AI Reliance and Decision Quality: Fundamentals, Interdependence, and the Effects of Interventions

Jakob Schoeffer   J.J.SCHOEFFER@RUG.NL
University of Groningen
Groningen, The Netherlands

Johannes Jakubik   JOHANNES.JAKUBIK@KIT.EDU
Karlsruhe Institute of Technology
Karlsruhe, Germany

Michael Vössing   MICHAEL.VOESSING@KIT.EDU
Karlsruhe Institute of Technology
Karlsruhe, Germany

Niklas Kühl   KUEHL@UNI-BAYREUTH.DE
University of Bayreuth
Fraunhofer FIT
Bayreuth, Germany

Gerhard Satzger   GERHARD.SATZGER@KIT.EDU
Karlsruhe Institute of Technology
Karlsruhe, Germany

Abstract

In AI-assisted decision-making, a central promise of having a human-in-the-loop is that they should be able to complement the AI system by overriding its wrong recommendations. In practice, however, we often see that humans cannot assess the correctness of AI recommendations and, as a result, adhere to wrong or override correct advice. Different ways of relying on AI recommendations have immediate, yet distinct, implications for decision quality. Unfortunately, reliance and decision quality are often inappropriately conflated in the current literature on AI-assisted decision-making. In this work, we disentangle and formalize the relationship between reliance and decision quality, and we characterize the conditions under which human-AI complementarity is achievable. To illustrate how reliance and decision quality relate to one another, we propose a visual framework and demonstrate its usefulness for interpreting empirical findings, including the effects of interventions like explanations. Overall, our research highlights the importance of distinguishing between reliance behavior and decision quality in AI-assisted decision-making.

1. Introduction

Decision-making increasingly leverages support from artificial intelligence (AI)-based systems with the goal of making better and more efficient decisions. Especially in high-stakes domains, such as lending, hiring, or healthcare, researchers and policymakers have often advocated for installing a human-in-the-loop as the "last line of defense against AI failures" (Passi & Vorvoreanu, 2022). This approach assumes that humans are capable of rectifying such AI failures when they occur. In AI-assisted decision-making, typically, an AI system generates an initial decision recommendation, which the human-in-the-loop may

either adhere to or override (see Figure 1). In order to complement the AI system, the human needs to adhere to AI recommendations if and only if these recommendations are correct, and override them otherwise. Empirical studies have shown, however, that humans are often not able to achieve this type of *appropriate reliance*[1] (Fok & Weld, 2024). Instead, we often observe that they over- or under-rely on AI recommendations, indicating a deficiency to distinguish correct from wrong AI advice. Even the introduction of additional means of decision support (e.g., explanations) has rarely produced the expected benefits in terms of human-AI complementarity (Schemmer et al., 2022; Schoeffer et al., 2024). A key concern is that root cause analyses often struggle due to a limited understanding of (*i*) how interventions influence human reliance on AI advice, and (*ii*) how reliance relates to relevant metrics of decision quality.

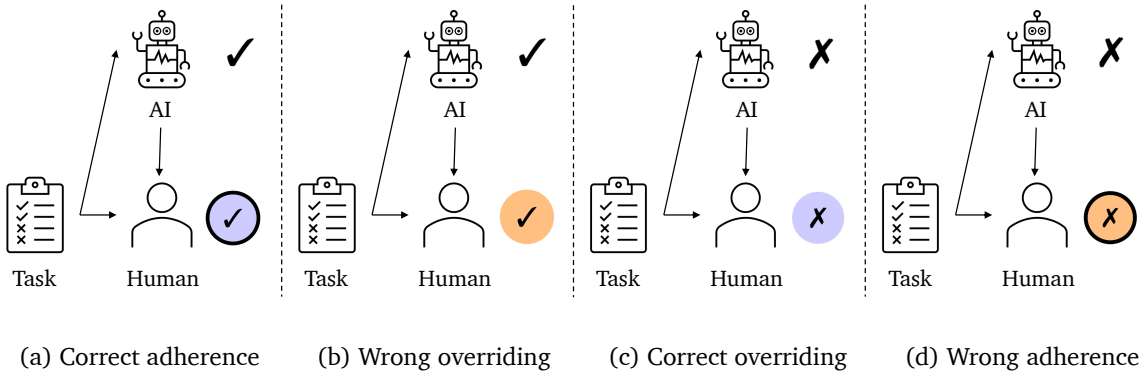(a) Correct adherence   (b) Wrong overriding   (c) Correct overriding   (d) Wrong adherence

Figure 1: We consider *concurrent* AI-assisted decision-making where a human-in-the-loop receives an AI recommendation that can either be correct (✔) or wrong (✗). The human can either adhere to (bordered circle) or override (no border) the AI recommendation. When the human adheres to a correct or overrides a wrong AI recommendation, the final decision will be correct (cases (a) and (c)); in the remaining cases, it will be wrong (cases (b) and (d)). The correctness of the final decision is indicated by either blue (correct) or orange (wrong) shading.

In this work, we conceptually study the interplay of AI reliance and decision quality, and we highlight the importance of disentangling them in studies on AI-assisted decision-making.[2] We propose a framework that differentiates between reliance *quantity* and *quality*, enabling us to examine how each—both independently and in combination—affects decision quality metrics, particularly *accuracy*. Additionally, we characterize the conditions under which human-AI complementarity occurs, where human involvement improves decision quality compared to scenarios where the AI system operates autonomously. Our framework is crafted to assist researchers and system designers in interpreting empirical findings, particularly the effects of interventions in AI-assisted decision-making. To support this, we also offer an open-source tool built on our framework, which is available at `https://github.com/jhnnsjkbk/accuracy-reliance`.

---

1. We use *reliance* as an umbrella term for humans' behavior of adhering to or overriding AI recommendations (Lai et al., 2023).
2. This article extends the HHAI2023 conference paper by the authors (Schoeffer et al., 2023). It includes additional empirical analyses, a software implementation, and relevant extensions of our framework.

This article is structured as follows. First, we provide relevant background and discuss related work in Section 2. After that, in Section 3, we formalize the relationship between AI reliance and decision-making accuracy, and we propose a visual framework to illustrate this relationship. In Section 4, we show how our framework enables a meaningful interpretation of intervention effects. Crucially, we show that interventions can influence decision quality through significantly different effects on human reliance. Without disentangling the effects on reliance and decision quality as we propose, two interventions may appear equally effective, when in reality, one leads to *more* frequent overriding of AI recommendations, while the other results in *fewer* overrides. We discuss relevant extensions of our framework in Section 5, and we conclude by discussing implications, limitations, and avenues for future work in Section 6.

2. Background and Related Work

In this section, we provide important background and discuss related work with respect to AI-assisted decision-making, reliance and trust, as well as relevant measures of decision quality and the effectiveness of interventions.

2.1 AI-Assisted Decision-Making

The utilization of AI-based systems in decision-making has been embraced across a multitude of critical domains (Lai et al., 2023). In hiring, recent studies show that more than 55% of human resource leaders in the United States use predictive algorithms to assist hiring activities (Reicin, 2021). The underlying motives of adopting AI systems for assisting decision-making are diverse, including cost-cutting, improving decision quality, and enabling more robust and objective decisions (Harris & Davenport, 2005; Kuncel et al., 2014; Lepri et al., 2018).

The degree of AI integration in decision-making processes may vary depending on the specific context. While many tasks may be well-suited for full automation through AI systems, others call for greater human oversight (De-Arteaga et al., 2020). Particularly in high-stakes domains, AI systems often serve as decision support tools that aid human experts, who retain discretionary power to override AI recommendations and ultimately bear responsibility for making final decisions. We refer to these human experts as the *human-in-the-loop*. For instance, in healthcare, AI systems can play a vital role in assisting clinicians with diagnoses or prognoses. Subsequently, the clinicians can utilize these insights to determine the most appropriate course of treatment (Leibig et al., 2022). Similarly, in the domain of criminal justice, judges might rely on AI-based risk assessment tools when determining bail (Lima et al., 2021).

> **Definition 1 (Human-in-the-loop)** *In the context of AI-assisted decision-making,* human-in-the-loop *refers to a human decision-maker that retains discretionary power to override initially generated AI recommendations.*

In human-in-the-loop settings, an AI system typically generates an initial decision recommendation, which the human may either adhere to or override (see Figure 1). In the taxonomy of Tejeda et al. (2022), this corresponds to *concurrent AI assistance*, where the

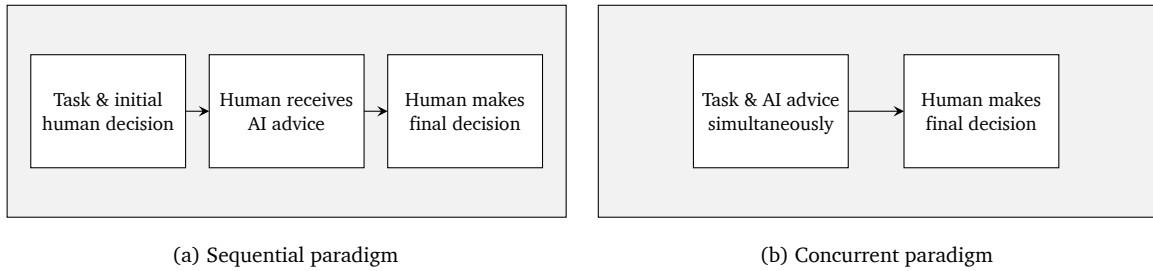(a) Sequential paradigm   (b) Concurrent paradigm

Figure 2: Two paradigms in AI-assisted decision-making: (a) sequential and (b) concurrent, taken from Tejeda et al. (2022). The focus of this work is on *concurrent* AI-assisted decision-making.

human-in-the-loop does *not* independently make a decision before AI assistance is provided (see Figure 2). Prior work has also studied decision-making setups where a human makes an initial decision without AI support and then a final decision after receiving AI advice. This *sequential* paradigm has been empirically studied by Schemmer et al. (2023), Dunning et al. (2024), and Van Dongen and Van Maanen (2013), among others. One supposed advantage of the sequential setup is its ability to isolate the effect of AI advice on human decision-making by allowing researchers to observe whether and how the presence of such advice may change the initial human guess. However, as pointed out by Tejeda et al. (2022), many decision-making scenarios in the real world do not include independent human-only decisions prior to AI involvement. Moreover, the sequential paradigm bears the risk of humans anchoring their second decision in their initial guess (Echterhoff et al., 2022). For these reasons, we build on prior work (Schoeffer et al., 2023) and focus on the *concurrent* paradigm, stressing that our work complements related work on sequential decision-making. Finally, we focus on binary decision-making as it is the most commonly studied type of decisions (Lai et al., 2023), e.g., in the realms of lending (loan/denial), hiring (offer/rejection), or recidivism prediction (release/detention).

2.2 Reliance and Trust in AI-Assisted Decision-Making

In concurrent AI-assisted decision-making, the human-in-the-loop is provided with an initial AI recommendation and may then leverage their discretionary power to either adhere to or override said recommendation. In line with prior work, we call this behavior *reliance* (Eckhardt et al., 2024; Vereschak et al., 2021).

