To manipulate explainability, one subgroup additionally received attribution-based and example-based explanations of the AI's predictions; for the educational intervention one subgroup received additional information on how the AI worked. We found that the group that received explanations outperformed that which did not and showed better calibrated trust levels. Contrary to our expectations, we found that the educational intervention, domain-specific (i.e., mushroom) knowledge, and AI knowledge had no effect on performance. We discuss practical implications and introduce the mushroom-picking task as a promising use case for XAI research.

1. Introduction

Imagine a tourist on vacation in Austria joining locals in mushroom picking to explore the country's nature. Since the tourist has not previously gone mushroom picking in Austria, they decide that some kind of classification aid would be helpful. The obvious choice would be to use a mushroom classification book. Due to the vast variety of mushrooms that grow in Austria, however, the tourist decides to use an app that is based on an artificial intelligence (AI) and aims to identify edible and inedible/poisonous mushrooms from images.[1] Discriminating mushrooms is a challenging task, as many poisonous mushrooms closely resemble edible ones. If a person is not knowledgeable about Austrian mushrooms, they might easily pick an inedible mushroom that was (for whatever reason) incorrectly classified as edible by the AI-driven app. In such a scenario the user trusts the system too much (*overtrust*), which results in a potentially dangerous error.

Image classification has long been a standard use case of AI (Lu & Weng, 2007; Rawat & Wang, 2017). While classification errors may be harmless in some contexts—such as the animal identification examples that are often used as a proof of concept—, such errors can have serious consequences in high-stakes domains, including mushroom picking (see reports of mushroom poisoning, e.g., Brandenburg & Ward, 2018) and medical image analysis (Jiao et al., 2021; Litjens et al., 2017). In order to avoid errors in AI-assisted decision-making tasks (i.e., too low acceptance of reliable classification systems and too high acceptance of wrong AI classifications), humans must understand how an AI arrives at a particular decision and must be able to evaluate whether a decision is reasonable or seems odd (e.g., Körber, Baseler, & Bengler, 2018; Kraus, Scholz, Stiegemeier, & Baumann, 2020; Long & Magerko, 2020; Parasuraman & Riley, 1997). It is thus important to understand the system's functionality and limitations for trust calibration. Trust evolves and adapts over time with growing knowledge until, ideally, "a user's trust in a system corresponds exactly with its objective capabilities and its actual performance" (Kraus et al., 2020, p. 719). Such a level of system comprehension can be achieved by (*i*) enabling AI systems to explain their decisions to users, which is often referred to as *explainable AI* or *XAI* (see e.g., Long & Magerko, 2020; Miller, 2019); or (*ii*) improving

* Corresponding author.
E-mail addresses: benedikt.leichtmann@jku.at (B. Leichtmann), christina.humer@jku.at (C. Humer), andreas.hinterreiter@jku.at (A. Hinterreiter), marc.streit@jku.at (M. Streit), martina.mara@jku.at (M. Mara).

[1] Note that this is an imaginary example for the purpose of illustration only. The safe identification of mushrooms is a difficult task that requires much expertise and training. Professional mycological associations therefore expressly warn against mushroom picking without the necessary expertise.

https://doi.org/10.1016/j.chb.2022.107539
Received 2 June 2022; Received in revised form 29 September 2022; Accepted 15 October 2022
Available online 26 October 2022

humans' AI literacy such that they better comprehend AI systems, their decisions, and limitations (see e.g., Long & Magerko, 2020; Moehring, Schroeders, Leichtmann, & Wilhelm, 2016; Ng, Leung, Chu, & Qiao, 2021). However, the evidence on the effects of these approaches in the context of human–AI interaction is still scarce. User research on XAI is in its infancy, and effects of respective measures are either unknown or varying, with potential influences of moderators that still need to be explored (Ehsan et al., 2021, 2022)—which, as discussed later, is also one of the takeaways of this study. Therefore, more research is required about user comprehension of and trust in AI-based classification systems—especially for high-risk decision tasks—and about how interventions such as user-centered system design and user training can affect these criteria of comprehension and trust. Researchers studying human-centered XAI (HCXAI) have called for more user studies about behavior and trust of end-users within different relevant contexts (e.g., conference workshops on HCXAI, Ehsan, Wintersberger et al., 2021; Ehsan et al., 2022).

Psychological research can help in this endeavor from two prominent angles and research traditions: *engineering psychology*, which is concerned with adequacy of system design (e.g., Carayon, 2006); and *psychology of learning and individual differences*, which is concerned with human knowledge, comprehension and problem-solving (e.g., Greiff, Kretzschmar, Müller, Spinath, & Martin, 2014; Moehring et al., 2016) and how these can be improved effectively by educational interventions (Pangrazio, Godhe, & Ledesma, 2020). Building on these research traditions, this article presents the findings of an exploratory study of the interactions between human users and an AI-based classification system in a high-risk decision task. In an experiment, we explored the effectiveness of an educational intervention converging how AI systems work (AI literacy research tradition) and the effect of visual explanations (XAI research tradition) on user decision-making, comprehension, and trust. Our study had the following objectives:

1. Explore the effects of an XAI interface with attribution-based and example-based visual explanations by comparing it with an interface that does not explain its recommendations.
2. Explore the effects of an educational intervention presented as texts about and illustrations of how machine learning works by comparing two groups—one with and one without such an intervention.
3. Understand the underlying psychological mechanisms by assessing users' prior knowledge of AI, trust, and domain-specific knowledge about mushrooms as major factors commonly used to explain differences in comprehension and system use.
4. Introduce mushroom picking as a new use case that can be employed for psychological research into comprehension and trust in AI-based classification systems as well as decision-making in high-risk contexts.

2. Theory and related work

The term "artificial intelligence" has a long history and its meaning has shifted throughout the last decades (Dick, 2019). Nowadays, AI is often associated with powerful *machine learning* models based on neural networks (Dick, 2019). Such models have been applied in a wide variety of contexts, ranging from healthcare (Davenport & Kalakota, 2019; Yu, Beam, & Kohane, 2018) and drug discovery (Vamathevan et al., 2019) to human resources (Votto, Valecha, Najafirad, & Rao, 2021) and education (Zawacki-Richter, Marín, Bond, & Gouverneur, 2019). For this study, we focused on image classification, one of the most prominent tasks for which machine learning models with high accuracy have been developed (Rawat & Wang, 2017).

The different research perspectives on AI are as diverse as its application areas. While works from computer science and application domains often focus on algorithmic accuracy, questions about the ethical or socio-technical ramifications of AI are gaining importance (Floridi et al., 2018; Morley et al., 2020). Our focus lies on understanding human behavior in an AI-assisted decision-making process from a user-centered perspective.

As noted in the introduction (Section 1), this work was concerned with researching the influence of increasing user comprehension of the AI system on performance in AI-assisted decision tasks. We did this coming from two different directions: educating the user on the one hand, and enabling systems to explain their decisions on the other.

Therefore, the first part of the theory section focuses on user comprehension. Discussing literacy concepts, we describe how domain-specific knowledge and knowledge about the technology being used to solve problems are important for task success. We sketch how knowledge about AI technology could be tackled by educational interventions.

In the second part, we describe how interface design can improve user understanding during human–technology interaction. In doing so, we discuss recent developments in visual explanations for AI-based decisions that aim to open up so-called "black-box" models, which are particularly prevalent in deep learning.

Finally, as trust is a key construct in understanding the dynamics in AI-assisted decision-making, we describe how comprehension and trust are related, and might therefore be affected by both educational and system-design interventions.

2.1. Comprehending system outcomes: Improving users' AI literacy

As previously mentioned, one possible intervention from a human factors perspective is to educate users to better comprehend AI technology (i.e., educational interventions). The general ability of users to comprehend AI technology and the ability to use it are often labeled *AI literacy*. While AI literacy is hard to define, it can be described as a complex construct based on the integration of various competencies, including, for example, knowing and understanding AI (e.g., knowing basic AI functions) (Long & Magerko, 2020; Ng et al., 2021).

Literacy constructs have not only been used in the context of AI. Several similar skill sets and closely related competencies in combination with (information) technology are conceptualized as partly overlapping literacies. Examples are information and communications technology (ICT) literacy (e.g., "the ability to understand and work with modern information technology, such as the Internet" Moehring et al., 2016, p. 171), digital literacy (e.g., "skill, knowledge, ethics, and creative outputs in the digital network environment" Covello & Lei, 2010, p. 3), and information literacy (e.g., "the ability to find sources, analyze, and synthesize the material, and evaluate the source credibility" Moehring et al., 2016, p. 171). Due to overlaps, a multitude of labels, and imprecise definitions, it is difficult to describe how these constructs may relate to each other. For example, while for Long and Magerko (2020) digital literacy is a prerequisite for AI literacy, Moehring et al. (2016) interpreted digital literacy more as an "overarching concept encompassing several disciplines" (p. 171 Moehring et al., 2016).

However, these constructs have in common their overlaps with traditional comprehension abilities and the associated cognitive requirements (Moehring et al., 2016). The traditional comprehension construct can be defined as "the process of developing mental representations, by which prior long-term knowledge is incorporated with the available information given through text, audio, or video through complex mental processes" (Moehring et al., 2016, p. 171).

It has bee shown (e.g., Moehring et al., 2016; Schroeders, Wilhelm, & Bucholtz, 2010) that comprehension ability can be explained to a large extent by underlying cognitive abilities, such as fluid intelligence and domain-specific knowledge, and as these constructs are independent of presentation of information, comprehension can be viewed as unidimensional. For example, Schroeders et al. (2010) reported that listening, reading and viewing comprehension are highly correlated and all three are influenced by crystallized intelligence (measured as declarative knowledge items). This can be explained by the theoretical

assumption that knowledge facilitates information retrieval (Moehring et al., 2016). It is thus reasonable to assume that such traditional comprehension constructs also apply to other literacies that require users to search and integrate necessary information in order to answer test questions or make decisions.

For example, Moehring et al. (2016) tested whether fluid intelligence and crystallized intelligence—or, more concretely, domain-specific knowledge—could explain variance in an ecological momentary assessment of digital literacy where information had to be retrieved from the internet—a task that requires more complex processes, such as evaluating, weighting, and integrating potentially contradicting information and evaluating the credibility of an information source (Moehring et al., 2016). Furthermore, it is conceivable that participants' computer usage could also influence digital literacy. However, results showed that computer usage had only a negligibly small influence on the comprehension construct, and variance was explained mostly by domain-specific knowledge and fluid intelligence (Moehring et al., 2016). In contrast to this, based on work by Greiff et al. (2014) it could be argued that the influence of such technology-related competencies or knowledge could vary depending on the level of complexity of human–computer interaction requirements of the task (Greiff et al., 2014).

Similarly, by analyzing the definitions of AI literacy (Long & Magerko, 2020; Ng et al., 2021) and its theoretical relations to other literacy constructs, it could be assumed that AI literacy can also be explained by domain-specific knowledge (i.e., mushroom knowledge when a mushroom classifier is used) and fluid intelligence. However, the use of AI as part of problem-solving or decision-making (e.g., declaring a mushroom as edible when it is in fact edible) also differs from traditional comprehension tests. While, as shown by Moehring et al. (2016), searching for information on the internet required additional cognitive processes such as evaluating the credibility of a source or integrating information from multiple documents, the task of correctly identifying mushrooms as edible, inedible, or poisonous with an AI-based assistant system also requires evaluating the system's output or the credibility of the system as an independent source of information in general. It is thus conceivable that users need a certain degree of AI background knowledge and knowledge about the system's functionality. This could mean that system knowledge, as a facet of AI literacy, is a variable (and a precondition) for system comprehension in interaction situations.

Especially in the human factors literature, the need for appropriate knowledge and understanding of a new technology, such as assistant systems and autonomous vehicles are essential for the development of trust and appropriate usage of the system (see Körber et al., 2018; Sauer, Chavaillaz, & Wastell, 2016). For example, Körber et al. (2018) showed that "introductory information" about an automated vehicle before a driving task had a moderate effect on participant trust levels and reliance behavior, as reflected in lower collision rates with an obstacle for an informed group of participants in a driving simulator study.

2.2. Explainable AI: Opening the black box

In addition to educational attempts to improve AI literacy, another means of establishing an adequate amount of system comprehension is human-centered design of AI systems that can explain their decisions to foster human comprehension.

Most state-of-the-art machine learning models—in particular deep neural networks (DNNs)—are based on vast amounts of data and deeply nested mathematical functions. Such models are considered to be "black boxes" because it is difficult for humans to understand how and why they arrive at particular decisions. Explainable artificial intelligence (XAI) methods are commonly used to make black-box models more understandable to human users (Alicioglu & Sun, 2022; Barredo Arrieta et al., 2020; Miller, 2019). While models exist that are intrinsically explainable (i.e., models with high degree of transparency; Barredo Arrieta et al., 2020; Došilović, Brčić, & Hlupić, 2018), they often come with the drawback of producing less accurate results. Therefore, post-hoc explanation techniques are employed to explain existing models (Došilović et al., 2018; Lipton, 2018).

