Effects of AI and Logic-Style Explanations on Users' Decisions Under Different Levels of Uncertainty

FEDERICO MARIA CAU, University of Cagliari
HANNA HAUPTMANN, Utrecht University
LUCIO DAVIDE SPANO, University of Cagliari
NAVA TINTAREV, Maastricht University

Existing eXplainable Artificial Intelligence (XAI) techniques support people in interpreting AI advice. However, although previous work evaluates the users' understanding of explanations, factors influencing the decision support are largely overlooked in the literature. This article addresses this gap by studying the impact of *user uncertainty*, *AI correctness*, and the interaction between *AI uncertainty* and *explanation logic-styles* for classification tasks. We conducted two separate studies: one requesting participants to recognize handwritten digits and one to classify the sentiment of reviews. To assess the decision making, we analyzed the *task performance, agreement* with the AI suggestion, and the user's *reliance* on the XAI interface elements. Participants make their decision relying on three pieces of information in the XAI interface (image or text instance, AI prediction, and explanation). Participants were shown one explanation style (between-participants design) according to three styles of logical reasoning (inductive, deductive, and abductive). This allowed us to study how different levels of AI uncertainty influence the effectiveness of different explanation styles. The results show that user uncertainty and AI correctness on predictions significantly affected users' classification decisions considering the analyzed metrics. In both domains (images and text), users relied mainly on the instance to decide. Users were usually overconfident about their choices, and this evidence was more pronounced for text. Furthermore, the inductive style explanations led to overreliance on the AI advice in both domains—it was the most persuasive, even when the AI was incorrect. The abductive and deductive styles have complex effects depending on the domain and the AI uncertainty levels.

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

Additional Key Words and Phrases: Explainable AI, user uncertainty, AI uncertainty, AI correctness, explanations, logical reasoning, MNIST, Yelp Reviews, neural networks, CNNs, intelligent user interfaces

F. M. Cau gratefully acknowledges CRS4.Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna for collaboration on the RIALE (Remote Intelligent Access to Lab Experiment) platform. This work was partially supported by the Sardinia Regional Government and by Fondazione di Sardegna, ADAM (FdS 2018) (CUP F74I19000900007) and ASTRID (FdS 2020) (CUP F75F21001220007) projects.

Authors' addresses: F. M. Cau and L. D. Spano, University of Cagliari, Palazzo Delle Scienze, Via Ospedale 72, 09124 Cagliari CA, Sardegna, Italia; emails: federico.cau@maastrichtuniversity.nl, davide.spano@unica.it; H. Hauptmann, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, The Netherlands; email: hjhauptmann@uu.nl; N. Tintarev, Maastricht University, Minderbroedersberg 4-6, 6211 LK Maastricht, Limburg, The Netherlands; email: n.tintarev@maastrichtuniversity.nl.

2160-6455/2023/12-ART22 $15.00
https://doi.org/10.1145/3588320

ACM Reference format:
Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Effects of AI and Logic-Style Explanations on Users' Decisions Under Different Levels of Uncertainty. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 22 (December 2023), 42 pages.
https://doi.org/10.1145/3588320

1 INTRODUCTION

In the past decade, **eXplainable Artificial Intelligence (XAI)** research has led to significant advances, introducing many explanation techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) [64], SHAP (SHapley Additive exPlanations) [51], LIME (Local Interpretable Model-Agnostic Explanation) [60], and DeepLIFT [66]. The "interpretability" of such approaches has been qualified as the degree to which a human can understand the cause of a decision made using an AI method [53]. However, understanding is rarely an end goal in itself, and it is often more meaningful to measure the effectiveness of explanations in terms of a specific notion of usefulness or explanatory goals, such as Task Performance (whether the user makes the correct decision or not), Agreement (whether the user follows the AI advice in making the decision or not), and Reliance on system advice [72] (which pieces of information in the XAI interface impact the final decision).

Largely overlooked in the literature are the factors that influence the effectiveness of explanations, such as User Uncertainty, AI Correctness, AI Uncertainty, and different Explanation Styles [2, 9, 34, 38, 45, 56, 61, 62, 87].

First, User Uncertainty occurs for many decision-making tasks, either due to the inherent (objective) complexity of the task, or more subjective and specific to the user and context (e.g., for previously unseen tasks). When user uncertainty is low, people may need less (or different) information than when they experience high uncertainty. People are used to making decisions under uncertainty, which may also result in overconfidence in their decisions when making decisions about AI predictions [92]. In this article, we define User Uncertainty as the distribution of users' accuracy per instance in a population of users, trying to capture the degree of difficulty in a subset of observations through users' performance results. We model user uncertainty by finding potential difficult instances to classify considering two approaches: (i) situations where the user does not have enough information *and* where observations can have different meanings across different groups of users, and (ii) methods applied to AI models that estimate potential noise in the data like *heteroscedastic* (*aleatoric*) uncertainty [36]. In contrast, we define the notion of User Confidence as the subjective measure indicating self-trust in the target task, as indicated by a specific user. Second, the prediction of AI is not always correct. AI Correctness of a specific prediction can influence the decision-making process in different ways. For example, a surprising incorrect prediction may cause the user to reflect more critically about the decision, or an intuitive but incorrect prediction could influence the user's decision to follow the system's (incorrect) advice. Previous work suggests that the correctness of the AI prediction has a direct influence on which information people use during decision making [38]. As recently pointed out by Kenny et al. [38], when the AI is correct, its prediction is sufficient for users in confirming their judgment or changing the decision. They also discovered that the perception of the overall system confidence decreases as the error rate increases—suggesting that other kinds of information or explanation might be necessary for systems with low AI correctness.

Third, a prediction may be correct, whereas the system has indications of high AI Uncertainty. Inspired by the findings of Kenny et al. [38], showing that explanations impact people's judgment of AI errors, we expect that explanations may indirectly expose the level of AI uncertainty to the users who, in turn, may change their decision accordingly. Indicating the uncertainty to the users may remind them to question the system advice more, but such a distinction is still underinvestigated in the literature. Only a few articles have evaluated the effects of AI uncertainty with users (e.g., [2, 76]). Further, conveying the AI uncertainty in a way that supports decision making without users losing too much trust in the system is a delicate challenge [52]. Therefore, we analyze the interaction between the AI Uncertainty and the Explanation Style on users' Task Performance and their Agreement with predictions.

In this article, we estimate the AI uncertainty of the model using what has been called the *epistemic* uncertainty, which is type of uncertainty that can be reduced by adding more data or improving the model [24].

Finally, the Explanation Style we present to users is also likely to influence the decision process. Different explanation styles can support different logical reasoning strategies (i.e., inductive, abductive, and deductive) [10].

To understand the influence of these four factors (User Uncertainty, AI Correctness, AI Uncertainty, Explanation Style), we conducted two separate user studies for two classification tasks: *handwritten digit recognition and sentiment analysis for reviews*. We had 659 participants for the image task and 665 for the text task, each assigned to one of the four explanation conditions (no explanation, inductive, abductive, deductive), further divided by the user and AI uncertainty (low, high) and AI correctness (right, wrong). To assess the quality of decision support, we measured the users' reliance, task performance, and agreement with the AI prediction. To the best of our knowledge, this is the first work in the literature studying the relationships among these four factors in AI-supported decision-making processes. This allows us to better characterize the situations in which people disagree with correct classifications, and agree with incorrect classifications.

We structured our work around the following research questions:

RQ1: Does the user Reliance on the explanation, instance, and AI prediction depend on the User Uncertainty level?

RQ2: Does users' Task Performance and Agreement with the AI depend on the User uncertainty, AI Correctness, and the interaction between the Explanation Style and AI Uncertainty?

The article is organized as follows. In Section 2, we give details on measuring user and AI uncertainty, XAI techniques involved when testing AI explanations with users, logical reasoning types, and evaluation metrics in user studies. In Section 3, we deepen the definition of reasoning types and identify some examples from the literature for both image and text data. In Section 4, we describe the method, hypotheses, and settings of our user studies on image and text data, whereas Section 5 presents the results. Section 6 proposes a discussion on the results by highlighting their implications and Section 7 their limitations. We conclude the article and mention future work in Section 8.

2 RELATED WORK

This section summarizes previous work on three relevant topics for our research, which ground the definition of the experimental setup in our two user studies. The first is measuring (AI) uncertainty in deep learning models and how to estimate uncertainty on the user side. The second topic is the categorization of XAI techniques, focusing on methods applicable to neural networks in the image and text domains, including user studies validating them with users (when available). The third topic is about task properties and evaluation metrics employed in evaluating XAI systems and interfaces with users, focusing on users' reliance, task performance, and agreement with the AI.

2.1 Uncertainty in Neural Network Models

Considering deep learning models for classification tasks, a possible strategy to capture uncertainty is to employ Bayesian deep learning approaches [24]. In Bayesian modeling, we can characterize two types of uncertainty: epistemic and aleatoric [14]. *Epistemic* (or model) uncertainty measures the uncertainty in the model parameters and can be explained away given enough data [36]. This type of uncertainty is "knowledge uncertainty" and sometimes referred to as "reducible uncertainty" [24]. A way to model epistemic uncertainty is using a type of variational inference called *Monte Carlo dropout sampling* at test time [25], including dropout layers in the model of interest and running it multiple times to create a distribution of outcomes. After that, we calculate the predictive entropy of this distribution [36]. We use epistemic uncertainty to distinguish AI Uncertainty as high or low.

Aleatoric (or data) uncertainty captures our uncertainty in the observations, like noisy or class overlapping input data. Aleatoric uncertainty is also known as irreducible uncertainty: it cannot be decreased by observing more data [36]. We can categorize aleatoric uncertainty into two further sub-categories: heteroscedastic and homoscedastic [37]. Heteroscedastic uncertainty depends on the input data and is predicted as a model output. One way to calculate this type of uncertainty is by using a modified loss function and letting the Bayesian deep learning model predict both softmax values and input (predicting) variance [24]. Homoscedastic or task-dependent uncertainty, on the contrary, stays constant for all input data and varies between different tasks such as semantic segmentation, instance segmentation, and depth regression [37]. Therefore, it is beneficial for applications such as multi-task learning, where multiple objectives need to be learned from a shared representation. Since heteroscedastic uncertainty captures potential noise in input data, we use this type of uncertainty for a preliminary categorization of low versus high User Uncertainty.

2.2 Uncertainty in Human Reasoning

In real life, people often make decisions under uncertainty concerning a given phenomenon in a given context. To model these decisions, it is necessary to define the causes and strategies they apply according to the context. Zimmermann [92] outlines a taxonomy of uncertainty properties, divided into causes of subjective uncertainty, available information, scale level of numerical information, and required information. Furthermore, he also proposes a definition of uncertainty: "Uncertainty implies that in a certain situation a person does not dispose about information which quantitatively and qualitatively is appropriate to describe, prescribe or predict deterministically and numerically a system, its behavior or other characteristica" [92]. With the term *system*, the author refers to a phenomenon about which judgments are to be made, such as part of physical reality or an artificial system. A more recent article by Schunn and Trafton [63] outlines a taxonomy of uncertainty sources mainly focused on functional magnetic resonance imaging (fMRI). The article divides sources of information uncertainty into four classes: physics uncertainty, computational uncertainty, visualization uncertainty, and cognitive uncertainty.

For our research purposes, we focus on two causes of uncertainty identified by Zimmermann [92], namely lack of information and ambiguity. The first describes a situation in which a user does not have enough information to describe an observable phenomenon accurately. The second identifies a situation where an observed phenomenon can have different meanings, or in which we have a one-to-many mapping. We use these two sources of uncertainty for the user evaluation on image data described in Section 4 to make a preliminary distinction between low and high user uncertainty, which we refined evaluating it with real users. If an image presents at least one of these two uncertainty sources, we assign it to a high user uncertainty and otherwise to a low user uncertainty. Figure 1 shows an example of high user uncertainty instances that contain these two sources. The top row of the image shows instances with a lack of information: we can notice

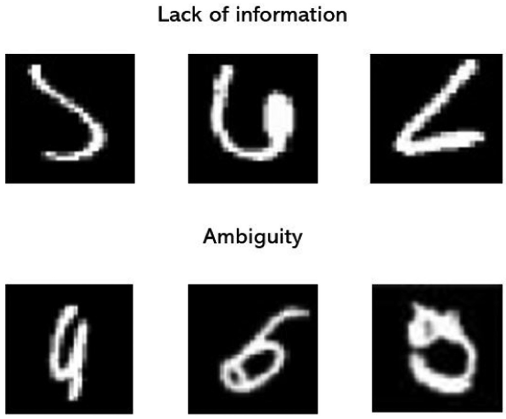


Fig. 1. Some of the MNIST [46] digits with a high user uncertainty: the top row presents instances with a lack of information, whereas the bottom row presents instances with ambiguity.

that these digits have missing strokes or a particular shape that differs from the well-known handwritten digits, making them more similar to symbols. The bottom row shows ambiguous instances, which capture characteristics belonging to more than one class. For example, the digit on the left resembles both a nine and a four without having strong evidence related to one particular digit.

For text data, we did not find a single validated general metric that captures uncertainty in text. We tested two reading difficulty metrics, Flesch-Reading-Ease [21] and Flesch-Kincaid metrics [41], but they were too sensitive to sentence length, finding "difficult" instances that were easy to classify for users. Hence, we use aleatoric uncertainty and the preliminary validation with users to capture User Uncertainty for both reviews and images.

2.2.1 User Confidence. User Uncertainty tries to capture the degree of difficulty in a subset of instances leveraging users' accuracy results in single observations, whereas user confidence represents a subjective measure indicating the self-trust on the target task depending on the current user. Some previous studies have found that explanations can increase user confidence [2, 56]. Most notably, Nourani et al. [56] studied how ordering biases can affect users' mental models and reliance/confidence formation in intelligent systems and the role the explanations play with such biases. They conducted experiments in an explainable video activity recognition tool in the cooking domain, exposing model weaknesses and strengths. They discovered that participants were more confident in their predictions when explanations were present, and users experienced overconfidence in their mental model either when explanations were present or when the model's strengths were observed earlier. Further, users who observed the model's strengths relied on the system more than they should (overreliance), and users who observed the model's weaknesses relied more on themselves than the system than they should (overconfidence). In our study, we asked users to express their confidence (low or high) in solving the classification task to assess the quality of our User Uncertainty labeling. What we are trying to capture with user uncertainty is a global and more objective view on User Confidence [2, 56].

2.3 XAI Techniques Overview

In the literature, we can find different taxonomies attempting a categorization of the different dimensions involved in generating explanations [77, 89]. For instance, we can distinguish explainability methods according to their scope. Local explanation methods focus on data and provide motivations on the model outcomes considering one or more instances. Instead, global methods explain the general mechanism by which the model works. Explainability methods can be intrinsic, approximating black-box models with interpretable ones, like decision trees or linear models. These methods also include models specifically created to render an explanation along with their output, like attention mechanisms [85]. Instead, post hoc methods extract information from an already learned model and provide some evidence for system outputs. We can also distinguish between model-specific methods, limited to specific model classes, or model agnostic, that apply to any already trained model.

We have categorizations of these methods according to the model type, data type, and explanatory goals [5, 85]. Considering the objectives in our studies, we discuss the methods that apply to neural network models in the image and text domains, focusing on articles that evaluated these methods with users.

A widely used method to explain an AI model is to provide example-based explanations. There are many ways to generate them: for instance, we can use similar or relevant training instances to explain the test ones from the target model [43, 90]. Another way is to leverage the Twin Systems approach [35] by coupling a black-box with case-based reasoning (CBR) system [71] and map the feature weights from the former to the latter to find cases that explain the networks' outputs. A recent study tested this technique with users [38], finding that example-based explanations impact users understanding of the AI model when it makes wrong predictions, letting them perceive the AI model acting in a more correct (or less incorrect) way.

Saliency methods are another popular class of tools that highlight relevant features in the input to explain predictions from a learned model. The large majority of these methods rely on gradients, such as Grad-CAM, [64], SmoothGrad [69], and GBP (Guided Back Propagation) [70]. Other famous methods that are not gradient based are SHAP [51], LIME [60], and LRP (Layer-wise Relevance Propagation) [7]. For image data, a remarkable study about the performance of gradient-based methods from Adebayo et al. [1] shows that some of them reflect the relationship between the model's weights on training examples and their labels. Besides, saliency maps methods better expose the trustworthiness of a classifier and inspire more trust in users [12, 64] if compared to other explanation techniques (i.e., [60]), but we do not have results on real decisions. Additionally, they may help users learn about some specific image features the system is sensitive to, although the maps do not increase their ability to predict the network's outcome for new images [3]. For text data, Lertvittayakumjorn et al. [47] reported that LIME is a good class discriminative method compared to other XAI techniques (Grad-CAM, LRP, DeepLIFT [67]), justifying predictions with relevant evidence.

Another way to explain network predictions is the attention mechanism, which approximates the human visual attention. For image data, they are widely employed in classification tasks or visual question answering systems [28, 80]. They have been tested with users showing that they provide good connections among visual explanations, textual explanations, and the visual question answering process [84]. For text data, the work of Bai et al. [8] highlights the pitfalls of attention mechanisms and proposes solutions on how to mitigate them, although no user evaluation is carried out on these methods. Further, attribute-based explainability methods for images provide rationales based on the input data, which uses class-relevant attributes to explain the prediction outcome [32, 73]. For text data, other methods provide explanations via concepts in the input sample and the training data, such as SELFEXPLAIN [58], perceived by users as more trustworthy, understandable, and adequate for explaining model decisions compared to saliency maps and influence attribution methods.