> **Definition 2 (AI reliance)** *In the context of AI-assisted decision-making,* AI reliance *refers to the behavior of the human-in-the-loop that manifests in adhering to or overriding AI recommendations.*

Measuring and manipulating the human reliance on AI recommendations has become a central pillar of research on AI-assisted decision-making (Fok & Weld, 2024; Schemmer et al., 2023; Lai et al., 2023). This is especially important as both humans and AI systems are commonly thought of as imperfect "decision-makers" with individual strengths and weaknesses (Kamar, 2016). For humans that are assisted by AI, it is therefore important to be able to identify in which cases an AI recommendation is correct vs. wrong. This ability

to detect AI errors has also been studied through the lens of signal detection theory, which describes how people distinguish signal from noise under uncertainty (Langer et al., 2024; Hautus et al., 2021). In the context of AI-assisted decision-making, the reliance behavior of overriding wrong AI recommendations and adhering to correct ones is typically referred to as *appropriate reliance* (Schemmer et al., 2023).

Empirically, however, prior work has shown that humans are often unable to achieve appropriate reliance, even in the presence of different types of decision aids, such as explanations (Fok & Weld, 2024; Schemmer et al., 2022). Instead, we often observe reliance behavior of the following types: (*i*) overriding *too many* AI recommendations (*under-reliance*), or (*ii*) overriding *too few* AI recommendations (*over-reliance*). For instance, Dunning et al. (2024) find that humans tend to over-rely on highly skilled AI agents but under-rely when the AI agent is lower skilled. Prior research, including Buçinca et al. (2021) and Kim et al. (2023), has identified different scenarios in which under- or over-reliance results in reduced decision quality. At the same time, prior work often lacks important nuance in discussing the effects of reliance on decision quality, such as when over-reliance vs. under-reliance might be preferable, and how instances of inappropriate reliance may still justify the presence of the human-in-the-loop. Our work addresses this gap and complements recent studies by Fok and Weld (2024) and Guo et al. (2024).

Also related to our work is prior work on trust calibration—i.e., achieving levels of *appropriate trust* in AI systems, as opposed to over- or under-trust (de Visser et al., 2020; Sheridan, 2019). Many prior studies have treated *reliance* and *trust* interchangeably, sometimes calling reliance a "behavioral trust measure" (Papenmeier et al., 2022). However, we stress that reliance and trust are different constructs: reliance is commonly defined as the behavior of adhering to or overriding AI recommendations (Eckhardt et al., 2024), whereas trust is a subjective attitude regarding the whole system, which builds up and develops over time (Parasuraman & Riley, 1997; Rempel et al., 1985; Schaefer et al., 2016; Glikson & Woolley, 2020). It has been argued that trust may impact reliance (Dzindolet et al., 2003; Lee & See, 2004; Shin & Park, 2019), but trust is neither necessary nor sufficient for reliance when other factors, such as time constraints, perceived risk, or self-confidence, impact decision-making (De-Arteaga et al., 2020; Lee & See, 2004; Riley, 2018). In this work, we propose a taxonomy specifically for reliance behavior to facilitate a meaningful interpretation of empirical results in AI-assisted decision-making.

2.3 Measuring Decision Quality

Measuring the quality[3] of AI-assisted decisions can be thought of as a two-stage process, in which the final decision quality depends on (*i*) the initial quality of AI recommendations, and (*ii*) the degree to which the human-in-the-loop can or cannot correct any AI mistakes. In other words, the quality of AI-assisted decisions immediately depends on the degree to which the human adheres to or overrides AI recommendations, and *how* they do so (Jakubik et al., 2023). For instance, even if the AI system issues 100% correct recommendations, the final decision quality will decrease unless the human adheres to *all* recommendations. On the other hand, a relatively low-performance AI recommender may still result in high-

---

3. We understand *decision quality* as an umbrella term that subsumes any metrics that are a function of correct and wrong decisions. This excludes measures like efficiency, among others.

quality decisions if the human-in-the-loop overrides all AI mistakes. A relevant line of prior work has studied how different degrees of reliability of automation support affect human decision-making (Dunning et al., 2024; Rovira et al., 2007; Wickens & Dixon, 2007; de Visser & Parasuraman, 2011). For instance, Dunning et al. (2024) show that humans tend to quickly over-rely on high-performance automation but under-rely if reliability is low. This shows that human reliance behavior may itself be a function of the initial quality of AI recommendations, stressing the complexity of the interplay between AI reliance and decision quality, which motivates our work.

The *accuracy* metric represents a key quality measure of AI-assisted decisions. *Accuracy* is defined as follows:

$$Accuracy = \frac{\#correct\ decisions}{\#decisions}.$$

Accuracy is frequently used for measuring the quality of AI-assisted decision-making (Lai et al., 2023) and evaluating the effectiveness of interventions (e.g., explanations) for decision support (Zhang et al., 2020; Lai et al., 2020; Cabrera et al., 2023; Lai et al., 2023). Note that we can also compute the accuracy of AI recommendations ("AI accuracy"). Regarding the accuracy of decisions, prior works have studied whether humans can complement AI systems, i.e., improve the final decision-making accuracy over the AI accuracy (see Lai et al. (2023) for an overview). In fact, a common motivation for providing humans with discretionary power is that they should be able to complement an AI system by overriding its wrong recommendations (De-Arteaga et al., 2020). When the human-in-the-loop achieves that, we refer to this as *complementarity*.[4]

Definition 3 (Complementarity) *We call* complementarity *the state when the human-in-the-loop relies on AI recommendations in a way that the final decision quality is higher than without human involvement.*

In the context of the interplay of reliance and accuracy, this means that when the human-in-the-loop adheres to all correct and overrides all wrong AI recommendations, we achieve a final decision-making accuracy of 100%. However, as stated earlier, empirical evidence shows that achieving appropriate reliance is difficult in practice (Fok & Weld, 2024; Schemmer et al., 2022). In cases where humans do *not* appropriately rely on AI recommendations, the relationship between reliance behavior and accuracy is less obvious. Crucially, there may be cases where *inappropriate* reliance may still lead to complementarity. These cases are not sufficiently well understood in the literature.

2.4 Measuring Effectiveness of Interventions

An emerging stream of research is studying decision support interventions that aim at facilitating complementarity (Schemmer et al., 2023; Cabitza et al., 2023). A prominent type of such interventions are explanations of various types (Gunning & Aha, 2019; Arrieta et al., 2020; Adadi & Berrada, 2018; Gilpin et al., 2018). However, prior work has often focused on assessing the effectiveness of interventions only with respect to *either* decision-making

---

4. Other works have also referred to this as the *team performance* of human and AI being greater than the performance of the individual agents alone (de Visser & Parasuraman, 2011; Hemmer et al., 2022).

accuracy (Lai et al., 2023; Kim et al., 2023) *or* the human reliance on AI recommendations (Buçinca et al., 2021; Lu & Yin, 2021).

Regarding *accuracy*, it is commonly claimed that explanations are an enabler for better decisions (Arrieta et al., 2020; Dodge et al., 2019; Gilpin et al., 2018; Rader et al., 2018). A meta-study by Schemmer et al. (2022), however, suggests that explanations in most empirical studies do not yield any significant benefits regarding decision-making accuracy; e.g., in the studies of Alufaisan et al. (2021), Green and Chen (2019), Liu et al. (2021), and Zhang et al. (2020). On the other hand, Lai and Tan (2019) find that explanations may enhance decision-making accuracy for the case of deception detection. Regarding *reliance*, several studies find that explanations may increase *or* decrease adherence of humans to AI recommendations regardless of their correctness (i.e., over- or under-reliance), such as the works by Bansal et al. (2021), Poursabzi-Sangdeh et al. (2021), van der Waa et al. (2021), and Schoeffer et al. (2024).

In this work, we argue that measuring *either* accuracy *or* reliance alone may provide an incomplete view when assessing AI-assisted decision-making generally and the effectiveness of interventions specifically. More concretely, our proposed framework shows that (*i*) focusing on accuracy alone is insufficient because an intervention may lead to an increase in decision quality by solely making a human adhere to more AI recommendations, and (*ii*) focusing on reliance alone is insufficient because both over- and under-reliance may under certain circumstances still lead to improvements in accuracy compared to the baseline without human involvement.

3. The Interdependence of Reliance Behavior and Accuracy

We consider binary decision-making tasks of $n \in \mathbf{N}$ instances with n AI recommendations. Let $Acc_{AI} \in (50\%, 100\%]$ be the AI accuracy[5], and $\mathcal{A} \in [0\%, 100\%]$ the degree of human adherence to AI recommendations—e.g., $\mathcal{A} = 70\%$ when the human-in-the-loop adheres to 70% of AI recommendations. As introduced in Figure 1, adherence can be correct ($\mathcal{A}_{correct}$) or wrong ($\mathcal{A}_{wrong}$). We measure the correct and wrong adherence in percent, and we have $\mathcal{A} = \mathcal{A}_{correct} + \mathcal{A}_{wrong}$. Similarly, we call the percentage of overrides $\mathcal{O} \in [0\%, 100\%]$ (correct: $\mathcal{O}_{correct}$ or wrong: $\mathcal{O}_{wrong}$), and we have $\mathcal{O} = \mathcal{O}_{correct} + \mathcal{O}_{wrong}$. We call the final decision-making accuracy, i.e., after human reliance behavior, Acc_{final}. While in practice humans can only adhere to or override a finite number of AI recommendations, we often consider $n \to \infty$ for our theoretical considerations, so as to avoid rounding. We summarize all notation in Table 1. Note that by definition we have:

$$\mathcal{A} + \mathcal{O} = \mathcal{A}_{correct} + \mathcal{A}_{wrong} + \mathcal{O}_{correct} + \mathcal{O}_{wrong} = 100\%$$
$$Acc_{AI} = \mathcal{A}_{correct} + \mathcal{O}_{wrong} \quad (1)$$
$$Acc_{final} = \mathcal{A}_{correct} + \mathcal{O}_{correct}.$$

3.1 Motivational Example

Consider the following motivational example: we have a task that consists of making $n = 10$ binary decisions. The AI system that is used for providing decision recommendations to the

---

5. Note that we only consider cases where the AI performs strictly better than chance.

Symbol	Description
Acc_{AI}	AI accuracy in %
Acc_{final}	(Final) decision-making accuracy in %, after human reliance
$\mathcal{A}$	Degree of human adherence to AI recommendations in %
$\mathcal{A}_{correct}$	Correct adherence in %: **adhering** to **correct** AI recommendation
$\mathcal{A}_{wrong}$	Wrong adherence in %: **adhering** to **wrong** AI recommendation
$\mathcal{O}$	Degree of human overriding of AI recommendations in %
$\mathcal{O}_{correct}$	Correct overriding in %: **overriding** of **wrong** AI recommendation
$\mathcal{O}_{wrong}$	Wrong overriding in %: **overriding** of **correct** AI recommendation
$n \in \mathbf{N}$	Number of AI recommendations
$Q \in [0,1]$	Quality of human reliance
✓	Correct AI recommendation
✗	Wrong AI recommendation

Table 1: Summary of important notation.

human has an accuracy of $Acc_{AI} = 70\%$; i.e., 7 out of 10 recommendations are correct (✓) and 3 are wrong (✗). Now, when the human-in-the-loop adheres to all AI recommendations ($\mathcal{A} = 100\%$), this leads to a decision-making accuracy of $Acc_{final} = 70\%$, equal to the AI accuracy. In terms of reliance behavior, this implies that the human-in-the-loop correctly adheres to 7 correct AI recommendations ($\mathcal{A}_{correct} = 70\%$), and wrongly adheres to the remaining 3 recommendations ($\mathcal{A}_{wrong} = 30\%$). In the other extreme case, where the human-in-the-loop overrides all AI recommendations ($\mathcal{O} = 100\%$), the resulting decision-making accuracy will be $100\% - 70\% = 30\%$, where the human correctly overrides 3 wrong AI recommendations ($\mathcal{O}_{correct} = 30\%$), and wrongly overrides 7 correct AI recommendations ($\mathcal{O}_{wrong} = 70\%$).

