



# Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations

Max Schemmer  
Karlsruhe Institute of Technology  
Karlsruhe, Germany  
max.schemmer@kit.edu

Niklas Kühn  
Karlsruhe Institute of Technology  
Karlsruhe, Germany  
niklas.kuehl@kit.edu

Carina Benz  
Karlsruhe Institute of Technology  
Karlsruhe, Germany  
carina.benz@kit.edu

Andrea Bartos  
Karlsruhe Institute of Technology  
Karlsruhe, Germany  
andrea.bartos@alumni.kit.edu

Gerhard Satzger  
Karlsruhe Institute of Technology  
Karlsruhe, Germany  
gerhard.satzger@kit.edu

## ABSTRACT

AI advice is becoming increasingly popular, e.g., in investment and medical treatment decisions. As this advice is typically imperfect, decision-makers have to exert discretion as to whether actually follow that advice: they have to “appropriately” rely on correct and turn down incorrect advice. However, current research on appropriate reliance still lacks a common definition as well as an operational measurement concept. Additionally, no in-depth behavioral experiments have been conducted that help understand the factors influencing this behavior. In this paper, we propose Appropriateness of Reliance (AoR) as an underlying, quantifiable two-dimensional measurement concept. We develop a research model that analyzes the effect of providing explanations for AI advice. In an experiment with 200 participants, we demonstrate how these explanations influence the AoR, and, thus, the effectiveness of AI advice. Our work contributes fundamental concepts for the analysis of reliance behavior and the purposeful design of AI advisors.

## CCS CONCEPTS

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

## KEYWORDS

Appropriate Reliance, Explainable AI, Human-AI Collaboration, Human-AI Complementarity

### ACM Reference Format:

Max Schemmer, Niklas Kühn, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *28th International Conference on Intelligent User Interfaces (IUI '23)*, March 27–31, 2023, Sydney, NSW, Australia. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3581641.3584066>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*IUI '23*, March 27–31, 2023, Sydney, NSW, Australia

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0106-1/23/03...\$15.00

<https://doi.org/10.1145/3581641.3584066>

## 1 INTRODUCTION

Most important decisions are made by calling upon advisors. While in the past advice was typically obtained from human experts, nowadays advisors based on artificial intelligence (AI) are becoming increasingly frequent in research and practice [33, 34]. For example, AI advises medical professionals in breast cancer screening [48], or in loan decisions [16].

In the past, research has predominantly focused on maximizing the reliance [35], trust [66], utilization [2], compliance [37], or acceptance [65] of AI advice [34]. With certain nuances, all these terms basically describe a concept that aims at maximizing the amount of AI advice that a human decision-maker eventually follows, i.e. maximizing the AI reliance. Recently, a new line of research has emerged that argues that maximising AI reliance does not fully exploit the potential of state-of-the-art human-AI decision making [6, 9, 62, 78]. We summarize the reasons for this line of thought in three main categories:

**Increasing usage of imperfect AI advisors.** First, prior research on AI advice often assumed “perfect” advice [34]. This makes sense if one considers narrow application spaces, e.g. performing deterministic algebra. But AI nowadays is used for more complex tasks [19] which increases both the number and severity of AI errors. Therefore, generally accepting AI advice would also include the acceptance of incorrect AI advice: Assume that without an AI advisor’s intervention, a physician would have diagnosed cancer, yet in the setting with AI advice had been misled by an incorrect AI advice and, thus, had failed to detect the cancer.

**Increasing alignment of the objectives of human decision-makers and AI advisors.** Second, in previous research, the goal of human or AI advisors was often inconsistent with the goal of the decision-maker, e.g., financial advisors wanted to persuade a client to purchase certain financial products that would yield the highest financial benefit for their own bonus. In this situation, AI designers do not want customers to differentiate good from bad advice but simply increase acceptance of advice. Today, however, AI advisors are often specifically designed to enhance human decision-making [41]. The advice seekers can design the advisor based on individual goals and with desired features such as “honest” explainability. This honesty might be missing if the inherent goal of the advisor is different from the advice seeker’s. If the AI is not designed in alignment with human goals, it is often not in the AI designer’s

interest to enable humans to critically question the advice they receive.

**Increasing potential for complementary team performance.** Third, as modern AI is not only more performant and offers more application areas, but also complements humans [27, 28, 40], there is an increasing potential to achieve complementary team performance (CTP), i.e., a performance that exceeds both—individual AI and human performance [6, 20, 28, 40, 50]. However, this level of performance can only be achieved by exploiting complementary capabilities. Even in the case of superior AI advice, it cannot be achieved simply by accepting AI advice, as it is then tied to AI performance without considering the potential additional strengths of the human.

In conclusion, human decision-makers should not simply rely on AI advice, but should be empowered to differentiate when to rely on AI advice and when to rely on their own, i.e. they should display **appropriate reliance** (AR) [6, 74, 75, 78]. Despite being a necessary condition for the effective use of AI, current research on AR on AI advice is still very ambiguous with regard to definition, measurement, and impact factors.<sup>1</sup>

First, we deal with the ambiguous concept of AR: Current research inconsistently uses the term both for a binary target state (“appropriate reliance is either achieved or not”) and a metric indicating a degree of appropriateness. To clarify this, we **introduce a two-dimensional metric**—the **appropriateness** of reliance (AoR)—to describe and measure reliance behavior. It is based on relative frequencies of correctly overriding wrong AI suggestions (self-reliance) and following correct AI suggestions (AI reliance), and reflects a metric understanding of AR. This metric can then be used to define different levels as target states of AR that mark the achievement of particular objectives like certain legal, ethical and performance requirements.

Second, we analyze how the provision of explanations influences AoR and the achievement of AR states. Existing literature is ambiguous with regard to effects of explanations [3, 6, 74]: While in some experiments, explanations support AR [74, 75], in others they cause “blind trust” [3, 6] in AI advice. To better understand and reconcile conflicting results, we consider additional constructs that may mediate the effect of explanations. More specifically, we hypothesize that explanations do not only influence the information available to the decision-maker, but also have an impact on trust toward AI and on self-confidence.

To test our hypotheses, we conduct a behavioral experiment with 200 participants. Our experiment underscores the advantages of AoR as a metric to examine in detail the factors that lead to changes in overall performance. Moreover, our results help explain the relationship between explanations and AoR by assessing the role of reliance and confidence as mediators, thus mitigating the ambiguity of previous research.

Our work provides researchers with a theoretical foundation of AR within human-AI decision-making research and provides guidance on how to design AI advisors. More specifically, our research contributes to research and practice by defining AR, developing a measurement concept (AoR), and analyzing how explainable AI

influences the AoR. Our definition should help researchers to more accurately describe whether they have achieved AR in their experiments. The AoR metric allows to precisely steer the development towards AR. Lastly, our experimental insights can be seen as a starting point for in-depth experimental evaluations of factors impacting AoR.

The remainder of this article is structured as follows: In Section 2, we first outline related work on AR in the context of human-AI decision-making. In Section 3, we define AR and develop a measurement concept, the AoR, to isolate different possible effects. Subsequently, we derive impact factors on AoR in Section 4. In Section 5, we describe the design of our behavioral experiment and summarize the results in Section 6. In Section 7, we discuss our results and provide ideas for future work. Section 8 concludes our work.

## 2 RELATED WORK

In the following, we introduce the related work of this article, structured along the topics of appropriate reliance in human advice, automation, and human-AI decision-making as well as the role of explainability.