2.3.1 Types of Reasoning. Explanations leverage human reasoning for showing why the AI suggests a decision. Therefore, one interesting factor involved in the decision process is the reasoning style the explanations support. Buçinca et al. [9] briefly discusses inductive and deductive reasoning, explaining how to integrate them in a user evaluation context, without an in-depth exploration. Furthermore, abductive reasoning is often (unintentionally) used to compare novel explanation techniques with state-of-the-art techniques where the user role is to identify the best-generated explanation during the evaluation. Another interesting article from Van der Waa et al. [75] compare rule-based and example-based contrastive explanations: the former consists of *"if... then..."* statements, whereas the latter reports past experiences similar to the one examined. These explanations recall logical reasoning styles. We can match the rule-based explanations with deductive reasoning and the example-based ones with inductive reasoning. Wang et al. [79] highlight that the AI role is to facilitate the user connection with its decisions, starting from the reasoning expressed through the AI explanations. They conclude that a reasonable choice is to deeply explore reasoning theories. In this article, we consider Peirce's syllogistic theory [20] and focus on logical ones: inductive, deductive and abductive. We define the reasoning styles associated to a given explanation technique by identifying the conclusion (or result), the major premise (the rule), and the minor premise (the cause). We investigate reasoning styles and how to couple them with existing explainability methods in Section 3.

2.4 Evaluating XAI Systems

2.4.1 Defining the Tasks. To evaluate XAI systems appropriately, we need to define what type of task the user will accomplish to collect the data. For this purpose, Doshi-Velez and Kim [16] define a taxonomy for evaluating XAI systems, highlighting three categories: the first is the application-grounded evaluation, which affects domain experts evaluated on actual tasks. The second is the human-grounded evaluation, which considers novice users evaluated on simplified tasks. The last is the functionally grounded evaluation, which requires no human experiments and uses some formal definition of interpretability as a proxy for explanation quality. We position our work in the second evaluation type. Further, Narayan et al. [55] consider tasks supported by an explanations generator system, evaluating with users whether the output is consistent with the input and explanation. They distinguish between intrinsic and extrinsic tasks. The first category relies on just the explanation alone and includes verification and counterfactual reasoning problems. The latter focuses on the explanation and other facts about the environment, including goals such as safety and trust. We evaluate the factor impacts on extrinsic tasks by asking participants to classify images or text. A more recent article [9] divides tasks into two types: proxy and real tasks. The first evaluates how well users can predict AI's outcomes or decision boundaries, leading users to focus on the AI and its explanations [59, 65, 88]. Conversely, studies that use real tasks evaluate the cooperation between users and AI: users mainly focus on the task and may decide to use the AI advice to complete it [9, 18, 86]. The main criticism of XAI evaluations based on proxy tasks is that their conclusions may not reflect the usage of the system while pursuing real goals [9]. Building on these results, and provided we are focusing on the cooperation between AI and the user for making the final decision, we consider real tasks in our user study (see Section 4).

2.4.2 Task Complexity. When evaluating users, an essential task characteristic that affects the evaluation results is its complexity. Li et al. [49] identify the most effective objective and subjective measures to estimate task complexity from users' perspective via six simulated task situations. These measures are related to the objective and subjective task complexity distinction: the first considers task complexity exclusively based on their characteristics and independently from the performer [83]. The latter is a variable to measure users perception of the task complexity degree [74].

In a follow-up article, Liu and Li [50] propose a six-component task model for identifying salient complexity contributory factors, also structuring task complexity in 10 dimensions. They suggest a definition of task complexity, leveraging on previous works on the same subject: “Task complexity is the aggregation of any intrinsic task characteristic that influences the performance of a task.” Moreover, they clarify the difference between task complexity and task difficulty: the first involves the objective characteristics of a task, whereas the latter considers the interaction between the task, its performer, and context characteristics. Harvey and Koubek [29] provide a relevant definition of uncertainty as one of the task complexity dimensions as follows:“the degree of predictability or confidence associated with a task.” Such a dimension received little attention in previous user studies. In our work, we want to investigate whether different levels of uncertainty in both the model and users impact the accomplishment of a classification task. In addition, we want to understand how uncertainty interacts with the different reasoning styles employed in the explanations.

2.4.3 XAI User Evaluation Metrics. To outline user evaluation metrics, we will build upon two relevant surveys that provide an exhaustive overview of user metrics and methods for evaluating XAI systems. Mohseni et al. [54] categorize between design goals for explainable interfaces in XAI and evaluation methods for machine learning explanations, considering different XAI evaluation measures and targeted XAI user types. Zhou et al. [89] propose a comprehensive overview of the state-of-the-art methods for the evaluation of machine learning explanations. Generally, there are two types of user evaluation metrics widely used in XAI research: subjective and objective. Subjective measures consider the personal experience of the user on tasks and AI explanations, such as trust, confidence, and preference [81, 91]. Instead, objective measures involve evidence measured on task and AI explanations, like the task completion time and task performance [43, 90].

Among the subjective measures, we focus on the Reliance factor, which is generally used in the literature to measure users’ trust in a system when it makes errors [17, 33], also considering the effect of different levels of user uncertainty. For example, Buçinca et al. [9] studied users’ experience of AI errors with explanations that elicit inductive and deductive reasoning in a nutrition-related task. In inductive reasoning, one infers general patterns from repeated observations, so the authors supported it through example-based explanations. In deductive reasoning, one starts from general rules for getting conclusions on a specific situation, and the authors created them by letting the AI provide general rules used to generate its recommendation. They discovered that users trusted the AI more with the inductive ones in the proxy task experiment; on the contrary, users trusted the AI more with the deductive ones in the actual decision-making task. Trust values were self-reported. Further, Alipour et al. [2] measured users’ reliance (on a Likert scale) on the AI’s explanation while predicting the system performance in a user-machine prediction task. They found a correlation between the reliance and users’ accuracy in those cases when the system was wrong.

Regarding the objective measures, we study how Task Performance and the Agreement between the user and the AI vary considering different levels of user uncertainty, AI correctness, and the interaction between the explanations and the AI uncertainty levels. Buçinca et al. [9] discovered that users achieved nearly identical task performance in inductive and deductive reasoning explanations using proxy tasks. However, in the actual decision-making task, users gave the correct answer significantly more with inductive explanations than deductive ones when the AI provided the wrong recommendation.

Another metric we consider is the agreement, already used in the literature to quantify response times in user studies, as well as the agreement between annotators and model predictions [34, 45, 61, 62]. In our study, we use the agreement to assess in which experimental settings users tend to follow AI advice in the task and identify potential persuasive behavior in explanations. As an example, Zhang et al. [87] collected the percentage of trials in which users’ final prediction agreed with the AI’s prediction in an income prediction task, showing that they switched to the AI’s predictions more often when AI’s confidence scores were displayed. Van der Waa [75] investigated the persuasiveness of rule-based and example-based explanations in the context of decision support in diabetes self-management. This shows that explanations are more persuasive than the no explanation condition. Moreover, the example-based ones are slightly more persuasive compared to rule-based ones, although results do not support a significant difference.

In summary, we study how user uncertainty and AI correctness impact users' reliance on the elements available in the XAI interfaces (the instance, AI prediction, and the explanation), and how users' task performance and agreement are affected by the user uncertainty, AI correctness, and the interaction between the AI uncertainty and the explanation styles. Further, we measure and compare users self-reported confidence in their decision and their confidence in the AI's suggestion with our uncertainty labeling described in Sections 2.1 and 2.2.

2.5 Summary of Research Gaps and Article Contributions

Thus far in Section 2, we discussed studies that offer an insight into typical users' evaluation metrics and XAI techniques effectiveness. We identified different research gaps that motivate our work. The first is a lack of studies evaluating the User Uncertainty impact on the users' Reliance on the information provided by XAI interfaces. Existing studies focus on users' confidence, and they evaluate it on proxy tasks [2, 56]. The second is a systematic investigation of the effects of the User Uncertainty, AI Uncertainty, the AI Correctness in the prediction, and the Explanation Style on the decision process. We focus on the Task Performance (i.e., making a correct decision) and the Agreement with the AI's advice. There are many studies that provide insights on some aspects, but they have limitations in terms of the factors they analyze [9, 34, 38, 45, 61, 62, 87], the domain they consider (we cover images and text) [9, 38, 87], or the evaluation on proxy tasks [2, 38, 56]. The third gap is the role of the logic reasoning styles (Explanation Style) in the effectiveness of explanations, considering Task Performance and the Agreement. We found comparisons limited to a subset of the logical reasoning styles, and they do not extend the result beyond a single domain [9, 75].

Given these premises, our contributions include the following:

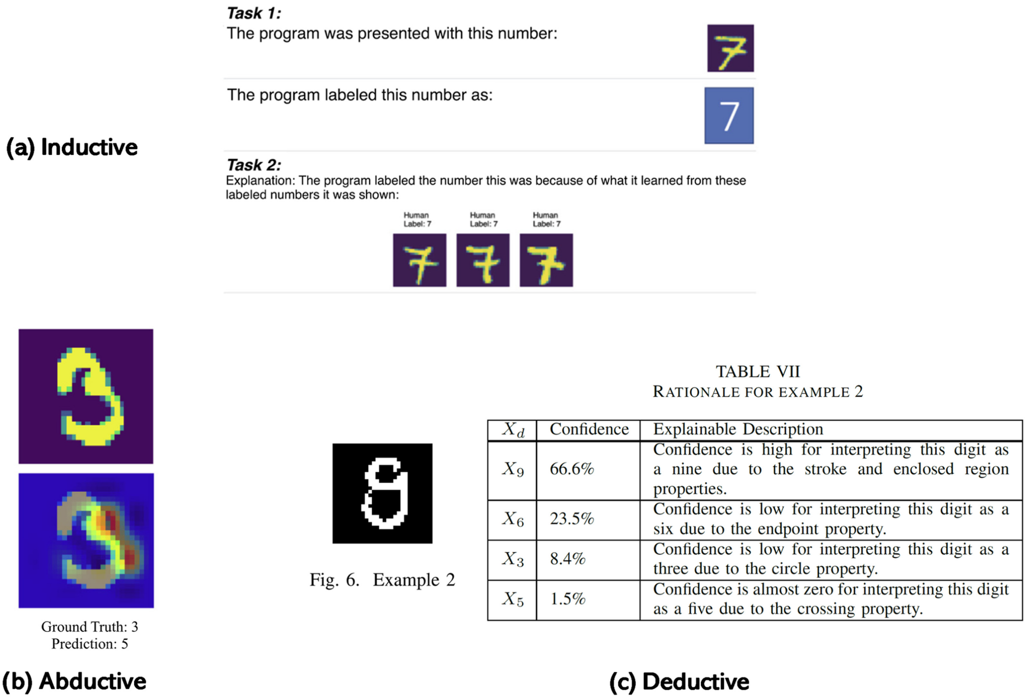
(1) We assess how User Uncertainty and AI Correctness impact users' Reliance on three pieces of information included in XAI interfaces: the instance, AI prediction, and explanations. We consider image and text classification tasks.
(2) We distinguish explanations according to the logical reasoning they support through Peirce's syllogistic theory [20]. We consider inductive, abductive, and deductive reasoning, identifying them in existing explanations techniques in the literature, involving neural networks (where applicable) in the image and text domains. We assess the impact of the Explanation Style in Task Performance and Agreement with the AI.
(3) We provide a comprehensive evaluation of how User Uncertainty, AI Correctness, and AI Uncertainty coupled with Explanation Styles impact Task Performance and Agreement with the AI prediction in the image and text domains.

3 IDENTIFYING LOGICAL REASONING STYLES IN XAI SYSTEMS

This section defines how to identify the reasoning types we discussed previously in relevant examples from the literature. For assigning a reasoning style, we need to identify the three components defined in Peirce's syllogistic theory [20]: one (or more) Cause (or Case/Explanation/Reason), Effect (or Observation/Result), and Rule (or Generalization/Theory). When applying this theory to XAI interfaces, we may have an *implicit or explicit* representation of such components. For example, a rule or a cause is implicit when it comes from the user's mental model and not from the AI. Instead, it is explicit when the XAI interface renders such components, for example,

Table 1. Illustrative Articles That Contain Examples of Logical Reasoning Explanations, Divided According to the Type of Data, Reasoning Style, Model, and Whether They Evaluated Explanations with Users

Year	Reference	Data Type	Reasoning style	Model Type	Evaluation with Users
2021	[38]	Image	Inductive	ResNets [31]	✓
2021	[39]	Image	Abductive	VGG-16 [68], ResNet50 [30]	✗
2021	[82]	Image	Deductive	Multi-layer perceptrons	✗
2021	[39]	Text	Inductive	VGG-16 [68], ResNet50 [30]	✗
2019	[48]	Text	Abductive	1D convolutional neural networks	✓
2021	[61]	Text	Deductive	Naïve Bayes	✓



Task 1:
The program was presented with this number:

The program labeled this number as:

Task 2:
Explanation: The program labeled the number this was because of what it learned from these labeled numbers it was shown:

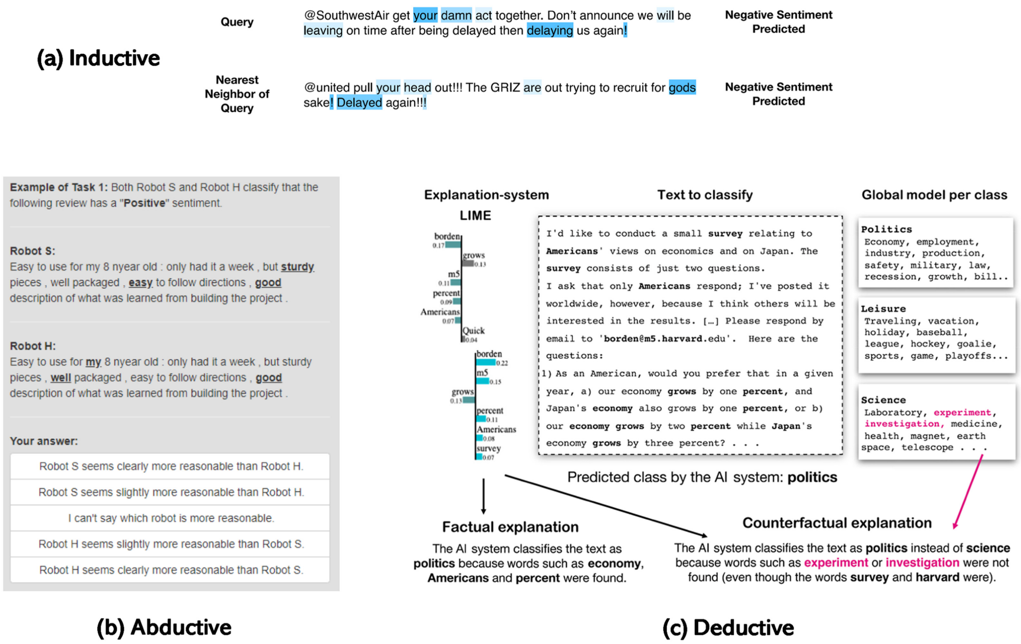
TABLE VII
RATIONALE FOR EXAMPLE 2

X_d	Confidence	Explainable Description
X_9	66.6%	Confidence is high for interpreting this digit as a nine due to the stroke and enclosed region properties.
X_6	23.5%	Confidence is low for interpreting this digit as a six due to the endpoint property.
X_3	8.4%	Confidence is low for interpreting this digit as a three due to the circle property.
X_5	1.5%	Confidence is almost zero for interpreting this digit as a five due to the crossing property.

Fig. 2. Reasoning styles for the image domain. (a) Inductive: the example-based explanations shown by the AI identify the cause, and the effect is that the AI did recognize the number as a 7; therefore, the user needs to understand the rule from the AI's examples [38]. (b) Abductive: the AI's prediction represents the effect, and the AI's highlight that supports its answer represents the rule; therefore, the user has to find the best cause, identifying the weights s/he considers more significant [39]. (c) Deductive: the AI's rationales that identify the number are the cause, and the rule is explicit and identified by the AI from the link between the rationales and the AI's answer; the AI answer is the effect [82].

by representing what the AI has learned in the training process and the explanations on its prediction. The work of Flach and Kakas [20] is the reference for the concepts we are going to describe. The list of the articles we chose for representing reasoning styles is shown in Table 1, although not all of them used a neural network architecture or carried out a user evaluation.

Induction: Given a Cause and an Effect, Induce a Rule. This style of reasoning involves drawing a general conclusion from a set of specific observations. It is alternatively referred to as "bottom-up" logic because it involves widening specific premises into broader generalizations. The work by Kenny and Keane [38] provides an example of inductive reasoning for the image domain, as depicted in Figure 2(a). The causes are the AI's example-based explanations. The effect is that AI was


(a) Inductive
Query
Negative Sentiment Predicted
Nearest Neighbor of Query
Negative Sentiment Predicted

@SouthwestAir get your damn act together. Don't announce we will be leaving on time after being delayed then delaying us again!

@united pull your head out!!! The GRIZ are out trying to recruit for gods sake! Delayed again!!!

Example of Task 1: Both Robot S and Robot H classify that the following review has a "**Positive**" sentiment.