If the human reliance behavior is mixed, i.e., when the human-in-the-loop adheres to some AI recommendations and overrides others, decision-making accuracy will depend on how well they can distinguish cases where the AI is correct from cases where it is wrong. To make this clear, consider the same AI system as above with an accuracy of 70%, and a human-in-the-loop that adheres to 7 out of 10 of its recommendations ($\mathcal{A} = 70\%$). This is illustrated in Figure 3. If the human-in-the-loop is able to perfectly distinguish between correct and wrong AI recommendations, they will adhere to all 7 correct AI recommendations ($\mathcal{A}_{correct} = 70\% = \mathcal{A}$) and override the 3 wrong ones ($\mathcal{O}_{correct} = 30\% = \mathcal{O}$). The resulting decision-making accuracy would then be $Acc_{final} = 100\%$ (case (a) in Figure 3). In this case, the human-in-the-loop is able to perfectly complement the AI by correcting for its mistakes. Cases (b)–(d) in Figure 3 show situations where the human-in-the-loop still adheres to 70% of AI recommendations but their ability to override wrong AI recommendations decreases. For instance, consider case (d), where the human-in-the-loop does not perform any correct overrides ($\mathcal{O}_{correct} = 0$). If the degree of human adherence to AI recommendations is fixed at 70%, this is, in fact, the worst possible reliance behavior with respect to accuracy, resulting in a decision-making accuracy of $Acc_{final} = 40\%$.

From Figure 3, we can also infer that if the human-in-the-loop overrides *more* than 3 AI recommendations, at least one of these overrides must be wrong (i.e., the human
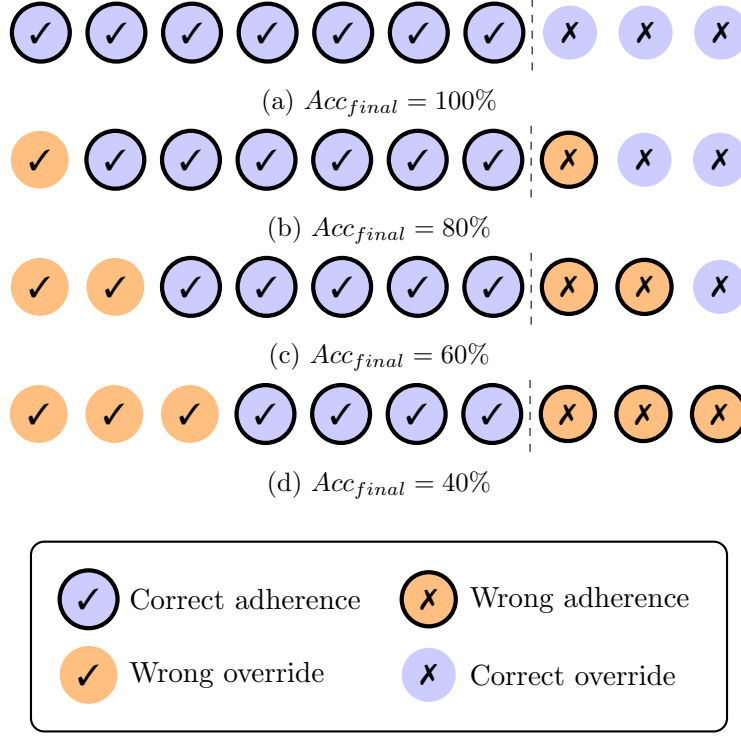
Figure 3: Possible scenarios of reliance behavior and associated decision-making accuracy, given an AI accuracy of $Acc_{AI} = 70\%$ and an adherence level of $\mathcal{A} = 70\%$. Correct AI recommendations (✓) and wrong AI recommendation (✗) are separated by a dashed line.

would override a correct AI recommendation), meaning that a decision-making accuracy of 100% would no longer be possible. We may think of such a reliance behavior as *under-reliance*. Similarly, when the human overrides *less* than 3 AI recommendations, there must be at least one instance of wrong adherence. This might be referred to as *over-reliance*. Generally, we may think of under-reliance as a behavior where $\mathcal{A} < Acc_{AI}$, and over-reliance as $\mathcal{A} > Acc_{AI}$. Note that there exists other work that has been thinking of these terms with respect to behavior at the level of individual decisions (Schemmer et al., 2023; Vasconcelos et al., 2023).

3.2 The General Case

Generally, any degree of adherence to AI recommendations is associated with a range of possible decision-making accuracy, based on how well the human-in-the-loop can override the AI recommendations when they are wrong and adhere to them when they are correct. In Figure 3, this range would be $Acc_{final} \in \{40\%, 60\%, 80\%, 100\%\}$ for $n = 10$, a given AI accuracy of $Acc_{AI} = 70\%$, and a degree of adherence to AI recommendations of $\mathcal{A} = 70\%$. As mentioned earlier, we generally consider $n \to \infty$, in which case the possible values in this range become continuous. We state the following proposition on the attainable decision-making accuracy as a function of the AI accuracy as well as the degree of human adherence to AI recommendations.

Proposition 1 *For $n \to \infty$, a given AI accuracy Acc_{AI}, and a degree of adherence to AI recommendations, $\mathcal{A}$, the range of attainable decision-making accuracy Acc_{final} is*

$$Acc_{final} \in \begin{cases} [100\% - Acc_{AI} - \mathcal{A}, 100\% - Acc_{AI} + \mathcal{A}] & \text{if } 0 \leq \mathcal{A} \leq 100\% - Acc_{AI} \\ [-100\% + Acc_{AI} + \mathcal{A}, 100\% - Acc_{AI} + \mathcal{A}] & \text{if } 100\% - Acc_{AI} < \mathcal{A} \leq Acc_{AI} \\ [-100\% + Acc_{AI} + \mathcal{A}, 100\% + Acc_{AI} - \mathcal{A}] & \text{if } Acc_{AI} < \mathcal{A} \leq 100\%. \end{cases}$$

The maximum of this accuracy range will be attained whenever the human-in-the-loop maximizes correct adherence and correct overrides given a degree of adherence $\mathcal{A}$, since $Acc_{final} = \mathcal{A}_{correct} + \mathcal{O}_{correct}$. Hence, in the ideal case, we would have $\mathcal{A}_{correct} + \mathcal{O}_{correct} = 100\%$, which immediately implies that $\mathcal{A}_{wrong} = \mathcal{O}_{wrong} = 0\%$. This would be case (a) in Figure 3. However, as we can see in Proposition 1, this is only possible when $\mathcal{A} = \mathcal{A}_{correct} = Acc_{AI}$, meaning that the human must adhere to AI recommendations if and only if they are correct, and override otherwise. In other words, to achieve a decision-making accuracy of $Acc_{final} = 100\%$, we need two conditions:

(i) The general degree of adherence to AI recommendations, $\mathcal{A}$, is equal to the AI accuracy Acc_{AI}, i.e., $\mathcal{A} = Acc_{AI}$.

(ii) The human-in-the-loop must be able to adhere to any correct AI recommendation and override any wrong one, i.e., $\mathcal{A}_{correct} = \mathcal{A}$ and $\mathcal{O}_{correct} = \mathcal{O}$.

However, in practice, it is likely that either (i) or (ii) are not satisfied and, hence, the decision-making accuracy is less than 100%. Even if (i) is satisfied, like in Figure 3, we see in cases (b)–(d) that Acc_{final} is negatively affected when humans adhere to wrong AI recommendations and override correct ones.

3.3 A Visual Framework

To make the general relationship between reliance behavior and decision-making accuracy more tangible, we visualize Proposition 1 in Figure 4 for (a) $Acc_{AI} = 70\%$ and (b) $Acc_{AI} = 90\%$. On the horizontal axes we have the human adherence to AI recommendations, $\mathcal{A} \in [0, 100\%]$. The vertical axes show the decision-making accuracy, $Acc_{final} \in [0, 100\%]$. The filled rectangular area in orange and blue combined constitutes the attainable decision-making accuracy for any given $\mathcal{A}$. We distinguish orange and blue to highlight areas where the human-in-the-loop complements the AI (blue, $Acc_{final} > Acc_{AI}$) or impairs it (orange, $Acc_{final} < Acc_{AI}$) regarding accuracy. The green dashed vertical line indicates the level of $\mathcal{A} = Acc_{AI}$, which corresponds to the degree of adherence where the maximum decision-making accuracy of 100% can be attained, as discussed previously. Note that as the AI accuracy increases (Figure 4 (a) $\to$ (b)), the colored area decreases; and for $Acc_{AI} = 100\%$ it becomes a line, in which case Proposition 1 collapses into $Acc_{final} = \mathcal{A}$. This means that the better the AI, the smaller the variance for different reliance behaviors.

Contrasting the orange and blue areas, we immediately see that up to a certain level of adherence $\mathcal{A}$ there is no possibility to reach the blue area of complementarity. We also see that the minimum level of $\mathcal{A}$ for which the human-in-the-loop may complement the AI increases as Acc_{AI} increases ($\mathcal{A} = 40\%$ in Figure 4 (a) $\to \mathcal{A} = 80\%$ in Figure 4 (b)). Finally,

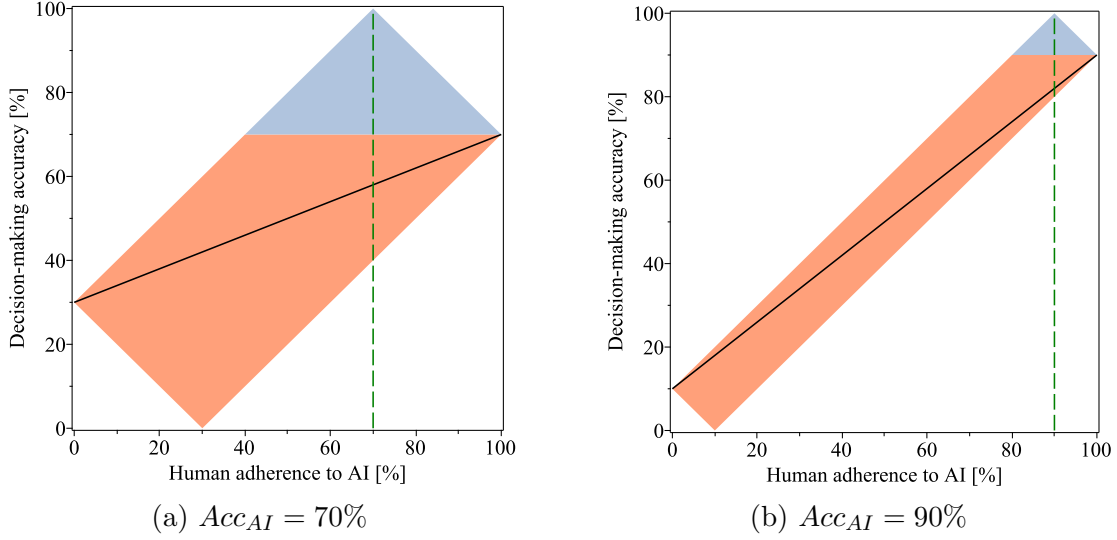(a) $Acc_{AI} = 70\%$

(b) $Acc_{AI} = 90\%$

Figure 4: The area of attainable decision-making accuracy for a given AI accuracy of (a) 70% and (b) 90%, and different levels of human adherence. The orange area indicates $Acc_{final} < Acc_{AI}$; blue indicates $Acc_{final} > Acc_{AI}$; the green dashed line indicates the level of adherence where $Acc_{final} = 100\%$ is attainable; the black line indicates the expected value of Acc_{final} when the human-in-the-loop cannot discern correct and wrong.

when $\mathcal{A} \geq Acc_{AI}$, attaining a decision-making accuracy in the blue area is always *possible*. We characterize this in the following corollary:

Corollary 1 *When the human-in-the-loop under-relies at a degree of $\mathcal{A} < 2 \cdot Acc_{AI} - 100\%$, we will always have $Acc_{final} < Acc_{AI}$. When $\mathcal{A} > 2 \cdot Acc_{AI} - 100\%$, achieving a decision-making accuracy greater than the AI accuracy, i.e., $Acc_{final} > Acc_{AI}$, is possible.*

From the visual framework, we can also formally infer that any $Acc_{final} \in (0, 100\%)$ can be associated with different degrees of adherence $\mathcal{A}$. In fact, due to the symmetric shape of the rectangle, when we think of Acc_{final} as a function of $\mathcal{A}$, the inverse $\mathcal{A}(Acc_{final})$ would be identical to the function itself. For instance, a decision-making accuracy of $Acc_{final} = 70\%$ may correspond to any $\mathcal{A} \in [40\%, 100\%]$ in Figure 4 (a).