**Appropriate reliance in human advice.** Historically, the use of (human) advice is generally discussed in the so-called judge-advisor system (JAS) research stream [26, 67, 76]. The term “judge” refers to the decision-maker who receives the advice and must decide what to do with it [8]. The judge is the person responsible for making the final decision. The “advisor” is the source of the advice [8]. The research stream mainly focuses on advice acceptance.<sup>2</sup>

**Appropriate reliance in automation.** In contrast, many researchers have worked on AR with regard to automation [45] and robotics [69]. In the following, we will provide an overview of the most common definitions. Fundamental work in the context of AR in automation has been laid by Lee and See [45]. The authors outline the relationship between “appropriate trust” and AR in their work. However, they do not define AR explicitly but provide examples of inappropriate reliance, such as “misuse and disuse are two examples of inappropriate reliance on automation that can compromise safety and profitability” [45, p. 50]. Wang et al. [72] define appropriate reliance as the impact of reliance on performance. For example, they discuss the situation in which automation reaches a reliability of 99%, and the human performance is 50%. In their opinion, it would be appropriate to always rely on automation as this would increase performance. Talone [69] follows the work by Wang et al. [72] and defines AR as “the pattern of reliance behavior(s) that is most likely to result in the best human-automation team performance” [69, p. 13]. Both see AR as a function of human-automation team performance.

**Appropriate reliance in human-AI decision-making.** Recent work in human-AI decision-making has started to discuss AR in the context of AI advice. Lai et al. [41] give an overview of empirical studies that analyze AI advice considering AR. For example, Chandrasekaran et al. [12] analyze whether humans can learn to predict the AI’s behavior. This ability is associated with an improved ability to appropriately rely on AI predictions. Moreover,

<sup>1</sup>It is worth to mention that several studies have examined AR in automation and robotics research [45, 72], but an agreed-upon definition of AR and a respective metric are still missing [41, 72].

<sup>2</sup>In this article, we use the term advice acceptance as a generic term to describe the behavior of following AI advice regardless of its quality.

Gonzalez et al. [25] measure the acceptance of incorrect and correct AI advice. Similarly, Poursabzi-Sangdeh et al. [57, p. 1] point out the idea of AR in the form of “making people more closely follow a model’s predictions when it is beneficial for them to do so or enabling them to detect when a model has made a mistake”. However, the authors do not explicitly relate this idea to the concept of AR. In this context, additional work uses the term “appropriate trust” with a similar interpretation as the behavior to follow “the fraction of tasks where participants used the model’s prediction when the model was correct and did not use the model’s prediction when the model was wrong” [74, p. 323]. Finally, also Yang et al. [75, p. 190] define “appropriate trust is to [not] follow an [in]correct recommendation”. All these articles have in common that they consider AR or appropriate trust on a case-by-case basis. Bussone et al. [10] assess how explanations impact trust and reliance on clinical decision support systems. The authors divide reliance into over- and self-reliance as part of their study. They use a qualitative approach to answer their research questions. Chiang and Yin [13] evaluate the impact of tutorials on AR and measure AR through team performance.

The related work highlights that current research inconsistently uses the term both for a binary target state (“AR is either achieved or not”) and a metric indicating a degree of appropriateness. Additionally, previous research does not provide a unified measurement concept that allows measuring the degree of appropriateness [41].

**Explainable AI and appropriate reliance.** Most researchers that studied AR so far have proposed to use explanations of AI as a means for AR [1, 43, 78]. We refer to AI that generates explanations as explainable AI (XAI). Explanations can be differentiated in terms of their scope, i.e., being global or local explanations [1]. **Global XAI** techniques address holistic explanations of the models as a whole. In contrast, **local explanations work** on an individual instance basis. Besides the scope, XAI techniques can also be differentiated with regard to being model specific or model agnostic, i.e., whether they can be used with all kinds of models [1]. The most commonly used model agnostic technique is feature importance [46, 47]. Feature importance can be used to generate saliency maps for computer vision tasks or highlight important words for text classification.

Several studies have evaluated whether different types of explanations can support humans’ understanding of the AI model with the goal of better relying on recommendations in the correct cases [3, 9, 11, 71]. However, it has also been shown that some types of explanations can lead people to rely too much on the AI’s recommendation, especially in cases where the AI advice is wrong [6, 57, 63]. In Table 1, we provide a comprehensive overview of the results that were found in the current studies on AR in XAI-assisted decision-making. Overall, we find mixed results regarding the effect of explanations.

To sum it up, related work is missing a precise definition of AR, a unified measurement concept, and a precise understanding of when and why explanations of AI advisors influence AR.

### 3 CONCEPTUALIZATION OF APPROPRIATE RELIANCE

Although several studies have examined AR in automation and robotics research [45, 72], an agreed-upon definition of AR and a

respective metric are still missing [41, 72]. We, therefore, initiate our research by deriving a definition of AR and a corresponding metric. To do so, we first discuss the terms reliance and appropriateness. Following that, we derive our metric and lastly define AR.

#### 3.1 Reliance and Appropriateness

Reliance itself is defined as a behavior [18, 45]. This means it is neither a feeling nor an attitude but the actual action conducted. This means reliance is directly observable. Scharowski et al. define reliance in the AI advisor context as “user’s behavior that follows from the advice of the system” [61, p. 3]. Defining reliance as behavior also clarifies the role of trust in this context, which is defined as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” [45, p. 51]. In general, research has shown that trust in AI increases reliance, but reliance can also take place without trust being present [45]. For example, we might not trust the bank advisor but consciously decide that following the advice is still the best possible decision. The other way around, we could also generally trust an advisor, but consciously decide that the given advice is not correct in a particular situation. Finally, reliance is influenced beyond trust by other attitudes such as perceived risk or self-confidence [59].

After establishing a common understanding of reliance, we proceed by defining “appropriateness”. Appropriateness depends on different types of AI errors. Current AI is imperfect, i.e., it may provide erroneous advice. This erroneous advice can be divided into systematic errors and random errors [69]. While humans can identify systematic errors, random errors have no identifiable patterns and can not be distinguished. These different types of errors allow differentiation between two cases. If all errors are random and cannot be detected, then, from a performance perspective, humans should always rely on the AI’s advice when the AI performs better on average, and never when the AI performs worse on average [69]. However, suppose there are some systematic errors. In that case, humans might be able to differentiate between correct and incorrect advice, which may even result in superior performance, i.e. complementary team performance (CTP) [29], compared to a scenario in which AI and humans conduct the task independently of each other.

This changes the overall discrimination to a case-by-case discrimination. In the presence of systematic errors, humans should evaluate each case individually. Since the solution approach in the presence of just random errors is relatively simple, as pointed out above, in this article, we focus on the more complicated setting when a significant proportion of task instances exhibit systematic errors. After having defined the terms reliance and appropriateness in the following, we derive a metric.

#### 3.2 Towards a Measurement Concept—Appropriateness of Reliance

Appropriateness is often measured by the percentage in which the decision-maker relies on correct AI advice and the percentage in which the decision-maker does not rely on incorrect AI advice [6, 25]. The major disadvantage of this measurement is that we cannot know whether the reliance on correct AI advice stems from a correct discrimination or simply an overlap of the human and

**Table 1: Related work on explainable AI (XAI) and Appropriate Reliance (AR).**

Study	AR Metric	Independent variable	Effect of XAI on AR
[6]	Accuracy on correct or incorrect AI advice	Local feature importance Adaptive explanations based on AI confidence	Negative Positive
[9]	Ratio of reliance on incorrect AI advice	Local feature importance	Negative
[32]	Ratio of reliance on correct or incorrect AI advice	Local feature importance Predictive outcomes	No effect Negative
[74]	Accuracy on correct or incorrect AI advice	Global feature importance Local feature importance Examples Counterfactuals	No effect Positive No effect No effect
[75]	Ratio of reliance on correct or incorrect AI advice	Local feature importance	Positive

the AI’s decision, i.e. instances where the AI advisor just confirms the human [70]. Especially from an ethical point of view, it makes a major difference whether the final decision is incorrect because an AI advisor “convinced” a human decision-maker to accept an incorrect AI advice or whether the human decision-maker would also not have been competent to solve it alone.