Robot S:
Easy to use for my 8 nyear old : only had it a week , but **sturdy** pieces , well packaged , **easy** to follow directions , **good** description of what was learned from building the project .

Robot H:
Easy to use for **my** 8 nyear old : only had it a week , but sturdy pieces , **well** packaged , easy to follow directions , **good** description of what was learned from building the project .

Your answer:

Robot S seems clearly more reasonable than Robot H.
Robot S seems slightly more reasonable than Robot H.
I can't say which robot is more reasonable.
Robot H seems slightly more reasonable than Robot S.
Robot H seems clearly more reasonable than Robot S.

I'd like to conduct a small **survey** relating to **Americans**' views on economics and on Japan. The **survey** consists of just two questions.
I ask that only **Americans** respond; I've posted it worldwide, however, because I think others will be interested in the results. […] Please respond by email to '**borden@m5.harvard**.edu'. Here are the questions:
1) As an American, would you prefer that in a given year, a) our economy **grows** by one **percent**, and Japan's **economy** also grows by one **percent**, or b) our **economy grows** by two **percent** while **Japan**'s economy **grows** by three percent? . . .

Politics
Economy, employment, industry, production, safety, military, law, recession, growth, bill..

Leisure
Traveling, vacation, holiday, baseball, league, hockey, goalie, sports, game, playoffs...

Science
Laboratory, experiment, investigation, medicine, health, magnet, earth space, telescope . . .

Predicted class by the AI system: **politics**

Factual explanation
The AI system classifies the text as **politics** because words such as **economy**, **Americans** and **percent** were found.

Counterfactual explanation
The AI system classifies the text as **politics** instead of **science** because words such as **experiment** or **investigation** were not found (even though the words **survey** and **harvard** were).

(b) Abductive

(c) Deductive

Fig. 3. Reasoning styles for the text domain. (a) Inductive: the example-based explanation shown by the AI identifying the cause, and the effect is that the AI did recognize the sentence with a negative sentiment; therefore, the user needs to understand the rule from the AI's example [39]. (b) Abductive: the AI's prediction represents the effect, and the AI's word highlight that supports its answer represents the rule; therefore, the user has to find the best cause, identifying the weights (words) s/he considers more significant [48]. (c) Deductive: the AI's words that identify the class are the cause, and the rule is explicit and identified by the AI from the link between the words and the AI's answer; the AI answer is the effect [61].

able to recognize the number as 7. The rule is that certain properties of the examples shown by the AI lead to the classification result of the number (implicit). For the text domain, another work by Kenny and Keane [39] presents an example of inductive reasoning (excluding the words highlighting), as depicted in Figure 3(a). The cause is the AI's example-based explanation. The effect is that AI was able to recognize the sentence with a negative sentiment. The rule is that certain properties of the example shown by the AI lead to the classification result of the sentence (implicit).

Abduction: Given an Effect and a Rule, Abduct a Cause. This style of reasoning typically begins with an incomplete set of observations and proceeds to the likeliest possible explanation. The work of Kenny and Keane [39] reports an example of abductive reasoning for the image domain, depicted in Figure 2(b). The AI's prediction gives the effect, whereas the highlight on the image provides the user with the intuition of the weights the AI uses for computing the answer, representing the implicit rule. The user must select the best cause, identifying the weights s/he considers more significant. For the text domain, Lertvittayakumjorn and Toni [48] provide an example of abductive reasoning, as depicted in Figure 3(b). The AI's prediction gives the effect, whereas the highlight on the text provides the user with the intuition of the weights (words) the AI uses for computing the answer, representing the implicit rule. The user must select the best cause, identifying the weights s/he considers more significant.

Deduction: Given a Cause and a Rule, Deduce an Effect. This style of reasoning starts with general rules and examines the possibilities to reach a specific, logical conclusion. Deductive reasoning is alternatively referred to as "top-down" logic because it usually starts with a general statement and ends with a narrower, specific conclusion. The work of Whitten et al. [82] contains an example of this reasoning for the image domain, as depicted in Figure 2(c). The causes are the rationales presented by the AI, which identify the number. The rule is the AI representation of how rationales contribute to the answer (explicit). The effect is given by the AI. The user only decides whether the AI's explanations are trustable or not. For the text domain, the work of Riveiro and Thill [61] presents an example of deductive reasoning, as depicted in Figure 3(c). The causes are the words presented by the AI in the explanation, which identify the class. The rule is the AI representation of how words contribute to the answer (explicit). The effect is given by the AI. The user only decides whether the AI's explanations are trustable or not.

4 METHOD

For evaluating the effect of the AI and the logic-style explanations under different levels of uncertainty, we carried out two user studies. The first considered image classification, whereas the second focused on text classification. In each study, we asked participants to make their decision providing them with one instance (image or text), the AI prediction, and one among the four styles of explanation considered (no explanation, inductive, abductive and deductive).

The two studies are quite similar. They share hypotheses, procedure, and statistical methods. For this reason, the two studies are merged in this section. Note that only the materials used in the two studies are distinct.

The independent variables we used for establishing the different conditions are the following:

- The User Uncertainty, which has two levels (low and high).
- The AI Uncertainty, which has two levels (low and high).
- The AI Correctness, which has two levels (wrong and right).
- The Explanation Style, which has four levels (no explanation, inductive, abductive, and deductive).

We considered binary levels for the independent variables for balancing the tradeoff between the number of participants and the statistical power of the study.

We measured their effect on three dependent variables:

- The users' Reliance on the different pieces of information provided to the user (the image/text, the AI prediction, and the explanation).
- The Task Performance, which is whether the label the user decides to assign to the current image/text is correct or not.
- The Agreement with the AI, which is whether the user confirms the AI prediction with his/her decision.

We also collected task completion time for assessing possible differences among the reasoning styles. Further, we collected feedback on the perceived AI uncertainty for understanding whether users can estimate the AI confidence in the prediction. Finally, we collected subjective feedback on user confidence to better understand the participant's perception of the decision task and find differences between their judgment and our uncertainty labeling on instances.

4.1 Hypotheses

Considering the literature we summarized in Section 2.2, we expect that the users' Reliance [54] on the instance, the AI prediction, and the explanation type depends on User Uncertainty. In general, users should rely on the instance as the primary source of information for making their decision. In particular, we expect this behavior when user uncertainty is low since users might not need additional information from the AI to accomplish the classification task. On the contrary, when the user uncertainty is high, users might find it challenging to decide based only on the instance, even if they should still consider it as the main information source. In case of high user uncertainty, we expect that users rely more on the explanation than the AI prediction. The reason for our expectation is that explanations reflect the AI decision process on the target instance, whereas the prediction does not provide any hint on this. Therefore, we believe that a high user uncertainty will result in a higher reliance on the instance and the explanation, supporting users in the decision-making task. In summary, we formulated hypothesis 1 as follows:

- *H1*: The user's Reliance on the information provided in XAI interfaces depends on the User Uncertainty level:
  - *H1a*: When the User Uncertainty is low, the user will rely more on the instance rather than on AI prediction and the explanation.
  - *H1b*: When the User Uncertainty is high, the user will rely more on the instance and the explanation rather than on the AI prediction.

According to the work of Buçinca et al. [9] and Kenny et al. [38], we expect that Task Performance depends on User Uncertainty, AI Correctness, and the interaction between AI Uncertainty and the Explanation Style. Users should successfully classify the instance when the user uncertainty is low, possibly identifying the AI correctness by considering its prediction and explanation. Instead, we do not expect such identification when the instance has a high user uncertainty. An incorrect prediction from the AI followed by a convincing explanation should lead users to follow AI's guidance, thus decreasing Task Performance. We expect a positive impact of AI correctness, increasing the performance when the prediction is correct. In addition, we expect that AI uncertainty affects the effectiveness of the explanations and, consequently, the Task Performance. For example, a high AI uncertainty could lead to unreliable or contrasting explanations that may confuse users during the classification process.

Therefore, we formulated hypothesis 2 as follows:

- *H2*: Users' Task Performance is moderated by the levels of User Uncertainty, AI Correctness, and the interaction between Explanation Style and AI Uncertainty:
  - *H2a*: Users will achieve a higher Task Performance with a low User Uncertainty than a high one.
  - *H2b*: Users will achieve a higher Task Performance with a right AI Correctness than with a wrong one.
  - *H2c*: AI Uncertainty moderates the effect of the Explanation Style on the Task Performance (positively or negatively).

Concerning the Agreement [34, 87], we expect that it depends on the User Uncertainty, AI Correctness, and the interaction between AI Uncertainty and the Explanation Style. We believe that users may have a higher agreement when the user uncertainty is high since they are not sure of the task outcome, and they may rely on the AI prediction. AI correctness will lead to higher agreement when it confirms the users' (correct) prediction. Explanations may persuade or dissuade the users to follow the AI, and this should depend on how the user perceives the AI's uncertainty through them.

We formulated hypothesis 3 as follows:

- *H3*: The user's Agreement with the AI prediction may be moderated by the User Uncertainty, AI Correctness, and the interaction between AI Uncertainty and the Explanation Style:
  - *H3a*: Users will achieve a higher Agreement with a high User Uncertainty than a low one.

H3b: Users will achieve a higher Agreement with a right AI Correctness than with a wrong one.

H3c: AI Uncertainty moderates the effect of the Explanation Styles on the Agreement (positively or negatively).

4.2 Materials

In this section, we briefly describe the materials we employed for implementing the studies for image and text data. Appendix A contains the technical details on how we built the classification models, how we selected the instances having low and high user uncertainty, and how we obtained the different explanation styles.

4.2.1 Image Data: Dataset. We selected the MNIST [46] dataset since digits recognition requires no prior knowledge for the participants. It contains both instances having low and high user and AI uncertainty.

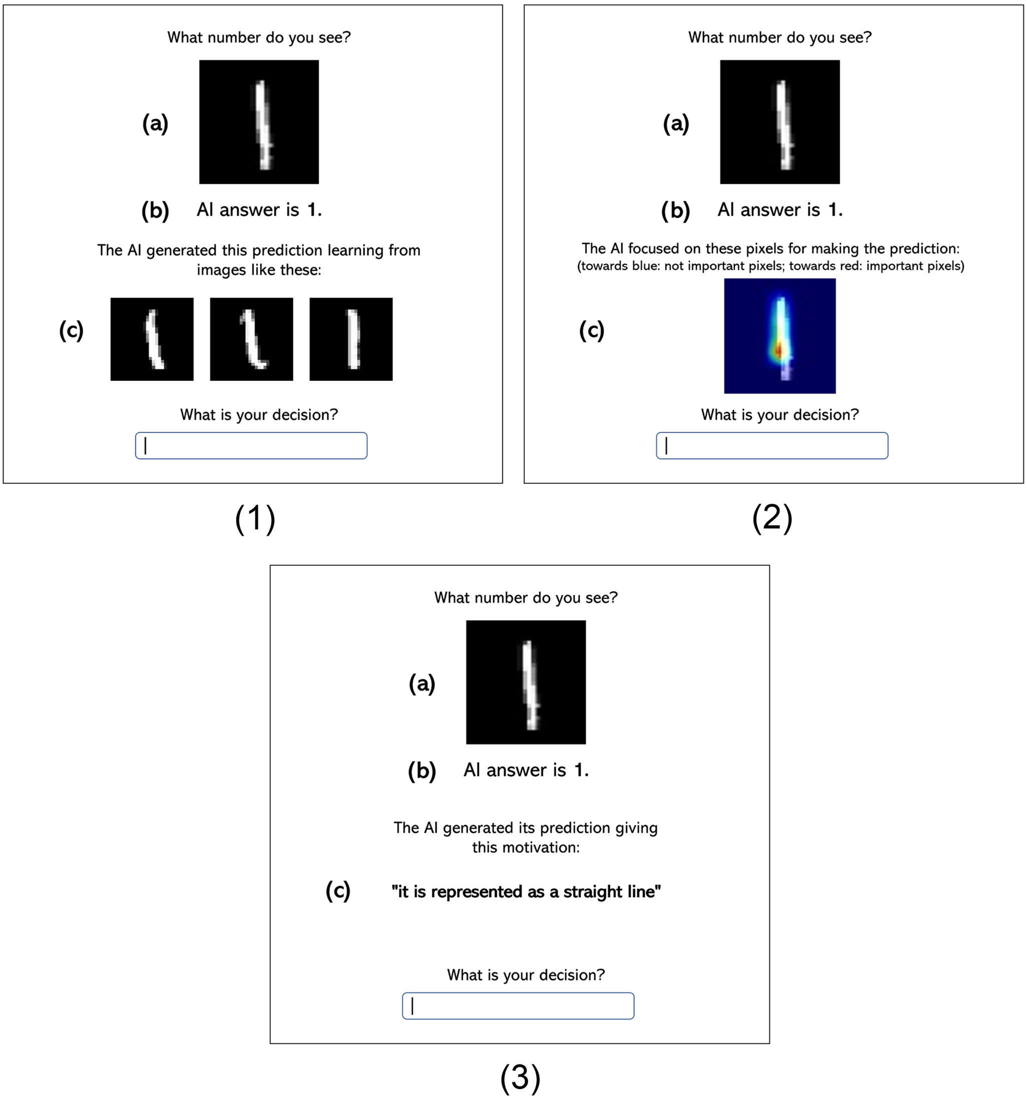
Classification Model. We adapted the AlexNet architecture [44] for supporting digits classification in MNIST and for calculating the AI uncertainty.

Instance Selection. For running the study, we identified 6 images for each pair in the combination of user and AI uncertainty levels (i.e., low-low, low-high, high-low, and high-high), 24 in total. For obtaining the User Uncertainty values, we run a preliminary study with users (details are available in Appendix A). We obtained the AI Uncertainty calculating the epistemic uncertainty on the adapted AlexNet classification model (see Appendix A). Then, we assigned 2 images from each group to an Explanation Style for pre-computing explanations. We obtain inductive explanations mapping the weights of the AlexNet model on a k-NN (k-nearest neighbor) model to obtain the three nearest neighbors. For abductive explanations, we use Grad-CAM [64], whereas for deductive explanations, we exploit an encoder-decoder architecture [78] generating natural language sentences (see Appendix A). Figure 4 shows an example of an image classification task for each explanation style. Finally, we included in each reasoning style sub-group an instance where the AI prediction is correct and one where it is wrong (AI Correctness).

4.2.2 Text Data: Dataset. We used the Yelp Reviews dataset [6], consisting of online reviews in free-form text and a star rating out of five stars. We chose this dataset since it contains both low and high user and AI uncertainty instances, adapting it to let users recognize sentiments (good, average, bad) in reviews instead of predicting the star rating.

Classification Model. We adapted the Yoon Kim **Convolutional Neural Network (CNN)**-multichannel [40] for the classification, united with Transformer Universal Sentence Encoder (USE_T) [11] for features extraction. We refer to this model as CNN-USE_T (see Appendix A).

Instance Selection. We replicated the approach we used for selecting the images, this time using the CNN-USE_T model for computing the prediction (AI Correctness) and the epistemic uncertainty (AI Uncertainty). We identified low and high User Uncertainty reviews through a dedicated pre-study, similar to the one we used for images (see Appendix A). As for the Explanation Style, for the inductive explanation we use the three nearest neighbors obtained by mapping the CNN-USE_T to a k-NN model, and for the abductive and deductive explanation we use the information and the visualizations provided by LIME [60]. Abductive explanations highlight the words that caused the AI's outcome by showing the importance of each word for the prediction through the opacity of its background color. Deductive explanations visualize the most important 10 words for each sentiment, showing their weights in supporting or opposing the assignment to a specific class (see Appendix A). Figure 5 shows an example of an text classification task for each explanation style.



What number do you see?

(a)

(b) AI answer is 1.

The AI generated this prediction learning from images like these:

(c)

What is your decision?

(1)

What number do you see?

(a)

(b) AI answer is 1.

The AI focused on these pixels for making the prediction:
(towards blue: not important pixels; towards red: important pixels)

(c)

What is your decision?

(2)

What number do you see?

(a)

(b) AI answer is 1.

The AI generated its prediction giving this motivation:

(c) **"it is represented as a straight line"**

What is your decision?

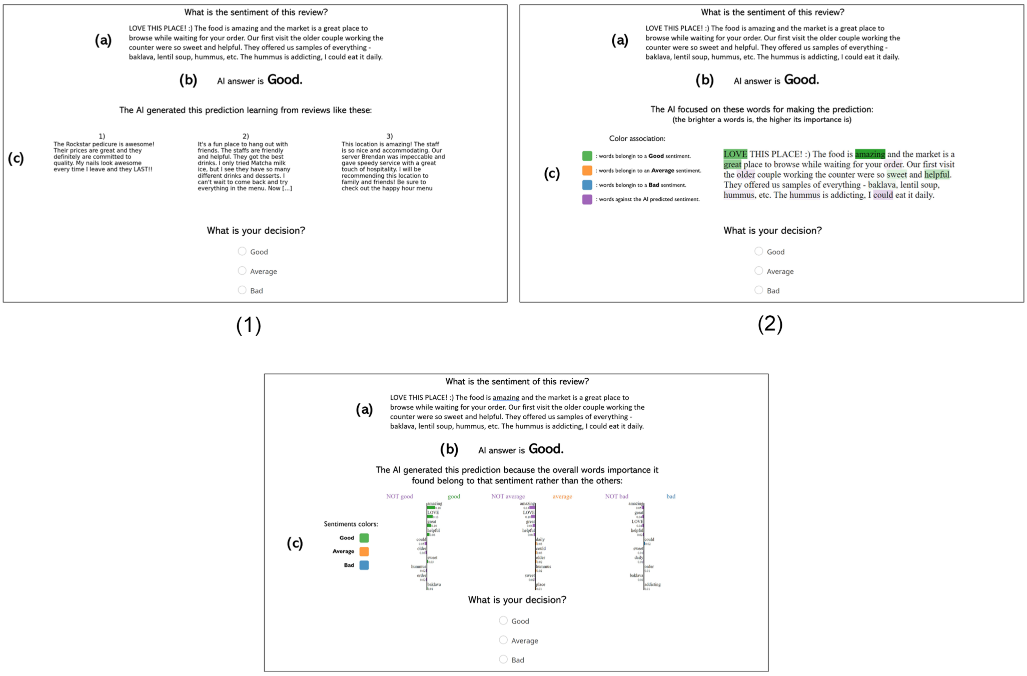
(3)

Fig. 4. An example of the image classification task with inductive (1), abductive (2), and deductive (3) explanations.