Proposition 2 *When $Acc_{final}(\mathcal{A}) \in [u, v]$ for a given $\mathcal{A}$, we have $\mathcal{A}(Acc_{final}) \in [u, v]$.*

However, when fixing Acc_{final} at 70% in Figure 4 (a), different levels of $\mathcal{A}$ correspond to different vertical positions within the rectangle: $\mathcal{A} = 40\%$ corresponds to a position at the very northern border of the rectangle, whereas any $\mathcal{A} > 40\%$ corresponds to a position on the horizontal line separating the orange and blue areas, which leaves room for accuracy improvements. This means that a given decision-making accuracy can be achieved through strikingly different qualities of reliance. We address this, as well as the role of the black separating lines in Figure 4, in more detail in the following.

3.4 Discerning Correct and Wrong AI Recommendations

While a horizontal movement in the framework constitutes a change in the *quantity* of adherence to AI recommendations, this information alone does not capture the *quality* of reliance—this information is captured in the vertical movements. To make this more concrete, consider again a task with AI recommendations that are 70% accurate. When the human-in-the-loop has no ability to distinguish correct from wrong AI recommendations, the likelihood of adhering to or overriding a given AI recommendation is the same regardless of whether that recommendation is correct or wrong. Hence, at an adherence of $\mathcal{A}$, we would expect the human-in-the-loop to adhere to $\mathcal{A}\%$ of correct AI recommendations and $\mathcal{A}\%$ of wrong AI recommendations. At $Acc_{AI} = 70\%$, this implies that $\mathcal{A}\%$ of 70% are correct adherences, $\mathcal{A}\%$ of 30% are wrong adherences, $(100 - \mathcal{A})\%$ of 70% are wrong overrides, and $(100 - \mathcal{A})\%$ of 30% are correct overrides. When we have $\mathcal{A} = 70\%$, this would imply $\mathcal{A}_{correct} = 49\%$, $\mathcal{A}_{wrong} = 21\%$, $\mathcal{O}_{correct} = 9\%$, and $\mathcal{O}_{wrong} = 21\%$, with a decision-making accuracy of $\mathcal{A}_{correct} + \mathcal{O}_{correct} = 58\%$. This corresponds to the intersection of the black line with the dashed green vertical line in Figure 4 (a). We generalize this in the following proposition.

Proposition 3 *When the human-in-the-loop cannot discern correct and wrong AI recommendations, the expected decision-making accuracy is linearly increasing in $\mathcal{A}$ and given by*

$$Acc_{final}(\mathcal{A}) = \mathcal{A} \cdot Acc_{AI} + (100\% - \mathcal{A}) \cdot (100\% - Acc_{AI})$$
$$= (100\% - Acc_{AI}) + \underbrace{(2 \cdot Acc_{AI} - 100\%)}_{>0} \cdot \mathcal{A},$$

for a given AI accuracy Acc_{AI}.

Note that the relationship from Proposition 3 equates to the black lines in Figure 4, which separate the respective rectangles in half. We immediately see the following:

Corollary 2 *When the human-in-the-loop cannot discern correct and wrong AI recommendations, the expected decision-making accuracy is always lower than or equal to the AI accuracy, i.e., $Acc_{final} \leq Acc_{AI}$.*

Having established the expected decision-making accuracy when the human-in-the-loop is not able to distinguish correct and wrong AI recommendations, we now turn to cases where they can—to different degrees. Such reliance behavior corresponds to points in the framework that are situated *above* the black line. While certainly less relevant in practice, we might also think of cases where a human adheres to and overrides AI recommendations worse than chance, which would correspond to points *below* the black line. Following up on Proposition 1, we now examine three cases based on different adherence levels, and we characterize the reliance behavior that is associated with the maximum and minimum decision-making accuracy for given $\mathcal{A}$.

Case: $0 \leq \mathcal{A} \leq 100\% - Acc_{AI}$ Since we assume that $Acc_{AI} > 50\%$, we have $\mathcal{A} < Acc_{AI}$ in this case. When the degree of adherence to AI recommendations is strictly smaller than

the AI accuracy, achieving a decision-making accuracy of $Acc_{final} = 100\%$ is no longer possible. This also implies that there must be at least one instance where the human-in-the-loop overrides a correct AI recommendation, i.e., $\mathcal{O}_{wrong} > 0$. From Proposition 1 we also see that the **maximum** achievable decision-making accuracy in that case is $100\% - Acc_{AI} + \mathcal{A}$, which is achieved when $\mathcal{A} = \mathcal{A}_{correct}$. Using the definition of $\mathcal{A}$ and relationships from Equation (1), this directly implies that $\mathcal{A}_{wrong} = 0$, $\mathcal{O}_{correct} = 100\% - Acc_{AI}$, and $\mathcal{O}_{wrong} = Acc_{AI} - \mathcal{A} > 0$. The **minimum** achievable decision-making accuracy, on the other hand, is attained when adherence only happens to wrong AI recommendations, hence, $\mathcal{A}_{wrong} = \mathcal{A}$. Similar to above, this implies that $\mathcal{A}_{correct} = 0$, $\mathcal{O}_{wrong} = Acc_{AI}$, and $\mathcal{O}_{correct} = 100\% - Acc_{AI} - \mathcal{A}$.

To illustrate this, let us reconsider the example from Figure 3, but with a degree of adherence to AI recommendations of $\mathcal{A} = 20\%$. The attainable decision-making accuracy in this case is, according to Proposition 1, $Acc_{final} \in [10\%, 50\%]$. To achieve $Acc_{final} = 50\%$, the human-in-the-loop would have to adhere to 2 correct AI recommendations ($\mathcal{A}_{correct} = 20\%$) and 0 wrong AI recommendations ($\mathcal{A}_{wrong} = 0$). The remaining 8 AI recommendations, 5 of which are correct and 3 wrong, are overridden (i.e., $\mathcal{O}_{wrong} = 50\%$ and $\mathcal{O}_{correct} = 30\%$). The minimum decision-making accuracy of 10%, on the other hand, is attained when the human-in-the-loop only adheres to wrong AI recommendations (i.e., $\mathcal{A}_{wrong} = 20\%$ and $\mathcal{A}_{correct} = 0$). The remaining AI recommendations, 7 correct and 1 wrong, are overridden, which implies $\mathcal{O}_{wrong} = 70\%$ and $\mathcal{O}_{correct} = 10\%$. Overall, we conclude the following:

> **Corollary 3** *When $0 \leq \mathcal{A} \leq 100\% - Acc_{AI}$, the decision-making accuracy is maximal when all adherence is to correct AI recommendations (i.e., $\mathcal{A}_{correct} = \mathcal{A}$), and it is minimal when all adherence is to wrong AI recommendations (i.e., $\mathcal{A}_{wrong} = \mathcal{A}$).*

Case: $100\% - Acc_{AI} < \mathcal{A} \leq Acc_{AI}$ With the same argument as in the previous case, the **maximum** decision-making accuracy is attained when $\mathcal{A}_{correct} = \mathcal{A}$, which directly implies $\mathcal{A}_{wrong} = 0$, $\mathcal{O}_{correct} = 100\% - Acc_{AI}$, and $\mathcal{O}_{wrong} = Acc_{AI} - \mathcal{A}$. As for the **minimum** decision-making accuracy, note that since $\mathcal{A} > 100\% - Acc_{AI}$, we must have $\mathcal{A}_{correct} > 0$, i.e., the human-in-the-loop must be adhering to at least one correct AI recommendation. The minimum accuracy is thus attained when the human-in-the-loop adheres to all wrong AI recommendations (plus at least one correct recommendation). This implies that all overrides must be of correct AI recommendations, i.e., we have $\mathcal{O}_{wrong} = \mathcal{O}$, $\mathcal{O}_{correct} = 0$, as well as $\mathcal{A}_{correct} = Acc_{AI} - \mathcal{O} > 0$, and $\mathcal{A}_{wrong} = 100\% - Acc_{AI}$.

> **Corollary 4** *When $100\% - Acc_{AI} < \mathcal{A} \leq Acc_{AI}$, the decision-making accuracy is maximal when all adherence is to correct AI recommendations (i.e., $\mathcal{A}_{correct} = \mathcal{A}$), and it is minimal when all overrides are of correct AI recommendations (i.e., $\mathcal{O}_{wrong} = \mathcal{O}$).*

Case: $Acc_{AI} < \mathcal{A} \leq 100\%$ While in the previous two cases we had $\mathcal{A} \leq Acc_{AI}$, we now consider the case where the human-in-the-loop over-relies on the AI recommendations, meaning that there must be at least one case where they adhere to a wrong AI recommendation, i.e., $\mathcal{A}_{wrong} > 0$. The **maximum** decision-making accuracy will thus be attained when all overrides are correct, i.e., $\mathcal{O}_{correct} = \mathcal{O}$, which immediately implies $\mathcal{O}_{wrong} = 0$, $\mathcal{A}_{correct} = Acc_{AI}$, and $\mathcal{A}_{wrong} = 100\% - Acc_{AI} - \mathcal{O} > 0$. The **minimum** decision-making accuracy, on the other hand, will be attained when all overrides are wrong, similar to the

previous case. Hence, we would also observe $\mathcal{O}_{wrong} = \mathcal{O}$, $\mathcal{O}_{correct} = 0$, $\mathcal{A}_{correct} = Acc_{AI} - \mathcal{O}$, and $\mathcal{A}_{wrong} = 100\% - Acc_{AI}$.