Therefore, to measure the degree of appropriateness in a more narrow sense, we follow the approach of the judge-advisor system (JAS) paradigm and include an initial human decision [67]. This approach requires participants to make a decision, receive advice, and then make a second, potentially revised decision. In general, if we do not consider the initial human decision, information about the human discrimination ability, including the consequent action, gets lost—it is not traceable how the human would have decided without the AI advice. Nevertheless, especially this interaction needs to be documented to research AR holistically.

We use a simple discrete decision case to highlight the different possible outcomes of reliance. Note that for simplicity, we refer to classification problems. However, the measurement concept can be extended to regression problems as well (see for example Petropoulos et al. [56]). We focus on a single task instance perspective and consider a sequential decision process which can be described as follows: First, the human makes a decision, then receives AI advice. Second, the human is asked to update the initial decision, i.e., either adopt or overwrite the AI advice. This allows measuring appropriateness in a fine-granular way. Figure 1 highlights the different combinations. Four of the eight combinations are cases where the human’s initial decision and the AI’s advice are the same, i.e. the AI confirms the human’s decision. In our reliance measure, we exclude these confirmation cases for two reasons: First, if the same decision is made in all three steps, it is impossible to objectively measure whether AI or self-reliance was present. Second, if the final decision differs from the advice and the initial human decision, it is questionable whether we can speak of a reliance outcome. For example, if both the human and the AI are initially incorrect, then arriving at a correct final decision is less a matter of reliance than of human-AI collaboration. While these cases of collaboration are relevant to CTP, they are beyond the scope of our work. Therefore,

we arrive at four different reliance outcomes, which we present below.

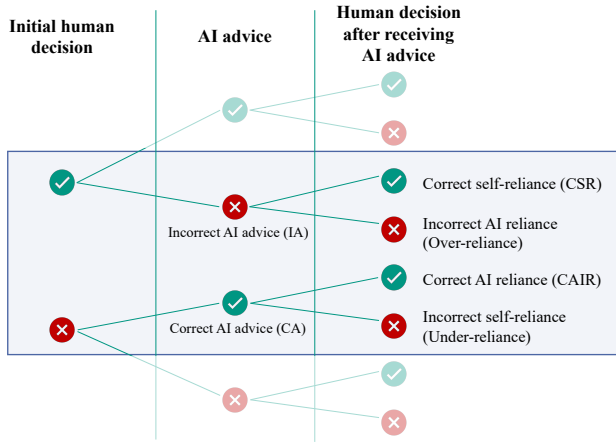
On a high level, we can cluster those four outcomes into either AI or self-reliance. *AI reliance* refers to cases in which the initial decision-maker’s decision is different from the AI advisor and the decision-maker relies on the AI’s advice. Likewise, *self-reliance* refers to cases in which the decision-maker is different from the AI advisor but finally relies on themselves. On a more detailed level, we can further differentiate whether the final decision is correct or incorrect which leaves us with the following four reliance outcomes: First, *correct AI reliance (CAIR)*, which describes the case when the human is initially incorrect, receives correct advice, and relies on that advice. Second, the case in which the human relies on the initial incorrect decision and neglects correct AI advice. This is denoted as *incorrect self-reliance* or under-reliance. Third, if the human is initially correct and receives incorrect advice, this can either result in *correct self-reliance (CSR)*, i.e., neglecting the incorrect AI advice, or relying on it, which is denoted as *incorrect AI reliance* or over-reliance.<sup>3</sup> Based on these cases, we propose a two-dimensional metric that transfers the instance perspective on a measurement for multiple task instances.

Let us consider a prediction task  $T = \{X_i, y_i\}_i^N$  as a set of  $N$  instances  $x_i \in X$  with a corresponding ground truth label  $y_i \in Y$ . On the first dimension, we calculate the ratio of the number of cases in which humans rely on correct AI advice and the decision was initially not correct, i.e., in which humans rightfully change their mind to follow the correct advice.

$$Relative\ AI\ reliance\ (RAIR) = \frac{\sum_{i=0}^N CAIR_i}{\sum_{i=0}^N CA_i} \quad (1)$$

$CAIR_i$  is one if, in this particular case, the original human decision was wrong, the AI recommendation was correct and the human decision after receiving the AI recommendation is correct, and zero otherwise.  $CA_i$  is one if the original human decision was wrong and the AI advice was correct, regardless of the final human

<sup>3</sup>This also clarifies our definition of over- and under-reliance. Both are errors on a task instance level, when humans do not rely appropriately.



**Figure 1: Combinatorics of initial human decisions, AI advice and human reliance for a single task instance in a sequential task setting.**

decision, and zero otherwise. On the second dimension, we propose to measure the relative amount of correct self-reliance in the presence of incorrect AI advice.

$$\text{Relative self-reliance (RSR)} = \frac{\sum_{i=0}^N \text{CSR}_i}{\sum_{i=0}^N \text{IA}_i} \quad (2)$$

$\text{CSR}_i$  is one if on this particular instance the initial human decision was correct, the AI advice was incorrect and the human decision after receiving AI advice is correct.  $\text{IA}_i$  is one, if the initial human decision for a task instance  $i$  was correct and the AI advice was incorrect.

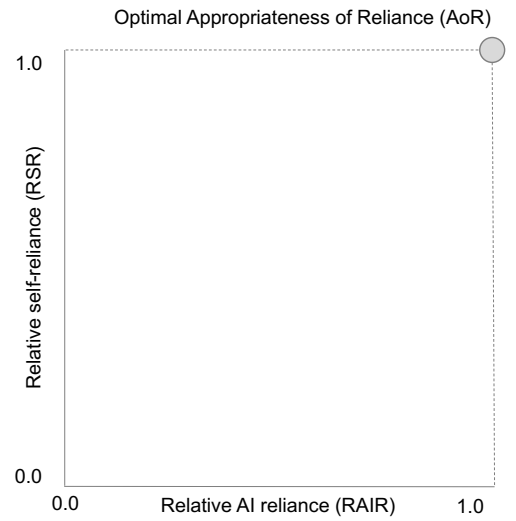
Figure 2 highlights both dimensions. On the x-axis, we depict the relative AI reliance ( $\text{RAIR}$ ), and on the y-axis, the relative self-reliance ( $\text{RSR}$ ). The figure highlights the properties of the measurement concept. It ranges between 0 and 1 along both dimensions. We call the tuple of  $\text{RAIR}$  and  $\text{RSR}$  *Appropriateness of Reliance (AoR)*.

$$\text{Appropriateness of Reliance (AoR)} = (\text{RSR}; \text{RAIR}) \quad (3)$$

We refer to the theoretical goal of having a  $\text{RSR}$  and a  $\text{RAIR}$  metric of “1” as *optimal AoR*. Most likely, this theoretical goal will not be reached in any practical context as humans will not always be able to perfectly discriminate on a case-by-case basis whether they should rely on AI advice. Furthermore, random errors will reduce AoR as they cannot be discriminated against. Therefore, optimal AoR will most likely be a theoretical goal.