The substantial body of work on XAI contains several attempts to classify the various techniques (Došilović et al., 2018; Du, Liu, & Hu, 2019; Guidotti et al., 2019; Molnar, 2022; Murdoch, Singh, Kumbier, Abbasi-Asl, & Yu, 2019), for example into global and local methods. Global explanation methods give insights into the model as a whole, whereas local methods yield insights about individual data items (i.e., instances). XAI techniques can also be grouped based on whether they are model-agnostic (i.e., apply to all models) or model-specific (i.e., apply only to specific types of models).

The AI model used in our study to identify different types of mushrooms is a deep convolutional neural network (CNN). Below we give an overview of existing domain-specific XAI techniques for CNNs, which we grouped into feature-based, attribution-based, and example-based techniques.

Feature-based. Deep neural networks transform their input into a numerical representation (also called *features*) that is derived automatically during training. Visualizations of these features have shown that early layers of image classification networks learn low-level concepts such as edges and simple patterns, while later layers specialize in detecting high-level concepts (Olah, Mordvintsev, & Schubert, 2017; Zeiler & Fergus, 2014). In a similar fashion, network dissection (Bau et al., 2020) aims to explain to which feature or concept a certain unit corresponds. In the case of a CNN, "unit" refers to a single convolution kernel, which can be thought of crudely as a pattern detector.

Attribution-based. In the domain of image classification, *attribution*-based explanations are commonly used to explain which regions of an image were important for a model's decision. The most important regions are typically visualized directly within the image. Examples of such techniques are saliency maps (Simonyan, Vedaldi, & Zisserman, 2014), deconvnet (Zeiler & Fergus, 2014), Grad-CAM (Selvaraju et al., 2017), LIME (Ribeiro, Singh, & Guestrin, 2016), and SHAP (Lundberg & Lee, 2017).

Example-based. In this type of explanation techniques, certain data items are picked out as examples and shown to users. Such example instances can be prototypes (Bien & Tibshirani, 2011)—typical representatives of particular classes—or counter-examples that differ from these prototypical instances. Nearest-neighbor techniques are used to extract training instances that are closest to a specific target data point (Cai, Jongejan, & Holbrook, 2019; Jeyakumar, Noor, Cheng, Garcia, & Srivastava, 2020). Prior work on example-based explanations suggests that humans prefer example-based methods to other explanations (Cai et al., 2019; Jeyakumar et al., 2020; Kim, Khanna, & Koyejo, 2016).

Most of these XAI techniques were originally developed by domain experts, such as AI model developers, in an attempt to better understand their own creations. Thus, they are typically not optimized for consumption by lay end users, and knowledge about their effectiveness when used by this target audience remains limited.

While a number of empirical studies have tested (*i*) the effectiveness of different XAI techniques with human participants and/or (*ii*) the underlying processes (e.g., Buçinca, Lin, Gajos, & Glassman, 2020; Cai et al., 2019; Kenny, Ford, Quinn, & Keane, 2021; van der Waa, Nieuwburg, Cremers, & Neerincx, 2021; Wang & Yin, 2021; Yang, Huang, Scholtz, & Arendt, 2020; Zhang, Liao, & Bellamy, 2020), the field is still in its infancy. Buçinca et al. (2020), for instance, explored the effects of an example-based explanation and an attribution-based explanation technique using a task in which participants had to decide whether the food shown in images had a specific amount of fat. They found that explanations led to higher trust levels, higher ratings of

understanding, and better performance (higher accuracy). However, they also showed that results were inconsistent when compared to a second experiment with different task instructions (proxy task).

Additionally, taking together results from other studies, no clear conclusions can be drawn, as studies (*i*) are scarce and recent, (*ii*) show mixed results, (*iii*) vary in quality, as also criticized by Kenny et al. (2021), (*iv*) vary in contexts, for instance artificial and abstract tasks such as number recognition (Kenny et al., 2021), nutrition (Buçinca et al., 2020), and video game play (Huber, Weitz, André, & Amir, 2021), and (*v*) vary in study design and the constructs measured. It is thus hard to draw summary conclusions—a general problem of behavioral studies as emphasized by Meehl (1990). One example of a potential moderator causing different results in the literature could be differences in AI error rates (Kenny et al., 2021), amongst many other possible candidates. Differences between study designs are not negative, but they make summaries difficult, which in turn shows that research is at an early stage. Additionally, generalizations are difficult, and effects need to be retested for every new use case. Clear theory and accurate predictions may not be possible at this point, which calls for exploration as a first step toward theory building. During this exploratory experimentation step, key variables, such as moderators of effects, can already be identified by comparing exploratory research results between studies, and on the basis of this, theoretical models with more complex variable relationships beyond simple linear models (e.g., on human behavior and trust) can eventually be formed in the future—a goal echoed in many behavioral disciplines (e.g. Leichtmann, Nitsch, & Mara, 2022; Miranda, Berente, Seidel, Safadi, & Burton-Jones, 2022; Scheel, Tiokhin, Isager, & Lakens, 2021).

2.3. Comprehension and trust calibration in human–machine interaction

The psychological construct of trust is a key variable in models of human–technology interaction such as automated driving, human–robot interaction and other assistant systems—and consequently also in AI-based systems (see Hoff & Bashir, 2015; Körber, 2019; Lee & See, 2004; Parasuraman & Riley, 1997). According to current literature, trust may be defined as an attitudinal construct that describes the willingness of users to be vulnerable to the actions of some automated system to achieve some goal (Hannibal, Weiss, & Charisi, 2021; Lee & See, 2004), and such trust may be dynamically adapted (Kraus et al., 2020). Note that a situation of vulnerability and uncertainty is key to studying trust, as they are a prerequisite for trust (Hannibal et al., 2021; Lee & See, 2004).

However, a high level of trust cannot be a goal in itself. Rather, an appropriate level of trust—so-called calibrated trust—that also corresponds with the competencies of the system should be achieved (e.g., Kraus et al., 2020). System designers thus aim to develop systems which can establish an amount of trust that is neither too low nor too high. Insufficient trust may cause humans to avoid using a system despite it benefitting the task outcome (distrust). Excessive trust may lead to errors and negative consequences when a human user trusts an automated system more than its actual competences and boundaries would justify (overtrust). In the domain of automated driving, this could mean trusting an automated system without intervening (e.g., breaking or steering manually), although a user intervention would be necessary to respond to system failure and/or errors in order to avoid an accident (e.g., Körber et al., 2018; Kraus et al., 2020).

In our mushroom-picking example, calibrated trust means knowing in what situations and how the system can add value by delivering additional information that can be trusted, without the user overtrusting and picking recommended mushrooms in situations in which a system failure should be considered (e.g., when low quality of the input image indicates that the AI is likely unable to classify the mushroom reliably). This requires understanding how the system works and where its boundaries are (Körber et al., 2018), which can be achieved by user education (Körber et al., 2018) or by system transparency (Kraus et al., 2020).

3. Research question

Based on the literature review in the previous section, this section derives open questions that we addressed by exploratory experimentation.

3.1. Exploring the effects of visual explanations and an educational intervention about AI on user performance, trust and comprehension

As stated in the previous sections, there are two main routes toward user comprehension of AI-based decisions to achieve improved performance and calibrated trust levels: improving users' general understanding of AI systems (i.e., AI literacy), and enabling the AI system to explain its decisions (i.e., AI explainability).

The goal of this study was to investigate the effects of (*i*) an educational intervention about AI systems, and (*ii*) visual explanations on user cognition and behavior in terms of performance, trust, and comprehension. While some initial studies on the effects of educational interventions already exist in the human factors literature (Körber et al., 2018) and some can be found on the effects of visual explanations (e.g., Buçinca et al., 2020; Kenny et al., 2021), overall, the literature remains sparse, and further research is required because of mixed results and varying quality, contexts, and study designs. Thus a robust prediction of effects is not possible at this point.

Due to this lack of theory and robust data, we as AI researchers might simply not be ready to test specific hypotheses for certain contexts—a conclusion that is not unheard of in other research fields, such as psychology (Scheel et al., 2021) and human–robot interaction (Leichtmann et al., 2022). Scheel et al. (2021) suggested that, before testing hypotheses confirmatorily, prior steps such as exploratory experimentation can help to discover relationships between variables and to explore boundaries on the basis of which more concrete theoretical assumptions can be made. Only then can specific hypotheses be tested with corresponding power. Therefore, we decided to choose exploratory experimentation as the basis of inquiry in our study.

To promote the development of stronger theory, we additionally included some of the most central variables known from the psychological literature. Aside from task performance as a central variable, trust in the AI system also plays an important role, especially in situations of vulnerability (Hannibal et al., 2021), as discussed in Section 2.3. Under particular circumstances, for instance in high-risk contexts with uncertainty, task performance might also reflect the construct of trust in the form of reliance. This is especially evident in situations where the AI makes an incorrect recommendation and overtrust causes the user to follow it. We thus explored the effects of visual explanations and of an educational intervention on task performance and trust; in order to explore overtrust, we paid particular attention to the situation in which the AI classified incorrectly.

Therefore, as a first step this work aims to explore (*i*) the effects (direction and size) of a concrete combination of XAI methods, and (*ii*) the effects of a concrete realization of a simple image- and text-based educational intervention on behavior and trust in a new high-risk decision task (see objectives 1. and 2. in Section 1).

As known from traditional comprehension studies and studies on other literacies (e.g., ICT literacy), domain-specific knowledge is a key predictor variable (see for example Moehring et al., 2016). Although the results of these studies on the influence of technology knowledge are mixed, Greiff et al. (2014) argued that the influence of such technology-specific knowledge (as included in the construct of ICT literacy) might have to be tested for every new use case. We thus included both domain-specific knowledge about mushrooms and AI knowledge in our exploratory model testing. Exploring the effects of underlying psychological constructs was therefore the third objective of this study (see Section 1).

Human behavior is determined by the context of use. Therefore, a detailed description of the use case and its characteristics is important,

as this study context determines to what extent the results are generalizable to other contexts. For our exploratory experimentation we chose the use case of mushroom picking, and the fourth contribution of our work is to provide necessary information and reflection on this novel use case for studying XAI in future research (see Section 1 and below).

3.2. In search of a use case: Why mushroom picking is a promising context for XAI

Researchers have argued that results and recommendations for human-centered interaction with AI might differ depending on the context of use. For example, Behnke et al. (2017) and Leichtmann et al. (2018) argued that plan linearization of an AI-based planner could vary depending on the problem space. In the field of XAI, the importance of careful selection of use cases has also been emphasized, as differences in contexts affect the transferability of study results (Buçinca et al., 2020; Kenny et al., 2021).

In order to set the boundaries that delineate the use cases between which a transfer of empirical findings is possible, contextual characteristics must be well described and defined. A generalization from a specific use case to another might then only be valid if it is similar in the most central characteristics. In the case of visual explanations of an AI recommendation, for example, use cases vary in the gravity of their consequences and thus engage different user values. Thus, a proper description of the use case is crucial.

In order to research the effects of visual explanations of AI decisions and the effects of an educational intervention about AI on human decision-making, comprehension, and trust, we chose the use case of mushroom picking with an AI classifier for the following reasons:

Relevance of decision Use cases differ with regards to whether it is of a particular value to users that the outcome of an AI-assisted decision is accurate. As stated in the introduction (Section 1), a wrong classification of cat and dog pictures will have little or no impact on human lives, whereas incorrectly classified mushrooms can have severe health consequences for the user. Mushroom poisoning leads to symptoms such as vomiting, dizziness, loss of consciousness, or even death (Brandenburg & Ward, 2018). These situations of vulnerability are especially important when researching trust in human–technology interaction, as vulnerability is a prerequisite for trust (Hannibal et al., 2021). By choosing the high-risk context of mushroom picking we ensured a degree of relevance to users.

Availability of target population Many use cases, such as medical image analysis (e.g., Jiao et al., 2021) and detection of dangerous objects in X-ray images of luggage (e.g., Hättenschwiler, Mendes, & Schwaninger, 2019), involve highly relevant decisions, but require specifically trained experts. Such expert users are often difficult to recruit for participation in scientific studies. For quantitative research (which typically requires large sample sizes to achieve high statistical power), it might be easier to have a more readily available target population. Since mushroom picking is part of the culture in many (European) countries (e.g., Aigner & Krisai-Greilhuber, 2016; Kaaronen, 2020; Svanberg & Lindh, 2019), we could draw participants from a larger population of people who engage in mushroom picking, or at least from a population of people who can easily imagine such a scenario.