> **Corollary 5** *When $Acc_{AI} < \mathcal{A} \leq 100\%$, the decision-making accuracy is maximal when all overrides are of wrong AI recommendations (i.e., $\mathcal{O}_{correct} = \mathcal{O}$), and it is minimal when all overrides are of correct AI recommendations (i.e., $\mathcal{O}_{wrong} = \mathcal{O}$).*

3.5 Measuring the Quality of Reliance for Given $\mathcal{A}$

In the previous subsection, we established the reliance behavior that is associated with the extreme cases of maximum and minimum decision-making accuracy for any given degree of adherence to AI recommendations. Now, we develop a metric $Q \in [0, 1]$ for the quality of reliance given Acc_{AI}, such that a value of $Q = 1$ corresponds to the maximum attainable decision-making accuracy, and $Q = 0$ to the minimum. First, we derive the following corollary from Proposition 1:

> **Corollary 6** *The width W of the range of attainable values of Acc_{final} is:*
>
> $$W(\mathcal{A}) = \begin{cases} 2 \cdot \mathcal{A} & if \ \ 0 \leq \mathcal{A} \leq 100\% - Acc_{AI} \\ 2 \cdot (100\% - Acc_{AI}) & if \ \ 100\% - Acc_{AI} < \mathcal{A} \leq Acc_{AI} \\ 2 \cdot (100\% - \mathcal{A}) & if \ \ Acc_{AI} < \mathcal{A} \leq 100\%. \end{cases}$$

Geometrically, W corresponds to the distance between the upper and lower vertical boundary of the rectangle (see, e.g., Figure 4) for a fixed $\mathcal{A}$. With that, we can define our metric Q as follows:

$$Q(\mathcal{A}, Acc_{final}) := \begin{cases} \dfrac{Acc_{final} - (100\% - Acc_{AI} - \mathcal{A})}{W(\mathcal{A})} & if \ \ 0 \leq \mathcal{A} \leq 100\% - Acc_{AI} \\ \dfrac{Acc_{final} + (100\% - Acc_{AI} - \mathcal{A})}{W(\mathcal{A})} & if \ \ 100\% - Acc_{AI} < \mathcal{A}. \end{cases} \tag{2}$$

If Acc_{AI} and $\mathcal{A}$ are fixed, maximizing the quality of reliance corresponds to maximizing $Acc_{final} = \mathcal{A}_{correct} + \mathcal{O}_{correct}$ given $\mathcal{A}$, and we have seen what this entails in terms of reliance behavior for any value of $\mathcal{A}$ in Section 3.4.

4. Applying the Framework to Assess Interventions

In the following, we leverage the presented framework to interpret the different effects that interventions exhibit on human reliance behavior and decision quality. We start with analyses based on hypothetical interventions to demonstrate the relevance of disentangling effects on reliance behavior and accuracy in general. Subsequently, we use our framework to assess and visualize the effects of interventions studied in prior work.

4.1 Understanding the Effects of Interventions

Our theoretical results and the visual framework can be used to assess empirical studies in AI-assisted decision-making and understand them better. Any such empirical finding

would be a static point in the colored rectangle, from which we can immediately infer interesting properties, such as the quantity and quality of reliance, the exact percentages of correct and wrong adherence and overrides, and the effectiveness of the human-in-the-loop in complementing the AI system.

Another key usage of the framework is its ability to disentangle the effects of interventions, such as explanations or other means of decision support (see Lai et al. (2023) for an overview of such interventions). For that, let us consider the following hypothetical example: through a randomized experiment, we have collected data where humans are relying on AI recommendations in the presence of two different types of explanations (blue • and purple • dots) vs. a baseline without explanations (black dot •). We can think of these interventions as movements in our visual framework, as illustrated by the colored arrows in Figure 5. The black dot corresponds to a situation where a human-in-the-loop cannot discern correct and wrong AI recommendations and adheres to $\mathcal{A} = 50\%$ AI recommendations.
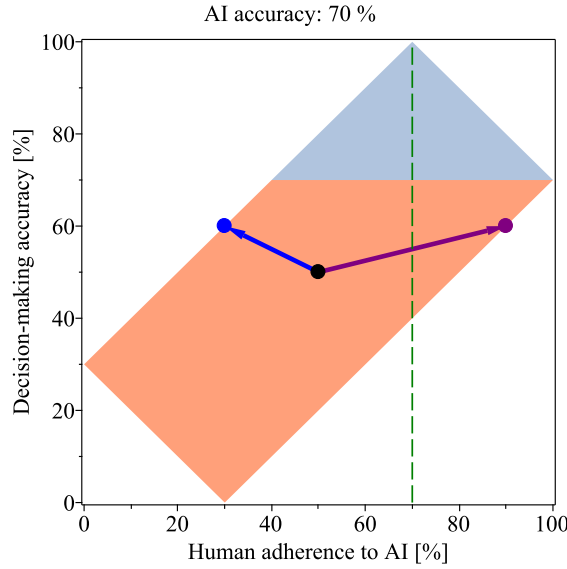

Figure 5: Visualizing the effects of two different interventions (blue • and purple •) on reliance behavior and decision-making accuracy.

In the case of the blue • intervention, we see that it leads to a decrease in the degree of adherence to AI recommendations, compared to the baseline ($\mathcal{A} = 50\% \rightarrow \mathcal{A} = 30\%$), but an increase in decision-making accuracy ($Acc_{final} = 50\% \rightarrow Acc_{final} = 60\%$) through a better reliance quality ($Q = 0.5 \rightarrow Q = 1$). In the case of the purple • intervention, we see the same effect with respect to accuracy but an entirely different effect on the reliance behavior: this intervention leads to an *increase* in adherence to AI recommendations ($\mathcal{A} = 50\% \rightarrow \mathcal{A} = 90\%$). At the same time, reliance quality drops from $Q = 0.5$ to $Q = 0$, which from Corollary 5 we know corresponds to a situation of over-reliance where any of the 10% overrides are of correct AI recommendations. Overall, this implies that different interventions can have seemingly similar effects on decision quality but drastically different effects on reliance behavior. Our framework enables us to disentangle these effects.

4.2 Application to Empirical Studies

To further emphasize its practical relevance, we now apply our framework to three empirical studies that investigate the effects of different interventions on reliance behavior and accuracy in AI-assisted decision-making. The studies cover both binary and multi-class decision-making tasks. We propose that the way we document and present the empirical findings can serve as a blueprint for researchers when reporting their own results.

4.2.1 INTRODUCTION OF STUDIES

In **study 1** by Liu et al. (2021), the authors investigate how interactive explanations can be utilized to facilitate human-AI complementarity. They also compare the performance of human and AI system for in-distribution and out-of-distribution setups originating from distribution shifts. Overall, the authors observe significant performance differences on in-distribution and out-of-distribution data. For interactive explanations, the authors observe mixed results; they suggest that interactive explanations can improve the perceived usefulness of AI assistance, but they may also reinforce human biases and lead to limited performance improvement.

Study 2 by Schoeffer et al. (2024) assesses the effects of feature-based explanations on the fairness of AI-assisted decisions, through the mediating role of reliance behavior. The research shows that feature-based explanations may foster or hinder fairness, depending on which type of features they highlight. Moreover, this study suggests that humans override (*i*) more AI recommendations when explanations highlight gendered features, and (*ii*) fewer AI recommendations when only task-relevant features are highlighted. At the same time, the authors find that feature-based explanations do not enhance people's ability to execute correct vs. wrong overrides compared to a baseline scenario without explanations.

In **study 3** by Tejeda et al. (2022), the authors compare differences in reliance behavior based on whether or not a human-in-the-loop provides independent decisions before they are provided with AI assistance (i.e., *sequential* vs. *concurrent* decision-making). This work develops a cognitive model to infer the latent reliance strategy of humans on AI assistance without asking them to make an independent decision beforehand. The predictions of this cognitive model are then validated through behavioral experiments for a multi-class classification task. Specifically, participants had to classify everyday images, such as different animals or types of vehicles.

4.2.2 IMPLEMENTATION

We apply our proposed framework to these studies to visualize the effects of the respective interventions and to gain a deeper understanding of how the interventions affect the ability of humans to discern between correct and wrong AI recommendations. To this end, we implemented a user interface that allows to customize the visual framework to any empirical study requiring only information on the AI accuracy (Acc_{AI}), the level of adherence ($\mathcal{A}$), and the final decision-making accuracy (Acc_{final}).[6] We leverage the developed tool to generate the figures in this section. Moreover, in addition to the visualization, our implementation solves the following system of equations to determine the proportions of correct and wrong

---

6. The tool is available at `https://github.com/jhnnsjkbk/accuracy-reliance`.

overrides as well as correct and wrong adherence:

$$\mathcal{A}_{correct} + \mathcal{O}_{correct} = Acc_{final}$$
$$\mathcal{A}_{correct} + \mathcal{A}_{wrong} = \mathcal{A}$$
$$\mathcal{A}_{correct} + \mathcal{O}_{wrong} = Acc_{AI}$$
$$\mathcal{A}_{correct} + \mathcal{A}_{wrong} + \mathcal{O}_{correct} + \mathcal{O}_{wrong} = 100\%.$$

Note that, at the moment, the tool is primarily developed for binary decision-making tasks. Finally, our implementation calculates the quality of reliance, Q. As defined in Equation (2) in Section 3.5, the quality of reliance represents the ability of the human-in-the-loop to discern correct from wrong AI recommendations given a level of adherence $\mathcal{A}$. Importantly, this measure is only dependent on $\mathcal{A}$ and the observed decision-making accuracy (Acc_{final}), but it is not influenced by the AI accuracy—facilitating comparisons of empirical findings across different employed AI models and thus across empirical studies.

We briefly explain the data aggregation process for the three studies. For **study 1**, we focus on the *in-distribution* condition on the ICPSR dataset and obtain scores for the two interventions called *static/static* and *interactive/interactive*. We obtain the adherence levels per intervention directly from the respective paper by Liu et al. (2021): $\mathcal{A} = 85.9\%$ (*static/static*) and $\mathcal{A} = 87.9\%$ (*interactive/interactive*). Similarly, we obtain the values for decision-making accuracy as $Acc_{final} = 61.5\%$ (*static/static*) and $Acc_{final} = 60.7\%$ (*interactive/interactive*). From that, and the reported *negative* accuracy gains from interventions, we infer the AI accuracy as follows: $Acc_{AI} = 61.5\% - (-6.2)\% = 67.7\%$ (*static/static*) and $Acc_{AI} = 60.7\% - (-6.3)\% = 67.0\%$ (*interactive/interactive*). Thus, we obtain an average AI accuracy of $Acc_{AI} = 67.4\%$. As we have access to the raw data for **study 2**, we directly use this data to determine accuracy and adherence values. Similarly, for **study 3**, we are able to directly infer the aggregated values for AI accuracy, adherence, and the final decision-making accuracy across the three conditions *weak AI*, *medium AI*, and *strong AI* from the publicly available raw data.

4.2.3 ANALYSIS OF STUDIES

In the following, we visualize the empirical findings from the selected studies with the help of our framework and present the results in Figure 6. The framework summarizes the key findings within a single visualization and extends them by additional observations. For instance, participants in **study 1** were performing worse than the AI alone in both cases (with static and interactive explanations). Notably, the participants' reliance behavior was similar to that of a random guess. The different types of explanations resulted both in very similar reliance behavior and decision-making accuracy. The degree of adherence to AI recommendations was slightly higher for intervention 2 (interactive explanations; light blue •) compared to intervention 1 (static explanations; dark blue •). However, in both cases, the level of adherence was significantly higher than the optimal level of adherence of 67.4%, as can be seen in Figure 6 (a). This means that participants *over-relied* on AI recommendations.