### 3.3 Definition of Appropriate Reliance

So far, we have discussed how to measure AoR and the theoretical upper boundaries of AoR. The challenge, however, is to define the level of  $\text{RAIR}$  and  $\text{RSR}$  that constitutes AR. In our work, we take an objective-oriented perspective and propose to define AR individually in the context of the task. Dependent on the task, different levels of  $\text{RAIR}$  and  $\text{RSR}$  might be appropriate. In this work,



**Figure 2: Two-dimensional depiction of the Appropriateness of Reliance (AoR).**

we focus on AR with respect to performance ( $P$ ) following Talone [69], Wang and Yin [74].<sup>4</sup> Thus, we define AR with regard to CTP:

$$\text{Appropriate Reliance (AR)} = \begin{cases} 1, & \text{if } P_{H\&AI} > \max(P_H, P_{AI}) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

With  $P_{H\&AI}$  being the performance after receiving AI advice,  $P_H$  the individual human performance and  $P_{AI}$  the individual AI performance. Essentially, this means any tuple of  $\text{RAIR}$  and  $\text{RSR}$  that leads to CTP is considered AR.

In summary, we derived a metric (AoR) and defined AR. In the next section, we will use this foundation to derive a research model to analyze the impact of AI recommendation explanations on AoR.

## 4 THEORY DEVELOPMENT AND HYPOTHESES

With the measurement concept (AoR) at hand, we now develop hypotheses on the effect of explanations on AoR. Research has already investigated in-depth the effect of explanations on AI advice acceptance [65]. However, research is missing theoretical and empirical evidence on how explanations influence AoR [5]. Thus, our work contributes a research model that is evaluated applying the AoR concept.

As a dependent variable, we use the before-defined two dimensions of AoR— namely  $\text{RAIR}$  and  $\text{RSR}$ . We believe that both dimensions need to be treated differently as there are inherent differences between the  $\text{RAIR}$  and the  $\text{RSR}$ . The  $\text{RAIR}$  essentially deals with cases where the human is initially incorrect, gets correct advice, and relies on this advice. In contrast,  $\text{RSR}$  focuses on cases where

<sup>4</sup>Note that besides this performance perspective, appropriateness could also be discussed from an ethical perspective. Since in high-stake decision-making humans have an oversight responsibility [64].

the human is initially correct and receives incorrect advice, which is then rightfully ignored. Thus, we formulate different hypotheses for the *RAIR* and the *RSR*.

The most central impact factor in the assessment of the AI advisor might be the explanations of its recommendations. These explanations should provide insights into the AI’s thought process. In the presence of incorrect advice, explanations might enable the human to detect whether the advice is incorrect, for example, by validating if the AI advisor violates some universal axioms of the task. Similarly, [6] hypothesize that if explanations do not “make sense”, humans will reject the AI advice. However, on the other hand, sometimes explanations are rather interpreted as a general sign of competence and thereby increase over-reliance [9]. Which effect exceeds the other is unclear. Therefore, we hypothesize, without specifying a direction of the effect, that the explanations have a general effect on the *RSR*:

**H1a:** *Providing explanations of the AI advisor influences the relative self-reliance (RSR).*

The second effect of explanations on AoR is through the *RAIR*. Essentially, the *RAIR* measures the percentage of times decision-makers follow the correct advice after initially being wrong about the task instance. This means they do not have enough domain knowledge to solve the task on their own. Thus, to increase the *RAIR*, human decision-makers need to extend their knowledge and simultaneously validate whether this knowledge extension makes sense. Here, explanations are needed to first get inspired to derive new knowledge and second to validate the knowledge. Figure 3 shows an illustrative example based on an animal classification task. Imagine that a child has just seen big dogs and then sees a very small dog. It might think that this animal is something else, like a rat. Next, the child receives AI advice that says the animal it sees is a dog, and provides additional justification by highlighting the part of the image that led to the AI’s decision. Now, the child’s first task is to figure out whether this advice makes sense in general, while building the knowledge base. In this illustrative case, it might understand that the animal has characteristics of a dog, but is only smaller and therefore relies on the AI, thereby increasing the *RAIR*. In the presence of correct advice, explanations might point humans towards new patterns they have not seen before and help discriminate these knowledge extensions. In the presence of correct advice, the convincing element of explanations would not have a negative effect as a higher overall reliance on the AI advice would simply increase the *RAIR*. We therefore hypothesize:

**H1b:** *Providing explanations of the AI advisor increases the relative AI reliance (RAIR).*

Beyond the provisioning of additional information, explanations might change the attitude toward the AI advisor. In 1992, Lee and Moray [44], already discussed the influence of self-confidence and trust as predominant attitudes for reliance decisions in the context of automation. Therefore, in the following, we discuss potential impacts on trust and the change in self-confidence induced through explanations and their impact on AoR.

Confidence is defined as a person’s degree of belief that their own decision is correct [55, 77]. Confidence in one’s own decision is a key mechanism underlying advice acceptance [14, 73]. So far, most research has discussed the influence of static human self-confidence on AoR, i.e. a confidence in doing the task instance without any



**Figure 3: Illustrative example of how humans can increase their relative AI reliance (RAIR) based on explanations of the AI advice.**

advisor [14]. However, we hypothesize that the absolute level of human confidence actually plays a minor role in comparison to the change of confidence after seeing the AI advice as it essentially reflects the combination of a self-assessment and the AI advisor’s assessment. Explanations should not influence the initial human self-confidence but the confidence level after seeing the AI advice. Thus, we hypothesize:

**H2** *Providing explanations of the AI advisors increases the change in self-confidence.*

We hypothesize that this change in self-confidence positively correlates with the discrimination capability and thus should increase the *RSR* and *RAIR*.

**H3a** *An increased change in human self-confidence increases the relative self-reliance (RSR).*

**H3b:** *An increased change in human self-confidence increases the relative AI reliance (RAIR).*

We hypothesize that also trust in the AI advisor influences AoR. There are different levels of trust, e.g. trust in AI in general, trust in a specific AI advisor [31] and some researchers even refer to the case-by-case discrimination as trust on a task instance level [74]. In this work, we focus on the specific trust in our developed AI advisor. In general, trust is a complex, multidisciplinary construct with roots in diverse fields such as psychology, management and information systems [51]. In our study, we define trust as a belief in the integrity, benevolence, trustworthiness, and predictability of the AI advisor following [15, 17, 21, 49]. Understanding is crucial in building trust [23]. Psychological research shows that in general explanations of humans increase trust [36]. Thus, we hypothesize:

**H4** *Providing explanations of the AI advisor increases trust in the AI advisor.*

Trust influences reliance, but does not fully determine it [45]. When people show a high level of trust in the advisor, they consider the advice to be high-quality advice from an advisor with good intentions, and they will give more weight to that advice [73]. In general, this should increase the acceptance of AI consulting. For systems’ most effective use, however, users must appropriately trust AI advisors [45]. Trust should be calibrated and match the AI’s capabilities [45]. Insufficient trust is called distrust and when trust exceeds capability it is called over-trust [45]. Research on automation has shown that over-trust can result in over-reliance