Closeness to participants' reality Scientific studies with human participants usually have a certain artificiality. When people are tasked to imagine a specific situation or to use some technology they would not normally use, they often become aware of the study situation. To mitigate this problem, the use case should be as close to a real user situation as possible. Participants can then easily imagine the situation, and their behavior in the study should be a better proxy for real-world behavior. A use case close to participants' reality is challenging to find. Financial decisions, for instance, usually take considerable time and critical reflection. If in a short study users are tasked to imagine how they would decide financially, their decisions might not reflect their real-world behavior. In contrast, the decision of whether to pick a mushroom or leave it is typically fast and made within seconds or minutes and is characterized as "fast and frugal" based on simple rules, such as avoiding spotted mushrooms (Kaaronen, 2020). Therefore, although the mushroom-picking behavior was studied in an abstract research situation, it might be easier for participants to imagine going mushroom picking than other contexts in which decision-making is done more thoughtfully and takes more time. We thus expect the results to be closer to true values compared with results in other artificially constructed situations.

Public interest Some use cases—such as building a home theater—might be close to people's reality and easily imaginable in laboratory experiments (Leichtmann et al., 2018), but they lack public interest. While mushroom picking may not be prominent in all cultures, it is an integral part of cultural heritage in many European countries. This is indicated by studies from Sweden (Svanberg & Lindh, 2019), Finland Kaaronen (2020), and Austria (Aigner & Krisai-Greilhuber, 2016). As an outdoor activity, mushroom picking also has a recreational aspect (see, e.g., Aigner & Krisai-Greilhuber, 2016) that relates it to health-trends such as "shinrin-yoku", the Japanese practice of "forest bathing" (Park, Tsunetsugu, Kasetani, Kagawa, & Miyazaki, 2010). Such activities are assumed to have positive effects on human health and well-being, which is echoed in studies on mindfulness (Howell, Dopko, Passmore, & Buro, 2011). Furthermore, mushroom picking is linked to values of culture and tradition (Aigner & Krisai-Greilhuber, 2016; Peintner et al., 2013; Svanberg & Lindh, 2019). Finally, as mushroom picking is usually associated with the intention to eat the collected mushrooms (Aigner & Krisai-Greilhuber, 2016; Kaaronen, 2020), the reliable identification of edible mushrooms is of particular interest.

Furthermore, the availability of mushroom image databases due to AI "challenges" (Visipedia, 2018) and community-driven projects (Danish Mycological Society, 2022) allows an actual model to be trained as a basis for the AI-assisted decision-making.

4 Methods

We conducted an online study in which participants were tasked to decide—for a set of mushroom pictures—whether a depicted mushroom is edible or poisonous[2] and whether they would pick the mushroom for cooking or leave it. For each decision task, participants were shown a mock-up of a screenshot of an app named *Forestly*. The app aided the participants in their decisions by showing AI classification results. While the screenshots were mockups, the information shown was based on a real machine learning model (see Section 4.5).

We chose an online method for this first exploratory investigation of XAI and educational intervention in the context of mushroom hunting, because research via the online database of a market research company allowed us to target a diverse sample and to recruit a large number of people (for advantages of online research, see Gosling & Mason, 2015). Large sample sizes are especially necessary for unknown effect sizes in under-explored contexts. Furthermore, past research in other fields of human–technology interaction, such as robotics, has shown that on-site experiments conducted at universities often have samples biased toward higher education levels, are thus more homogeneous, and consisted of only low sample sizes (Leichtmann et al., 2022). Thus, it is expected that online experimentation can be advantageous in these aspects.

In the following section, we describe the methods used in our exploratory experimentation, including study design, sample size justification and sample description, the mushroom-picking task, descriptions

[2] For this study, we did not distinguish between poisonous and inedible mushrooms.

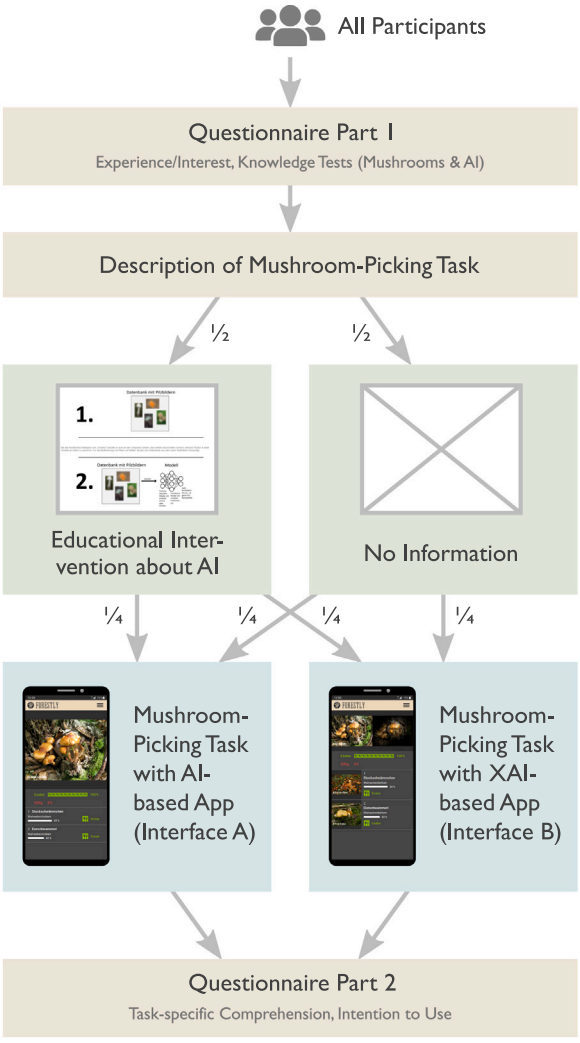
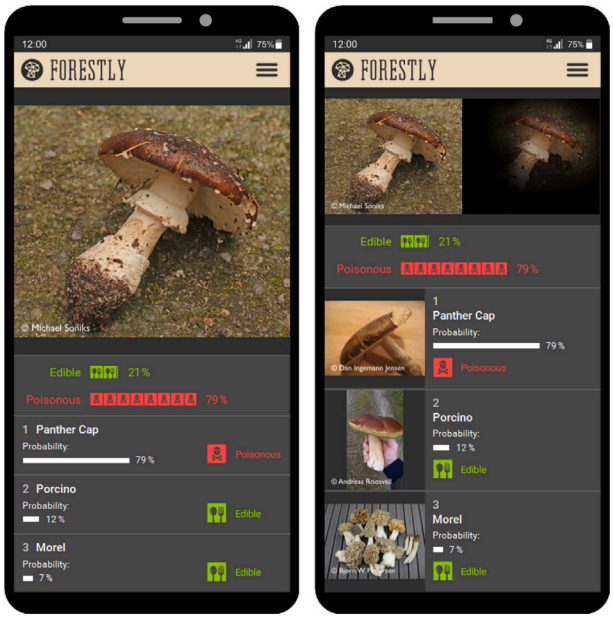
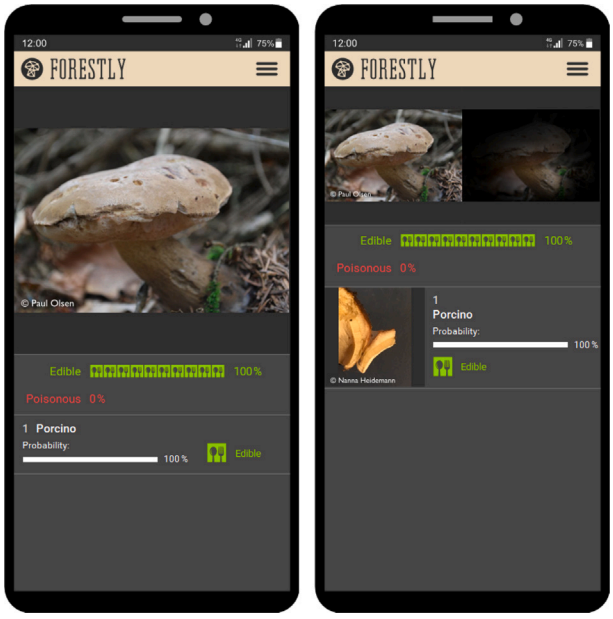


Fig. 1. Participant flow in the 2 × 2 study design, including questionnaires, the educational intervention, and the mushroom-picking tasks with or without visual explanations.



(a) Plain interface (b) XAI interface

Fig. 2. Examples of the two variants of the *Forestly* app interface for a correct prediction, with a "reasonable" explanation in the XAI variant (translated version).



(a) Plain interface (b) XAI interface

Fig. 3. Examples of the two variants of the *Forestly* app interface for a wrong prediction, with a "questionable" example-based explanation in the XAI variant (translated version).

of all variables used and of the procedure and materials. This research complied with the tenets of the Declaration of Helsinki and adhered to ethical guidelines of the APA Code of Conduct. Informed consent was obtained from each participant. As no profound prior information for the domain of mushroom picking existed, this study was exploratory, and thus no preregistration before data collection was possible. Because values of transparency are important to us in the context of Open Science, we share additional study details in the Appendices A, B, C, as well as key replication materials, and data and R code for analysis as supplementary material on OSF (newly developed test items, the R code used for data analysis, raw data with variable descriptions, and code used to train the AI model and to generate the explanations can be found under https://osf.io/tqbgh/files/).

4.1 Independent variables

We used a 2 × 2 between-subjects study design, varying the presence of (*i*) an educational intervention about the AI functionality, and (*ii*) visual explanations of the AI classification, as indicated in Fig. 1. To test the effectiveness of an educational intervention on two levels, one group of participants was given a short introduction to machine learning (see Appendix A.2). This description included high-level concepts of typical image classification networks such as the model used in the study. Explainability was also manipulated on two levels: one group received the classification result only, and one group additionally received two visual explanations (see Figs. 2 and 3). The explanation methods we chose were (*i*) Grad-CAM as an attribution-based technique, and (*ii*) an adapted version of ExMatchina as an example-based technique, where we showed the most similar training image of each predicted class.

4.2 Dependent and control variables

For extensive exploration, we chose a large number of dependent and control variables, as listed below.

4.2.1 Human classification performance and picking behavior

In our study, we broke down the mushroom-picking activity into two components: an assessment of a mushroom's edibility, and the final decision of whether to pick it (i.e., take-home decision). Both could be correct or incorrect independently of each other. A correct edibility assessment was the identification either of an edible mushroom as edible or of a poisonous mushroom as poisonous; an incorrect assessment was the identification either of an edible mushroom as poisonous or of a poisonous mushroom as edible. A correct take-home decision was to pick up an edible mushroom or to leave a poisonous one; and an incorrect take-home decision was to pick up a poisonous mushroom or to leave an edible one. Both of these decisions were made 15 times (as 15 pictures were presented).

To reflect these different outcomes, we recoded each of the variables—edibility assessment and take-home decision—into performance measures. Obviously, particular incorrect decisions pose greater risks than others. For example, not picking up an edible mushroom merely leads to missing out on the enjoyment, while incorrectly picking up a poisonous mushroom might have severe health consequences. Thus, it might be interesting to not just understand correct decisions in general, but also subsets of decisions with specific characteristics. This is even more important in the light of the potentially wrong or misleading classifications made by the AI system, which participants may or may not trust.

4.2.2 Trust in AI and self-reported comprehension of AI classifications

In addition to the edibility assessment and the take-home decision, participants were asked to rate how much they trusted ("I TRUST this mushroom identification of the AI") and how much they comprehended ("I UNDERSTAND how the AI arrives at this mushroom classification") the mushroom identification/classification of the AI on a 5-point Likert scale ranging from "do not agree at all" to "fully agree".

4.2.3 AI knowledge

Participant knowledge of AI technology was measured with 8 items. Each item consisted of a question and four possible responses, only one of which was correct (the other three were distractors). All items were evaluated by at least three computer scientists in order to ensure content validity of the test. The test can be viewed in the supplementary material in German (original) and in English translation (see https://osf.io/tqbgh/files/). For example, one question was "A newly developed AI is supposed to detect foxes in images. What is usually a prerequisite for such an AI?" with possible responses "(A) A large collection of fox images from which similarities and dissimilarities are inferred.", "(B) A few high-quality images of foxes in which characteristic traits are marked.", "(C) Fixed code written by programmers, that tells the AI exactly how to distinguish foxes from other animals.", and "(D) Knowledge bases on the internet which the AI accesses actively via search engines". (Here, option A is correct.)

4.2.4 Domain knowledge about mushrooms

The domain knowledge about Austrian mushrooms was assessed with 11 declarative knowledge items. The test was developed in close cooperation with mycologists of the Mycological Working Group Linz at the Biology Centre Linz, Austria to ensure content validity. Again, each item was constructed as a multiple-choice item with four possible responses, only one of which was correct.

For example, one question was "The caps of cultivated button mushrooms are…" with possible responses "(A) yellowish", "(B) fuzzy along the edges", "(C) tapered and pointy", and "(D) roundish and hemispherical". (Here, option D is correct.)

