HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230074

On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making

Jakob SCHOEFFER [a,*,1], Johannes JAKUBIK [a,*], Michael VOESSING [a], Niklas KUEHL [b] and Gerhard SATZGER [a]
[a] *Karlsruhe Institute of Technology (KIT), Germany*
[b] *Universität Bayreuth, Germany*

Abstract. In AI-assisted decision-making, a central promise of putting a human in the loop is that they should be able to complement the AI system by adhering to its correct and overriding its mistaken recommendations. In practice, however, we often see that humans tend to over- or under-rely on AI recommendations, meaning that they either adhere to wrong or override correct recommendations. Such reliance behavior is detrimental to decision-making accuracy. In this work, we articulate and analyze the interdependence between reliance behavior and accuracy in AI-assisted decision-making, which has been largely neglected in prior work. We also propose a visual framework to make this interdependence more tangible. This framework helps us interpret and compare empirical findings, as well as obtain a nuanced understanding of the effects of interventions (e.g., explanations) in AI-assisted decision-making. Finally, we infer several interesting properties from the framework: (i) when humans under-rely on AI recommendations, there may be no possibility for them to complement the AI in terms of decision-making accuracy; (ii) when humans cannot discern correct and wrong AI recommendations, no such improvement can be expected either; (iii) interventions may lead to an increase in decision-making accuracy that is solely driven by an increase in humans' adherence to AI recommendations, without any ability to discern correct and wrong. Our work emphasizes the importance of measuring and reporting *both* effects on accuracy *and* reliance behavior when empirically assessing interventions.

Keywords. AI-assisted decision-making; human-AI complementarity; reliance behavior; explanations; framework

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 having a human in the loop as the "last line of defense against AI failures" [1]. This assumes that humans can correct such AI failures in the first place. In human-in-the-loop settings, typically, an AI system generates an initial decision rec-

[1]Corresponding author: Jakob Schoeffer, jakob.schoeffer@kit.edu; * denotes equal contribution.

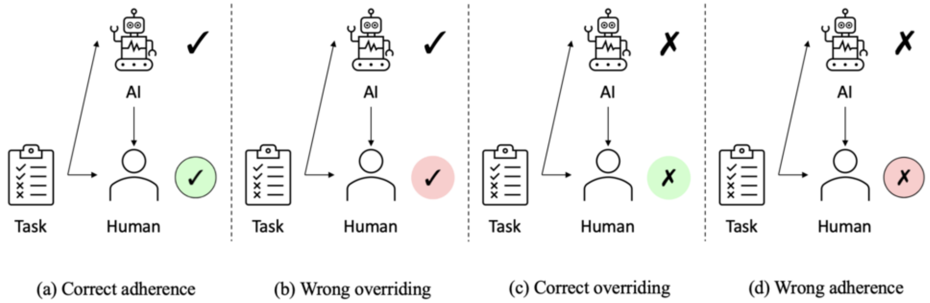


Figure 1. We consider *concurrent* AI-assisted decision-making setups where a human decision-maker receives a task and corresponding AI recommendation that can either be correct (✓) or wrong (✗), as indicated by the respective symbol next to the AI. The human can then either adhere to (bordered circle) or override (no border) the AI recommendation. When the human adheres to 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 green or red shading.

ommendation, which the human may either adhere to or override (see Figure 1). In the taxonomy of Tejeda et al. [2], this corresponds to *concurrent AI assistance*, where the human does *not* independently make a decision before AI assistance is provided. In order to complement the AI system, the human would have to adhere to its 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*[2] [1, 4, 5, 6]. Instead, we often observe that humans either over- or under-rely on AI recommendations, or simply cannot calibrate their reliance. Even the introduction of additional means of decision support (e.g., explanations) has often not shown the expected benefits in terms of enabling humans to complement AI systems. Worryingly, any root cause analyses are hindered by the fact that the mechanisms through which such interventions affect humans' reliance behavior are not well understood.

In this work, we make explicit and analyze the interplay of human reliance behavior on AI recommendations and decision-making accuracy, and we highlight the importance of assessing and reporting *both* in empirical studies on AI-assisted decision-making. To this end, we develop a framework that disentangles reliance *quantity* and *quality*, and lets us understand how both—individually and in conjunction—affect decision-making accuracy. We also visualize these interdependencies geometrically, which aims at making them easier to grasp. The visual framework is ultimately intended to serve researchers for interpreting empirical findings, including the effects of interventions, in AI-assisted decision-making. It may also be used by practitioners to reflect on their reliance behavior when interacting with AI-assisted decision-making systems.