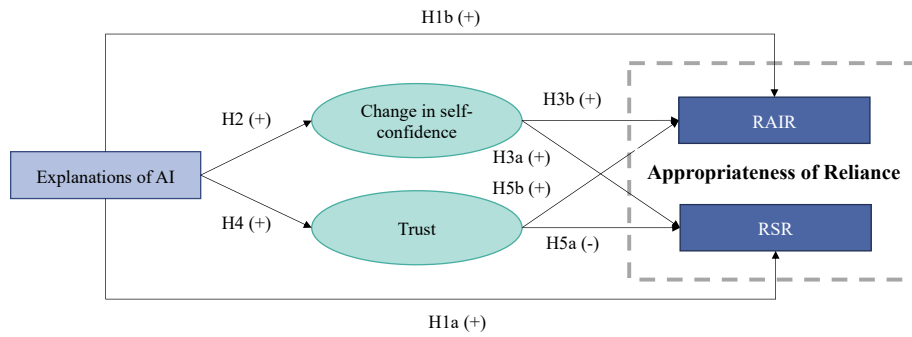


Figure 4: Research model on the effect of explanations on Appropriateness of Reliance (AoR).

on automation [4, 24, 54] and, therefore, should decrease the *RSR*. Thus, we hypothesize:

**H5a:** *Trust decreases the relative self-reliance (RSR).*

Since trust increases reliance it should also increase the *RAIR*. Therefore, we hypothesize:

**H5b:** *Trust increases the relative AI reliance (RAIR).*

Figure 4 highlights all our hypotheses and combines them into one integrated research model.

## 5 EXPERIMENTAL DESIGN

In this section, we present our study task, the AI model, and the corresponding explanations. We then explain the study procedure and measurements.

### 5.1 Task, Model, and Explanations

As an experimental task, we have chosen a deceptive hotel review classification. Humans have to differentiate whether a given hotel review is deceptive or genuine. Ott et al. [52, 53] provide the research community with a data set of 400 deceptive and 400 genuine hotel reviews. The deceptive ones were created by crowd-workers, resulting in corresponding ground truth labels.

The implemented AI advisor is based on a Support Vector Machine with an accuracy of 86%, which is a performance that is similar to the performance in related literature [42]. For the explanations, we use a state-of-the-art explanation technique, LIME feature importance explanations [58], as it is the most common one for textual data. Feature importance aims to explain the influence of an independent variable on the AI’s decision in the form of a numerical value. Since we deal with textual data, a common technique to display the values is to highlight the respective words according to their computed influence on the AI’s decision [42]. We additionally provide information on the direction of the effect and differentiate the values into three effect sizes following the implementation of Lai et al. [42] (see step 2 in Figure 5).

### 5.2 Study Design and Procedure

The research model is tested in an online experiment with a between-subject design. We tested two different conditions. First, a *control* condition in which the human receives AI advice without feature importance explanations and second, a *feature importance* condition.

In each condition, participants are provided with 16 reviews. We incorporate an advanced sampling strategy to isolate the effects of discriminating AI advice. We have a test set of 32 reviews to which we apply stratified sampling and select four reviews of each class of a confusion matrix (True Positive, False Positive, True Negative, False Negative), two with a positive sentiment and two with a negative sentiment. This approach allows us to ensure a high-performing AI that should provide good explanations but also the potential for incorrect AI advice.

**Task flow.** The online experiment is initiated with an attention control question that asks participants to state the color of grass. To control for internal validity, participants are randomly assigned to the condition groups. Then, both condition groups receive an introduction to the task and either AI alone or AI including feature importance. We provide the participants with a general intuition of the AI but not with specific performance information. Then, the participants conduct two training tasks, to familiarize the participants with the task and the AI and, depending on the condition, with its explanations. Additionally, the participants receive feedback on the training tasks. After the two training reviews, the participants are provided with the 16 main tasks. For the AoR measurement concept, sequential task processing is essential. In our study, this means the human first receives a review without any AI advice, i.e., just plain text, and classifies whether the review is deceptive or genuine (see step 1 in Figure 5). Then the participant is asked to classify the review and provide a confidence rating. Following that, the human either receives a simple AI advice statement, e.g. “the AI predicts that the review is fake” or the AI advice and additional explanations (see step 2 in Figure 5). After receiving the AI advice the participant is able to change the initial decision and provide a new self-confidence assessment. This sequential two-step decision-making allows us to measure AoR. During the main tasks, the participants do not receive feedback on their performance. After classifying the hotel reviews, we collect data on trust and demographic variables.

**Reward.** To incentivize the participants, they were informed that for every correct decision, they get an additional 12 Cents in addition to a base payment of 5.83 Euro. Hereby, the two training classifications do not count for the final evaluation.

**Participant information.** The participants are recruited using the platform “Prolific.co”. We note that crowd workers might limit

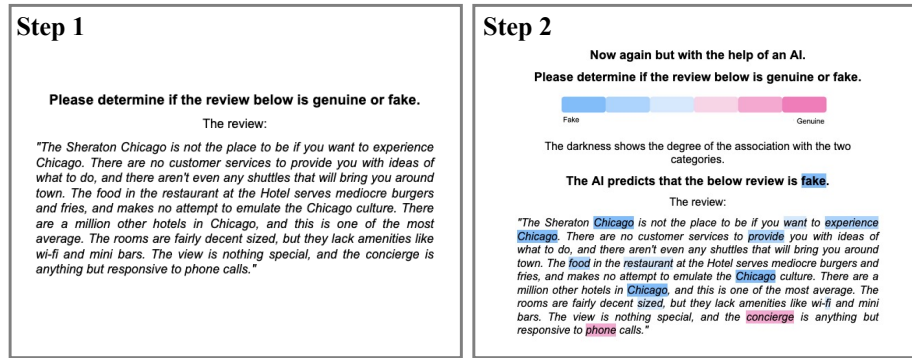


Figure 5: Online experiment graphical user interface for the feature importance condition. The ground truth of the exemplarily shown hotel review is “fake”. The design of the interface is adapted from Lai et al. [42].

Table 2: Summary of Participants’ Characteristics.

Number per condition	Control = 100 Feature importance = 99
Age	$\mu = 27.5, \sigma = 8.5$
Gender	46 % Male 54 % Female
Education	32 % High school 38 % Bachelor 14 % Master 16 % Other

the generalizability of our results. However, deception detection of digital information is often done in online communities. Future work could analyze the effects in professional deception detection screening services. In total, we conducted the experiment with 200 participants. We excluded one participant in the feature importance condition because of a failed manipulation check. Table 2 shows the age, gender, and education distribution of the participants.

### 5.3 Evaluation Measures

To measure AoR, we use the upfront derived measurements *RAIR* and *RSR*. We measure the *change in self-confidence* for all task instances per participant:

$$\text{Change in self-confidence} = \sum_{i=1}^{16} \text{Conf}_{H_2, i} - \text{Conf}_{H, i} \quad (5)$$

$\text{Conf}_H$  refers to the human self-confidence when doing the task instance alone and  $\text{Conf}_{H_2}$  to the human self-confidence after receiving AI advice. Both are measured with a 7-point Likert scale (“How confident do you feel in your decision?”). We measure the trust in the AI advisor as a subjective latent construct with a 7-point Likert scale. We use four items based on [15, 17, 21, 22]. The items were: “I think I can trust the AI”, “The AI can be trusted to provide reliable support”, “I trust the AI to keep my best interests in mind.” and “In my opinion, the AI is trustworthy.”. Cronbach’s

Alpha was 0.89 (high). The original scales were validated. As both classes are equally distributed, task performance was measured by the percentage of correctly classified images, i.e., accuracy. To measure the *human accuracy*, we calculated this measure for both conditions based on the initial human decision across all 16 task instances. Furthermore, we calculate the *AI-assisted accuracy* based on the revised human decision after receiving AI advice.