4.2.5 Task-specific AI comprehension

In addition to the more general items used to measure the participants' AI domain knowledge, we created a test to measure how well participants comprehended the mushroom identification AI presented in this study. The test consisted of 5 multiple-choice items with 4 possible responses each. Again, only one response was correct in each case. For example, one question was "What does a percentage of 77.5% shown in the *Forestly* app for a mushroom species mean?". The possible responses in this example were "(A) That the AI's prediction is correct in 77.5% of the cases", "(B) That the AI is 77.5% sure that the mushroom belongs to this species", "(C) That 77.5% of the users trust the decision of the AI", and "(D) That the AI was able to recognize and categorize 77.5% of the image area". (Here, option B is correct.)

4.2.6 Other measures

For sample description, we collected data on participant gender, age, educational level, and their self-reported domain-specific experience and interest (e.g., "Have you ever picked mushrooms?" and "How high would you rate your interest in mushroom picking?"). We also asked participants about their intention to use an AI-based system for mushroom picking before and after the identification tasks ("How high would you rate your willingness to use an app or other software that suggests classifications of a mushroom based on photographs?" and "How high would you rate your willingness to use the *Forestly* app shown in the study in a real mushroom hunt?", rated on a 5-point scale from "very low willingness" to "very high willingness"). Finally, we asked participants to evaluate the presented AI system ("How would you rate the *Forestly* app shown in the study?" rated on two 5-point bipolar scales from "bad" to "good" and from "incomprehensible" to "comprehensible").

As online questionnaires are especially vulnerable to inattentive responding (which could be increased even further by the monetary compensation), we used attention check items, as recommended by Maniaci and Rogge (2014). We used a total of 4 items throughout the questionnaire to screen out inattentive participants to improve data quality (e.g., attention check item placed within the mushroom knowledge test: "Several edible mushrooms can be mistaken for similar-looking poisonous ones. This question checks your attention. Please simply select 'Yellow foot.'", with possible responses "(A) Common ink cap", "(B) Pinecone cap", "(C) Snowy waxcap", and "(D) Yellow foot").

4.3 Mushroom-picking task

In order to study human mushroom picking supported by an AI classifier within an online experiment, we developed a new mushroom-picking task. The task consisted of 15 items in total. Participants were presented with mushroom pictures taken from a database containing mushroom photographs (Danish Mycological Society, 2022). They were asked to imagine going mushroom picking in a forest with the intention to cook a meal from the picked mushrooms (see full description in Appendix A.1). In this imaginary scenario, the mushroom pictures they were presented showed the mushrooms they found. The participants had to decide for each of the mushrooms shown whether they would classify it as edible or poisonous/inedible (see also Appendix A.4), and whether they would ultimately take this mushroom home for consumption or rather leave it in the imaginary forest (see Section 4.2).

For each item, participants were additionally presented with one of two versions of an interface of the fictitious app *Forestly* showing the classification recommendation of a real AI. This classifier takes as input an image of a mushroom and predicts the most likely from a set of 18 species. Details on how we implemented the classifier can be found in Section 4.5.

The classification result was presented to the participants in the form of the top three mushroom species as predicted by the AI, along with bars indicating the percentages of certainty for each species. Additionally, the total predicted probability of the mushroom being

edible or poisonous/inedible was indicated by a bar chart and numbers. This total probability was determined using the AI's species predictions. Depending on the version of the interface presented to the participants (see manipulations in Section 4.1), the interface additionally included visual explanatory elements: example pictures of each of the top three predicted species, and an attribution map indicating important regions in the image. Examples of the *Forestly* interfaces are shown in Figs. 2 and 3. For a more detailed description of the explanations, see Section 4.5.

We selected 15 representative images of the training dataset (Danish Mycological Society, 2022) as the final item pool for the mushroom-picking task. This selection of the final set of 15 stimuli was based on the following criteria: (*i*) Both, edible and poisonous/inedible mushrooms were to be included in the set of items (as the value participants assign to an outcome might differ for the various scenarios discussed in Section 4.2.1). (*ii*) The rating of the AI's certainty of its classification results (ranging from 0% to 100% certainty) was required to vary, as this could be a potential confounding variable (see e.g., Kunze, Summerskill, Marshall, & Filtness, 2019). (*iii*) The AI classifier was required to give false recommendations for at least 5 items in order to allow overtrust to be measured (i.e., a person's incorrect assessment of a mushroom's edibility based on an incorrect AI prediction).

For each task, the mushroom picture (stimulus) and the corresponding screenshot of a fictitious app, as well as the dichotomous items on edibility classification and picking behavior, the single-item Likert-scale on trust in the system's recommendation, and a single-item Likert-scale of self-reported recommendation understanding were presented on one page. Each task—and thus each stimulus—was presented on a separate page without the option to return to a previous task. The order of the stimuli was randomized to control for order effects.

4.4 Procedure

After being screened for eligibility of participation (see Section 4.6), participants were asked to read general information on purpose, course, and duration of the study, and on data protection regulations and contact information. We then sought informed consent from every participant. Participation was voluntary and could have been terminated at any point without consequences.

After giving informed consent, participants filled out a questionnaire about their experience and interest in mushroom picking (see Section 4.2.6). Participants then filled out the knowledge tests on Austrian mushrooms and on artificial intelligence (see Sections 4.2.4 and 4.2.3, respectively).

The next phase of the study was the AI-assisted mushroom-picking task. Here, participants were first given brief instructions on the imaginary scenario of mushroom picking: "Imagine that you want to cook a tasty mushroom dish for dinner. To this end, you want to collect fresh mushrooms in the forest..." (the full instructions are given in Appendix A.1). Participants then received brief instructions on the fictitious app *Forestly* (which was powered by the AI). In addition, subjects from the experimental group with the educational intervention received further information on how image classification by AI works (see Section 4.1 on independent variables and Appendix A.2 for a screenshot of the educational intervention).

In the mushroom-picking task, participants were presented static screenshots from one of two different versions of the *Forestly* app, depending on whether they were in the group that received explanations of the AI's predictions. The instructions on the *Forestly* app mentioned above were tailored to the version shown in the picking task (see Appendix A.3).

After the mushroom-picking task, participants were asked to complete the task-specific AI comprehension test (see Section 4.2.5). Finally, participants were asked to evaluate the *Forestly* app and answer questions about their intentions to use such an app (see Section 4.2.6). After completion of the study, each person received €2 reward for their participation.

4.5 (X)AI implementation

For the purpose of this study, we developed a simple image classification network based on the ResNet50 network (for a detailed discussion on ResNet50 see He, Zhang, Ren, & Sun, 2016a) with pre-trained weights from the ImageNet benchmark dataset (Deng et al., 2009). We adapted the model architecture of ResNet50 by removing its top layers (global average pooling and final linear layer) and adding a flattening layer, followed by two dropout and linear layers. We set the dropout rate for both dropout layers to 50%. The activation function for the first linear layer was a rectified linear unit (ReLU), and the activation function of the final linear layer was softmax. A sketch of the network architecture is given in Fig. C.4 in Appendix C. We trained the network to classify images of mushrooms into one of 18 mushroom species.

Although we did not conduct an exhaustive hyperparameter search, we tried a range of settings by varying (*i*) the number of top layers (one, two, and three sets of dropout and linear layers), (*ii*) the number of layers (none vs. 15) for which the weights were fixed to the pre-trained ImageNet weights during training, (*iii*) the input resolution (224×224 vs. 512×512), and (*iv*) the pre-trained base models (InceptionV3 by Szegedy, Vanhoucke, Ioffe, Shlens, and Wojna (2016); VGG16 by Simonyan and Zisserman (2015); ResNet50V2 by He, Zhang, Ren, and Sun (2016b); ResNet50 by He et al. (2016a)) to obtain satisfactory results.

For fine-tuning of the model for our mushroom classification task, we chose the Adam optimization algorithm (Kingma & Ba, 2017) with a categorical cross-entropy loss function, and used a held-out validation dataset for early stopping of the training.

For training data, we acquired mushroom images from various sources. Most images came from the fungal records database by the Danish Mycological Society (2022). We randomly split these images into train, test, and validation datasets. Since certain species were underrepresented in our training data, we then added images from Google Images and Wikimedia, to have at least 80 training images per species. This resulted in a total of 3480 training images, 241 test images and 219 validation images ($0.88 : 0.06 : 0.06$). Each mushroom species was represented by between 80 and 450 images. To adjust for this imbalance, each species' contribution to the loss function was weighted accordingly.

Images were preprocessed with the same function that is typically used for the ImageNet dataset (Chollet et al., 2021). During training, data augmentation was performed by randomly varying the rotation, zoom, width and height shift, and horizontal flip of the images.

The accuracy of the chosen classifier was 94% for the training dataset, 82% for the validation dataset, and 71% for the test dataset.[3]

To obtain estimates of the certainty with which the AI predicted a particular species, we chose Monte Carlo dropout (Gal & Ghahramani, 2016) during inference. We repeated the prediction of a single input 100 times and measured how often each class was predicted. We showed the top three predicted classes and their corresponding percentages to users.

The attribution-based XAI method we decided to use was Grad-CAM (Selvaraju et al., 2017), since it is widely used and passes specific "sanity checks" (Adebayo et al., 2020). Grad-CAM calculates regions in the input image that were particularly important for the AI's prediction. These regions can then be highlighted directly in the image and shown to the user. For our visual explanations, we calculate the Grad-CAM for the top species predicted for an image with respect to the last convolutional layer of the network (with dropout deactivated).

[3] Note that it was never our intention to create an actual mushroom-picking assistant to be used in the real world. For such a system, an accuracy of 71% would be far too low. For this study, however, the low accuracy gave rise to numerous incorrect predictions, whose use by humans in their own decision-making processes is interesting from a research perspective.

Our example-based XAI method of choice was an adapted version of a technique proposed by Jeyakumar et al. (2020). We retrieved the training image for a certain species that was closest to the input image in terms of the network's internal representation. We retrieved the closest example image for each of the three most likely species. To determine the similarity between two images, we calculated the Euclidean distance between their representation vectors after the flattening procedure.

The python code for model training and calculations of the explanations can be found in the supplementary material along with the trained model that we used in this study (see https://osf.io/tqbgh/).

4.6 Sample size justification and sample description

Based on the study design, sample size was calculated in an a priori power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) for the interaction effect of the two independent variables of educational intervention and explainability. A low to medium effect size was assumed ($f = 0.19$), and the global α-error probability was set to 0.01. A sample size of $N = 327$ was estimated to be necessary to achieve a minimum statistical power of 80%. To be on the safe side, we planned to recruit at least $N = 360$ participants, that is $n = 90$ for each of the 4 groups based on the 2×2 study design. Data collection started on June 14, 2021, and was set to terminate at the end of the day on which we reached the threshold of $N = 360$ participants. All participants were drawn from an online panel by a market research company and were paid for participation (€2). The market research company invited participants based on information from their online panel database and ensured diversity in terms of gender, age, and educational level.

Furthermore, we defined the following inclusion criteria: (*i*) Since the occurrence of fungal species differs regionally, and cultural differences in mushroom-picking and consumption habits exist, only people with main residency in Austria were surveyed. (*ii*) Only people who eat or would eat mushrooms in general were surveyed to control particular error variables. (*iii*) For legal and ethical reasons, only people at least 18 years of age were allowed to participate.

In total, 617 persons started the questionnaire, 33.5% terminated the questionnaire early or had to be excluded due to low data quality (as detected by certain attention check items; see Section 4.2.6). The total sample size for data analysis was $N = 410$; of these, $n = 200$ were in the group with educational intervention and $n = 210$ in the one without the intervention; $n = 202$ participants received visual explanations while $n = 208$ did not.

From these 410 participants, 213 persons identified as female, 193 as male, 2 persons as diverse (third gender option in German-speaking countries), and 2 persons did not indicate their gender. The mean age of the subjects was $M = 44.58$ years ($SD = 15.29$ years). Only 15.9% of participants held a university degree, while most held a professional but non-academic educational diploma (53.8%). When asked about their experience with mushroom-picking, 80% of participants said that they had been mushroom picking at least once in their life, and 98% of those had intended to eat the mushrooms picked; about 15% had previously used an app to identify mushrooms in the context of mushroom picking. The general interest in mushroom picking, however, was moderate ($M = 3.04$, $SD = 1.10$, indicated on a 5-point Likert scale ranging from 1 = "very low interest" to 5 = "very high interest"), and the willingness to use a corresponding identification app before study participation was medium to high ($M = 3.69$, $SD = 1.14$, indicated on a 5-point Likert scale ranging from 1 = "very low willingness" to 5 = "very high willingness").

5 Results

For data analysis, we used the open source statistic software RStudio (R version 4.1.1) (RStudio Team, 2020). The R Code used for this analysis can be found in the supplementary material (https://osf.io/tqbgh/).