4.3 Procedure

We ran the two studies online, using the Prolific[1] platform. At first, participants read a document containing a brief description of the study and expressed their informed consent. Next, each participant completed a single classification task based on an assigned evaluation data type (image or text) in one of the four Explanation Styles—that is, no explanation, inductive, abductive, and deductive—which were equally randomized across participants. An example of the classification task is shown in Figure 4 for image data and Figure 5 for text data. Each of these conditions was further randomized on the value of the other independent variables, namely User Uncertainty,

[1] https://www.prolific.co.



What is the sentiment of this review?

(a) LOVE THIS PLACE! :) The food is amazing and the market is a great place to browse while waiting for your order. Our first visit the older couple working the counter were so sweet and helpful. They offered us samples of everything - baklava, lentil soup, hummus, etc. The hummus is addicting, I could eat it daily.

(b) AI answer is **Good.**

The AI generated this prediction learning from reviews like these:

(c)

1) The Rockstar pedicure is awesome! Their prices are great and they definitely are committed to quality. My nails look awesome every time I leave and they LAST!!

2) It's a fun place to hang out with friends. The staffs are friendly and helpful. They got the best drinks. I only tried Matcha milk ice, but I see they have so many different drinks and desserts. I can't wait to come back and try everything in the menu. Now [...]

3) This location is amazing! The staff is so nice and accommodating. Our server Brendan was impeccable and gave speedy service with a great touch of hospitality. I will be recommending this location to family and friends! Be sure to check out the happy hour menu

What is your decision?

○ Good

○ Average

○ Bad

(1)

What is the sentiment of this review?

(a) LOVE THIS PLACE! :) The food is amazing and the market is a great place to browse while waiting for your order. Our first visit the older couple working the counter were so sweet and helpful. They offered us samples of everything - baklava, lentil soup, hummus, etc. The hummus is addicting, I could eat it daily.

(b) AI answer is **Good.**

The AI focused on these words for making the prediction:
(the brighter a words is, the higher its importance is)

(c)

Color association:

🟩 : words belongin to a **Good** sentiment.

🟧 : words belongin to an **Average** sentiment.

🟦 : words belongin to a **Bad** sentiment.

🟪 : words against the AI predicted sentiment.

LOVE THIS PLACE! :) The food is amazing and the market is a great place to browse while waiting for your order. Our first visit the older couple working the counter were so sweet and helpful. They offered us samples of everything - baklava, lentil soup, hummus, etc. The hummus is addicting, I could eat it daily.

What is your decision?

○ Good

○ Average

○ Bad

(2)

What is the sentiment of this review?

(a) LOVE THIS PLACE! :) The food is amazing and the market is a great place to browse while waiting for your order. Our first visit the older couple working the counter were so sweet and helpful. They offered us samples of everything - baklava, lentil soup, hummus, etc. The hummus is addicting, I could eat it daily.

(b) AI answer is **Good.**

The AI generated this prediction because the overall words importance it found belong to that sentiment rather than the others:

(c)

Sentiments colors:

Good 🟩

Average 🟧

Bad 🟦

What is your decision?

○ Good

○ Average

○ Bad

(3)

Fig. 5. An example of the text classification task with inductive (1), abductive (2), and deductive (3) explanations.

AI Uncertainty, and AI Correctness (see Section 4.1), which are not visible to participants. We placed two attention checks among the study questions. For both of them, the correct answer was clearly reported in the question text. The protocol has been formally approved by the Ethics Commission of the University of Cagliari.[2]

We recruited about 900 participants for each user evaluation via Prolific. Such a number derives from the power analysis indicating the need for 597 people (see Section 4.4), and considering that, in previous studies, about 30% of users failed attention checks. We paid each participant £0.63 for image data and £0.75 for text data. The task lasted 3 minutes on average for images and 5 minutes for text, so the reward per hour respectively was £8.56 and £10.72, higher than £7.80, which is the recommended payment in the platform. After filtering submissions failing one or both attention checks, we considered 659 submissions for the image data and 665 for the text data. Demographic data and other information about participants, like age, sex, level of education, and task completion time, were gathered via the Prolific platform.

We collected the following information:

- *User confidence*: We asked participants how confident they were about their decision on the classification task. Participants responded to the statement: "How confident are you about your decision?" The available answers were "Very confident" and "Not very confident."
- *Perceived AI uncertainty*: We asked the participants to evaluate the AI confidence the prediction on the instance. Participants responded to the statement: "In your opinion, how confident is the AI about its decision?" The available answers were "Very confident" and "Not

[2] Received on January 9, 2022, Prot. 955.

very confident." This question was not available in the no explanation condition since, in this experimental setting, participants cannot evaluate the AI uncertainty relying only on its predictions.

- *Reliance*: A ranking of the information included in the XAI interface, namely the instance (image or text), AI prediction, and explanation. Participants responded to the statement: "Please rank the following information in terms of how much it helped you in making a decision: (a) image (or text), (b) AI prediction, (c) explanation."
- *Task performance*: Whether the participant's final decision is correct or not. The possible values are "correct" when the participant's answer is correct and "wrong" otherwise.
- *Agreement*: Whether the participant's final decision agrees with the AI prediction or not. The possible values are "yes" when the decision matches the AI prediction and "no" otherwise.

4.4 Analytical Approaches

In this section, we discuss the analytical approaches we follow for each of the hypotheses and assess the number of participants required for catching medium effects through a power analysis.

For H1 (reliance), we assess the results with the Friedman test [22, 23], analyzing user uncertainty values (low and high) separately to find significant differences in the factors' distributions. We conduct the Nemenyi post hoc analysis when we discover significant factors in the Friedman test. For assessing the number of participants required to validate this hypothesis, we carried out a power analysis using G*Power3 [19]. We set the analysis for medium effects (effect size = 0.3), an alpha of .05, and power of .85 for hypothesis 1. We used the Friedman test and a within-factors design, using two levels for the uncertainty (low and high) on the three ranked measurements (image, explanation, and AI prediction). The results showed that we needed a sample size of 52 people to catch medium effects.

For H2 (task performance) and H3 (agreement), we used logistic regression. The model includes these factors: user uncertainty (low, high), AI correctness (wrong, right), and the interaction between the explanation (noexp, inductive, abductive, deductive) and the AI uncertainty (low, high). The baselines for the logistic regression factors are "noexp" for the explanation, "high" for the AI and user uncertainty, and "wrong" for the AI correctness. In case we find a significant interaction in the model, we proceed to determine the marginal means of the model and the pairwise comparison using Bonferroni's p-value adjustment.

For both H2 and H3, the results of power analysis showed that we needed a sample size of 597 people for medium effects (*a priori* χ^2 test with effect size = 0.15, alpha = 0.05, power = 0.85, Df = 4). To sum up, we needed 597 participants for catching medium effects in each study, considering for H1 that 52 people are sufficient.

5 RESULTS

This section details the results obtained in the image and text evaluations. First, we discuss some details about participants, their confidence on the decision, and the perceived AI uncertainty. After that, we go into the test details for each hypothesis, reporting charts and statistical analyses.

5.1 Participants

For the image study, 659 users successfully passed the attention checks and were included in the evaluation. The participant set consists of 333 females and 326 males, aged between 18 and 71 years ($\bar{x}$ = 27.2, $\tilde{x}$ = 24, s = 8.4). For the text study, 665 users successfully passed the attention checks and were included in the evaluation. We ensured that participants had a good level of English for reading the reviews through the pre-screening supported by the Prolific platform. Participants consisted of 315 females and 350 males, aged between 18 and 79 years ($\bar{x}$ = 36.2, $\tilde{x}$ = 32, s = 33.1).

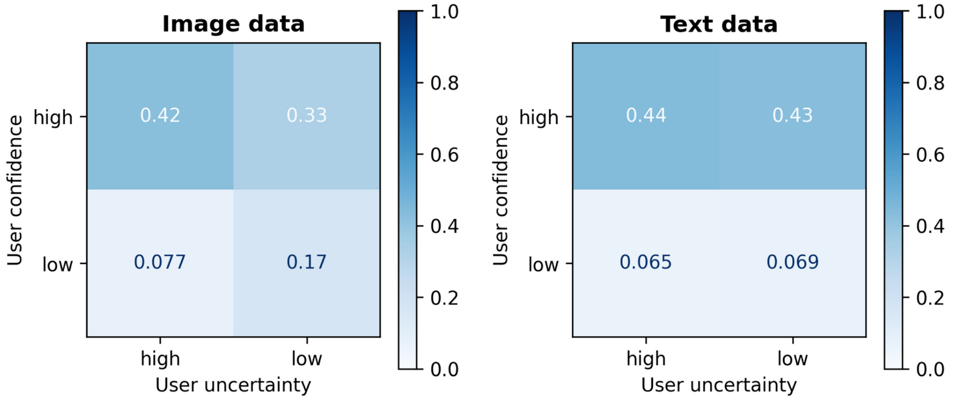


Fig. 6. User confidence vs user uncertainty results for image (left) and text data (right). Please note that a high self-reported user confidence should correspond with a low user uncertainty (as we represent it) and vice versa, but in fact it does not. Rather, users tend to report high user confidence, potentially overestimating their abilities.

5.2 User Confidence

We asked participants' explicit feedback on the confidence on their decision for confirming the different uncertainty levels. Figure 6 shows the confusion matrix for image and text data. In both studies, the participants' confidence consistently confirmed the labeling for the instances we obtained through the pre-study when the user uncertainty is low: 84.6% of low-uncertainty images correspond to a high confidence in the decision, and the same applies for 87.2% of low-uncertainty texts.

Instead, the instances in which participants reported a low confidence are much fewer than those we labeled with a high user uncertainty value. Users only reported a low confidence on 34.5% of high-uncertainty images and 14.0% of high-uncertainty texts. This seems to indicate a difference between our (high) user uncertainty labeling and the participant's confidence.

Although this could be due to errors or inconsistencies in our labeling, we can analyze whether this is overconfidence in a different way. Namely, we consider the case when users who are confident pick different prediction classes for the same instance. This results in a high user uncertainty *between* these groups. In this case, people would report a high confidence, but at most one group eventually makes the correct decision (the one selecting the right class, if any). We can identify a part of them by counting wrong decisions on instances we assigned as high uncertainty made by highly confident participants. They are 92 out of 215 in the image study and 211 out of 284 in the text. Summing these instances—92 for images and 211 for text—to those in which users correctly marked as high uncertainty in both domains, we confirm an overall high uncertainty of 62.5% for images and 77.9% for texts, respectively.

5.3 Perceived AI Uncertainty

To assess the participants' perception of the AI Uncertainty in the classification task, we asked them for explicit feedback in the survey. In this way, we collected the users' perception of the AI uncertainty, hinted by the combination of prediction and explanation. Figure 7 shows the confusion matrix for both image and text data. In both studies, the participants correctly labeled instances with low AI uncertainty: they identified 84.3% of low-uncertainty images and 85.8% of low-uncertainty texts.

Instead, the instances participants correctly perceived as having a high AI uncertainty are much fewer: participants identified 22.2% of high-uncertainty images and 21.7% of high-uncertainty text.

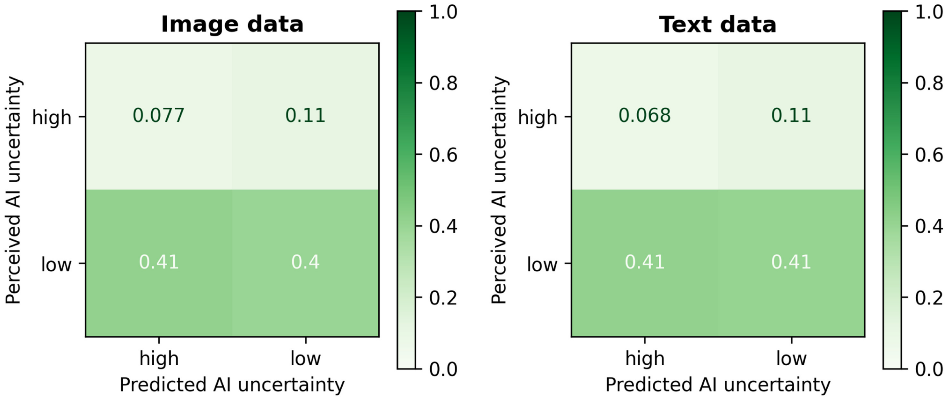


Fig. 7. Perceived vs actual AI uncertainty results. Confusion matrix for image data (left). Confusion matrix for text data (right).

These results consistently indicate that users had difficulties in identifying high AI uncertainty instances in both domains, showing that AI's prediction and explanation may not be enough to expose the AI uncertainty.

5.4 Task Completion Time

We obtained the following results for the task completion time on the image data: no explanation (noexp) ranged between 4.89 and 450.65 seconds ($\bar{x}$ = 102.52s, $\tilde{x}$ = 74.86s, s = 103.1s), inductive ranged between 3.39 and 703.27 seconds ($\bar{x}$ = 209.47s, $\tilde{x}$ = 189.78s, s = 115.2s), abductive ranged between 10.6 and 926.41 seconds ($\bar{x}$ = 208.97s, $\tilde{x}$ = 189.27, s = 129.1), and deductive ranged between 11.67 and 907.87 seconds ($\bar{x}$ = 212.75s, $\tilde{x}$ = 191.65s, s = 152.6). The ANOVA analysis highlights a significant difference between the no explanation condition and the other three reasoning styles.

For text data, we obtained the following results: no explanation ranged between 22.90 and 506.65 seconds ($\bar{x}$ = 98.65s, $\tilde{x}$ = 83.28s, s = 65.8s), inductive ranged between 56.36 and 793.04 seconds ($\bar{x}$ = 247.39s, $\tilde{x}$ = 215.31s, s = 130.4s), abductive ranged between 54.28 and 1132.41 seconds ($\bar{x}$ = 274.0s, $\tilde{x}$ = 229.49s, s = 248.2), and deductive ranged between 88.52 and 1,338.87 seconds ($\bar{x}$ = 262.66s, $\tilde{x}$ = 230.30s, s = 141.1). As for image data, the ANOVA analysis only highlights a significant difference between the no explanation condition and all the others.

5.5 H1: Reliance

For assessing H1, we investigate the effects of varying the level of User Uncertainty on the Reliance of the user on the different pieces of information in a XAI interface: the instance (image or text), the AI prediction, and explanation. In the analysis, we excluded participants who did not see an explanation (i.e., assigned to the no explanation condition), resulting in 478 users for image data and 454 users for text data.

Image Data. In the image study, the Friedman test for the Reliance shows a significant difference between the instance, AI prediction, and explanation when the User Uncertainty is low (H1a, $\chi^2(2) = 234.76$, df = 2, $p < .05$). The same happens when the User Uncertainty is high (H1b, $\chi^2(2) = 189.3$, df = 2, $p < .05$).

The pairwise comparisons using Nemenyi post hoc test for mean rank considering a low user uncertainty (H1a) shows a significant difference between the image and the AI prediction, and

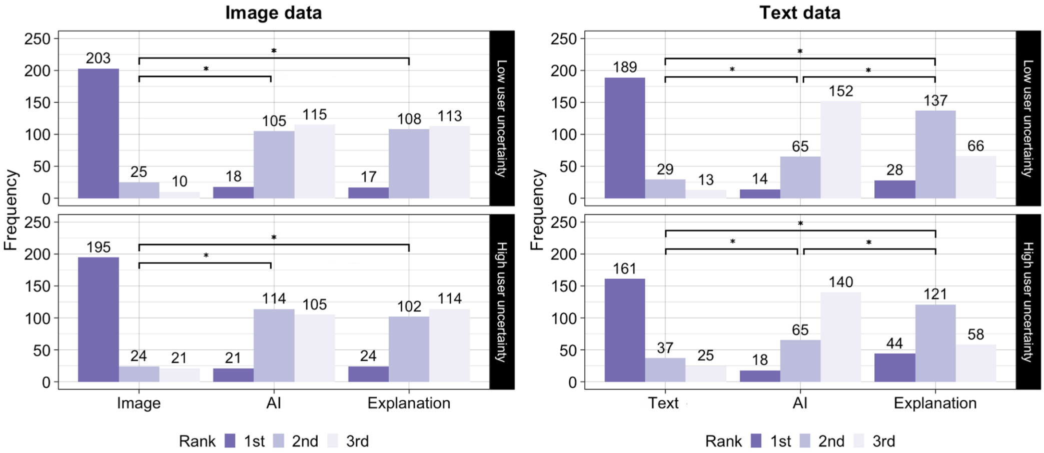


Fig. 8. Rank frequencies for a user's information Reliance in the image and text domain, considering a low and a high User Uncertainty.

between the image and explanation type (both with $p < .05$). Thus, we reject the null hypothesis and conclude that users rely most on the instance data (H1a).[3] The left side of Figure 8 highlights the dominance of the image compared to the AI prediction and explanation when the user uncertainty is low.