## 6 RESULTS

In the following, we present the results of our behavioral experiment. We start by presenting descriptive results, followed by the results with respect to AoR and AR. Following that, we analyze our full research model, including mediations, by applying Structural Equation Modelling (SEM).

### 6.1 Descriptive analysis

Descriptive results of our study can be found in Table 3. They are split according to the experimental condition. We evaluate the significance of the results using t-tests after controlling for normality. The participants’ *RAIR* is significantly higher in the explanation group compared to the control group ( $t = -1.95, p = 0.05$ ). The change in confidence is statistically significant ( $t = -2.33, p = 0.02$ ) which means that on average people feel more confident after receiving AI advice including explanations. Neither AI-assisted nor human accuracy is significantly different between conditions. However, the difference between the human and AI-assisted performance is significant (feature importance condition = 2.45 pp; control condition = -1.56 pp ;  $t = 2.29, p = 0.02$ ) which means that explanations not only improve the *RAIR* but also as a consequence the overall performance.

### 6.2 Appropriateness of Reliance & Appropriate Reliance

We depict our AoR results of the experiment in Figure 6. They highlight in the control condition a high *RSR* of 71.87% ( $\pm 3pp$ ) and a relatively low *RAIR* of 29.59% ( $\pm 3pp$ ). This indicates that humans in the setting were able to differentiate between wrong AI advice and self-rely to a high degree. The *RAIR* of 29.59% shows that we can observe a severe share of under-reliance on AI.

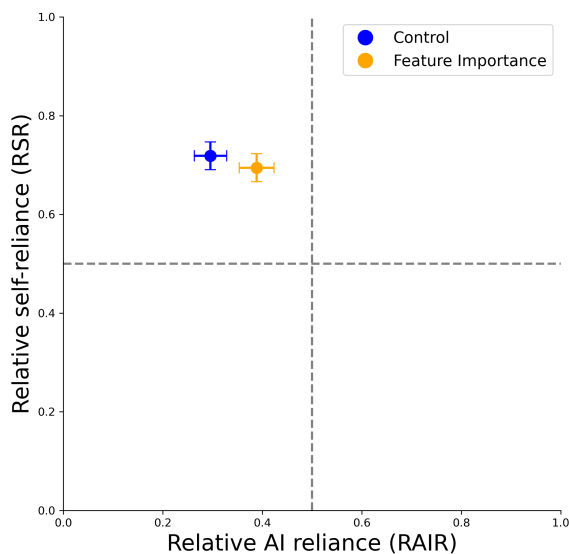


**Table 3: Descriptive outcomes.**

Condition	<i>RAIR</i>	<i>RSR</i>	Trust (SD)	Change in Self-Confidence (SD)	AI-assisted Accuracy	Human Accuracy
Control	29.59 %	71.87%	4.45 (1.17)	0.19 (0.42)	53.94 %	55.50 %
Feature Importance	38.87%	69.45 %	4.4 (1.18)	0.06 (0.35)	56.30 %	53.85 %

In the XAI condition, we can observe a significant increase ( $t = -1.95, p = 0.05$ ) in *RAIR* from 29.59% ( $\pm 3pp$ ) to 38.87% ( $\pm 3pp$ ) while the *RSR* does not change significantly ( $71.87\% \pm 3pp$  for the control condition and  $69.45\% \pm 3pp$  for the feature importance condition,  $t = 0.61, p = 0.54$ ). This means explanations of AI decisions can reduce the share of under-reliance. It is important to highlight that *RAIR* is not increased simply by relying more often on AI advice, as this would have also reduced the *RSR* significantly. Thus, our experiment indicates that feature importance on textual data can have a positive effect on human-AI decision-making.

Following our AR definition, to evaluate whether the participants display AR, we need to calculate whether we reached CTP. Therefore, we compare the individual human and AI performance with the human-AI team performance. The down-sampled AI performance is 50 % for both conditions. The human accuracy varies depending on the condition. The human-AI team performance is not significantly different from the human accuracy which means we do not reach CTP and therefore AR is not displayed. Further means would be necessary to reach AR.



**Figure 6: Illustration of Appropriateness of Reliance (AoR) including standard errors. Explanations increase the *RAIR* significantly. Differences in *RSR* are not significant.**

### 6.3 Structural Equation Model

In addition to analyzing the direct effect of explanations on the *RAIR* and *RSR*, we use SEM analysis to test our hypothesized research model. Before fitting our SEM, we conducted missing data analysis, outlier detection, a test for normality, and the selection of an appropriate estimator. We observe no missing data and no outliers. However, one participant failed our attention check leaving us with a final sample size of 199 for both conditions. Shapiro’s test for normality indicates that several variables of interest deviate significantly from normal distributions. As a result, we conducted the analysis with an estimator that allows for robust standard errors and scaled test statistics [38]. Therefore, we use the MLR estimator [39].

Our dependent variables are the *RAIR* and *RSR*. Since these dependent variables are between 0 and 1, we employed a logistic model in the Lavaan package, version 0.6-9, in R [60]. This model has an excellent overall fit (see Table 4). The results for each independent variable are discussed below and visualized in Figure 7.

We do not find any significant relationship between providing explanations of the AI advisor and the *RSR* (**H1a**). However, also in the SEM the effect of explanations on the *RAIR* is significant (**H1b**). Additionally, we find a significant positive relationship between providing explanations of the AI advisor and the change in self-confidence which confirms **H2**. We also observe a positive correlation between the change in self-confidence and the *RAIR* (**H3b**). Since the absolute strength of the relationship also decreased with including the change in confidence, we can conclude that the positive effect of explanations on the *RAIR* is partially mediated by the change in confidence. Interestingly, we do not find a relationship between the change in self-confidence and the *RSR* (**H3a**). Additionally, we find no effect of our explanations on trust (**H4**). However, both **H5a** and **H5b** are confirmed, i.e. increasing trust increases *RAIR* but also decreases *RSR*. We display a summary of the hypotheses results in Table 5. In the following Section, we discuss our results.

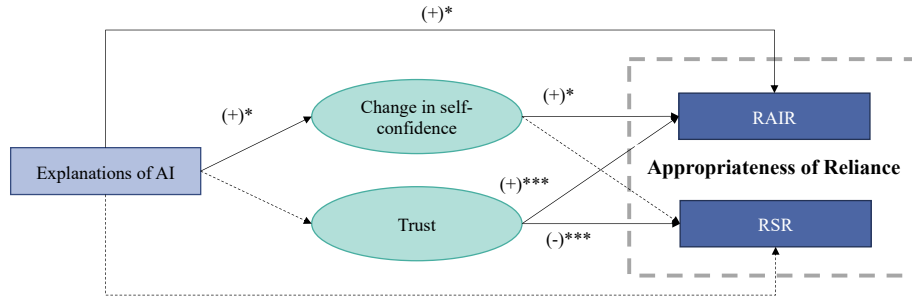
## 7 DISCUSSION

In this article, we first defined AR and conceptualized a measurement concept (AoR). Following that, we derived a research model regarding the impact of explanations on AoR which we subsequently tested on a deception detection task. Though conducted in a limited scope, our findings should help to guide future work on AR.