Table 1
Values of model fits for the mushroom and AI knowledge tests and for the task-specific AI comprehension test.

	k	df	χ^2	*p*	CFI	TLI	RMSEA	SRMR	ω
Mushroom knowledge	6	9	17.81	.04	.90	.83	.05	.04	.44
General AI knowledge	5	5	5.42	.37	~1.00	.99	.01	.03	.50
Situational AI knowledge	4	2	2.67	.26	.99	.96	.03	.02	.42

Note: *k* is the number of items in a test, and df is the degrees of freedom.

5.1 Measurement models and item selection

We developed new tests to measure levels of expertise in fields that were likely to influence the results of the study. As mentioned in Section 4.2, we measured participants' prior general knowledge about mushrooms and AI, and we measured the task-specific AI comprehension after using the *Forestly* app. For all new measurement models, the goodness of fit must be tested and the psychometric properties analyzed.

To ensure a good model fit and good properties for further testing, we excluded items if they did not fit the model well or if they were too difficult (i.e., items with $M \leq .25$).

We chose the weighted least squares mean and variance adjusted (WLSMV) estimator due to the dichotomous nature of the coded items and its superiority to the maximum likelihood (ML) estimator (Beaudu-cel & Herzberg, 2006). Furthermore, we chose cut-off values of $\geq .96$ for the comparative fit index (CFI), and $\leq .05$ for the root mean squared error of approximation (RMSEA) (Yu, 2002).

Table 1 summarizes the results of the test analysis after item exclusion. All tests showed a moderate to good model fit, but low internal consistency.Tables B.1, Table B.2, and Table B.3 (all in Appendix B) give an overview of item characteristics.

In contrast to the single-factor measurement models we created for the knowledge tests from Section 5.1, for the mushroom identification tasks we calculated only the sums of correct items per participant. To enable efficient measurement in future studies, we reduced the number of items from 15 to 10 by excluding items with extreme difficulties (i.e., items that only a few persons answered correctly or items that most persons answered correctly), since these contain little information. For example, one picture of a chanterelle was correctly identified by 99% of the participants, and 96% of participants would have correctly picked this mushroom. For this item we assumed that there would not be much difference between groups, which rendered it uninformative in this study. We ensured that at least 4 items were kept that (*i*) depicted poisonous mushrooms, and (*ii*) depicted mushrooms with an incorrect AI prediction, since these subsets of items are of special interest in mushroom-picking scenarios. A list of all 15 items and their levels of difficulty is given in Table 2 (note that item selection was performed before conducting further analysis to avoid confirmation bias).

5.2 Effects of educational intervention and explanations

For the statistical tests, we used a global alpha-level of $\alpha < .05$. Multiple hypotheses were tested, and, as this study was exploratory in nature, a large number of tests were performed. Usually, when conducting multiple tests, the family-wise error rate (FWER) is controlled, which means that the probability of at least one Type-I error is adjusted. However, with a large number of tests, the FWER control might be too restrictive. Benjamini and Hochberg thus suggested that "a desirable error rate to control may be the expected proportion of errors among the rejected hypotheses" (Benjamini & Hochberg, 1995, p.290), which is termed false discovery rate (FDR). For our statistical analysis, we therefore controlled the FDR according to the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995), rather than the FWER.

Table 2
Mushroom classification items used in the study. *Ed.* is the ground truth of whether the mushroom is edible ( ) or inedible/poisonous ( ). *AI prediction* lists the top 3 results of the classification model, with certainty values for each prediction. *AI corr.* indicates whether the AI prediction was correct ( ) or incorrect ( ). *Ed. (AI)* is the probability of edibility as predicted by the AI. M_{pick} and M_{ed} are the mean performance scores for the picking and edibility prediction tasks, respectively, where 0 means that no one answered correctly and 1 means that all answers were correct. For each item, *Incl.* indicates whether the item was used for analysis ( ) or not ( ).

ID	Species	Ed.	AI prediction	AI corr.	Ed. (AI)	M_{pick}	M_{ed}	Incl.
1	Chanterelle		Chanterelle 100 %		100 %	.99	.96	
2	Bitter bolete		Porcino 100 %		100 %	.19	.27	
3	Bitter bolete		Porcino 92 % Bitter bolete 6 % Sheathed woodtuft 1 %		94 %	.19	.27	☑
4	Porcino		Bitter bolete 46 % Porcino 25 % Autumn skullcap 17 %		31 %	.16	.13	☑
5	Blusher		Panther cap 85 % Blusher 10 % Death cap 2 %		12 %	.20	.12	☑
6	Autumn skullcap		Autumn skullcap 68 % Chanterelle 26 % Death cap 5 %		26 %	.98	.99	
7	Panther cap		Panther cap 79 % Porcino 12 % Morel 7 %		21 %	.91	.94	☑
8	Porcino		Porcino 85 % Panther cap 15 %		85 %	.89	.80	☑
9	Sheathed woodtuft		Sheathed woodtuft 60 % Chanterelle 40 %		100 %	.76	.60	☑
10	Funnel chanterelle		Funnel chanterelle 62 % Jelly baby 38 %		62 %	.21	.09	
11	Field mushroom		Field mushroom 94 % Destroying angel 4 % Death cap 2 %		94 %	.84	.71	
12	Bitter bolete		Bitter bolete 40 % Death cap 31 % Porcino 29 %		29 %	.76	.83	☑
13	Amethyst deceiver		Amethyst deceiver 66 % Sheathed woodtuft 27 % Lilac bonnet 3 %		94 %	.40	.23	☑
14	Blusher		Blusher 70 % Death cap 27 % Porcino 3 %		73 %	.43	.25	☑
15	False chanterelle		Chanterelle 100 %		100 %	.30	.38	☑

To test the main effects of educational intervention and visual explanation on identification, behavioral intentions and task-specific AI comprehension, we used non-parametric Brunner–Munzel tests (Brunner & Munzel, 2000). The effects were tested (*i*) with all items in the mushroom-picking task, (*ii*) with a subset of items that included a poisonous mushroom, and (*iii*) with a subset of items for which the AI classification was wrong (e.g., classifying a mushroom picture as edible when it was in fact poisonous or inedible). Thus, we performed a total of 14 Brunner–Munzel tests.

In the edibility assessment, only the visual explanations had a significant effect when considering either all 10 test items ($BM(398.68) = 4.65$, $p < .001$, $d = .44$, $CI_{95} = [.25, .64]$) or the subset of items for which the AI recommended a false classification ($BM(401.73) = 3.48$, $p = .004$, $d = .35$, $CI_{95} = [.15, .54]$). Participants who received the additional explanations by the AI (all 10 items: $M = 5.29$, $SD = 1.36$; 4 items with incorrect AI: $M = 1.00$, $SD = .97$) performed better than those who did not receive them (all 10 items: $M = 4.72$, $SD = 1.22$; 4 items with incorrect AI: $M = .69$, $SD = .84$). Tests showed that the educational intervention had no significant effects—neither for all test items, nor for items depicting a poisonous mushroom or mushrooms incorrectly classified by the AI (e.g., for all 10 items: $BM(403.89) = .59$, $p = .56$, $d = -.08$, $CI_{95} = [-.28, .11]$).

When testing the effects of the educational intervention and explainability on participant intention to pick the mushroom in this imaginary situation, again only the visual explanation had a statistically significant effect, but not the educational intervention (e.g., for all 10 items: $BM(407.75) = -0.53$, $p = .77$, $d = .01$, $CI_{95} = [-.18, .20]$). Visual explanations had a significant effect only when considering all 10 items ($BM(400.71) = -2.69$, $p = .03$, $d = .26$, $CI_{95} = [.07, .46]$), but not when only poisonous mushrooms or AI misclassifications were considered. Again, participants were more likely to correctly pick or leave mushrooms when the recommendation of the AI system included

Table 3
Values of model fits for self-reported trust and self-reported comprehension single-factor models and a model with correlated factors.

	k	df	χ^2	p	CFI	TLI	RMSEA	SRMR	ω
Self-reported trust	10	35	117.15	<.001	.95	.93	.08	.04	.88
Self-reported comprehension	10	35	181.52	<.001	.93	.91	.10	.04	.91
Correlated trust and comprehension	20	169	2729.11	<.001	.59	.53	.19	.09	–

Note: k is the number of items in a test, and df is the degrees of freedom.

visual explanations ($M = 4.71$, $SD = 1.23$) than when it did not ($M = 4.40$, $SD = 1.09$).

None of the manipulations had a significant effect on the task-specific AI comprehension test after the main task. This means that, in the comprehension test, neither did participants with visual explanations ($M = 2.34$, $SD = 1.14$) perform differently from participants without visual explanations ($M = 2.35$, $SD = 1.19$), nor did participants in the educational intervention group ($M = 2.34$, $SD = 1.19$) from those without intervention ($M = 2.34$, $SD = 1.14$).

5.3 Does domain-specific knowledge or AI knowledge predict task success?

To test the influence of domain-specific knowledge (i.e., mushroom knowledge) and AI knowledge on task success—that is, correctly identifying a mushroom as edible or poisonous and correctly picking an edible mushroom or leaving a poisonous one—we performed two multiple linear regression analyses. The two independent variables educational intervention and explainability (operationalized via visual explanations), and mushroom knowledge and AI knowledge were included as predictors.

However, both the model for edibility classification performance ($F(4, 405) = 5.73$, $p < .001$, $R^2 = .05$) and the model for mushroom-picking behavior ($F(4, 405) = 2.43$, $p = .05$, $R^2 = .02$) described the dependent variables poorly. Neither mushroom knowledge ($\beta = .06$, $SE = .05$, $p = .25$) nor AI knowledge ($\beta = .05$, $SE = .05$, $p = .30$) seem to predict participants' mushroom identification performance with the AI recommendations. Like the results reported above, this model also reflects the significant influence of explainability ($\beta = .22$, $SE = .13$, $p < .001$), but not of the educational intervention ($\beta = -.10$, $SE = -.04$, $p = .44$).

Similar results were obtained when using the intention to pick a mushroom as a dependent variable. Neither mushroom knowledge ($\beta = .02$, $SE = .04$, $p = .75$) nor AI knowledge ($\beta = .08$, $SE = .04$, $p = .11$) seem to have significant influence. Again, explainability ($\beta = .13$, $SE = .11$, $p = .007$), but not educational intervention ($\beta = .01$, $SE = .11$, $p = .82$) are significant predictors in this model.

5.4 Self-reported trust in the AI system

As reported in Section 4.2.2, participants' trust in, and self-reported comprehension of, the AI classification were measured for each item of the mushroom-picking task. Based on the final 10 selected items from the mushroom-picking task, self-reported trust achieved a good internal consistency and an overall acceptable model fit when trust was considered a single-factor model (see Table 3). Self-reported comprehension also achieved good internal consistency and acceptable model fit (see Table 3). In a model including both factors, the latent constructs correlated with $r = .74$, but the fit of the model to the data was poor (see Table 3).

Multi-group comparisons for visual explanations and educational intervention (corrected based on the Benjamini–Hochberg procedure) showed that only explainability had an effect on self-reported trust ($BM(403) = 3.91$, $p < .001$, $d = -.36$, $CI_{95} = [-.56, -.17]$) and on self-reported comprehension ($BM(404) = 2.49$, $p = .02$, $d = -.21$, $CI_{95} = [-.41, -.02]$). Participants without explanation trusted ($M = 3.64$, $SD = .74$) and reported to comprehend ($M = 3.57$, $SD = .87$) the system more than participants who received further explanation (trust: $M = 3.36$, $SD = .81$; self-reported comprehension: $M = 3.39$, $SD = .76$). The educational intervention had no statistically significant effect.

These exploratory results could indicate that participants without additional visual explanations tended to overtrust, following AI recommendations uncritically. We therefore tested in a linear model whether trust ratings predicted poorer user classification performance for items that the AI incorrectly classified as edible. These are the cases in which trusting the AI recommendations too much can lead to potential harm (rather than missed pleasure). Indeed, the model ($F(1, 408) = 74.76$, $p < .001$, $R^2 = .15$) showed that participants reporting higher trust in the AI were also significantly more likely to incorrectly classify a poisonous or inedible mushroom as edible ($\beta = -.40$, $SE = .04$, $p < .001$). Furthermore, they were also more likely to *pick* such an incorrectly classified mushroom ($F(1, 408) = 96.6$, $p < .001$, $R^2 = .19$; $\beta = -.44$, $SE = .04$, $p < .001$).

5.5 Effects of educational intervention and explanation on AI system evaluation

Finally, we also explored the main effects of explainability and educational intervention on users' overall evaluation of the system (i.e., the *Forestly* app). Participants who did not receive any visual explanations evaluated the app as better ($M = 4.12$, $SD = .82$) than those who received the interface with visual explanations ($M = 3.77$, $SD = 1.03$; $BM(390) = -3.44$, $p = .003$, $d = .38$, $CI_{95} = [.18, .57]$). The group without visual explanations also reported higher intentions to use the app ($M = 3.58$, $SD = 1.10$) than that given explanations ($M = 3.24$, $SD = 1.23$; $BM(404) = -2.73$, $p = .01$, $d = .29$, $CI_{95} = [.09, .48]$). Instructions did not have significant effects on evaluation and intention to use.