We repeated the pairwise comparisons considering a high user uncertainty (H1b), and we obtained the same results. Users rely most on the instance, and there is no significant difference between the AI prediction and the explanation. So, we fail to reject the null hypothesis for H1b. Figure 8 shows a good similarity between the rankings with low and high user uncertainty, highlighting no differences between the AI prediction and the explanation type.

Alipour et al. [2] found an increased reliance on the explanations when the AI was wrong. Therefore, we analyze the impact of the AI Correctness factor in our study to get a deeper understanding of the Reliance. We found a significant difference between factors considering correct ($\chi^2(2)$ = 234.2, df = 2, $p < .05$) and wrong predictions ($\chi^2(2)$ = 202.03, df = 2, $p < .05$).

We proceeded with a Nemenyi post hoc test, and we found a significant difference among the three factors ($p < .05$) in both conditions. When the AI makes correct predictions, users rely more on the image (rank 1) and the AI prediction (rank 2) rather than the explanation (rank 3). When the AI predictions are wrong, users rely more on the image (rank 1), explanation (rank 2), and AI prediction (rank 3). This evidence is visible on the left side of Figure 9, confirming the results in the work of Alipour et al. [2].

In summary, for the image domain, when we have a low User Uncertainty, users rely more on the image (first rank) than on the AI prediction and explanation type, as we expected in H1a. Instead, we cannot validate hypothesis H1b since there are no significant differences between the AI prediction and explanation in the case of a high User Uncertainty. The image is still ranked first. When the AI makes correct predictions (i.e., the AI Correctness is right), the user relies on the AI prediction as to the second source of information and the explanation as the third. When the AI makes wrong predictions (i.e., the AI Correctness is wrong), the second source of information is the explanation and the AI prediction is the third.

[3] There is no significant difference between the AI prediction and the explanation type.

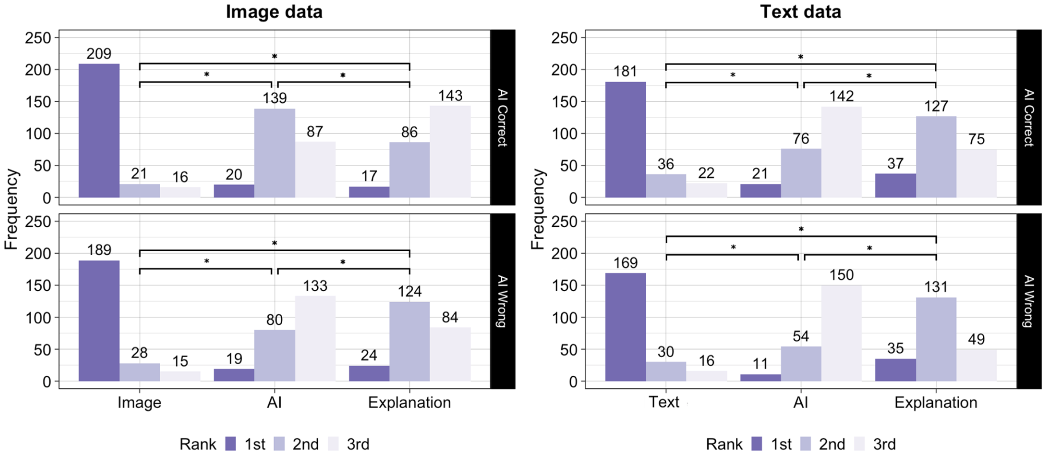


Fig. 9. Rank frequencies for a user's information Reliance by AI Correctness in the image domain on the left and the text domain on the right.

Text Data. In the text study, the Friedman test for the Reliance shows a significant difference between the instance, AI prediction, and explanation when the User Uncertainty is low (H1a, $\chi^2(2) = 222.79$, df = 2, $p < .05$). The same happens when the user uncertainty is high (H1b, $\chi^2(2) = 150.57$, df = 2, $p < .05$).

When the user uncertainty is low (H1a), the Nemenyi post hoc test for mean rank shows a significant difference between (i) the text and the AI prediction, (ii) the text and explanation type, and (iii) the AI prediction and the explanation type (all with $p < .05$), resulting in the text that ranks as the information having the highest reliance, the explanation as the second, and the AI prediction as the third. The right part of Figure 8 summarizes such results, which allow us to reject the null hypothesis for H1a and conclude that users rely on the instance data as the most important source of information. We repeated the pairwise comparisons considering a high user uncertainty (H1b), and we obtained the same results. Therefore, we reject the null hypothesis for H1b and conclude that, in the text domain, participants consider the explanation after the instance as the second source of information. The right part of Figure 8 highlights the differences between the AI prediction and the explanation.

We repeated the analysis including the AI Correctness factor to examine its impact on the Reliance. As for images, we found a significant difference between factors for both levels of the AI Correctness (Right: $\chi^2(2) = 173.08$, df = 2, $p < .05$; Wrong: $\chi^2(2) = 199.66$, df = 2, $p < .05$).

We proceeded with the Nemenyi post hoc test considering AI correctness and found a significant difference among the review text, the AI prediction, and the explanation ($p < .05$) in both levels of the AI Correctness, resulting in the same rankings: users preferred the text (rank 1) and explanation (rank 2) over the AI prediction (rank 3), as highlighted in the right part of Figure 9. This result extends the findings of Alipour et al. [2] to the text domain since users do rely more on the explanations than the AI prediction when the AI is wrong. We registered the same rankings also in case of a correct AI prediction.

In summary, in the text domain, users relied more on the text (first rank) rather than on the AI prediction and explanation when the User Uncertainty is low (H1a). Users rely more on the text (rank 1), followed by the explanation (rank 2) when the User Uncertainty is high (H1b). This ranking persisted also when analyzing the AI Correctness as a factor, considering correct and wrong predictions.

Table 2. Logistic Regression Results on Task Performance (H2, Image Domain)

Predictor	Log-Odds	Std. Error	*z*-Value	*p*
User uncertainty [low]	***2.379	0.239	9.959	<.001
AI correctness [right]	***2.073	0.229	9.021	<.001
Explanation style [inductive]	*1.057	0.432	2.444	.015
Explanation style [abductive]	0.2954	0.392	0.754	.451
Explanation style [deductive]	*0.7791	0.395	1.971	.048
AI uncertainty [low]	***1.4423	0.411	3.512	<.001
Explanation style [inductive] * AI uncertainty [low]	***−2.579	0.612	−4.217	<.001
Explanation style [abductive] * AI uncertainty [low]	−0.418	0.589	−0.711	.477
Explanation style [deductive] * AI uncertainty [low]	***−1.949	0.585	−3.330	<.001

*$p < .05$; **$p < .01$; ***$p < .001$.

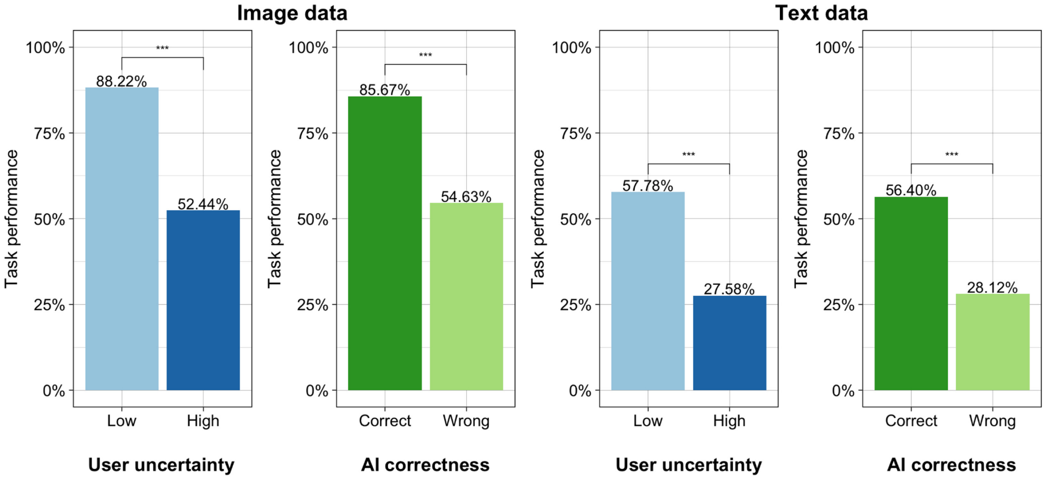


Fig. 10. Task Performance comparison on the different levels of User Uncertainty and AI Correctness for the image domain on the left and the text domain on the right (H2a and H2b).

5.6 H2: Task Performance

We now know which information participants looked at. The question is how this influenced Task Performance—that is, whether the user makes the correct decision or not. When User Uncertainty is low, we expect users to achieve a higher Task Performance than a high User Uncertainty level (H2a). Symmetrically, when AI predictions are correct (i.e., AI Correctness is right), we expect a higher Task Performance than when the AI is wrong (H2b). Further, we expect differences in how the Explanation Style conveys the information about the AI Uncertainty, improving Task Performance in some combinations and decreasing it in others (H2c).

Image Data. Table 2 shows the results of the logistic regression analysis of the Task Performance in the image study. We found a significant main effect for the User Uncertainty, which positively affects Task Performance when it is low. Symmetrically, we found a significant main effect for AI Correctness on Task Performance: a correct AI prediction positively affects Task Performance. Thus, we reject the null hypothesis for H2a and H2b. The left side of Figure 10 shows the effects of User Uncertainty and AI Correctness on Task Performance in the image domain.

Table 3. Pairwise Comparison Between the Levels of AI Uncertainty and Explanation Style on Task Performance in the Image Domain (H2)

AI Uncertainty = low				
Contrast (Explanation Style)	Log-Odds	Std. Error	*z*-Value	*p*
Inductive vs noexp	−1.522	0.421	−3.618	.002
Abductive vs inductive	1.400	0.436	3.208	.008
Deductive vs noexp	−1.170	0.423	−2.765	.034

We only report the significant pairs in the comparison.

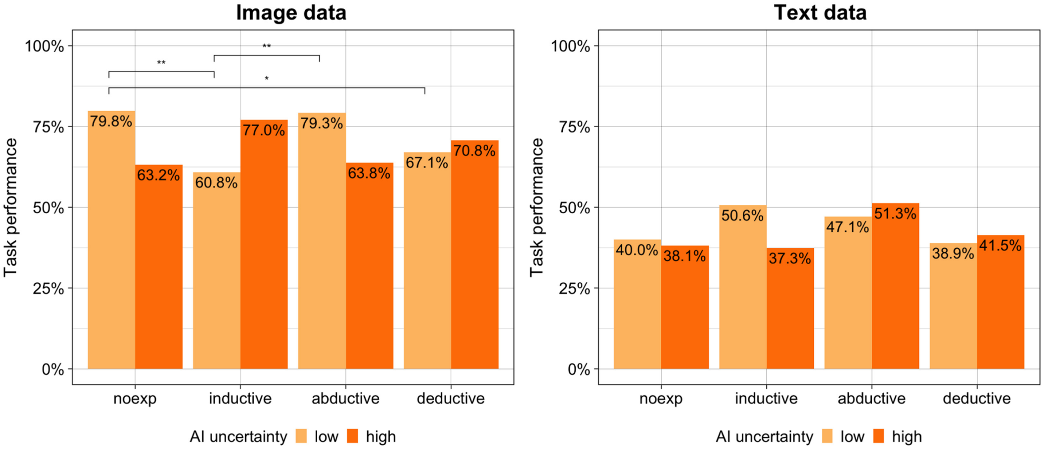


Fig. 11. Task Performance comparison among the different Explanation Styles split into the AI Uncertainty levels for the image and text domains (H2c).

We also found a significant interaction between the Explanation Style and the AI Uncertainty, so we proceeded with the post hoc analysis through a pairwise comparison, using Bonferroni's *p*-value adjustment. We report the significant pairs in Table 3. We found significant differences only in the case of low AI uncertainty. The inductive Explanation Style has a significant negative impact on Task Performance if compared against the no explanation and the abductive conditions. The deductive style has a significant negative impact on Task Performance if compared to the no explanation condition. We depicted the results of the pairwise comparisons on the left side of Figure 11. When the AI Uncertainty is low, inductive and deductive Explanation Styles have a negative effect on Task Performance. Comparatively, we do not register such a negative effect for abductive explanations.

For investigating the negative impact on Task Performance of inductive and deductive explanations with a low AI Uncertainty, we split the results according to AI Correctness factor, as depicted in Figure 12. The plot clearly shows that inductive explanations have a significant negative impact on Task Performance when AI makes wrong predictions, if compared against the no explanation (*Log-Odds* = −2.09, *Std. error* = 0.50, *z-value* = −4.148, $p < .05$) and the abductive Explanation Style (*Log-Odds* = −3.63, *Std. error* = 0.71, *z-value* = −4.148, $p < .05$) .

In summary, for image data, low User Uncertainty and right AI Correctness positively impacted the Task Performance (H2a and H2b). Further, inductive and deductive Explanation Styles negatively impact the performance in case of a low AI Uncertainty (H2c), especially when the AI is wrong.

Text Data. Table 4 summarizes the results of the logistic regression analysis in the text study. Similarly to the image domain, we found a significant main effect of the User Uncertainty, which

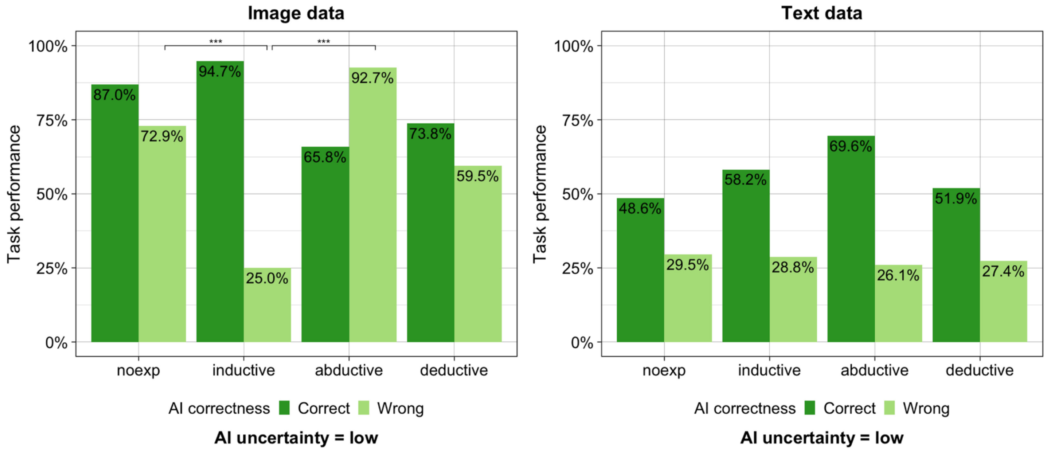


Fig. 12. Task Performance comparison among the different Explanation Styles with a low AI Uncertainty and split into the AI Correctness levels for the image (left part) and the text (right part) domains.

Table 4. Logistic Regression Results on Task Performance (H2, Text Domain)

Predictor	Log-Odds	Std. Error	z-Value	p
User uncertainty [low]	***1.455	0.179	8.090	<.001
AI correctness [right]	***1.379	0.180	7.654	<.001
Explanation style [inductive]	−0.125	0.343	−0.365	.715
Explanation style [abductive]	0.565	0.332	1.701	.089
Explanation style [deductive]	0.020	0.331	0.062	.950
AI uncertainty [low]	0.064	0.312	0.207	.836
Explanation style [inductive] * AI uncertainty [low]	0.578	0.480	1.203	.229
Explanation style [abductive] * AI uncertainty [low]	−0.207	0.480	−0.432	.666
Explanation style [deductive] * AI uncertainty [low]	−0.024	0.478	−0.051	.959

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

positively affects the Task Performance when it is low. The effect of the AI Correctness is significant, and a correct prediction yields an increase in the Task Performance. Such results allow us to reject the null hypotheses for H2a and H2b in the text domain, as expected. Figure 10 (right part) shows the variation in the performance according to the levels of these two factors.

In contrast with the results we obtained in the image domain, we do not find a significant interaction between the AI Uncertainty and the Explanation Style. Therefore, we fail to reject the null hypothesis for H2c (see Figure 11).

In summary, in the text domain, a low User Uncertainty and the right level of the AI Correctness positively impact Task Performance (H2a and H2b), but there are no significant interactions between the AI Uncertainty and the Explanation Style factors (H2c).