**Theoretical foundation of appropriate reliance.** The main contribution of our work is the theoretical development of AR. So far, terms like “appropriate trust”, “calibrated trust” and AR

**Table 4: Structural equation model fitting index using a chi-squared test ( $\chi^2$ ), root mean square error of approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Standardized Root Mean Squared Residual (SRMR).**

	$\chi^2$	RMSEA	CFI	TLI	SRMR
Measurement criteria	> 0.05 [7]	< 0.05 [30]	> 0.96 [30]	> 0.95 [30]	< 0.08 [30]
Value	0.083	0	1	1.02	0.02



**Figure 7: Structural equation modeling results. Significance: \*\*\* p < 0.01, \*\* p < 0.05, \* p < 0.1.**

**Table 5: Analysis results of the structural equation model (\*\*\* p < .01, \*\* p < .05, \* p < 0.1)**

X	Y	z-value	Standardized regression coefficient	Result
H1a: Explanations ->	<i>RSR</i>	0.04	-0.03	not supported
H1b: Explanations ->	<i>RAIR</i>	1.73	0.08 *	supported
H2: Explanations ->	Change in self- confidence	2.30	0.13 *	supported
H3a: Change in confidence ->	<i>RSR</i>	0.02	0.00	not supported
H3b: Change in confidence ->	<i>RAIR</i>	1.84	0.11 *	supported
H4: Explanations ->	Trust	-0.21	-0.04	not supported
H5a: Trust ->	<i>RSR</i>	-3.08	-0.05 ***	supported
H5b: Trust ->	<i>RAIR</i>	3.19	0.06 ***	supported

were often used interchangeably in prior research. We provide clarity by defining AR and putting the terms in perspective. Second, we derive a granular, two-dimensional measurement concept—Appropriateness of Reliance (AoR). Most prior work neglected the initial human decision that would have been made without any AI advice.<sup>5</sup> Taking this initial human decision into account allows to differentiate between the effects of advice (correct and incorrect) and confirmation. Without the human initial decision, we cannot say for sure whether a final wrong human decision is due to over-reliance on a wrong AI advice or due to the human and the AI being wrong.

**Implications for appropriate reliance and explainable AI.** We further have investigated the effect of AI explanations on AoR. We confirm the results of prior research [6, 25, 74] by finding an effect of explanations on the *RAIR*. We believe that the reason for this could be that the AI’s explanations increase people’s knowledge of the task [25, 68]. Maybe in such cases, the human should be seen less as a “judge” but instead more as a student of the AI. On the

other hand, our results show that explanations do not influence the *RSR*. While this may sound disappointing at first, it also shows that the claim that explanations would reduce overreliance [6, 9] does not seem to hold for all kinds of tasks.<sup>6</sup> It further suggests that new techniques must be developed to distinguish incorrect AI advice. Moreover, our study is the first one that analyzed mediators of the effect of explanations on AR. Interestingly, in our study, we find no effect of explanations on trust. However, prior research has shown that it depends on a lot of confounding factors. We find significant effects of trust on *RAIR* as well as *RSR*. Additionally, we show that the effect of explanations on *RAIR* partially depends on the change in confidence after receiving AI advice.

**Appropriate reliance and complementary team performance (CTP).** Lastly, we want to elaborate on the relationship between AoR and CTP. To reach CTP, the task needs to have instances where the AI is better than the human and vice versa, i.e. a certain amount of complementarity potential needs to be present

<sup>5</sup>The one exception is the work of [9] who measure over-reliance in the same way that we do but do not consider under-reliance.

<sup>6</sup>An earlier study by this research team [63] found initial signs of a reduced *RSR* in a pretest. However, this study with more participants shows that the effect is not significant in a larger sample.

[28]. CTP essentially depends on the relationship between  $RSR$  and  $RAIR$  and this complementarity potential.

The human impact on CTP is given by multiplying the  $RSR$  and the share of incorrect AI advice. However, simultaneously involving a human might reduce the AI performance ( $1 - RAIR$  multiplied by the share of correct advice). That means to reach CTP, the gain through human involvement needs to be larger than the loss through discounting correct AI advice or more formally<sup>7</sup>:

$$CTP = \begin{cases} 1, & \text{if } RSR * IA > (1 - RAIR) * CA \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Here  $IA$  is the total number of task instances in a test set where the human is initially correct and receives incorrect advice.  $CA$  refers to the number of task instances where the human is initially incorrect and receives correct advice. If this condition is not fulfilled, AR depends on the relationship between human and AI performance. If the human performs worse than the AI advisor and has a low  $RSR$  and  $RAIR$ , one should always favor AI advice. If the human decision-maker performs better on average than the AI, one can argue that from a performance perspective the AI should not be used.

**Limitations.** No research is without limitations. We would like to emphasize that deception detection is a difficult task for humans [42, 43]. Humans on average just perform slightly better than by chance on this particular hotel review task [42]. This makes AR difficult as the task of discrimination requires human domain knowledge. Additionally, the generalizability of our experimental findings is limited due to the choice of explanations. However, we have deliberately chosen the most modern form of explanation to maximize the impact of our results. Our concept is limited to classification tasks but will be extended in future work. First approaches can be found in the work of Petropoulos et al. [56].

Furthermore, the sequential task setup necessary for our measurement concept has some disadvantages as it changes the task itself. Since conducting the same task initially alone before receiving AI advice, the human is already mentally prepared and might react differently than after directly receiving AI advice. More specifically, research has shown that letting humans conduct the task alone before receiving AI advice might reduce over-reliance [9]. The sequential task setup could induce an anchoring effect which prevents the human to more actively take the AI into account [9]. This could have led to the overall low  $RAIR$  in our experiment. Moreover, sequentially conducted tasks with AI advice might not always be possible or desired in real-world settings. Therefore, the measurement should be seen as an approximation of real human behavior. Instead of having a sequential task setup, one alternative option could be to simulate a human model based on a data set of task instances solved by humans without AI advice. This simulation model could approximate the initial human decision within a non-sequential task setting. However, this approach is also an approximation of real human behavior. In other work, a latent construct has been derived to measure reliance behaviour [70]. Future work should compare the approaches.

<sup>7</sup>Provided that nothing changes in the cases where the initial decision of the human and the AI advice are the same.

**Future Work.** Essentially, the improvement in  $RAIR$  depends on the knowledge gain of the human. Future work could therefore extend our research model by adding newly learned knowledge as a mediator. Empirically, this could be measured by asking humans before collaborating with AI to do a couple of task instances on their own and afterwards [68]. The performance improvement can be interpreted as learned knowledge.

Most importantly, future research needs to investigate the impact of different design features of AR. We initiate our research with state-of-the-art feature importance but many other ones can be thought of, e.g. counterfactuals, global explanations. Future research should evaluate these potential design features to provide practitioners with a toolkit for effective use of AI.

## 8 CONCLUSION

Appropriate reliance in AI advice is the next milestone after a decade of research focused on AI adoption and acceptance. Nowadays, many AI applications are deployed and used on a daily basis. While adoption and acceptance remain important, we argue that a perspective shift is necessary. In the use phase of AI, researchers need to find ways to ensure appropriate reliance and, thus, effective use of AI. In this article, we provide guidance for future research on appropriate reliance by providing a definition and a measurement concept—Appropriateness of Reliance (AoR). Furthermore, we generate initial insights how explanations influence the Appropriateness of Reliance. We hope that our research will inspire researchers and practitioners for future research on appropriate reliance, resulting in effective human-AI collaboration.