This is different in **study 2**, visualized in Figure 6 (b), where both the baseline condition and the interventions resulted in levels of adherence that are more aligned with the optimal level of adherence of 57.1%. Compared to the baseline condition (dark blue •), intervention 2
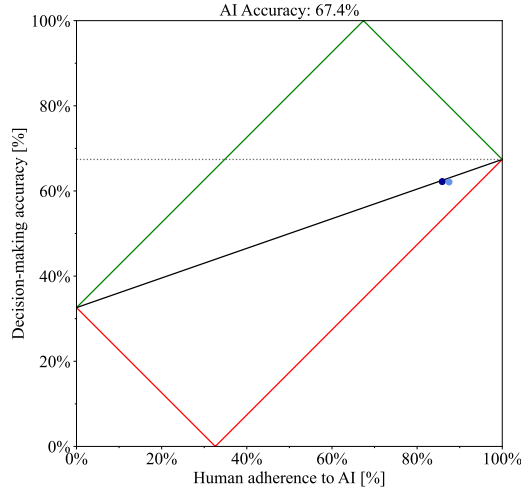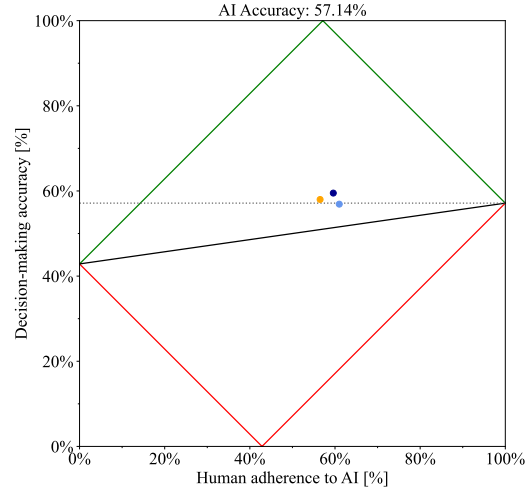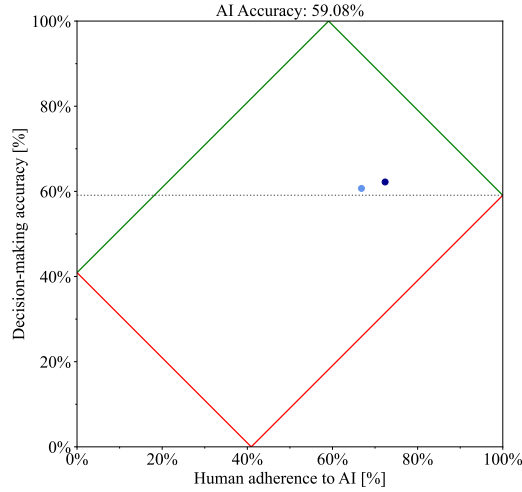
(a) **Study 1** (Liu et al., 2021).

(b) **Study 2** (Schoeffer et al., 2024).

(c) **Study 3** (Tejeda et al., 2022).

Figure 6: Visualization of empirical findings on the effects of different interventions (colored dots) on human reliance and decision-making accuracy using our proposed framework. Note that we do not show the level of random guessing (black line) in the multi-class setting (c).

(*gendered*; orange •) reduced the reliance quantity while intervention 1 (*task-relevant*; light blue •) increased it. We observe that intervention 2 resulted in a level of adherence closer to the optimum. That is, adherence was reduced from 59.6% in the baseline condition to 56.5% based on intervention 2, while the optimal level of adherence is given by 57.1%. Although both intervention 1 and 2 resulted in a reduced decision-making accuracy compared to the

baseline, our framework shows that intervention 2 improves over intervention 1 both in *quantity* of reliance and in *quality.*

For **study 3**, we observe that sequential decision-making (light blue •) resulted in a reduced decision-making accuracy compared to concurrent decision-making (dark blue •), as shown in Figure 6 (c). This is similar to the effect of intervention 2 in **study 2**. Additionally, adherence to AI recommendations decreased when moving from concurrent to sequential decision-making. Interestingly, this moves the degree of adherence closer to the optimal level of 59.1%. Hence, sequential decision-making improved the level of adherence but reduced the quality of reliance.

After having obtained a general overview of the empirical findings through Figure 6, we further utilize our framework to gain a more detailed understanding of the reliance behavior induced by different interventions. We summarize these insights per intervention for **study 1** and **study 2** in Table 2.[7] First, based on Equation (1) in Section 3, we derive information on the proportions of correct and wrong adherence as well as correct and wrong overrides. Overall, we observe that interventions affect both the quantity and quality of reliance across studies. For instance, as noted, the level of adherence is significantly higher in **study 1** than in **study 2**; however, the ratio of correct adherence to overall adherence is comparable: for intervention 1 of **study 1**, we have $\mathcal{A}_{correct}/\mathcal{A} = 66.8\%$, compared to 63.9% in the baseline condition of **study 2**. This implies that when participants followed AI recommendations, they relied on incorrect advice in roughly one-third of the cases. We may interpret the ratio $\mathcal{A}_{correct}/\mathcal{A}$ as the *precision* of adherence, and we further explore this and related metrics in Section 5.2.

We also see differences in overriding behavior across studies. Interestingly, participants in **study 1** overrode more correct than wrong AI recommendations (i.e., $\mathcal{O}_{wrong} > \mathcal{O}_{correct}$), suggesting a deficient ability to distinguish correct from wrong AI advice. This is also reflected in the quality of reliance, Q, which we calculate based on Equation (2) in Section 3.5.[8] Here, we observe that the quality of reliance is significantly lower in **study 1** compared to **study 2**. Since these differences in reliance quality persist across interventions, this suggests that general factors such as task complexity or the participants' backgrounds may vary between the two studies.

5. Extensions of the Framework

In the preceding sections, we outlined the key aspects of our proposed framework and applied it to study the effects of interventions. In this section, we derive relevant extensions of the framework. Specifically, we leverage our framework to estimate the likelihood of achieving human-AI complementarity (Section 5.1), and we introduce a possible evaluation of decision quality beyond accuracy based on F_β-scores (Section 5.2).

---

7. For multi-class decisions, our framework does not currently allow inferring whether adherence and overriding was correct or wrong.

8. Recall that our measure of reliance quality, Q, is independent of AI accuracy, enabling meaningful comparisons of empirical findings across different studies.

Study 1 (Liu et al., 2021)		Study 2 (Schoeffer et al., 2024)	
Intervention 1 (dark blue ●)	$\mathcal{A}_{correct} = 57.4\%$ $\mathcal{A}_{wrong} = 28.5\%$ $\mathcal{O}_{correct} = 4.1\%$ $\mathcal{O}_{wrong} = 10.0\%$ $Q = 0.29$	**Baseline** (dark blue ●)	$\mathcal{A}_{correct} = 38.1\%$ $\mathcal{A}_{wrong} = 21.5\%$ $\mathcal{O}_{correct} = 21.4\%$ $\mathcal{O}_{wrong} = 19.0\%$ $Q = 0.53$
Intervention 2 (light blue ●)	$\mathcal{A}_{correct} = 58.0\%$ $\mathcal{A}_{wrong} = 29.9\%$ $\mathcal{O}_{correct} = 2.7\%$ $\mathcal{O}_{wrong} = 9.4\%$ $Q = 0.22$	**Intervention 1** (light blue ●)	$\mathcal{A}_{correct} = 37.5\%$ $\mathcal{A}_{wrong} = 23.5\%$ $\mathcal{O}_{correct} = 19.4\%$ $\mathcal{O}_{wrong} = 19.6\%$ $Q = 0.50$
		Intervention 2 (orange ●)	$\mathcal{A}_{correct} = 35.8\%$ $\mathcal{A}_{wrong} = 20.7\%$ $\mathcal{O}_{correct} = 22.2\%$ $\mathcal{O}_{wrong} = 21.3\%$ $Q = 0.52$

Table 2: Empirical findings from Liu et al. (2021) and Schoeffer et al. (2024) interpreted through our framework.

5.1 Estimating the Likelihood of Achieving Complementarity

It is sometimes important to estimate the likelihood of realizing a decision-making accuracy that is greater than the AI accuracy, i.e., $Acc_{final} > Acc_{AI}$, meaning an ability of the human-in-the-loop to complement the AI. For that, let us recall from Corollary 1 that such complementarity cannot be achieved if the human under-relies on AI recommendations at $\mathcal{A} \leq 2 \cdot Acc_{AI} - 100\%$. For instance, when we have an AI accuracy of $Acc_{AI} = 70\%$, this adherence threshold is at $\mathcal{A} = 40\%$ (see, e.g., Figure 4 (a) in Section 3.3). That is, the human-in-the-loop can theoretically improve the decision-making accuracy over the initial AI accuracy if (*i*) they adhere to more than 40% of AI recommendations (or, equivalently, override *less than* 60%), and (*ii*) their reliance quality is sufficiently high.

To estimate the likelihood of achieving complementarity (i.e., $Acc_{final} > Acc_{AI}$), we can simulate the human reliance behavior as an urn model, where we have $n \in \mathbf{N}$ marbles that correspond to individual AI recommendations. The colors of marbles indicate their correctness (blue vs. orange), and drawing a marble corresponds to adhering to a given AI recommendation. This is illustrated for the case of $n = 10$ in Figure 7. Here, an AI issues $n = 10$ recommendations, 3 of which are wrong ($Acc_{AI} = 70\%$); i.e., we have 3 orange and 7 blue marbles in the urn. In this example, the human-in-the-loop then selects 7 marbles to indicate adherence to the respective AI recommendations ($\mathcal{A} = 70\%$), one of which is wrong. The remaining 3 marbles are left in the urn and imply that the corresponding AI recommendations (2 wrong and 1 correct) are overridden.

If the human-in-the-loop cannot distinguish correct from wrong AI recommendations, which means their reliance behavior is *independent* of the correctness of an AI recommen-
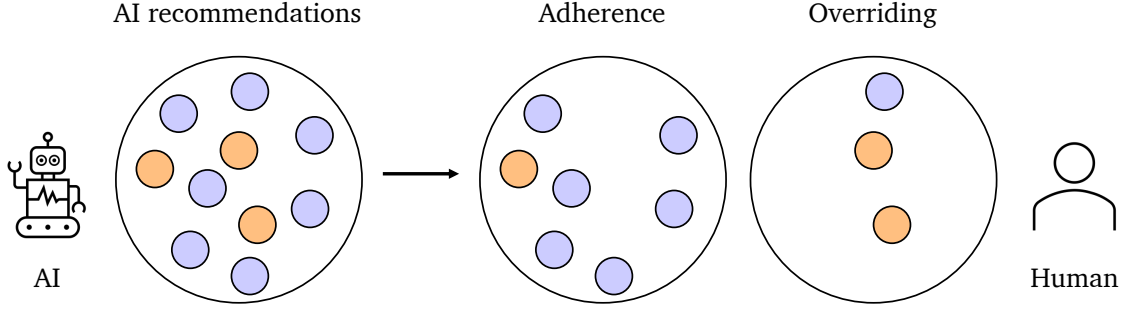
Figure 7: An exemplary human reliance on AI recommendations illustrated as marbles in an urn. In this example, we have $n = 10$, $Acc_{AI} = 70\%$, $\mathcal{A} = 70\%$, and a resulting decision-making accuracy of $Acc_{final} = 80\%$, consisting of 6 instances of correct adherence and 2 instances of correct overriding. Blue marbles are correct AI recommendations; orange marbles are wrong.

dation, we expect a decision-making accuracy of

$$Acc_{final}(\mathcal{A}) = (100\% - Acc_{AI}) + (2 \cdot Acc_{AI} - 100\%) \cdot \mathcal{A},$$

as stated in Proposition 3. We also know from Corollary 2 that this value is always lower than the AI accuracy, except for the case of $\mathcal{A} = 100\%$ where they are equal. However, even if the human-in-the-loop cannot distinguish correct from wrong AI recommendations, a decision-making accuracy of $Acc_{final} > Acc_{AI}$ may in practice be realized by sheer luck—especially when n is small. This has important implications for the interpretation of empirical findings. We now simulate different scenarios and infer the likelihood of observing complementarity.