6 Discussion

In this exploratory experiment, we manipulated both (*i*) the educational intervention about how the AI-based classifier works, with the goal to enable users to better comprehend AI-based technology, and (*ii*) the explainability of the AI-based system with attribution-based, example-based explanations. In the mushroom-picking use case, we tested the effects of our manipulations on task performance (i.e., a user's assessment of the edibility of a depicted mushroom and their behavioral intention to pick or leave it), and on user comprehension of the AI assistant, user trust, user evaluation and their intention to use. Furthermore, the influences of domain-specific knowledge (i.e., mushroom knowledge) and AI knowledge were measured for exploratory modeling.

6.1 Summary of results

The results showed that the educational intervention had no statistically significant effect on any of the dependent variables. This does not mean that educational interventions have generally no effect, as we used only one of many possible implementations of "educational interventions". This absence of effect may thus be attributed to the relative shortness of the intervention used in our study, which might not have led participants to properly understand the AI. Maybe other, longer interventions would be more effective. For example, Körber et al. (2018) showed study participants videos about automated driving that included functionality, sensors, and trajectory planning—a more

intense and vivid intervention than text and images. Furthermore, the literature on AI literacy suggests the use of educational elements in which users could explore the boundaries of an AI system (e.g., Long & Magerko, 2020).

Explanations of the AI's predictions, in contrast, were found to be effective, leading to a statistically significantly better performance in the mushroom-picking task. Better performance here means a higher accuracy both in assessing a mushroom's edibility and in picking an edible mushroom but leaving an inedible or poisonous one. However, results were clearer for edibility assessment than for actual intentions to pick a mushroom. While edibility assessment may represent a measure of maximal effort, behavioral intentions to actually take mushrooms may also reflect personality traits, as it might represent typical behavior (for maximal effort and typical behavior see Cronbach, 1970 as cited by Olderbak & Wilhelm, 2020), such as a greater propensity for risk aversion, which could lead to smaller overall effect sizes.

Domain-specific (i.e., mushroom) knowledge and AI knowledge—while shown to be key variables for decision-making with technological support—did not affect performance in the mushroom-picking task. Several reasons may account for this null result. First, we had to construct new measurement instruments (i.e., knowledge tests) for domain knowledge about mushrooms and AI. Although we carefully constructed the items for the tests (e.g., by consulting a group of mycologists), it may be that the measurement instruments are only of limited validity in this use case. For example, most people are probably not mushroom experts and know only a few common species. However, in the mushroom-picking task they were also confronted with less common ones. While the test might be able to detect differences between laypersons and experts, it might not be able to detect differences in expertise at a higher knowledge level, which is often required for mushroom picking. The lack of evidence might therefore be a consequence of too little variance at this higher level of mushroom knowledge. Similarly, the AI knowledge test might also be of limited validity in this study due to the study design. It is unclear how a higher level of AI knowledge would have affected performance in this task, since the AI-powered app did not allow exploration. Only fixed, predefined pictures were presented, and there was no opportunity for users to test the AI or particular limits of the app. For example, a knowledgeable person might have taken several pictures of the same mushroom from various angles to test the model. Therefore people with greater AI knowledge who know how to test the quality of an AI-based system were not able to apply their knowledge the way they usually would. Effects of domain-specific and AI knowledge may thus have been eliminated by the artificiality of the experimental design with its inherent restrictions, which may, however, well play a role in a real-world application setting.

Finally, the results show that participants with visual explanations trusted the AI classification significantly less and reported to understand it less. This could mean that the visual explanations prevented overtrust and thus led to more adequate trust. Participants better understood the limits of the AI's performance and seemed to have a more adequate understanding of the system's fallibility. This can be seen in more adequate trust ratings—especially for items for which the AI's prediction is incorrect. Participants performed better with visual explanations, and showed lower trust and lower intentions to use the app. At first glance, these results appear to contradict those of other studies, such as that by Buçinca et al. (2020). This indicates that the effects have a wide variance, which suggests that specific variables moderate such effects and that more complex models are necessary. For example, trust may also be influenced by AI classification error rates (Kenny et al., 2021), of which participants in our study were not informed. For some of our items, the visual explanations clearly indicated that the classification must be flawed, but were this not the case, and a visual explanation did not clearly flag an error, such explanations could reassure the user that the classification was correct when it was not. Our results are therefore only valid for this composition of task characteristics. Experiments with other implementations of the task, such as other error rates, may yield different results. It is therefore important that studies with other designs are conducted and the influences of various moderators on effect sizes and directions are calculated in subsequent meta-analyses.

6.2 Limitations

As in every study, a number of limitations need to be considered and can be addressed in future work. First, our study considered only a small subset of possible XAI techniques for the visual explanations. Although we tried to include two practically relevant techniques, this provides a limited view on the subject.

Second, although online experimentation has advantages—such as allowing to recruit a diverse sample that might not be reached in field studies and laboratory experiments, and allowing for larger sample sizes more easily—online studies come with their own limitations. While we sought a task that is relevant to people and has serious implications if a wrong decision is made, participants might not have been properly motivated to optimize their decisions. The experimental control to ensure this motivation is limited in online studies, leading to uncertainty about the validity of results with respect to real-world mushroom-picking behavior. Apart from conducting the experiment on-site, motivation could be increased either by offering a bonus reward or by adding a gamification aspect to the study. In non-online experiments (in a laboratory or even in the field), it would also be possible to implement a study setup closer to the real situation of mushroom hunting.

Furthermore, the experimental task might have been too restrictive, leading to null results for AI knowledge and our educational intervention. Maybe more knowledgeable users would have gathered more meta-information about the AI training or the quality of the underlying data, or tested the AI with a sample of test images before actual usage in a mushroom hunt. However, the study was restricted, usage was only simulated, and users were not able to test the app on their own. Thus, it might not have been possible for users with better AI knowledge to actually use their knowledge in order to evaluate the quality of the AI's decision outcomes. Future studies will have to (*i*) allow users to use the *Forestly* app more freely and (*ii*) additionally apply qualitative methods of user observation or log file analyses in order to understand how users might use the app differently depending on technology knowledge. Such artificiality in experiments has been criticized both in the literature on comprehension (Moehring et al., 2016) and in the XAI literature, for example, by Ehsan, Liao, Muller, Riedl and Weisz (2021). The latter emphasized that the interactions of humans with AI are socially situated and that XAI research must therefore take the socio-technical perspective into account, which was echoed by other researchers calling for XAI studies that are closer to real-world contexts (Kenny et al., 2021).

Ghassemi, Oakden-Rayner, and Beam (2021) argued that local explanations are unreliable or give only superficial information. In our study, explanations led to better performances in correctly picking edible mushrooms because the explanation communicated that the AI might be incorrect (e.g., giving questionable example pictures or only highlighting areas that do not show the actual mushroom). However, this need not be the case. Explanation is neither necessary nor sufficient for making a correct decision. According to Ghassemi et al. (2021), such explanations are useful only for model development, but not for individual decisions. As previously mentioned, the results of our study also need to be carefully examined by considering the specific study and material design. For example, effects of explainability could vary greatly depending on the characteristics of the items chosen. Therefore, study characteristics should be varied in future studies depending on different potential moderators, such as error rates. Additionally, trust dynamics and behavioral change should be studied for long-term usage

in order to test boundaries and robustness of these effects. However, although the results in this study might be specific and tied to the context and concrete implementations, explanations led to more appropriate trust in this situation; hence, the mushroom-picking task presented here could be used to educate people in exploring the boundaries of AI, and might thus improve general AI literacy. As discussed by Ng et al. (2021), such exploration of AI can then be seen as an educational intervention.

As mentioned before, the null results for AI knowledge and domain-specific knowledge contradicting previous research (e.g., Greiff et al., 2014; Moehring et al., 2016) could be due to inadequate measurement instruments (e.g., they might not have had the power to measure knowledge within the spectrum needed to predict study outcomes; i.e., they were not sensitive at this level of knowledge). Although not shown by the results, we would still argue that domain-specific knowledge is an important aspect for explanations: In order to know whether the AI looked at appropriate parts of the picture, users need domain knowledge about mushrooms. Thus, future studies need to explore various levels of expertise to test under what circumstances domain-specific knowledge is a driver of effects.

Finally, we must consider limitations of the statistical analyses we chose. While the use of FDR control was reasonable in this exploratory study due to the high number of statistical tests (Benjamini & Hochberg, 1995), FDR control is not a strict control of errors. This means, to test the robustness of these results, this study must be replicated with stricter error control. Since this study establishes which effect sizes can be expected approximately, only the most promising effects (that would be a considerably smaller number of tests) can then be tested specifically with strict error control like FWER in further studies.

6.3 Mushroom-picking task: a use case for further research

In this article we presented the use case of mushroom picking with an AI-based assistant system as a means of studying XAI and AI literacy research questions. Mushroom picking is an interesting use case for this purpose because it represents a decision-making situation of relevance to users due to the high stakes of the domain, availability of the target population, closeness to participants' reality, and public interest (e.g., in European countries such as Austria, Finland, and Sweden). It offers great potential for future research to explore other aspects of XAI and AI literacy.

A first important step could be replication of this study with the same manipulations but with stricter error control. Since it is an easy task to be implemented in online studies, results can be tested for their robustness.

Furthermore, of course other XAI techniques and various combinations of techniques could be tested with this use case. The educational intervention had no effect in this specific study, but this could be different for other implementations.

Another important aspect would be to replicate such studies in a study situation closer to the real-world context of mushroom picking to show generalizability beyond online behaviors. For this purpose, two approaches could be used: one could perform a laboratory experiment in which users are instructed to scan specific mushrooms in a restricted area; alternatively a field study with limited control could be conducted. In such study contexts, other (socio-technical) dynamics may become visible that are not apparent in restricted online studies (e.g., see studies on socio-technical effects, such as work by Ehsan, Liao et al., 2021).

Furthermore, it is important to confirm the results with other sets of items and characteristics (e.g., different error rates in AI decisions) in order to explore boundary conditions and robustness of effects more deeply.

For more holistic theory development, future studies could also include other relevant variables, such as personality traits (e.g., propensity to trust or risk aversion).

We used a specific pool of participants (e.g., mostly laypersons from Austria). Cross-cultural comparisons and testing of effects of other user characteristics, such as domain expertise, would be interesting.

Finally, mushroom picking is of interest in some countries, such as Austria and Sweden, but might not be of interest in others. However, this use case might share features with other use cases, for instance, because it is from a high-risk domain or has similar distributions of domain-specific knowledge in the population, and thus results could—to some extent—be generalized to other, similar use cases. To test the generalizability of effects, it would therefore be interesting to compare the results with those from other domains that are similar in the most central characteristics.

With this article, we provide a first basis for further research. We have shown that mushroom picking is an interesting use case for XAI that offers several advantages. We have provided first estimates and developed first measurement instruments of AI knowledge and domain-specific knowledge (i.e., mushroom knowledge), and a possible implementation of a mushroom-picking task for simple online experiments for other researchers to build upon in future studies.

6.4 Theoretical implications

In our study we showed that the educational intervention did not work as intended and that visual explanations of AI-based decisions had positive effects on human behavior and trust calibration. However, the directions and the size of effects vary compared to other studies. This shows that a specific effect cannot simply be assumed, but must be reconsidered depending on the context, and this empirically observed variance in turn provides information for theory-building. For example, while Körber et al. (2018) reported the effectiveness of an introductory video about a technology's limitation, our educational intervention had no effect. However, these educational attempts vary considerably in content (limitations during usage vs. abstract description of concepts) and medium (video vs. images and text). This indicates that more intensive and immersive training may be needed for educational interventions instead of simple, short text and image-based interventions. A simple general explanation of how AI works therefore does not seem to be sufficient to improve the interaction of users with AI, but may need to be more targeted in pointing out limitations in concrete examples or through exploration. As noted in the previous section, the effects of visual explanations on decision-making and trust can be compared to studies by Buçinca et al. (2020) or Kenny et al. (2021). As there are differences in effects, differences in study designs allow to point out possible moderators: error-rates, severity of negative consequences through wrong decisions, and other context-specific factors. Studies therefore need to be well contextualized and these differences in contexts from other studies need to be highlighted to understand how variances occur and to avoid over-generalizations. The theoretical contribution for both aspects—educational intervention and XAI methods—is therefor the reported effect sizes for concrete realizations and detailed descriptions of task characteristics. This will (*i*) allow future researchers to estimate expected effect sizes for power analyses, (*ii*) raise awareness about potential moderators in XAI studies, and (*iii*) constitute data for running moderator analyses in future meta-analyses. While a whole model and effects of various moderators cannot be investigated in single studies, a larger body of studies varying in different factors can be analyzed in future meta-analyses if well-documented. This body of effects in different contexts and with different user groups and technological characteristics will finally allow to build more complex theoretical models that take into account such moderator effects and dynamic changes based on the complex interplay of different factors. While this step is beyond the scope of the present work, the comparison of a broader body of studies will help to enlarge and refine current theoretical frameworks (e.g., on trust Körber, 2019). For example, whether XAI methods lead to higher or lower trust in an AI-based system compared to no explanations might be a function of the error-rate of the system and other factors. The results from this study are thus already a promising start for this overarching goal of theory building.