## REFERENCES

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.
- [2] Majdi Alnowami, Fouad Abolaban, Hussam Hijazi, and Andrew Nisbet. 2022. Regression Analysis of Rectal Cancer and Possible Application of Artificial Intelligence (AI) Utilization in Radiotherapy. *Applied Sciences* 12, 2 (2022), 725.
- [3] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does explainable artificial intelligence improve human decision-making?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6618–6626.
- [4] Nathan R Bailey and Mark W Scerbo. 2007. Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science* 8, 4 (2007), 321–348.
- [5] Gagan Bansal, Alison Marie Smith-Renner, Zana Bućinca, Tongshuang Wu, Kenneth Holstein, Jessica Hullman, and Simone Stumpf. 2022. Workshop on Trust and Reliance in AI-Human Teams (TRAIT). In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.
- [6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [7] Peter M Bentler. 1990. Comparative fit indexes in structural models. *Psychological bulletin* 107, 2 (1990), 238.
- [8] Silvia Bonaccio and Reeshad S Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101, 2 (2006), 127–151.
- [9] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
- [10] Adrian Bussone, Simone Stumpf, and Dymna M. O’Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics* (2015), 160–169.
- [11] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don’t Help People Detect Misclassifications of Online Toxicity. *Aaai.Org* 2020

- (2020). Issue Icwsm. [www.aaai.org](http://www.aaai.org)
- [12] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *EMNLP*.
  - [13] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.
  - [14] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2022), 107018.
  - [15] Lawrence A Crosby, Kenneth R Evans, and Deborah Cowles. 1990. Relationship quality in services selling: an interpersonal influence perspective. *Journal of marketing* 54, 3 (1990), 68–81.
  - [16] Lara Marie Demajo, Vince Vella, and Alexiei Dingli. 2020. Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749* (2020).
  - [17] Patricia M Doney and Joseph P Cannon. 1997. An examination of the nature of trust in buyer–seller relationships. *Journal of marketing* 61, 2 (1997), 35–51.
  - [18] 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.
  - [19] Carl Benedikt Frey and Michael A Osborne. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change* 114 (2017), 254–280.
  - [20] Andreas Fügner, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2021. Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *Management Information Systems Quarterly (MISQ)-Vol 45* (2021).
  - [21] Shankar Ganesan. 1994. Determinants of long-term orientation in buyer-seller relationships. *Journal of marketing* 58, 2 (1994), 1–19.
  - [22] David Gefen, Elena Karahanna, and Detmar W Straub. 2003. Trust and TAM in online shopping: An integrated model. *MIS quarterly* (2003), 51–90.
  - [23] Leilani H Gilpin, Cecilia Testart, Nathaniel Fruchter, and Julius Adebayo. 2019. Explaining explanations to society. *arXiv preprint arXiv:1901.06560* (2019).
  - [24] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.
  - [25] Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer. 2020. Human evaluation of spoken vs. visual explanations for open-domain qa. *arXiv preprint arXiv:2012.15075* (2020).
  - [26] Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes* 70, 2 (1997), 117–133.
  - [27] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. 2022. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. 2478.
  - [28] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2022. On the Effect of Information Asymmetry in Human-AI Teams. *arXiv preprint arXiv:2205.01467* (2022).
  - [29] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS 2021 Proceedings* (2021).
  - [30] Li-tze Hu and Peter M Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal* 6, 1 (1999), 1–55.
  - [31] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.
  - [32] Johannes Jakubik, Jakob Schöffner, Vincent Hoge, Michael Vössing, and Niklas Kühl. 2023. An Empirical Evaluation of Predicted Outcomes as Explanations in Human-AI Decision-Making. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part I*. Springer, 353–368.
  - [33] Dominik Jung, Verena Dörner, Florian Glaser, and Stefan Morana. 2018. Robo-advisory. *Business & Information Systems Engineering* 60, 1 (2018), 81–86.
  - [34] Ekaterina Jussupov, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. 2021. Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* 32, 3 (2021), 713–735.
  - [35] Charalampia Xaroula Kerasidou, Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. 2021. Before and beyond trust: reliance in medical AI. *Journal of Medical Ethics* (2021).
  - [36] Derek J Koehler. 1991. Explanation, imagination, and confidence in judgment. *Psychological bulletin* 110, 3 (1991), 499.
  - [37] Niklas Kühl, Jodie Lobana, and Christian Meske. 2020. Do you comply with AI?—Personalized explanations of learning algorithms and their impact on employees' compliance behavior. *ICIS* (2020).
  - [38] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
  - [39] Keke Lai. 2018. Estimating standardized SEM parameters given nonnormal data and incorrect model: Methods and comparison. *Structural Equation Modeling: A Multidisciplinary Journal* 25, 4 (2018), 600–620.
  - [40] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. 1–18.
  - [41] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).
  - [42] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
  - [43] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.
  - [44] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.
  - [45] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
  - [46] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).
  - [47] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
  - [48] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutun Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.
  - [49] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The journal of strategic information systems* 11, 3-4 (2002), 297–323.
  - [50] Giang Nguyen, Mohammad Reza Taesiri, and Anh Nguyen. 2022. Visual correspondence-based explanations improve AI robustness and human-AI team accuracy. *arXiv preprint arXiv:2208.00780* (2022).
  - [51] Bethany Niese and Monica Adya. 2022. Who Can I Trust? Use and Disuse of AI-Based Systems Under Conditions of High and Low Stakes. (2022).
  - [52] Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*. 497–501.
  - [53] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557* (2011).
  - [54] Raja Parasuraman, Robert Molloy, and Indramani L Singh. 1993. Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology* 3, 1 (1993), 1–23.
  - [55] Dane K Peterson and Gordon F Pitz. 1988. Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14, 1 (1988), 85.
  - [56] Fotios Petropoulos, Robert Fildes, and Paul Goodwin. 2016. Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research* 249, 3 (2016), 842–852.
  - [57] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.
  - [58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).
  - [59] Victor Riley. 2018. Operator reliance on automation: Theory and data. In *Automation and human performance: Theory and applications*. 19–35.
  - [60] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. <https://doi.org/10.18637/jss.v048.i02>
  - [61] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI—Distinguishing Between Attitudinal and Behavioral Measures. *arXiv preprint arXiv:2203.12318* (2022).
  - [62] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making. *ACM Conference on Human Factors in Computing Systems (CHI '22), Workshop on Trust and Reliance in AI-Human Teams (trAlt)* (2022).
  - [63] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the Influence of Explainable AI on Automation Bias. In *Proceedings of the 30th European Conference on Information Systems (ECIS), Timișoara, RO, June 18-24, 2022*.

- [64] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *arXiv preprint arXiv:2209.11812* (2022).
- [65] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.
- [66] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31, 2 (2018), 47–53.
- [67] Janet A Sniezek and Lyn M Van Swol. 2001. Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes* 84, 2 (2001), 288–307.
- [68] Philipp Spitzer, Niklas Kühl, and Marc Goutier. 2022. Training Novices: The Role of Human-AI Collaboration and Knowledge Transfer. *Workshop on Human-Machine Collaboration and Teaming (HM-CaT 2022). The 39th International Conference on Machine Learning, Baltimore, Maryland, USA* (2022).
- [69] Andrew Talone. 2019. The effect of reliability information and risk on appropriate reliance in an autonomous robot teammate. (2019).
- [70] Heliodoro Tejeda, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2022. AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior* (2022), 1–18.
- [71] 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>
- [72] Lu Wang, Greg A Jamieson, and Justin G Hollands. 2008. Selecting methods for the analysis of reliance on automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. 287–291.
- [73] Xiuxin Wang and Xiufang Du. 2018. Why does advice discounting occur? The combined roles of confidence and trust. *Frontiers in Psychology* 9 (2018), 2381.
- [74] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.
- [75] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.
- [76] Ilan Yaniv. 2004. Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes* 93, 1 (2004), 1–13.
- [77] Paul Zarnoth and Janet A Sniezek. 1997. The social influence of confidence in group decision making. *Journal of Experimental Social Psychology* 33, 4 (1997), 345–366.
- [78] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*. 295–305.