5.7 H3: Agreement

We have seen so far that task performance is influenced by AI correctness. We also have seen that participants rely a lot on the instance information rather than the prediction. So when do they agree with the AI prediction? We investigate whether the Agreement is affected by the levels of User Uncertainty, AI Correctness, and the combined effect of the AI Uncertainty and the Explanation Style. When user uncertainty is high, we expect users to rely on the prediction and

Table 5. Logistic Regression Results on Agreement (H3, Image Domain)

Predictor	Log-Odds	Std. Error	z-Value	p
User uncertainty [low]	−0.186	0.198	−0.936	.349
AI correctness [right]	***2.685	0.219	12.237	<.001
Explanation style [inductive]	**1.296	0.406	3.192	.001
Explanation style [abductive]	***1.662	0.412	4.028	<.001
Explanation style [deductive]	0.360	0.372	0.969	.333
AI uncertainty [low]	0.282	0.367	0.769	.441
Explanation style [inductive] * AI uncertainty [low]	0.045	0.570	0.080	.935
Explanation style [abductive] * AI uncertainty [low]	***−2.896	0.570	−5.078	<.001
Explanation style [deductive] * AI uncertainty [low]	−0.384	0.531	−0.724	.469

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

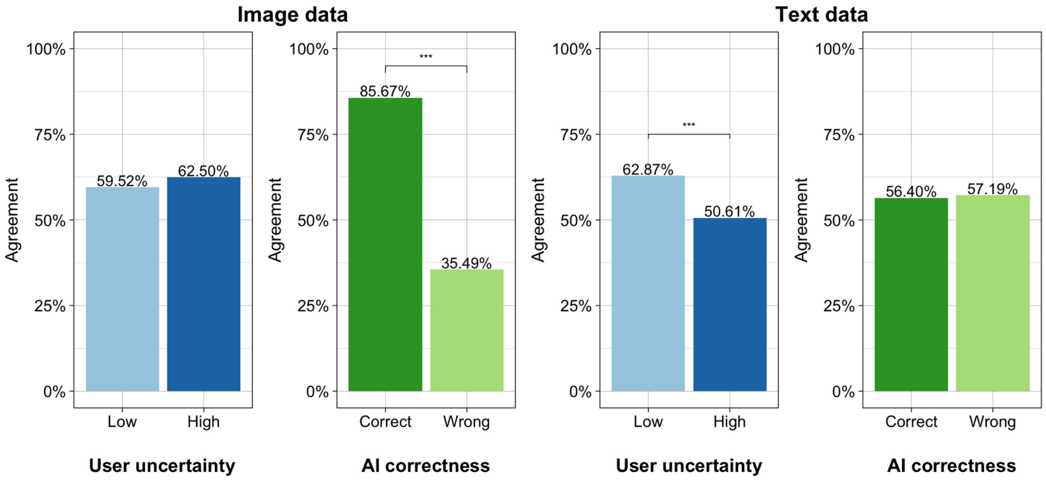


Fig. 13. Agreement comparison on the different levels of User Uncertainty and AI Correctness for the image domain on the left and the text domain on the right (H3a and H3b).

achieve a higher agreement than a low user uncertainty (H3a). When the AI is correct, we also expect users to rely on the system prediction and achieve a higher agreement than when they are wrong (H3b). Further, we expect differences in how the explanation type conveys the information about the AI uncertainty, convincing the user to follow the AI in some combinations and rejecting its suggestion in other ones (H3c).

Image Data. Table 5 shows the results of the logistic regression analysis of the Agreement in the image domain. We have not found any significant effect of the User Uncertainty, so we fail to reject the null hypothesis for H3a. We did find a significant main effect of the AI Correctness, which increases the Agreement when the AI Correctness level is right. Therefore, we reject the null hypothesis for H3b. We depicted the resulting values of the agreement according to the user uncertainty and the AI correctness on the left side of Figure 13.

We found a significant interaction between the AI Uncertainty and the Explanation Style (see Table 5), rejecting the null hypothesis for H3c. We proceeded to the post hoc analysis through a pairwise comparison using Bonferroni's p-value adjustment (Table 6). When the AI Uncertainty is low, the results highlight that inductive explanations significantly increase the

Table 6. Pairwise Comparison Between the Levels of AI Uncertainty and Explanation Style on the Agreement in the Image Domain (H3)

AI Uncertainty = low

Contrast (Explanation Style)	Log-Odds	Std. Error	*z*-Value	*p*
Inductive vs noexp	1.342	0.407	3.301	.006
Abductive vs noexp	−1.233	0.381	−3.240	.007
Abductive vs inductive	−2.576	0.434	−5.937	<.001
Deductive vs inductive	−1.366	0.424	−3.223	.008
Deductive vs abductive	1.209	0.396	3.054	.014

AI Uncertainty = high

Contrast (Explanation Style)	Log-Odds	Std. Error	*z*-Value	*p*
Inductive vs noexp	1.296	0.406	3.192	.008
Abductive vs noexp	1.663	0.413	4.028	<.001
Deductive vs abductive	−1.302	0.406	−3.205	.009

We only report the significant pairs in the comparison.

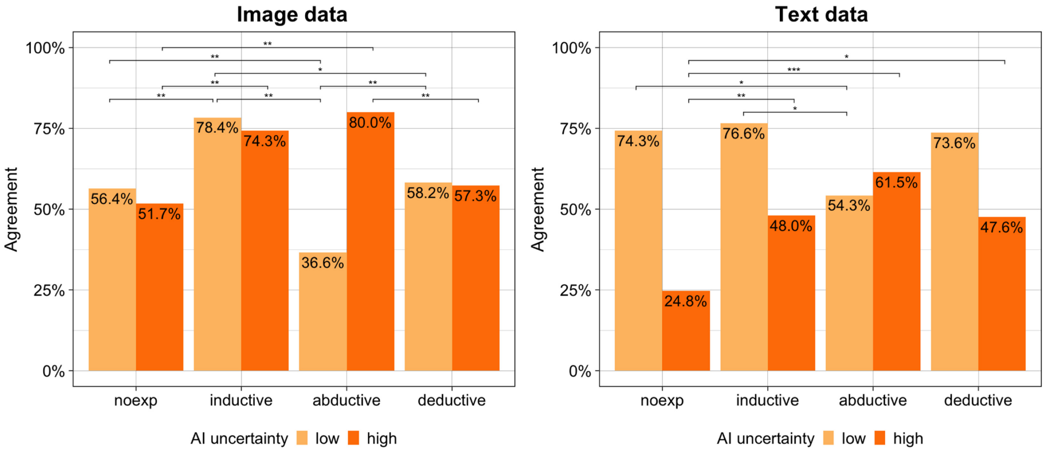


Fig. 14. Agreement comparison among the different Explanation Styles split into the AI Uncertainty levels for the image and text domains (H3c).

Agreement compared to all the other explanation styles. We registered an opposite effect for abductive explanations, which significantly decrease the Agreement compared to all the other styles. Such effects change when the AI Uncertainty is high. We registered a significant positive effect of inductive and abductive explanations compared to the no explanation condition. Further, abductive explanations have a significant positive effect if compared against deductive ones. We depicted the results of the pairwise comparisons on the left side of Figure 14. Such results confirm the results reported in the work of Cian et al. [12], which were assessed on proxy tasks, but only if we consider a high AI Uncertainty.

In summary, the User Uncertainty was not significant in the model (H3a), whereas the right level of AI Correctness positively impacted the Agreement (H3b). Further, we registered a significant interaction between the AI Uncertainty and the Explanation Style (H3c). When the AI Uncertainty is low, abductive explanations seem to dissuade the user from accepting the AI suggestion, but inductive explanations persuaded users to accept the suggestion. When

Table 7. Logistic Regression Results on Agreement (H3, Text Domain)

Predictor	Log-Odds	Std. Error	z-Value	p
User uncertainty [low]	***0.578	0.170	3.397	<.001
AI correctness [right]	−0.031	0.169	−0.184	.853
Explanation style [inductive]	**1.039	0.326	3.185	.001
Explanation style [abductive]	***1.623	0.327	4.960	<.001
Explanation style [deductive]	**0.997	0.319	3.124	.002
AI uncertainty [low]	***2.208	0.321	6.879	<.001
Explanation style [inductive] * AI uncertainty [low]	−0.940	0.48061	−1.957	.050
Explanation style [abductive] * AI uncertainty [low]	***−2.523	0.46559	−5.419	<.001
Explanation style [deductive] * AI uncertainty [low]	*−1.011	0.47452	−2.131	.033

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Table 8. Pairwise Comparison Between the Levels of AI Uncertainty and Explanation Styles on Agreement in the Text Domain (H3)

AI Uncertainty = low

Contrast (Explanation Style)	Log-Odds	Std. Error	z-Value	p
Abductive vs noexp	−0.899	0.331	−2.720	.039
Abductive vs inductive	−0.999	0.364	−2.746	.036

AI Uncertainty = high

Contrast (Explanation Style)	Log-Odds	Std. Error	z-Value	p
Inductive vs noexp	1.039	0.326	3.185	.009
Abductive vs noexp	1.623	0.327	4.960	<.001
Deductive vs noexp	0.997	0.319	3.124	.011

We only report the significant pairs in the comparison.

AI Uncertainty is high, inductive explanations have a positive effect on Agreement compared to the baseline condition without explanations. Abductive explanations have a positive effect on Agreement when the AI Uncertainty is high if compared to deductive explanations and the no explanation condition.

Text Data. Table 7 shows the results of the logistic regression analysis on the Agreement in the text domain. We found a significant main effect of the User Uncertainty, which has a positive impact when its level is low. Since the Agreement was higher with a low level of User Uncertainty than a high level, we fail to reject the null hypothesis for H3a. Further, we did not find any significant effect of the AI Correctness, so we fail to reject the null hypothesis for H3b. We depicted the results of the User Uncertainty and the AI Correctness on the right side of Figure 13.

As in the image domain, we found a significant interaction between the AI Uncertainty and the Explanation Style (see Table 7), which allows rejecting the null hypothesis for H3c. We proceeded to the post hoc analysis through a pairwise comparison using Bonferroni's p-value adjustment (Table 8). When the AI Uncertainty is low, abductive explanations have a significant negative effect if compared against inductive explanations and the baseline no explanation condition. Instead, when the AI Uncertainty is high, all explanation types have a positive effect if compared against the no explanation condition. We depicted the results of the pairwise comparisons on the right side of Figure 14.

In summary, in the text domain, the Agreement is higher with a low User Uncertainty than a high uncertainty (H3a). The AI Correctness factor does not impact the Agreement

Table 9. Hypotheses Results Summary for Image and Text Domains

Hypotheses	Image	Text
H1: Reliance		
H1a: When the user uncertainty is low, the user will rely more on the instance rather than on AI prediction and the explanation.	✓	✓
H1b: When the user uncertainty is high, the user will rely more on the instance and the explanation rather than on the AI prediction.	✗	✓
H2: Task Performance		
H2a: Users will achieve a higher task performance with a low user uncertainty than a high one.	✓	✓
H2b: Users will achieve a higher task performance with correct AI predictions than with wrong ones.	✓	✓
H2c: AI uncertainty moderates the effect of the explanations types on the task performance (positively or negatively).	✓	✗
H3: Agreement		
H3a: Users will achieve a higher agreement with a high user uncertainty than a low one.	✗	✗
H3b: Users will achieve a higher agreement with correct AI predictions than with wrong ones.	✓	✗
H3c: AI uncertainty moderates the effect of the explanations types on the agreement (positively or negatively).	✓	✓

(H3b). Further, a low AI Uncertainty negatively impacted the Agreement on the abductive explanations (H3c). When the AI Uncertainty is high, all Explanation Styles increase the Agreement if compared against the no explanation condition.

6 DISCUSSION

In this section, we discuss the findings in the studies' results, summarizing the impact of the factors we analyzed and comparing similarities and differences in the decision process for classifying images and texts. As we better discuss in Section 7, the impact of the findings we present are limited to the considered tasks. Table 9 summarizes the hypotheses results for both image and text domains.

6.1 Modeling User Uncertainty: Lesson Learned

The study results show that User Uncertainty is a key factor to consider for understanding the decision-making process. It has a significant effect on most of the metrics we collected in both studies, excluding the Agreement in the image domain. Therefore, modeling the user uncertainty would provide XAI systems with important information for selecting the information to provide users.

However, there are two important points that our study highlights in this regard. The first is that we needed to run a preliminary study to assign the instances to a level of uncertainty (see Appendix A). We performed a pre-screening for finding the characteristics identified by Zimmermann [92], but we needed users' feedback for correct classification, especially in the text domain, where such criteria are not available. We lack computational models for predicting the User Uncertainty. We are optimistic that the research community will eventually build such models and that they will be a powerful tool to use in XAI interfaces, given the results we present in this study.

The second important point that clearly emerges from our study is that we cannot rely on the on the User Confidence for replacing a model of the User Uncertainty. Indeed, our assessment shows that the self-evaluation correctly indicates only when the User Uncertainty is low (and user confidence is high). For capturing when the User Uncertainty is high, the confidence of a single user is not enough, and this is probably related to a process of overconfidence in the perception. In the study responses, we identified groups of participants indicating a high confidence but disagreeing on the class to assign to the same instance. So, each group is sure about their decision, but the disagreement on the class clearly indicates a high User Uncertainty. This means that we must model the User Uncertainty at a global level, and subjective measurements are unreliable.

6.2 Users Are Overconfident About Their Decisions

A common theme in our findings is that users are, in general, overconfident about their decisions. The first evidence comes from the results of the Reliance analysis, which are consistent in the image and text domains. Users identified the instance to classify as the most important factor for making the decision, regardless of the level of User Uncertainty. Such confidence in the instance explains the significance of the User Uncertainty on the Task Performance, showing that users' judgment on the instance impacts the final decision, as we expected before running the studies.

Text. In the text domain, overconfidence is particularly pronounced. The analysis of the User Confidence shows that 211 out of 284 people (about 72%) assigned a low uncertainty to an instance we marked as high, but they got the classification task wrong. This happened independently from the AI advice: 45% of users were assigned to a correct prediction and 55% to a wrong one. In addition, we registered a higher Task Performance for the image data, highlighting an underestimation of the tasks' difficulty in text classification.

Other evidence relates to the impact of the information provided by the AI. The factors explaining the Agreement for text data are (i) the User Uncertainty and (ii) the interaction between AI Uncertainty and Explanation Style. Explanations can persuade the user to agree or disagree with the AI, depending on their style and the level of AI uncertainty. We discuss this point in more detail in Section 6.4. The AI Correctness does not have a significant effect on the Agreement since it has similar rates when the AI is correct and when it is wrong. So, when users agree with the AI, they have a similar chance of eventually making a correct or wrong decision. As a consequence, the Agreement between the user and AI turns into a correct Task Performance only if the AI Correctness is actually right. This explains why AI Correctness is significant for performance and is not for the Agreement.

Images. There is a trend toward user overconfidence in the image domain, but the effect is less pronounced. We registered 92 out of 215 (about 43%) people reporting a high User Confidence on instances we assigned to a high level of User Uncertainty and eventually failing the classification task. The value is still high (and shows overconfidence), but it is lower than we registered for text. Participants considered the AI prediction and explanation as comparable sources of information, both in a low and high User Uncertainty setting. They consistently relied more on the AI prediction when it was correct and less when it was wrong. Overall, they seem to perceive when the AI prediction is reliable, and as a consequence, the AI Correctness is significant for the Task Performance.

The User Uncertainty is also significant for the Task Performance in the image domain, reflecting the reliance on the instance as the primary source for taking the decision. Instead, the User Uncertainty is not significant for the Agreement, showing that the AI prediction received a comparable consideration for both levels of uncertainty. The AI Correctness has a significant positive effect on both Task Performance and Agreement

when the prediction is correct, in line with the discussion of the Reliance results.

Overall, all these results indicate that users take advantage of the AI suggestion in the image domain to confirm the initial user's judgment when correct and for changing opinion when it was wrong. This confirms a lower users' overconfidence in the image domain if compared to text.

The impact of the interaction between the AI Uncertainty and the Explanation Style on Task Performance is more complex. It negatively affects the performance with inductive and deductive explanations when the AI uncertainty is low. We have a significant interaction on the Agreement resulting in mixed effects for different combinations of AI Uncertainty and Explanation Style.

Inductive explanations increase the Agreement independently from AI Uncertainty, in line with the results in the work of Kenny et al. [38]. The abductive explanations change their effect according to AI Uncertainty level: positive when the uncertainty is high (agree when AI is uncertain) and negative when the uncertainty is low (disagree when AI is certain). In brief, the Agreement data shows that users decided to follow AI suggestions in the image domain, relying on the information provided by the AI. They considered the AI advice more than in the text domain.

6.3 Mitigating the Overconfidence

The findings provide some indications for mitigating users' overconfidence in their abilities and increasing the consideration for AI suggestions. Of course, out with the experimental setting (in real-world tasks), we cannot indicate the AI Correctness and the User Uncertainty on the considered instance since they are unknown. However, we are able to compute the AI Uncertainty on a given instance. We also see that AI Uncertainty has an effect on the effectiveness of explanations in terms of Agreement with AI predictions. So, we can use this information to help users make informed decisions. We therefore propose that XAI interfaces should include information about AI Uncertainty.