6.5 Practical implications

In our study, users tended to rate the AI-powered app more negatively when they received additional visual explanations. We believe that this feedback reflects the relatively poor performance of our AI, which became more apparent to the users due to the explanations. This might tempt developers of real-world applications to hold back explanations in favor of receiving more positive initial user feedback. However, in the long term any application will be judged based on the actual performance users achieve, especially when consequences of wrong decisions are as severe as in the context of mushroom classification (e.g., food poisoning after relying on an incorrect AI classification). We thus encourage app developers to consider including explanations or to indicate potential flaws in their system in a different way.

This also has direct business implications. At first it may seem that not including explanations in an AI-powered app may lead to more users, as suggested by the higher intention to use reported by the participants who saw the non-XAI version of the app. However, the overtrust caused by the absence of the explanations might lead to erroneous decisions (such as the takeaway of an inedible or poisonous mushroom). This would in turn lead to customers massively distrusting the software company—and trust repair after damage is usually very difficult (see for example Kim, Ferrin, Cooper, & Dirks, 2004). Thus, in addition to the ethical obligation, a direct insight into possible misclassifications through explanations could be recommended from a business perspective. When possible misclassifications are more apparent to users, user numbers might be lower first. However, the transparency of the system and the avoidance of trust repair after damage caused by misclassifications might be beneficial in the long term.

Implementing the educational intervention about AI with a brief text and information graphics had no significant impact on performance, comprehension, or trust calibration in our study. Reasons for this may be manifold, as previously discussed, but assuming that educational measures are ineffective would be wrong. Rather, the study shows that short, simple educational measures via classic text-picture combinations may not be as effective as hoped, and more elaborate interactive interventions are needed—for example, to allow exploration by users. This could mean that companies need to consider to invest in more elaborate educational interventions about how to use their product appropriately, such as shown by the introductory videos by Körber et al. (2018).

7 Conclusion

In this exploratory experiment we showed that the newly conceptualized mushroom-picking task is a promising use case for XAI research that allows the exploration of a variety of research questions within a high-risk decision context. We explored the effects of visual explanations on typical usage characteristics, such as performance, trust, and understanding. Furthermore, we explored the influence of common AI literacy predictors (i.e., AI knowledge and domain-specific knowledge) on models, which brought up new questions about the roles of potential moderators and boundary conditions. In this initial study, we investigated mushroom picking within a simple online environment. However, the use case of mushroom picking also raises exciting questions about generalizability to actual collecting behavior in more natural, less restrictive environments. The materials and results presented here can be a first starting point for such inquiries.

CRediT authorship contribution statement

Benedikt Leichtmann: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Writing – original draft. **Christina Humer:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. **Andreas Hinterreiter:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. **Marc Streit:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Martina Mara:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and Code are shared online under the OSF link provided in the article.

Acknowledgments

We thank Dr. Otto Stoik and the members of the Mycological Working Group (MYAG) at the Biology Center Linz, Austria, who supported us in the development of items for the mushroom knowledge test and in the collection of suitable mushroom images for this study. We also thank the members of the German Mycological Society (DGfM) for providing additional images. Additionally, we would like to thank Alfio Ventura for formatting the tables in the Supplementary Material.

Funding

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).

Appendix A. Detailed study information

A.1. Introductory text for the mushroom-picking task

The following text was presented as an introduction to the mushroom-picking task. The description was illustrated with copyrighted pictures.

> Imagine that you want to cook yourself a tasty mushroom dish for dinner. For this you want to pick fresh mushrooms in the forest, from which you will prepare this meal with the intention of eating it. Of course, it is important that you use only edible mushrooms and not inedible or even poisonous ones. To support you identifying mushrooms, an app called *Forestly* has been developed that uses artificial intelligence to analyze photos of mushrooms and decide (*i*) what type of mushroom it might be, and (*ii*) whether it is edible or inedible/poisonous. In the following, you will be shown a series of photos of mushrooms taken while being on the mushroom hunt. Your task is then to decide for each photo whether you would classify the mushroom shown in the photo as edible or inedible/poisonous, and whether you would take and eat that mushroom. As a possible decision aid, you will see the decision of the *Forestly* app, which analyzes the photo using artificial intelligence. For further development of the app, each decision also asks questions about how much you trust this app.

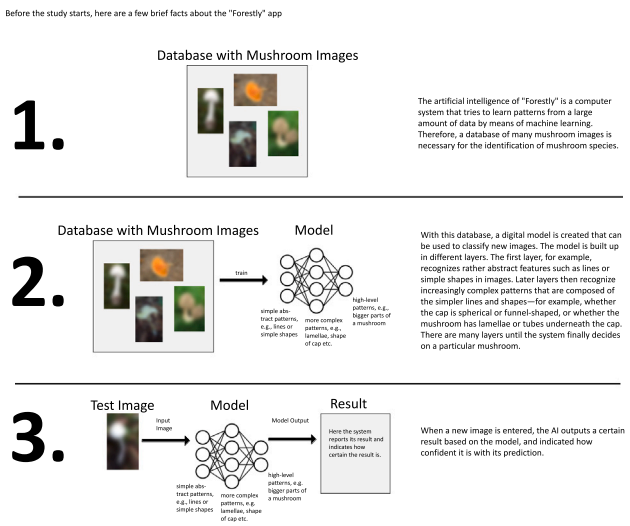


Fig. A.1. Educational intervention, describing briefly the requirements for, and inner workings of, a typical convolutional neural network (translated version).

Table B.1
Difficulties, standard deviations and item-total correlations for the mushroom knowledge items.

Item	M	SD	r_{it}
Agaricus	**.93**	**.26**	**.24**
Location	.68	.47	−.03
Imleria badia	**.31**	**.46**	**.40**
Funnel form	.64	.48	.15
Mushroom type	**.43**	**.50**	**.40**
Porcino location	**.43**	**.50**	**.42**
Sweetbread mushroom	.22	.42	−.01
Sheathed woodtuft	.23	.42	.03
Parasol mushroom	**.42**	**.49**	**.15**
Porcino	**.27**	**.44**	**.37**
Poison	.16	.37	.14

Note: **Bold items** were included in the measurement model.

A.2. Educational intervention

See Fig. A.1.

A.3. Forestly App Description

See Fig. A.2.

A.4. User interface for edibility assessment

See Fig. A.3.

Appendix B. Item characteristics for the mushroom knowledge, AI knowledge, and task-specific AI comprehension test

See Tables B.1–B.3

Appendix C. Network architecture

See Fig. C.4.

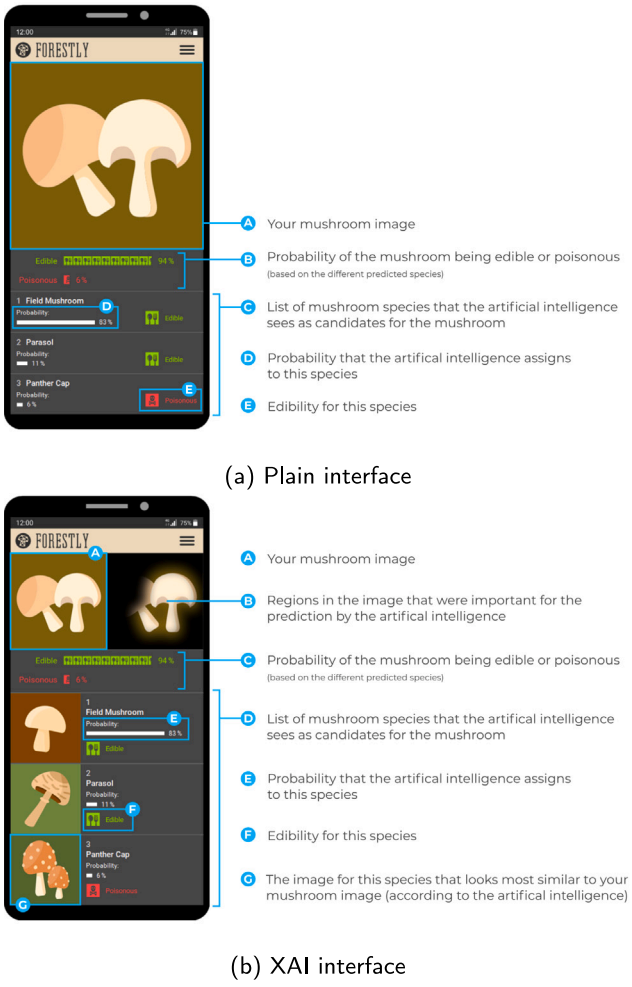


Fig. A.2. Information about how the *Forestly* app displays the AI's prediction for two interface variants (translated versions).

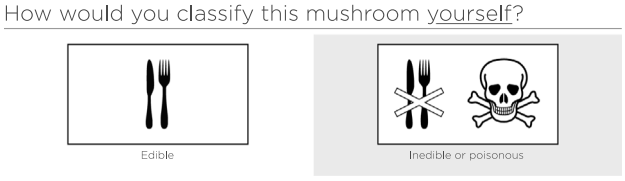


Fig. A.3. User interface for the edibility assessment task (translated version).

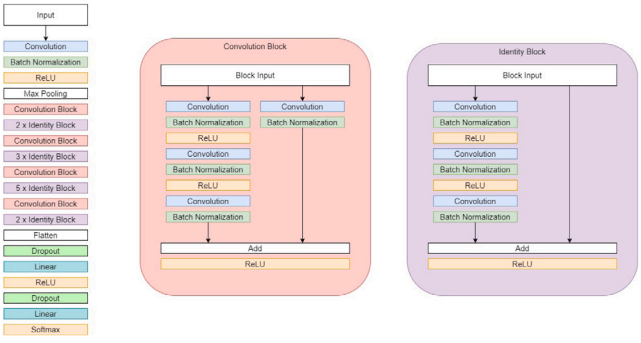


Fig. C.4. Network architecture of the classifier used to generate the data for the user study. The architecture is based on ResNet50 (He et al., 2016a), with the last layers being replaced by custom linear layers. The architecture (left) consists of several **Convolution** (center) and **Identity Blocks** (right).

Table B.2
Difficulties, standard deviations and item-total correlations for the AI knowledge items.

Item	*M*	*SD*	r_{it}
Fox	**.53**	**.50**	**.44**
Basis	**.27**	**.45**	**.48**
Improving	**.64**	**.48**	**.34**
Racism	**.67**	**.47**	**.23**
Statements	.26	.44	.10
Comprehension	.77	.42	.19
Tasks	.58	.49	.18
Logistics	**.61**	**.49**	**.41**

Note: **Bold items** were included in the measurement model.

Table B.3
Difficulties, standard deviations and item-total correlations for the task-specific AI comprehension items.

Item	*M*	*SD*	r_{it}
Percent	**.70**	**.46**	**.29**
Error	**.43**	**.50**	**.45**
Class	**.55**	**.50**	**.33**
Expert	**.66**	**.47**	**.37**
Single class	.21	.41	−.04

Note: **Bold items** were included in the measurement model.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.chb.2022.107539. Additional supplementary material can be found under https://osf.io/tqbgh/. This includes newly developed test items, the R code used for data analysis, raw data with variable descriptions, and python code used to train the AI model and to generate the explanations.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2020). Sanity Checks for Saliency Maps. arXiv:1810.03292 [cs, stat]. Retrieved April 27, 2021, from http://arxiv.org/abs/1810.03292.

Aigner, L., & Krisai-Greilhuber, I. (2016). Eine ethnomykologische studie über das pilzwissen in der bevölkerung des waldviertels. *Österreichische Zeitschrift für Pilzkunde*, *25*, 209–224.

Alicioglu, G., & Sun, B. (2022). A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, *102*, 502–520. http://dx.doi.org/10.1016/j.cag.2021.09.002.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. http://dx.doi.org/10.1016/j.inffus.2019.12.012.

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, *117*(48), 30071–30078. http://dx.doi.org/10.1073/pnas.1907375117.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186–203. http://dx.doi.org/10.1207/s15328007sem13022.

Behnke, G., Leichtmann, B., Bercher, P., Höller, D., Nitsch, V., Baumann, M., & Biundo, S. (2017). Help me make a dinner! Challenges when assisting humans in action planning. In *2017 international conference on companion technology (ICCT)* (pp. 1–6). http://dx.doi.org/10.1109/ICCT42709.2017.9151907.