For that, we model the human reliance behavior as sampling without replacement and without order from an urn that contains a finite number n of blue ($b \in \mathbf{N}$) and orange ($o \in \mathbf{N}$) marbles,[9] with $n = b + o$. The marbles are indistinguishable to the human. We know that the probability of drawing $k \in \mathbf{N}$ blue marbles out of $a = \lfloor \mathcal{A} \cdot n \rceil \in \mathbf{N}$ attempts is described by a hypergeometric distribution with parameters (n, b, a) and probability mass function

$$\Pr(k) = \frac{\binom{b}{k}\binom{n-b}{a-k}}{\binom{n}{a}}. \tag{3}$$

Coming back to our example from Figure 7, the relevant values are: $n = 10$, $b = 7$, $o = 3$, and $a = 7$. Recall that a is fixed ex ante. In order to achieve complementarity, the human-in-the-loop would have to choose either $k = 6$ or $k = 7$ blue marbles. Hence, the probability of observing complementarity given the marbles are not distinguishable is

$$\Pr(k \geq 6) = \Pr(k = 6) + \Pr(k = 7) = \frac{\binom{7}{6}\binom{3}{1}}{\binom{10}{7}} + \frac{\binom{7}{7}\binom{3}{0}}{\binom{10}{7}} = \frac{7}{40} + \frac{1}{120} = \frac{11}{60} \approx 18.3\%.$$

---

9. Note that the values of b and o follow immediately from a given AI accuracy: $b = \lfloor Acc_{AI} \cdot n \rceil$; $o = n - b$, where we use $\lfloor \cdot \rceil$ to denote rounding to the nearest integer.

$\mathcal{A}$ Acc_{AI}	30% $(a=3)$	40% $(a=4)$	50% $(a=5)$	60% $(a=6)$	70% $(a=7)$	80% $(a=8)$	90% $(a=9)$
60% $(b=6)$	16.7%	7.1%	26.2%	11.9%	33.3%	13.3%	40.0%
70% $(b=7)$			8.3%	3.3%	18.3%	2.2%	30.0%
80% $(b=8)$					6.7%	2.2%	4.4%
90% $(b=9)$							10.0%

Table 3: The likelihood of observing complementarity (i.e., $Acc_{final} > Acc_{AI}$) by chance, as a function of Acc_{AI} and $\mathcal{A}$, given $n = 10$. Empty cells correspond to cases where no complementarity is possible (see Corollary 1).

Hence, the likelihood of observing $Acc_{final} > Acc_{AI}$ by sheer chance in the given example is approximately 18.3%. This is also reflected in Table 3, which contains the likelihood of randomly achieving complementarity for all relevant cases of $n = 10$.

Based on this example, we now conduct a follow-up study to understand the impacts of Acc_{AI} and $\mathcal{A}$ on the likelihood of randomly observing complementarity. For that, we first establish in general how many blue marbles (i.e., correct adherence) the human-in-the-loop needs to choose in order to achieve complementarity.

Proposition 4 *Given* $\mathcal{A} \geq 2 \cdot Acc_{AI} - 100\%$, *achieving (strict) complementarity, i.e.,* $Acc_{final} > Acc_{AI}$, *is equivalent to* (*i*) *and* (*ii*):

(*i*) $0 \leq \mathcal{O}_{wrong} < \mathcal{O}_{correct}$

(*ii*) $Acc_{AI} - 50\% + 0.5 \cdot \mathcal{A} < \mathcal{A}_{correct} \leq \min\{Acc_{AI}, \mathcal{A}\}$.

Proposition 4 follows from Corollary 1 as well as Equation (1) in Section 3. Part (*i*) means that the human-in-the-loop *must* override at least one AI recommendation, and more overrides must be of wrong recommendations vs. correct ones. Part (*ii*) implies that given an AI accuracy Acc_{AI} and a sufficiently high level of adherence $\mathcal{A}$, the lower bound on the share of correct adherence that still leads to complementarity is $Acc_{AI} - 50\% + 0.5 \cdot \mathcal{A}$. For instance, if we have an AI accuracy of $Acc_{AI} = 70\%$ and a level of adherence $\mathcal{A} = 70\%$, then more than 55% of this adherence must be to correct AI recommendations (i.e., $\mathcal{A}_{correct} > 55\%$); which directly implies $\mathcal{A}_{wrong} < 15\%$, $\mathcal{O}_{wrong} < 15\%$, and $\mathcal{O}_{correct} > 15\%$. Proposition 4 also allows us to determine the values of k in Equation (3) for which complementarity is achieved.

We can now analyze the effects of Acc_{AI} and $\mathcal{A}$ on the likelihood of observing complementarity by chance. For clarity of exposition as well as computational tractability, we focus on $n = 100$. This means that our hypothetical urn contains 100 marbles, and we can treat one marble as one percentage point. The results are visualized in Figures 8 and 9. First, in Figure 8, we see that the likelihood of achieving complementarity increases in the degree of human adherence to AI recommendations, and that the likelihood is generally lower as the AI accuracy increases. Because the likelihoods are oscillating, we also provide moving averages to enhance readability. Importantly, in cases where the human-in-the-loop adheres to AI recommendations very often, the likelihood of achieving complementarity by

chance becomes significant. For instance, when we have $Acc_{AI} = 60\%$ and an adherence of $\mathcal{A} = 95\%$, this likelihood is approximately 31.4%. In other words, if humans cannot distinguish correct from wrong AI recommendations, they may increase their chances of achieving $Acc_{final} > Acc_{AI}$ by just adhering to more AI recommendations;[10] and these chances can become relatively high—especially when $Acc_{AI} \ll 100\%$. In such cases, it may sometimes seem that a human-in-the-loop is adding value to the system when in reality they are just over-relying on AI advice with no increased ability to distinguish correct from wrong AI recommendations. In Figure 9, we further see that the likelihood of randomly achieving complementarity decreases as the AI accuracy increases, and this decrease happens faster for low levels of adherence $\mathcal{A}$.
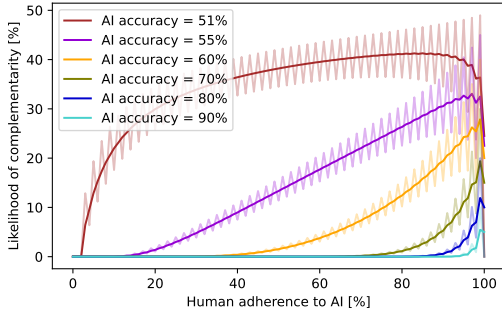

Figure 8: Likelihood of achieving complementarity (i.e., $Acc_{final} > Acc_{AI}$) when the human-in-the-loop cannot distinguish correct from wrong AI recommendations, as a function of the degree of human adherence, for different levels of AI accuracy. We also include moving averages.
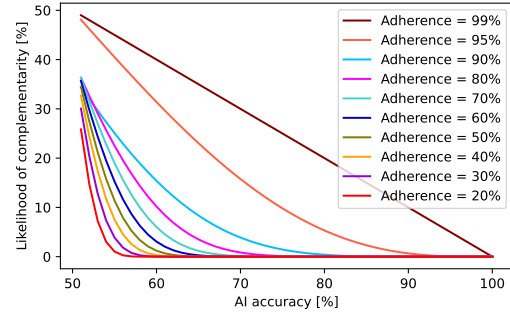
Figure 9: Likelihood of achieving complementarity when the human-in-the-loop cannot distinguish correct from wrong AI recommendations, as a function of the AI accuracy, for different levels of human adherence to AI recommendations.

5.2 Assessing Decision Quality with F_β-Scores

In the previous sections, we have primarily assessed decision quality using the accuracy metric, as it is the most widely adopted metric and allows for a clear and concise overview of our framework. However, the interdependence of reliance behavior and decision quality does not require an assessment based on accuracy but can be extended to other metrics that are a function of correct and wrong decisions, which we detail in the following.

When the underlying class distribution of the data is imbalanced, the accuracy metric is prone to misleading interpretations (Maratea et al., 2014). For this reason, more sophisticated metrics have evolved for evaluating the quality of predictions. One such metric, which involves a weighted combination of the *precision* and *recall* metrics (Goutte & Gaussier, 2005), is formally defined for $\beta \geq 0$ as follows:

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall},$$

---

10. Recall that we assume $Acc_{AI} > 50\%$, i.e., the AI performs better than chance.

where

$$precision = \frac{true\ positive}{true\ positive + false\ positive}, \quad recall = \frac{true\ positive}{true\ positive + false\ negative}.$$

The special case of $\beta = 1$, called F_1-score, defines a harmonic mean of precision and recall and is frequently used to assess the performance of machine learning models and associated systems on imbalanced datasets. The F_1-score is defined as

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

In the following, we demonstrate how the F_1-score can serve as a quality metric in AI-assisted decision-making and how it relates to reliance behavior. While, in general, we could compute the F_1-score also for the AI performance (alongside the final decision quality), it is not necessary for the subsequent considerations. Instead, we reinterpret the adherence to correct AI recommendations ($\mathcal{A}_{correct}$) as *true positive*, the overriding of wrong AI recommendation ($\mathcal{O}_{correct}$) as *true negative*, the adherence to wrong AI recommendations ($\mathcal{A}_{wrong}$) as *false positive*, and the overriding of correct AI recommendations ($\mathcal{O}_{wrong}$) as *false negative*. With these definitions in place, we can interpret *precision* as the fraction of correct AI recommendations among those that a human adheres to ($\mathcal{A}_{correct}/\mathcal{A}$); and we can think of *recall* as the fraction of correct AI recommendations that a human adheres to ($\mathcal{A}_{correct}/Acc_{AI}$).

For an illustration of the interdependence of reliance and decision quality based on the F_1-score, let us reconsider the example from Figure 3 in Section 3.1, with an AI accuracy of $Acc_{AI} = 70\%$. At an adherence level of $\mathcal{A} = 100\%$, the human-in-the-loop is correctly adhering to the AI recommendations in 7 out of 10 cases (i.e., *true positives*) and wrongly adhering to the recommendations in 3 out of 10 cases (i.e., *false positives*). Thus, for an AI accuracy of $Acc_{AI} = 70\%$ and adherence $\mathcal{A} = 100\%$, the corresponding final decision quality in terms of the F_1-score is

$$F_1 = \frac{2 \cdot 0.7 \cdot 1}{0.7 + 1} = 82.35\%.$$

With a level of adherence that is strictly lower than 100%, we again observe intervals that represent all attainable levels of decision quality. For instance, for an adherence level of $\mathcal{A} = 70\%$, we obtain $F_1 \in [57.14\%, 100\%]$. Thus, as with the assessment based on the accuracy metric, the best possible F_1-score can be obtained when the adherence level equals the accuracy of the AI system—making our conceptualization consistent across different metrics of decision quality. However, in contrast to the assessment based on accuracy, we observe non-linear boundaries of the area encompassing all attainable levels of decision quality resulting from the assessment via the F_1-score. For instance, for an adherence level of $\mathcal{A} = 60\%$, we have $F_1 \in [46.15\%, 92.31\%]$, and for an adherence level of $\mathcal{A} = 50\%$, we obtain $F_1 \in [33.33\%, 83.33\%]$. If the boundaries were linear, given the intervals for $\mathcal{A} = 50\%$ and $\mathcal{A} = 70\%$, we would expect $F_1 \in [45.24\%, 91.67\%]$ for $\mathcal{A} = 60\%$, which is in contrast to the actual calculated values $[46.15\%, 92.31\%]$. We visualize these non-linearities resulting from the assessment of decision quality based on the F_1-score in Figure 10 (a).