Further, while training the model, we can measure the accuracy on both instances having high and low AI Uncertainty in the test set, for deriving an indication of the relationship between the AI Uncertainty and AI Correctness. We can exploit these measurements for selecting the information to be presented to users. Depending on the accuracy for both low and high AI Uncertainty, we may want to (i) persuade the user to follow the AI (i.e., increasing the Agreement in case of high accuracy) or (ii) critically analyze the prediction (i.e., lowering the Agreement in case of low accuracy). To persuade the user to agree with the AI in a low AI Uncertainty setting, we should either avoid explanations or exploit an inductive style in the image domain. Using the latter solution, we expect the Task Performance to degrade when the AI prediction is wrong, but when the model is accurate enough, such a case is not frequent. Instead, with a high AI Uncertainty, inductive and abductive explanations guarantee a higher Agreement for both image and text domains. In the text domain, any explanation has a positive effect (higher Agreement). We also have means for dissuading the user from following the AI. We can do it in both domains by inserting abductive explanations when the AI Uncertainty is low. When it is high, it is sufficient to avoid explanations.

6.4 Effects of the Explanation Styles

Another interesting aspect covered by our studies is the impact of the Explanation Styles on Task Performance and Agreement. Although most of the existing literature focuses on the presence or absence of a given explanation method, our work shows differences caused by the reasoning style grounding the explanation. In this thread of research, we deepen the results of other works [9, 38]. We can summarize our findings on this topic as follows.

Inductive explanations have a negative effect on Task Performance in image recognition when the AI Uncertainty is low. They have no significant effect when in the other conditions (i.e., high

AI Uncertainty in the image domain, high or low AI Uncertainty in the text domain). They persuade the user to follow the AI suggestion consistently in the image and the text domains, regardless of the AI Correctness. This confirms the result in the work of Kenny et al. [38] for the MNIST image dataset, whereas our study adds evidence of the same effect in the text domain. Therefore, the cause of such an increase in the Agreement seems related to their structure: inductive explanations present other instances similar to the current, which belongs to the same predicted class. Such a selection in the same class defines a similarity concept shared between users and AI. So, presenting similar instances belonging to the same predicted class is a good way for convincing users, even if the prediction is wrong.

In contrast, *abductive explanations* do not affect Task Performance but have a mixed effect on the Agreement with the AI prediction, depending on the AI Uncertainty. Such effect is negative (i.e., less agreement) when the AI Uncertainty is low and positive when it is high, and it is consistent for both the image and text domains. Therefore, abductive explanations seem to make people disagree with a confident AI, and to make people agree with an uncertain AI. In general, this is not in line with what interface designers expect from explanations. Our results only allow us to highlight such an effect, but further research is needed to better understand this phenomenon.

Finally, *deductive explanations* have a negative effect on Task Performance when the AI Uncertainty is low, similar to the inductive style. On the Agreement, the effect is not significantly different from the no explanation condition when the AI Uncertainty is low. When it is high, the effect is negative (i.e., less agreement) on image data and positive on text. Therefore, it is difficult to exploit the study results for providing guidance on such an explanation style. We need more research to assess the origin of its effects, investigating the available generation and the presentation techniques. Considering that generating deductive explanations requires model architecture adjustments and specific training, a deeper understanding might establish if and when such effort is worth it.

7 LIMITATIONS

This section discusses some limitations in our work, which may lead to further research.

The first limitation to the generalization of the results is the tasks selected for the experiments. Image and sentiment classification require decisions a user can make without the need of the AI, considering that they are "easy" tasks for human beings. The reliance, the agreement, and the effects of the user uncertainty may be different if we consider more complex decisions, such as buying or selling stocks and selecting treatments. We will cover this point in future work. The second point limiting the results generalization concerns the selection of representative elements in our experiment among the many available options. This includes specific datasets, the participant sample, the classification model architecture, and the explanation methods. For each of these options, we selected a representative balancing the study's relevance and feasibility. For instance, we selected the datasets and the tasks (MNIST handwritten digit database [46] and Yelp Reviews) [6] since they are well known in the research community, are publicly available, and contain easy and difficult instances to classify, and the classification tasks do not require specific training for the average user. On the one hand, such selection eases finding enough participants to reach the appropriate power in the statistical analysis (about 600 per study). On the other hand, the proposed task may be perceived as easy enough to be solved without AI, which may explain the overconfidence we discussed previously for both domains. Even though we consider the proposed tasks as relevant for XAI, we are also aware that we require further studies considering other datasets and tasks, which may lead to results different from those reported in this article. Similar reasoning led to the choice of the specific explanation methods (e.g., Grad-CAM or

LIME) selected as representative of a broader explanation type (e.g., inductive, deductive, or abductive). Naturally, specific characteristics of other explanation techniques may lead to different results.

The third limitation concerns the generation of logical reasoning explanations. We used well-known methods in the literature, employed in previous evaluations with users, which are generalizable to different tasks involving image and text classification. Nevertheless, we acknowledge that there is space for developing diverse and novel XAI methods, which leverage logical reasoning differently compared to the explanations we included in the study. This may lead to different interactions with the AI uncertainty. We selected the techniques that, to the best of our knowledge, represent each reasoning type at the current state of the research in this field.

The fourth limitation regards the generation of deductive reasoning explanations for image data, which were in textual form and might have had a disadvantage in the evaluation compared to inductive and abductive styles. A possible solution for creating deductive visual explanations could be the solution in the work of Goh et al. [26], supporting users in exploring the intermediate layers and neurons activations of the neural network. However, understanding and using such explanations effectively may go beyond the ability of the participants in this study. Further research is needed for finding viable solutions in visually presenting deductive explanations, and to assess the differences between visual and textual representations in the image domain.

The final limitation concerns the way we establish AI and user uncertainty levels. For instance, we can use Evidential Deep Learning [4] for calculating epistemic and aleatoric uncertainty, so further studies are needed to assess the robustness of the different methods computing the AI uncertainty. For the user uncertainty, we based the distinction on the work of Zimmermann [92] and on further validation with 30 users for each dataset (see Appendix A). Although this strategy was the most suitable for our specific classification tasks, other studies are needed to cover a wider range of uncertainty properties and to validate them with users, considering more elaborate scenarios and diverse data types. We provide empirical evidence that our selection is valid for performing the study, but we require a more general (and hopefully computational) approach.

8 CONCLUSION AND FUTURE WORK

This article discussed the effect of four factors involved in the AI-supported decision process, that currently are overlooked in the literature. The first is the User Uncertainty, which indicates how unsure a user may be on assigning a class based on the information provided by the instance. The second is AI Correctness, which indicates whether the AI predicts the correct or wrong class. The third factor is the AI Uncertainty, representing how unsure the AI is about its prediction. The final factor is the Explanation Style, and how it interacts with AI Uncertainty. The XAI interfaces convey the information on AI uncertainty to the user through explanations, so we studied how its levels (low and high) interact with different explanation types based on logical reasoning, including inductive, abductive, and deductive explanation styles. We registered their effects on the Reliance, Task Performance, and Agreement with the AI collecting data in two user studies, focusing on image and text classification. Table 9 summarizes the hypothesis we formulated on the effects and the validity according to the study results.

Unsurprisingly, our results show that participants rely on the instance as the primary source of information for making a decision. In general, we noticed that they use the same pattern across instances with low or high user uncertainty, hinting at overconfidence in users' abilities. We find confirmation for this trend in the analysis of self-reported confidence and Task Performance. The Reliance on AI prediction and explanations depends on AI Correctness. When the prediction is wrong, users consider the explanation more than the AI prediction, suggesting that explanations are helpful for decision making and avoiding overreliance.

As expected, the Task Performance is positively influenced by low User Uncertainty and correct AI predictions (AI Correctness = right) in both domains. This demonstrates the importance of considering the user uncertainty and an accurate prediction for making the correct decision.

There is also a difference between domains on Task Performance. The interaction between AI Uncertainty and Explanation Style has no significant effect in the text domain, whereas it is significant in the image domain. Our results suggest that this is due to a more pronounced overconfidence by users in the text domain. To understand the difference between the two tasks, we need to link the results on Task Performance and Agreement. In the image domain, people agree with the AI more when it is correct and less when it is wrong, with similar rates for both user uncertainty levels, resulting in good use of the AI advice. In contrast, in the text domain, people decide according to their own judgment, regardless of AI advice. Jointly, these results indicate a more pronounced overconfidence by users in the text domain.

As for the Explanation Style, the inductive explanations persuade users in agreeing with the AI consistently in both studies. Most notably, participants agreed more with wrong AI predictions with inductive explanations than the other explanation styles. This confirms the results in the work of Kenny et al. [38], extending the effect in the text domain. Such a persuasive effect is independent from the AI correctness.

Instead, the effect on the Agreement of the abductive explanations changes from negative when the AI Uncertainty is low to positive when it is high. Deductive explanations have a negative effect on performance when the AI Uncertainty is low. On the Agreement, deductive explanations have a significant effect only when the AI Uncertainty is high. Furthermore, the interaction between AI Uncertainty and Explanation Style for the deductive is different between the two domains: negative for images and positive for text. Such complex mixed effects of deductive and abductive explanations require further research for being properly described.

In future work, we aim at investigating open questions not covered by the results of this study. First, we will deeply investigate the effects of abductive and deductive explanations in the decision-making process to identify the source of the mixed effects we registered considering different levels of AI Uncertainty. Additionally, we will try to understand if other XAI techniques (including counterfactual reasoning) leveraging on the same reasoning style have similar effects. Second, we intend to extend our view of logical reasoning types to different models architectures among neural networks and testing different ways to quantify the AI uncertainty. Third, given the importance of XAI in supporting human decision making, we intend to explore other ways to measure user uncertainty based on the target domain, user experience, and data type, possibly specifying a computational definition for employing it on XAI interfaces. Finally, we aim to explore this approach with other datasets. In particular, we will focus on applying and evaluating logical reasoning explanations in tabular and video data to study the generalizability of the findings.

APPENDIX

A STUDY REPRODUCIBILITY DETAILS

In this appendix, we report the technical details on the classification models, the explanation generation, and the identification of the User Uncertainty levels through a pre-study for each dataset (images and text). They should contribute to understand the details of the two studies presented in the article and allow their reproduction for further research. The final AI uncertainty values (epistemic and aleatoric) were calibrated using temperature scaling [27].

A.1 Image Data

For defining the image classification tasks, we decided to use the MNIST [46] dataset since digits recognition requires no prior knowledge for the participants. Furthermore, we choose this dataset because it contains both instances having low and high user and AI uncertainty.

Classification Model. For the classification task, we adapted the AlexNet architecture [44] for digits classification in the MNIST dataset and the uncertainty calculation. Since AlexNet was outlined to discriminate among 1,000 different classes from the ImageNet database [13], we needed to scale it appropriately for our use case to avoid overfitting by design.

First, we changed the input size of the network from 224 × 224 × 3 to 28 × 28 × 1 to match the shape of the MNIST digits and halved all the convolutional filter size dimensions, further adjusting the kernel size to 3 × 3 for each convolutional level. Then, we reduced the last two fully connected layers dimension from 4096 to 1024 and 512, also changing the softmax layer number from 1,000 to 10 classes distribution. For calculating the epistemic uncertainty, we added dropout layers with a 0.25 rate after each MaxPool layer using 100 Monte Carlo dropout samples once we trained the model. Further, we used a modified loss function as outlined in the work of Kendall and Gal [36] to calculate the heteroscedastic uncertainty. Since this kind of uncertainty is a function of the input data and predicted as part of the training process, our AlexNet model has two outputs: the categorical cross-entropy loss for the classification and the input variance to get the heteroscedastic uncertainty (see the work of Kendall and Gal [36] for an in-depth description). Finally, we trained this model using the Adam [42] optimizer with a batch size of 128 and 10 epochs, splitting the data between 60,000 train instances and 10,000 test ones and achieving about 99% accuracy on the test set.

Selecting Images Corresponding to the Uncertainty Levels. After defining the model architecture, we proceeded with selecting the images to include in the user study. The goal is to identify 6 groups of images for each pair in the combination of user and AI uncertainty levels (i.e., low-low, low-high, high-low, and high-high), 24 in total. We assign 2 images from each group to a reasoning style (inductive, abductive, deductive) for pre-computing the explanations. Each reasoning style sub-group includes one instance where the AI prediction is correct and one is wrong. As for the no explanation condition, we randomly pick the instances among those assigned to a reasoning style, maintaining the balance on all the experimental conditions. For selecting the instances, we proceeded as follows. First, we calculated the *epistemic uncertainty* values on the modified AlexNet model to find candidates for each level of AI uncertainty in the MNIST [46] test set. Then, we picked 20 images having the *highest* epistemic uncertainty values and 10 having the *lowest* epistemic uncertainty. After that, we repeated the process calculating the *aleatoric uncertainty* in the MNIST [46] test set. Again, we picked 20 images with the *highest* aleatoric uncertainty and 10 with *lowest* aleatoric uncertainty. The rationale behind selecting more images with high aleatoric or epistemic uncertainty is to increase the chance of finding instances with a high user uncertainty, which are more difficult to find than low-uncertainty images. A high aleatoric or epistemic uncertainty does not guarantee a high user uncertainty.

Next, we considered the causes of uncertainty described in Section 2.2 (lack of information and ambiguity). We identified 10 images in the test set where such causes are present. Summing them to the instances selected through the epistemic and aleatoric uncertainty, we end with a set of 70 images. This number derives from a rough estimation of the number of instances we require for the study.

After that, we estimated the user uncertainty on the pre-selected image set with real users. We asked a group of 30 people to recognize the digit depicted in each of them. The participant group consisted of 13 females and 17 males, aged between 18 and 59 years ($\bar{x} = 28.5$, $\tilde{x} = 25$, $s = 9.8$). Eleven of them had a high school degree, 14 a bachelor's degree, and 5 a master's degree. We computed the accuracy for each image as the number of participants who correctly classified the digit out of the total number of participants. Afterward, we computed the accuracy distribution and used the first quartile to establish the threshold for high versus low user uncertainty. We assigned each image having an accuracy of Q_1 = 40% or less to a high user uncertainty and the others to a low user uncertainty.

Finally, having collected the required information, we assigned the images to each group. In summary, the user uncertainty resulted from the pre-study, the AI uncertainty from the epistemic uncertainty, and the AI correctness from the AlexNet model prediction.

Generating the Explanations. We created *inductive explanations* by extracting the AlexNet model contributions for the train and test sets by leveraging a local feature-weighting method technique called the *COLE-Hadamard product* (COLE-HP, see the work of Kenny and Keane [39] for more details). With this technique, we map the weights of the features of the modified AlexNet model into a *k*-NN classifier, reaching 100% agreement on predictions on the test set between the twinned systems. The agreement metric measures the fidelity of the *k*-NN (transparent) model compared to the modified AlexNet black-box one. Therefore, we use the *k*-NN model to obtain three nearest-neighbor explanations of the task image. An example of this type of explanation is shown on the top left of Figure 4. We decided to use the COLE-HP method instead of choosing instances manually because it provides a very high agreement between the CNN and the *k*-NN, giving some insight into the model prediction through nearest neighbors.

For *abductive explanations*, we decided to use the Grad-CAM [64] technique to generate a visual explanation, applying it to the last convolutional layer of the modified AlexNet model. The result is a heatmap whose pixels range from the blue color indicating a low-class activation for that specific instance to the red that indicates a high-class activation. An example of this explanation is depicted in the top right of Figure 4. We decided to use Grad-CAM saliency maps explanations because they rely on the trained model's weights and the relationship between training examples and their labels, passing the sanity checks discussed in the work of Adebayo et al. [1].

To generate *deductive explanations*, we used an encoder-decoder architecture, similar to the one in the work of Vinyals et al. [78]. The original model is a generator of descriptive sentences matching an input image. It used a pre-trained CNN as an encoder for the input image, representing it as a fixed-length vector, corresponding to the last layer of the CNN. The vector is the input of the decoder model, embedding sentences associated with each image. The decoder model is a **Long-Short Term Memory (LSTM)** that generates sentences based on the CNN features and the embedded sentences related to each image. Consequently, we used this image captioning architecture to find significant characteristics of the image and use them as deductive explanations. We created a train set with three relevant sentences for each specific digit, without having the correct caption for the test set ones, which would require manual labeling. In general, other methods exist that require already labeled datasets, additional textual information for the training procedure, and model architecture modifications. For example, we became inspired by justification systems such as those of Hendricks et al. [32] and Hassan et al. [73], for which the authors generate sentences to justify the class prediction for a given image. We represent deductive explanations in a textual format since we did not find suitable candidates using a visual representation. For instance, the visualization in the work of Goh et al. [26] corresponds to deductive reasoning but requires users to interact with intermediate layers and neurons, and this goes beyond the abilities of our target study participants.

For our case study, we used the AlexNet COLE-HP train set contributions obtained in the inductive explanation as image features and an LSTM as the decoder to generate a caption for the image. The sentence generation part works as follows. First, we generated three different captions for all the MNIST training images using sentences with valuable attributes related to each digit. We empirically established that three captions were enough to cover the distinct representations of each number. For example, a possible caption for zero digits could be “it has a ring-like shape” or “it has a round shape.” Since we already have our image encoding from the COLE-HP contributions, we concatenate these features with the embedding layer and the LSTM memory, having a dimension of 512. Next, we initialized the matrix weights for the embedding layer by mapping all the words in our captions to 200-dimension GLoVe [57] vectors. Thus, we do not retrain these weights during the training process. We trained the caption generation model using the categorical cross-entropy loss and the Adam optimizer, with a batch size of 256 and 20 epochs, adopting the BeamSearch strategy with a beam of size 3 to generate captions for the 24 task images. To evaluate the goodness of the resulting captions, we calculated the agreement between the AlexNet model and the generated captions, obtaining 99.95%. An example of a generated caption for number one is depicted in the bottom of Figure 4, which reports “it is represented as a straight line.”