Benjamini, Y., & Hochberg, Y. (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), 289–300. http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics*, *5*(4), http://dx.doi.org/10.1214/11-AOAS495.

Brandenburg, W. E., & Ward, K. J. (2018). Mushroom poisoning epidemiology in the United States. *Mycologia*, *110*(4), 637–641. http://dx.doi.org/10.1080/00275514.2018.1479561.

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, *42*(1), 17–25. http://dx.doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (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* (pp. 454–464). ACM, http://dx.doi.org/10.1145/3377325.3377498.

Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 258–262). Association for Computing Machinery, http://dx.doi.org/10.1145/3301275.3302289.

Carayon, P. (2006). Human factors of complex sociotechnical systems. *Applied Ergonomics*, *37*(4), 525–535. http://dx.doi.org/10.1016/j.apergo.2006.04.011.

Chollet, F., Zhu, Q. S., Ayala-Acevedo, A., Zhang, Y., Majumdar, S., Zabluda, O., Gajare, N., Lee, T., Wood, L., Rasul, K., Kawamura, H., Tencé, F., & Zámečník, B. (2021). TensorFlow core v2.8.0: imagenet_utils.py. [Accessed August 4, 2022]. https://github.com/keras-team/keras/blob/r2.8/keras/applications/imagenet_utils.py.

Covello, S., & Lei, J. (2010). *A review of digital literacy assessment instruments*. Syracuse University, https://www.academia.edu/7935447.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). Harper & Row.

Danish Mycological Society (2022). Danish fungal records database. www.svampeatlas.dk. Accessed March 29, 2022.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94. http://dx.doi.org/10.7861/futurehosp.6-2-94.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). http://dx.doi.org/10.1109/CVPR.2009.5206848, ISSN: 1063-6919.

Dick, S. (2019). Artificial Intelligence. *Harvard Data Science Review*, *1*(1), http://dx.doi.org/10.1162/99608f92.92fe150c.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st international convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). IEEE, http://dx.doi.org/10.23919/MIPRO.2018.8400040.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77. http://dx.doi.org/10.1145/3359786.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–19). http://dx.doi.org/10.1145/3411764.3445188.

Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., Riener, A., & Riedl, M. O. (2021). Operationalizing human-centered perspectives in explainable AI. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1–6). ACM, http://dx.doi.org/10.1145/3411763.3441342.

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé, H., III, Riener, A., & Riedl, M. O. (2022). Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. In *CHI conference on human factors in computing systems extended abstracts* (pp. 1–7). New Orleans LA USA: ACM, http://dx.doi.org/10.1145/3491101.3503727.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. http://dx.doi.org/10.3758/BF03193146.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. http://dx.doi.org/10.1007/s11023-018-9482-5.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv:1506.02142 [cs, stat]. Retrieved March 30, 2022, from http://arxiv.org/abs/1506.02142.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. http://dx.doi.org/10.1016/S2589-7500(21)00208-9.

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology*, *66*(1), 877–902. http://dx.doi.org/10.1146/annurev-psych-010814-015321.

Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B., & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology*, *106*(3), 666–680. http://dx.doi.org/10.1037/a0035426.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. http://dx.doi.org/10.1145/3236009.

Hannibal, G., Weiss, A., & Charisi, V. (2021). "The robot may not notice my discomfort" – examining the experience of vulnerability for trust in human-robot interaction. In *2021 30th IEEE international conference on robot and human interactive communication (RO-MAN)* (pp. 704–711). IEEE, http://dx.doi.org/10.1109/RO-MAN50785.2021.9515513.

Hättenschwiler, N., Mendes, M., & Schwaninger, A. (2019). Detecting bombs in X-Ray images of hold baggage: 2D versus 3D imaging. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *61*(2), 305–321. http://dx.doi.org/10.1177/0018720818799215.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90, ISSN: 1063-6919.

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. arXiv:1603.05027 [cs]. Retrieved March 30, 2022, from http://arxiv.org/abs/1603.05027.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. http://dx.doi.org/10.1177/0018720814547570.

Howell, A. J., Dopko, R. L., Passmore, H.-A., & Buro, K. (2011). Nature connectedness: Associations with well-being and mindfulness. *Personality and Individual Differences*, *51*(2), 166–171. http://dx.doi.org/10.1016/j.paid.2011.03.037.

Huber, T., Weitz, K., André, E., & Amir, O. (2021). Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, *301*, Article 103571. http://dx.doi.org/10.1016/j.artint.2021.103571.

Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., & Srivastava, M. (2020). How can I explain this to you? An empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, *33*.

Jiao, Z., Choi, J. W., Halsey, K., Tran, T. M. L., Hsieh, B., Wang, D., Eweje, F., Wang, R., Chang, K., Wu, J., Collins, S. A., Yi, T. Y., Delworth, A. T., Liu, T., Healey, T. T., Lu, S., Wang, J., Feng, X., Atalay, M. K., ... Bai, H. X. (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), e286–e294. http://dx.doi.org/10.1016/S2589-7500(21)00039-X.

Kaaronen, R. O. (2020). Mycological rationality: Heuristics, perception and decision-making in mushroom foraging. *Judgment and Decision Making*, *15*(5), 630–647.

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (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*, Article 103459. http://dx.doi.org/10.1016/j.artint.2021.103459.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- Versus Integrity-Based Trust Violations. *Journal of Applied Psychology*, *89*(1), 104–118. http://dx.doi.org/10.1037/0021-9010.89.1.104.

Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in Neural Information Processing Systems*, *29*, 2288–2296.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. arXiv:1412.6980 [cs]. Retrieved March 30, 2022, from http://arxiv.org/abs/1412.6980.

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Advances in intelligent systems and computing*, *Proceedings of the 20th congress of the international ergonomics association (IEA 2018), Vol. 823* (pp. 13–30). Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-96074-6_2.

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31. http://dx.doi.org/10.1016/j.apergo.2017.07.006.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human Factors*, *62*(5), 718–736. http://dx.doi.org/10.1177/0018720819853686.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345–360. http://dx.doi.org/10.1080/00140139.2018.1547842.

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. http://dx.doi.org/10.1518/hfes.46.1.5030392.

Leichtmann, B., Bercher, P., Höller, D., Behnke, G., Biundo, S., Nitsch, V., & Baumann, M. (2018). Towards a companion system incorporating human planning behavior: A qualitative analysis of human strategies. In R. Weidner, & A. Karafillidis (Eds.), *Dritte transdisziplinäre konferenz "technische unterstützungssysteme, die die menschen wirklich wollen" 2018* (pp. 89–98).

Leichtmann, B., Nitsch, V., & Mara, M. (2022). Crisis ahead? Why human-robot interaction user studies may have replicability problems and directions for improvement. *Frontiers in Robotics and AI*, *9*, Article 838116. http://dx.doi.org/10.3389/frobt.2022.838116.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43. http://dx.doi.org/10.1145/3233231.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. http://dx.doi.org/10.1016/j.media.2017.07.005.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In R. Bernhaupt (Ed.), *ACM digital library*, *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–16). Association for Computing Machinery, http://dx.doi.org/10.1145/3313831.3376727.

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, *28*(5), 823–870. http://dx.doi.org/10.1080/01431160600746456.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, https://proceedings.neurips.cc/paper/2017.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. http://dx.doi.org/10.1016/j.jrp.2013.09.008.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. http://dx.doi.org/10.2466/pr0.1990.66.1.195.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. http://dx.doi.org/10.1016/j.artint.2018.07.007.

Miranda, S., Berente, N., Seidel, S., Safadi, H., & Burton-Jones, A. (2022). Editor's comments: computationally intensive theory construction: a primer for authors and reviewers. 46 (2).

Moehring, A., Schroeders, U., Leichtmann, B., & Wilhelm, O. (2016). Ecological momentary assessment of digital literacy: Influence of fluid and crystallized intelligence, domain-specific knowledge, and computer usage. *Intelligence*, *59*, 170–180. http://dx.doi.org/10.1016/j.intell.2016.10.003.

Molnar, C. (2022). *Interpretable machine learning: a guide for making black box models explainable* (2nd ed.). christophm.github.io/interpretable-ml-book/.

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: a mapping review. *Social Science & Medicine*, *260*, Article 113172. http://dx.doi.org/10.1016/j.socscimed.2020.113172.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080. http://dx.doi.org/10.1073/pnas.1900654116.

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509. http://dx.doi.org/10.1002/pra2.487.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), Article e7. http://dx.doi.org/10.23915/distill.00007.

Olderbak, S., & Wilhelm, O. (2020). Overarching principles for the organization of socioemotional constructs. *Current Directions in Psychological Science*, *29*(1), 63–70. http://dx.doi.org/10.1177/0963721419884317.

Pangrazio, L., Godhe, A.-L., & Ledesma, A. G. L. (2020). What is digital literacy? A comparative review of publications across three language contexts. *e-Learning and Digital Media*, *17*(6), 442–459. http://dx.doi.org/10.1177/2042753020946291.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. http://dx.doi.org/10.1518/001872097778543886.

Park, B. J., Tsunetsugu, Y., Kasetani, T., Kagawa, T., & Miyazaki, Y. (2010). The physiological effects of shinrin-yoku (taking in the forest atmosphere or forest bathing): evidence from field experiments in 24 forests across Japan. *Environmental Health and Preventive Medicine*, *15*(1), 18–26. http://dx.doi.org/10.1007/s12199-009-0086-9.

Peintner, U., Schwarz, S., Mešić, A., Moreau, P.-A., Moreno, G., & Saviuc, P. (2013). Mycophilic or mycophobic? Legislation and guidelines on wild mushroom commerce reveal different consumption behaviour in European countries. *PLoS One*, *8*(5), Article e63926. http://dx.doi.org/10.1371/journal.pone.0063926.

Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, *29*(9), 2352–2449. http://dx.doi.org/10.1162/neco_a_00990, Conference Name: Neural Computation.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *KDD'16: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). http://dx.doi.org/10.1145/2939672.2939778.

RStudio Team (2020). *RStudio: Integrated development environment for R*. Boston, MA: RStudio, PBC, http://www.rstudio.com/ [Accessed March 29, 2022].

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics*, *59*(6), 767–780. http://dx.doi.org/10.1080/00140139.2015.1094577.

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. http://dx.doi.org/10.1177/1745691620966795.

Schroeders, U., Wilhelm, O., & Bucholtz, N. (2010). Reading, listening, and viewing comprehension in english as a foreign language: One or more constructs? *Intelligence*, *38*(6), 562–573. http://dx.doi.org/10.1016/j.intell.2010.09.003.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV)* (pp. 618–626). Retrieved January 30, 2018, from http://arxiv.org/abs/1610.02391.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In Y. Bengio, & Y. LeCun (Eds.), *International conference on learning representations (ICLR) workshop track proceedings*. Retrieved January 30, 2018, from http://arxiv.org/abs/1312.6034.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs]. Retrieved March 30, 2022, from http://arxiv.org/abs/1409.1556.

Svanberg, I., & Lindh, H. (2019). Mushroom hunting and consumption in twenty-first century post-industrial Sweden. *Journal of Ethnobiology and Ethnomedicine*, *15*(1), 42. http://dx.doi.org/10.1186/s13002-019-0318-z.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2818–2826). http://dx.doi.org/10.1109/CVPR.2016.308, ISSN: 1063-6919.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., & Spitzer, M. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, *18*(6), 463–477. http://dx.doi.org/10.1038/s41573-019-0024-5.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, Article 103404. http://dx.doi.org/10.1016/j.artint.2020.103404.

Visipedia (2018). 2018 FGCVx fungi classification challenge. https://www.kaggle.com/competitions/fungi-challenge-fgvc-2018/. Accessed March 29, 2022.

Votto, A. M., Valecha, R., Najafirad, P., & Rao, H. R. (2021). Artificial intelligence in tactical human resource management: A systematic literature review. *International Journal of Information Management Data Insights*, *1*(2), Article 100047. http://dx.doi.org/10.1016/j.jjimei.2021.100047.

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th international conference on intelligent user interfaces* (pp. 318–328). Association for Computing Machinery, http://dx.doi.org/10.1145/3397481.3450650.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 189–201). ACM, http://dx.doi.org/10.1145/3377325.3377480.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Dissertation), Los Angeles: University of California.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, *2*(10), 719–731. http://dx.doi.org/10.1038/s41551-018-0305-z.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1–27. http://dx.doi.org/10.1186/s41239-019-0171-0.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Lecture notes in computer science*: *8689*, *Computer vision – ECCV 2014* (pp. 818–833). Springer, http://dx.doi.org/10.1007/978-3-319-10590-1_53.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *FAT* '20*, *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 295–305). Association for Computing Machinery, http://dx.doi.org/10.1145/3351095.3372852.