Additionally, in Figure 10 (b), we depict the attainable decision quality in terms of the $F_{0.5}$-score, which puts more weight on the precision than the recall (i.e., cases where false

positives are considered worse than false negatives). Thus, the $F_{0.5}$-score would consider adhering to wrong AI recommendations more problematic than overriding correct ones. We observe that based on the unequal weighting of precision and recall, the shape of the boundaries changes compared to the F_1-score, with wider intervals of attainable decision quality.
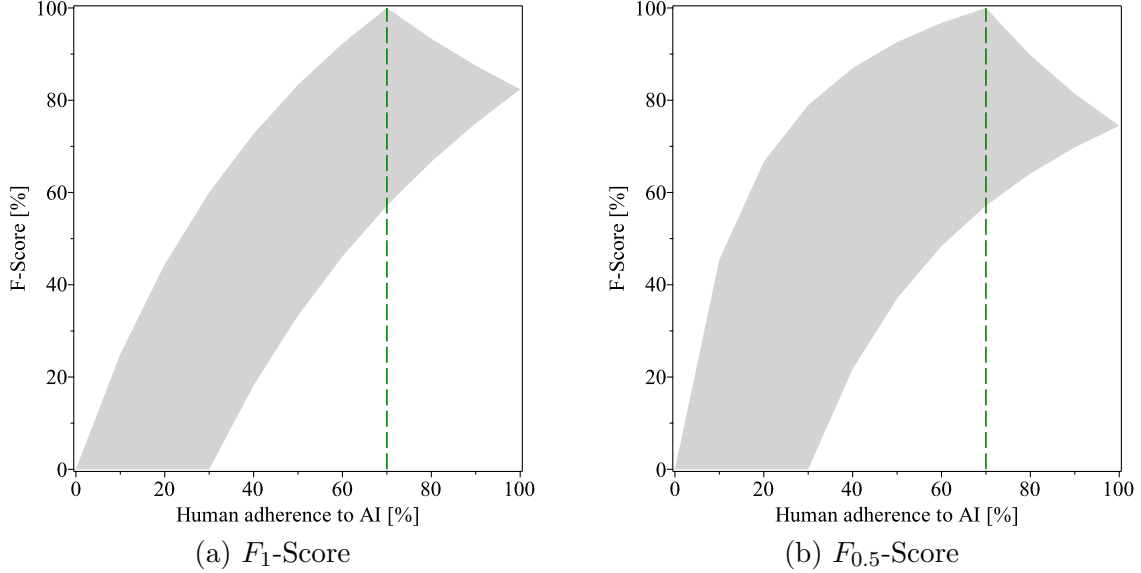

(a) F_1-Score   (b) $F_{0.5}$-Score

Figure 10: The area of attainable (a) F_1-scores and (b) $F_{0.5}$-scores for a given AI accuracy of 70% and different levels of human adherence.

6. Discussion and Conclusion

Summary of Findings and Implications   In this work, we study the relationship between reliance behavior and decision quality in AI-assisted decision-making. We show that any given *quantity* of human adherence to AI recommendations is associated with a specific range of attainable decision-making accuracy, depending on the *quality* of reliance, i.e., the ability of humans to adhere to AI recommendations if and only if they are correct. Vice versa, we also show that any accuracy level can be achieved through fundamentally different reliance behavior. This has implications for assessing the effectiveness of interventions, such as explanations, in AI-assisted decision-making. In particular, our work highlights the importance of disentangling and assessing *both* effects on accuracy *and* reliance behavior in order to derive meaningful implications on how interventions affect decision-making. For instance, by not capturing effects on reliance, we may conclude that an intervention led to an increase in accuracy, without understanding that this increase was driven solely by human over-reliance on AI advice.

We also characterize the conditions under which human-AI complementarity is achievable. In particular, we show that under- and over-reliance[11] are not symmetrical regarding their implications for complementarity. While complementarity is possible in the case

---

11. Recall that we define *under-reliance* globally as $\mathcal{A} < Acc_{AI}$, and *over-reliance* as $\mathcal{A} > Acc_{AI}$.

of over-reliance, there is no hope when the human-in-the-loop under-relies past a certain threshold. Notably, this threshold may be very high when the AI performs well. For instance, at an AI accuracy of 90%, any adherence to AI recommendations of less than 80% can *never* lead to a decision-making accuracy that is better than the AI baseline. Especially when the human-in-the-loop is not aware of such high AI performance, it might be unrealistic to expect complementarity. Conversely, our work demonstrates that complementarity can still occur in empirical studies even when humans cannot distinguish correct from incorrect AI recommendations—and the likelihood of this can be significant. This should be considered when interpreting empirical findings.

Finally, we propose a visual framework and demonstrate its practical relevance by applying it to prior empirical studies. Our framework manages to capture a range of previous findings within a single visualization and extends them by additional observations. For instance, it offers helpful additional information by disentangling reliance behavior in its basic components of correct/wrong adherence and overriding. The proposed framework also allows us to compare empirical findings across studies by contrasting the human ability to distinguish correct from wrong AI recommendations. Taken together, our work offers a blueprint and helpful metrics to guide the evaluation and design of effective interventions for AI-assisted decision-making.

Limitations and Future Work We acknowledge several assumptions that limit the scope of this work and allude to different areas that merit follow-up work. First, some of the key theoretical results in our study hinge on the assumption of a binary decision-making task. Although many critical decisions in the real world can be framed as binary, such as lending, hiring, or predicting recidivism, we recognize that certain tasks may involve more than two decision alternatives or even resemble regressions—e.g., in situations where a bank agent must decide on specific loan amounts to issue. While we demonstrate that our visual representation of the interplay between reliance and decision quality remains useful for interpreting empirical results in multi-class decision-making, a logical extension would be to generalize our theoretical findings to cases with more than two decision-making alternatives. In such scenarios, adjustments to our reliance taxonomy would be necessary to address situations where overriding a mistaken AI recommendation could still result in an incorrect decision. Moreover, extending our framework to regression tasks may warrant further exploration, albeit in a more comprehensive manner.

Additionally, the evaluation of reliance behavior can be either sequential or concurrent, meaning that a human-in-the-loop may sometimes make an initial guess before receiving AI advice (sequential) or make the final decision concurrently with receiving AI advice (concurrent), as discussed by Tejeda et al. (2022). We focus on the concurrent paradigm in this work, recognizing that sequential setups require a different conceptualization of reliance, as introduced, e.g., by Schemmer et al. (2023). As such, our work is complementing related literature on sequential AI-assisted decision-making, such as the works by Dunning et al. (2024) or Van Dongen and Van Maanen (2013). Finally, the decision-making setup that we consider is such that a human is relying on AI recommendations, i.e., either adhering to or overriding them. To this point, it could be interesting to build on previous research on "artificial trust" (Azevedo-Sa et al., 2021) and reverse this relationship between human and AI agent, such that AI agents are relying on human input.

It is also important to note that many of our theoretical results rely on the assumption that AI recommendations are more accurate than chance. This is motivated by prior work that has shown that unreliable automation may be worse than no automation at all (Wickens & Dixon, 2007); and an AI system that performs worse or at most as good as random guessing is an extreme case of unreliability. While we consider it a reasonable assumption for AI systems in deployment to perform better than chance, we acknowledge that in certain cases, this cannot be guaranteed ex ante.

Our visual framework is also limited in its ability to compare empirical findings across studies with different AI accuracy. Expanding it to account for varying AI accuracy would necessitate a three-dimensional visualization, introducing a third axis for Acc_{AI}, a task we defer to future work. Lastly, although it is often convenient to attribute initial recommendations to an AI system, such as a supervised machine learning model, our work makes no assumptions about the specific origin of these recommendations. They could originate from any form of technology, recommender system, or even another human being.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 6618–6626.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Azevedo-Sa, H., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2021). A unified bi-directional model for natural and artificial trust in human–robot collaboration. *IEEE Robotics and Automation Letters*, *6*(3), 5913–5920.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (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*, pp. 1–16.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (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), 1–21.

Cabitza, F., Campagner, A., Angius, R., Natali, C., & Reverberi, C. (2023). AI shall have no dominion: On how to measure technology dominance in AI-supported human decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20.

Cabrera, Á. A., Perer, A., & Hong, J. I. (2023). Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–21.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

de Visser, E., & Parasuraman, R. (2011). Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231.

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, *12*(2), 459–478.

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 275–285.

Dunning, R. E., Fischhoff, B., & Davis, A. L. (2024). When do humans heed AI agents' advice? When should they?. *Human Factors*, *66*(7), 1914–1927.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Echterhoff, J. M., Yarmand, M., & McAuley, J. (2022). AI-moderated decision-making: Capturing and balancing anchoring bias in sequential decision tasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–9.

Eckhardt, S., Kühl, N., Dolata, M., & Schwabe, G. (2024). A survey of AI reliance. *arXiv, arXiv:2408.03948*.

Fok, R., & Weld, D. S. (2024). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*, *45*(3), 317–332.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89. IEEE.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Goutte, C., & Gaussier, E. (2005). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval*, pp. 345–359. Springer.

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), 44–58.

Guo, Z., Wu, Y., Hartline, J. D., & Hullman, J. (2024). A decision theoretic framework for measuring AI reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 221–236.

Harris, J. G., & Davenport, T. H. (2005). Automated decision making comes of age. *MIT Sloan Management Review*, *46*(4).

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Hemmer, P., Schellhammer, S., Vössing, M., Jakubik, J., & Satzger, G. (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*, pp. 2478–2484.

Jakubik, J., Schoeffer, J., Hoge, V., Voessing, M., & Kuehl, N. (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*, pp. 353–368. Springer.

Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 4070–4073.

Kim, A., Yang, M., & Zhang, J. (2023). When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction*, *30*(1), 1–36.

Kuncel, N. R., Klieger, D. M., & Ones, D. S. (2014). In hiring, algorithms beat instinct. *Harvard Business Review*, *May 2014*.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1369–1385.

Lai, V., Liu, H., & Tan, C. (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*, pp. 1–13.

Lai, V., & Tan, C. (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*, pp. 29–38.

Langer, M., Baum, K., & Schlicker, N. (2024). Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. *Minds and Machines*, *35*(1), 1–30.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., et al. (2022). Combining the strengths of radiologists and AI for breast cancer screening: A retrospective analysis. *The Lancet Digital Health*, *4*(7), e507–e519.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*(4), 611–627.

Lima, G., Grgić-Hlača, N., & Cha, M. (2021). Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–17.

Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–45.

Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16.

Maratea, A., Petrosino, A., & Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences*, *257*, 331–341.

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *29*(4), 1–33.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Passi, S., & Vorvoreanu, M. (2022). Overreliance on AI: Literature review. Tech. rep., Microsoft.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–52.

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Reicin, E. (2021). AI can be a force for good in recruiting and hiring new employees. *Forbes*, *Leadership*.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95.

Riley, V. (2018). Operator reliance on automation: Theory and data. In *Automation and human performance*, pp. 19–35. CRC Press.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*(1), 76–87.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377–400.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 617–626.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 410–422.

Schoeffer, J., De-Arteaga, M., & Kuehl, N. (2024). Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–18.

Schoeffer, J., Jakubik, J., Voessing, M., Kuehl, N., & Satzger, G. (2023). On the interdependence of reliance behavior and accuracy in AI-assisted decision-making. In *HHAI 2023: Augmenting Human Intellect*, pp. 46–59. IOS Press.

Sheridan, T. B. (2019). Extending three existing models to analysis of trust in automation: Signal detection, statistical parameter estimation, and model-based control. *Human Factors*, *61*(7), 1162–1170.

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284.

Tejeda, H., Kumar, A., Smyth, P., & Steyvers, M. (2022). AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, *5*(4), 491–508.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404.

Van Dongen, K., & Van Maanen, P.-P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies*, *71*(4), 410–424.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–38.

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–39.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (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*, pp. 295–305.