A.2 Text Data

For defining the text classification tasks, we used the Yelp Reviews dataset [6], consisting of online reviews in free-form text and a star rating out of five. We chose this dataset since it contains both low and high user and AI uncertainty instances and can be adapted to let users recognize sentiments in reviews instead of predicting the star rating. Since associating the exact number of stars for each review would be hard for users, we decided to reduce the classes from five to three by considering 1 and 2 stars as bad reviews, 3 stars as average reviews, and 4 and 5 stars as good reviews. We use a subset of the original dataset by random sampling and balancing the new three-class sentiments, selecting 36,000 instances (12,000 for each class) in the English language. The adapted dataset contains 24,000 reviews for training, 6,000 for validation, and 6,000 for testing. Each sentiment is represented in about one-third of the instances in each set. Further, we tested removing/keeping stopwords from the dataset and noticed that the stopwords removal increased the model’s accuracy, and we decided to remove them.

Classification Model. We started testing two models that perform well in text classification tasks: Bidirectional Encoder Representations from Transformers (BERT) [15] with one additional output layer (same settings as in the work of Kenny and Keane [39]), and Yoon Kim CNN-multichannel [40] united with Transformer Universal Sentence Encoder (USE_T) [11] for features extraction. After the model modifications needed to obtain epistemic and aleatoric uncertainty, we decided to keep the Yoon Kim model since we obtained about 78% accuracy in the test and validation sets compared to BERT, which achieved about 74%. From now on, we refer to Yoon Kim’s CNN-multichannel as CNN-USE_T.

For calculating the epistemic uncertainty, we added dropout layers (with 0.25 rate) after the MaxPooling operation, one for each channel (three channels) using 100 Monte Carlo dropout samples once we trained the model. We set three filter region sizes (3, 4 and 5) with 200 feature maps for each one. As for images, we used a modified loss function as outlined in the work of Kendall and Gal [36] to calculate the heteroscedastic uncertainty. Since this kind of uncertainty is a function of the input data and predicted as part of the training process, our CNN-USE_T model has two outputs: the categorical cross-entropy loss for the classification and the input variance to get the heteroscedastic uncertainty (see the work of Kendall and Gal [36] for more details). Finally, we trained this model using the Adam [42] optimizer with a batch size of 128 and 20 epochs, obtaining about 78% accuracy in the test and validation sets.

Selecting Reviews Corresponding to the Uncertainty Levels. For selecting the reviews to include in the user study, we used a procedure similar to the one we followed for the image data. The goal is the same: we need to find 24 reviews for assigning 6 of them to all the combinations of AI and user uncertainty.

We created the two pre-selection sets again calculating the epistemic and aleatoric uncertainty on the CNN-USE_T model. We discarded reviews with more than 100 words for avoiding excessively long texts. Since we do not have theoretical criteria for identifying user uncertainty sources on text as we did for images [92]. We selected more instances through the epistemic and aleatoric uncertainty, respectively 25 with a high level and 10 with a low level for each set, ending up with 70 pre-selected instances.

After that, we proceeded to evaluate the user uncertainty with real users. We asked a group of 30 people to recognize the sentiment in the pre-selected instances among good, average, and bad. Participants consisted of 16 females and 14 males aged between 18 and 59 years ($\bar{x}$ = 30.9, $\tilde{x}$ = 28 , s = 11.3). Ten of them had a high school degree, 12 a bachelor's degree, and 8 a master's degree. We computed the accuracy for each review as the number of participants who correctly classified the digit out of the total number of participants, using the first quartile of its distribution as the threshold for high versus low user uncertainty. We assigned each review with an accuracy of (Q_1= 37%) or less to a high user uncertainty and the others to a low user uncertainty. Finally, we assigned the reviews to each group using the user uncertainty resulting from the pre-study, the AI uncertainty from the epistemic uncertainty, and the AI correctness from the CNN-USE_T model prediction.

Generating the Explanations. We created *inductive explanations* following the same procedure for images: we extracted the CNN-USE_T model contributions for train and test sets by leveraging a local feature-weighting method technique called the *COLE-Hadamard product* [39]. We map the weights of the features of the modified CNN-USE_T model into a k-NN classifier, reaching 100% agreement on predictions on the test set between the twinned systems. We use the k-NN model to obtain three nearest-neighbor explanations of the task review. For neighbor reviews that have more than 50 words, we decided to add three dots after the last word to avoid adding cognitive effort to users. We thought that 50 words were enough for users to catch the review's sentiment. An example of this type of explanation is depicted in the top left of Figure 5.

For *abductive* and *deductive* explanations, we decided to use LIME [60]. LIME is a technique that can be applied to any black-box model to interpret its predictions by perturbing the input of data samples and understanding how the predictions change, proposing human-friendly explanations. For text inputs, LIME returns a list of words in the text, ranked in decreasing order of importance, which can be interpreted as the importance of the word according to the model for a given prediction. We generated the explanations using 5,000 sizes of the neighborhood to learn the linear model and using 10 features to be presented in the explanation of the top three classes (good, average, bad). Further, the generated explanations follow the CNN-USE_T model predictions.

For the abductive explanations, we decided to show the importance of these 10 words on the text via sentence highlighting with words belonging to the model predicted class and those against that class. The importance of a word is represented by its background color: the more opaque a word background is, the higher its importance. An example of this type of explanation is depicted in the top right of Figure 5.

To generate the deductive explanations, we decided to show 10 features for each of the top three classes using the built-in visualization of LIME. For each class, we visualize 10 features comprehending those in favor of that class and those against. We used this visualization for presenting significant words in the review and use them as deductive explanations. An example of this type of explanation is depicted in the bottom part of Figure 5.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18).* 9525–9536.

[2] Kamran Alipour, Jürgen P. Schulze, Yi Yao, Avi Ziskind, and Giedrius Burachas. 2020. A study on multimodal and interactive explanations for visual question answering. *CoRR abs/2003.00431* (2020).

[3] Ahmed Alqaraawi, M. Schuessler, Philipp Weiß, Enrico Costanza, and N. Bianchi-Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: A user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces.*

[4] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep Evidential Uncertainty. Retrieved March 22, 2023 from https://openreview.net/forum?id=S1eSoeSYwr.

[5] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv:1909.03012* [cs.AI] (2019).

[6] Nabiha Asghar. 2016. Yelp Dataset Challenge: Review rating prediction. *arXiv:1605.05362* [cs.CL] (2016).

[7] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE* 10, 7 (2015), 1–46. https://doi.org/10.1371/journal.pone.0130140

[8] Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why attentions may not be interpretable? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'21).* ACM, New York, NY, 25–34. https://doi.org/10.1145/3447548.3467307

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

[10] Federico Maria Cau, L. D. Spano, and N. Tintarev. 2020. Considerations for applying logical reasoning to explain neural network outputs. In *Proceedings of the 2020 Italian Workshop on Explainable Artificial Intelligence (XAI.it@AI*IA'20).*

[11] Daniel Cer, Yinfei Yang, Sheng Yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, et al. 2018. Universal sentence encoder. *arXiv:1803.11175* [cs.CL] (2018).

[12] David Cian, Jan van Gemert, and Attila Lengyel. 2020. Evaluating the performance of the LIME and Grad-CAM explanation methods on a LEGO multi-label image classification task. *arXiv:2008.01584* [cs.CV] (2020).

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

[14] Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural Safety* 31 (2009), 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805* [cs.CL] (2019).

[16] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608* (2017).

[17] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[18] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, et al. 2015. From captions to visual concepts and back. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15).* https://doi.org/10.1109/cvpr.2015.7298754

[19] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2013. G*Power 3.1.7: A flexible statistical power analysis program for the social, behavioral and biomedical sciences. *Behavior Research Methods* 39, 2 (2013), 175–191.

[20] Peter Flach and Antonis Kakas. 2000. Abductive and inductive reasoning: Background and issues. In *Abduction and Induction.* Applied Logic Series, Vol. 18. Springer, 1–27. https://doi.org/10.1007/978-94-017-0606-3-1

[21] Rudolf Franz Flesch. 1979. *How to Write Plain English: A Book for Lawyers & Consumers.* HarperCollins.

[22] Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32, 200 (1937), 675–701. https://doi.org/10.1080/01621459.1937.10503522

[23] Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics* 11 (1940), 86–92.

[24] Yarin Gal. 2016. *Uncertainty in Deep Learning.* Ph.D. Dissertation. University of Cambridge.

[25] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning—Volume 48 (ICML'16)*. 1050–1059.

[26] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*. Retrieved March 22, 2023 from https://doi.org/10.23915/distill.00030. https://distill.pub/2021/multimodal-neurons.

[27] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* 70 (2017), 1321–1330. https://proceedings.mlr.press/v70/guo17a.html.

[28] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. 2021. Towards general purpose vision systems. *arXiv:2104.00743* [cs.CV] (2021).

[29] Craig M. Harvey and Richard J. Koubek. 2000. Cognitive, social, and environmental attributes of distributed engineering collaboration: A review and proposed model of collaboration. *Human Factors and Ergonomics in Manufacturing & Service Industries* 10, 4 (2000), 369–393. https://doi.org/10.1002/1520-6564(200023)10:4<369::AID-HFM2>3.0.CO;2-Y

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv:1512.03385* [cs.CV] (2015).

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

[32] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, B. Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *Computer Vision—ECCV 2016*. Lecture Notes in Computer Science, Vol. 9908. Springer, 3–19.

[33] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88. https://doi.org/10.1109/MIS.2013.24

[34] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (April 2011), 141–154. https://doi.org/10.1016/j.dss.2010.12.003

[35] Mark T. Keane and Eoin M. Kenny. 2019. How case-based reasoning explains neural networks: A theoretical analysis of XAI using *post-hoc* explanation-by-example from a survey of ANN-CBR twin-systems. In *Case-Based Reasoning Research and Development*. Lecture Notes in Computer Science, Vol. 11680. Springer, 155–171.

[36] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Red Hook, NY, 1–11. https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

[37] Alex Kendall, Y. Gal, and R. Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7482–7491.

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

[39] Eoin M. Kenny and Mark T. Keane. 2021. Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. *Knowledge-Based Systems* 233 (2021), 107530. https://doi.org/10.1016/j.knosys.2021.107530

[40] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv:1408.5882* [cs.CL] (2014).

[41] J. Peter Kincaid, Robert P. Fishburne, R. L. Rogers, and Brad S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. IST Technical Report. Institute for Simulation and Training, University of Central Florida.

[42] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2015).

[43] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. *Proceedings of the 34th International Conference on Machine Learning* 70 (2017), 185–1894. http://proceedings.mlr.press/v70/koh17a.html.

[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Red Hook, NY, 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[45] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, New York, NY, 1675–1684. https://doi.org/10.1145/2939672.2939874

[46] Y. LeCun and C. Cortes. 2010. The MNIST Database of Handwritten Digits. Retrieved March 22, 2023 from http://yann.lecun.com/exdb/mnist/.

[47] Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. *arXiv:1908.11355* [cs.CL] (2019).

[48] Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded evaluations of explanation methods for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. 5198–5208. https://doi.org/10.18653/v1/D19-1523

[49] Yuelin Li, Yu Chen, Jinghong Liu, Yuan Cheng, Xuan Wang, Ping Chen, and Qianqian Wang. 2011. Measuring task complexity in information search from user's perspective. *Proceedings of the American Society for Information Science and Technology* 48, 1 (2011), 1–8. https://doi.org/10.1002/meet.2011.14504801092

[50] Peng Liu and Zhizhong Li. 2012. Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics* 42, 6 (2012), 553–568. https://doi.org/10.1016/j.ergon.2012.09.001

[51] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. 4768–4777. https://doi.org/10.5555/3295222.3295230

[52] Sean McNee, Shyong Lam, Catherine Guetzlaff, Joseph Konstan, and John Riedl. 2003. Confidence displays and training in recommender systems. In *Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'03)*.

[53] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[54] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv:1811.11839* (2018).

[55] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv:1802.00682* [cs.AI] (2018).

[56] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R. Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI'21)*. ACM, New York, NY, 340–350. https://doi.org/10.1145/3397481.3450639

[57] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GLoVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543. https://doi.org/10.3115/v1/D14-1162

[58] Dheeraj Rajagopal, Vidhisha Balachandran, Eduard Hovy, and Yulia Tsvetkov. 2021. SelfExplain: A self-explaining architecture for neural text classifiers. *arXiv:2103.12279* [cs.CL] (2021).

[59] Nazneen Fatema Rajani and Raymond J. Mooney. 2017. Ensembling visual explanations for VQA. In *Proceedings of the NIPS 2017 Workshop on Visually-Grounded Interaction and Language (ViGIL'17)*. http://www.cs.utexas.edu/users/ai-labpub-view.php?PubID=127684.

[60] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics Demonstrations*. 97–101. https://doi.org/10.18653/v1/N16-3020

[61] Maria Riveiro and Serge Thill. 2021. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence* 298 (2021), 103507. https://doi.org/10.1016/j.artint.2021.103507

[62] Philipp Schmidt and Felix Biessmann. 2019. Quantifying interpretability and trust in machine learning systems. *arXiv:1901.08558* [cs.LG] (2019).

[63] Christian D. Schunn and J. Gregory Trafton. 2012. The psychology of uncertainty in scientific data analysis. In *Handbook of the Psychology of Science*, Gregory Feist and Michael Gorman (Eds.). Springer, 461–485. http://d-scholarship.pitt.edu/23034/.

[64] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128, 2 (Oct. 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[65] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Dhruv Batra, and Devi Parikh. 2019. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV'19)*. 2591–2600.

[66] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*. 3145–3153.

[67] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2019. Learning important features through propagating activation differences. *arXiv:1704.02685* [cs.CV] (2019).

[68] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556* [cs.CV] (2015).

[69] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: Removing noise by adding noise. *arXiv:1706.03825* [cs.LG] (2017).

[70] Jost Tobias Springenberg, A. Dosovitskiy, T. Brox, and Martin A. Riedmiller. 2015. Striving for simplicity: The all convolutional net. *CoRR abs/1412.6806* (2015).

[71] Frode Sørmo, Jörg Cassens, and Agnar Aamodt. 2005. Explanation in case-based reasoning—Perspectives and goals. *Artificial Intelligence Review* 24, 10 (2005), 109–143. https://doi.org/10.1007/s10462-005-4607-7

[72] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*. Springer, 353–382.

[73] M. Ul Hassan, P. Mulhem, D. Pellerin, and G. Quénot. 2019. Explaining visual classification using attributes. In *Proceedings of the 2019 International Conference on Content-Based Multimedia Indexing (CBMI'19)*. 1–6. https://doi.org/10.1109/CBMI.2019.8877393

[74] P. Vakkari. 1999. Task complexity, problem structure and information actions—Integrating studies on information seeking and retrieval. *Information Processing & Management* 35 (1999), 819–837.

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

[76] Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen, and Mark Neerincx. 2020. Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies* 144 (2020), 102493. https://doi.org/10.1016/j.ijhcs.2020.102493

[77] Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: A systematic review. *arXiv:2006.00093* [cs.AI] (2020).

[78] Oriol Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 3156–3164.

[79] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. https://doi.org/10.1145/3290605.3300831

[80] Wenguan Wang and Jianbing Shen. 2018. Deep visual attention prediction. *IEEE Transactions on Image Processing* 27, 5 (May 2018), 2368–2378. https://doi.org/10.1109/tip.2017.2787612

[81] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do you trust me?": Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA'19)*. ACM, New York, NY, 7–9. https://doi.org/10.1145/3308532.3329441

[82] Paul Whitten, Francis Wolff, and Chris Papachristou. 2021. Explainable artificial intelligence methodology for handwritten applications. In *Proceedings of the IEEE National Aerospace and Electronics Conference (NAECON'21)*. 277–282. https://doi.org/10.1109/NAECON49338.2021.9696413

[83] Robert Wood. 1986. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* 37, 2 (1986), 60–82. https://doi.org/10.1016/0749-5978(86)90044-0

[84] Jialin Wu and Raymond Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. https://doi.org/10.18653/v1/w19-4812

[85] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. 2020. Explainable deep learning: A field guide for the uninitiated. *arXiv:2004.14545* [cs.LG] (2020).

[86] Ming Yin, Jennifer Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. 1–12. https://doi.org/10.1145/3290605.3300509

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

[88] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*.

[89] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593. https://doi.org/10.3390/electronics10050593

[90] J. Zhou, Huaiwen Hu, Z. Li, K. Yu, and F. Chen. 2019. Physiological indicators for user trust in machine learning with influence enhanced fact-checking. In *Machine Learning and Knowledge Extraction: The 3rd IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference (CD-MAKE'19)*. 94–113.

[91] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. 2019. Effects of influence on user trust in predictive decision making. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA'19)*. ACM, New York, NY, 1–6. https://doi.org/10.1145/3290607.3312962

[92] Hans-Jürgen Zimmermann. 2000. Application-oriented view of modeling uncertainty. *European Journal of Operational Research* 122, 4 (2000), 190–198. https://doi.org/10.1016/S0377-2217(99)00228-3

Received 4 March 2022; revised 21 October 2022; accepted 17 February 2023