From our theoretical analyses, we infer several interesting properties: **first**, we show that over- and under-reliance are not symmetric with respect to their effects on decision-making accuracy. Specifically, when humans adhere too little to recommendations from an AI system that performs better than chance, it is impossible to improve decision-making accuracy over the AI baseline. **Second**, when humans are unable to distinguish correct from wrong AI recommendations, i.e., when their reliance behavior is indepen-

[2]We use *reliance* as an umbrella term for humans' behavior of adhering to or overriding AI recommendations [3].

dent of the correctness of AI recommendations, we cannot expect humans to complement an AI system, either. In such cases, we also see that "blindly" adhering more to AI recommendations increases the expected decision-making accuracy—without any improved ability to discern correct and wrong recommendations. Finally, **third**, we show that interventions may affect decision-making accuracy through drastically different effects on reliance. For instance, two different interventions may lead to an identical increase in accuracy, but one may do so through *decreasing* human adherence to AI recommendations, whereas the other may lead to an *increase* in adherence. Both interventions may look identically effective when not analyzing effects on reliance behavior at the level that we propose. These insights are crucial for designing meaningful decision support measures.

2. Background

Measuring and calibrating the human reliance on AI recommendations has become a central pillar of research on AI-assisted decision-making [1, 3, 4]. This is especially important as both humans and AI systems are imperfect "decision-makers" with individual strengths and weaknesses [6, 7, 8]. For humans that are assisted by AI, it is therefore essential to be able to identify strengths and weaknesses of the AI system (i.e., in which cases it is correct and in which wrong, see [9]). In this setting, latest research distinguishes three cases of reliance behavior: (*i*) relying on AI recommendations in too few cases (i.e., *under-reliance*, see [10, 11], e.g., by underestimating AI performance), (*ii*) relying on AI recommendations in too many cases (i.e., *over-reliance*, see [1, 12, 13], e.g., by overestimating AI performance), and (*iii*) relying *appropriately* on AI recommendations (i.e., adhering to AI recommendations when correct and overriding when wrong, see [5, 9, 14]). Thus far, research has identified many scenarios in which under-reliance or over-reliance results in reduced decision-making performance (e.g., [12, 15]). In an emerging effort, more works are developed around achieving an appropriate level of reliance, which is a prerequisite for the human decision-maker to complement the AI system and ultimately improve the overall decision-making accuracy over the AI baseline [4]. In this work, we develop a framework that aims at improving our understanding of how human reliance behavior translates to decision-making accuracy.

Accuracy of AI-assisted decision-making (i.e., the number of correct decisions given the overall number of decisions) represents a key metric that may indicate the utility of an AI-assisted decision-making system—apart from other metrics such as fairness [5]. The accuracy metric is hence frequently used for measuring the performance of AI-assisted decision-making systems [3] and evaluating the effectiveness of interventions (e.g., explanations) for decision support [3, 16, 17, 18]. Overall, we observe that research has typically focused on either the performance in terms of accuracy [3, 15] or on the human behavior in terms of their reliance on AI recommendations [11, 12], when assessing effects of interventions. However, in AI-assisted decision-making, accuracy is immediately influenced by the degree to which humans adhere to or override AI recommendations, and *how* they do so [19]. In this work, we show that the relationship between reliance behavior and accuracy follows clear mathematical patterns, and that measuring either decision-making accuracy or the level of reliance alone may provide an incomplete view when assessing AI-assisted decision-making generally and the effects of interventions specifically.

Table 1. We distinguish four cases of human reliance behavior in binary AI-assisted decision-making.

	Correct AI	Wrong AI
Adherence to AI	Correct adherence ($\mathscr{A}_{correct}$)	Wrong adherence ($\mathscr{A}_{wrong}$)
Overriding of AI	Wrong override ($\mathscr{O}_{wrong}$)	Correct override ($\mathscr{O}_{correct}$)

3. The Interdependence of Reliance Behavior and Accuracy

For clarity of exposure, 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[3], and $\mathscr{A} \in [0\%, 100\%]$ the degree of human adherence to AI recommendations—e.g., $\mathscr{A} = 70\%$ when the human adheres to 70% of AI recommendations. As introduced in Figure 1, adherence can be correct ($\mathscr{A}_{correct}$) or wrong ($\mathscr{A}_{wrong}$), and we have $\mathscr{A} = \mathscr{A}_{correct} + \mathscr{A}_{wrong}$. Similarly, we call the number of overrides $\mathscr{O} \in [0\%, 100\%]$ (correct: $\mathscr{O}_{correct}$ or wrong: $\mathscr{O}_{wrong}$), and we have $\mathscr{O} = \mathscr{O}_{correct} + \mathscr{O}_{wrong}$. While in practice humans can only adhere to or override an integer number of AI recommendations, we often consider $n \to \infty$ for our theoretical considerations, so as to avoid rounding. We summarize the possible cases of adhering and overriding AI recommendations in Table 1. Note that by definition we also have:

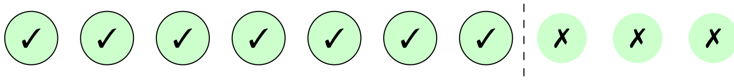
$$\begin{aligned}
\mathscr{A} + \mathscr{O} &= \mathscr{A}_{correct} + \mathscr{A}_{wrong} + \mathscr{O}_{correct} + \mathscr{O}_{wrong} = 100\% \\
Acc_{AI} &= \mathscr{A}_{correct} + \mathscr{O}_{wrong} \\
Acc_{final} &= \mathscr{A}_{correct} + \mathscr{O}_{correct}.
\end{aligned} \tag{1}$$

3.1. Motivational Example

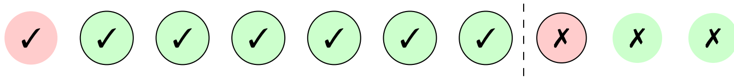
Consider the following motivational example: we have a task that consists of making $n = 10$ binary decisions. The AI system that is used for providing decision recommendations to the human has an accuracy of $Acc_{AI} = 70\%$; i.e., 7 out of 10 recommendations are correct (✓) and 3 are wrong (✗). Now, when the human adheres to all AI recommendations ($\mathscr{A} = 100\%$), 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 correctly adhered to 7 correct AI recommendations ($\mathscr{A}_{correct} = 70\%$), and wrongly adhered to the remaining 3 recommendations ($\mathscr{A}_{wrong} = 30\%$). In the other extreme case where the human overrides all AI recommendations ($\mathscr{O} = 100\%$), the resulting decision-making accuracy will be $100\% - 70\% = 30\%$, where the human correctly overrides 3 wrong AI recommendations ($\mathscr{O}_{correct} = 30\%$), and wrongly overrides 7 correct AI recommendations ($\mathscr{O}_{wrong} = 70\%$).

If the human reliance behavior is mixed, i.e., when the human adheres to some AI recommendations and overrides others, decision-making accuracy will depend on how well the human can distinguish cases where the AI is correct from cases where it is wrong. To make this clear, consider the same AI as above with an accuracy of 70%, and a human that adheres to 7 out of 10 of its recommendations ($\mathscr{A} = 70\%$). This is illustrated in Figure 2.

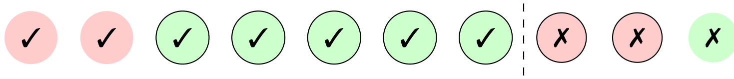
[3]Note that we only consider cases where the AI performs strictly better than chance.



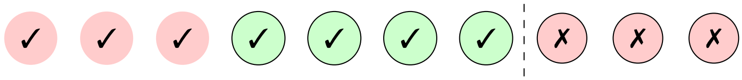
(a) $Acc_{final} = 100\%$



(b) $Acc_{final} = 80\%$



(c) $Acc_{final} = 60\%$



(d) $Acc_{final} = 40\%$

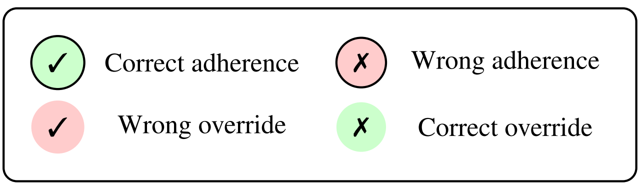


Figure 2. Possible scenarios of reliance behavior and associated decision-making accuracy, given an AI accuracy of $Acc_{AI} = 70\%$ and an adherence level of $\mathscr{A} = 70\%$. Correct AI recommendations (✓) and wrong AI recommendation (✗) are separated by a dashed line.

If the human is able to perfectly distinguish between correct and wrong AI recommendations, they will adhere to all 7 correct AI recommendations ($\mathscr{A}_{correct} = 70\% = \mathscr{A}$) and override the 3 wrong ones ($\mathscr{O}_{correct} = 30\% = \mathscr{O}$). The resulting decision-making accuracy would then be $Acc_{final} = 100\%$ (case (a) in Figure 2). In this case, the human is able to perfectly complement the AI by correcting for its mistakes. Cases (b)–(d) in Figure 2 show situations where the human still adheres to 70% of AI recommendations but their ability to override wrong AI recommendations decreases. For instance, consider case (d), where the human does not perform any correct overrides ($\mathscr{O}_{correct} = 0$). When the human degree of 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 2, we can also infer that if the human overrides *more* than 3 AI recommendations, at least one of these overrides must be wrong (i.e., the human 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*. In the general case, we may think of under-reliance as a behavior where $\mathscr{A} < Acc_{AI}$, and over-reliance as $\mathscr{A} > Acc_{AI}$. Note that there exists other work that has been thinking of these terms with respect to behavior at the level of individual decisions [4, 13].

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 can override the AI recommendations when they are wrong and adhere to them when they are correct. In Figure 2, 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 $\mathscr{A} = 70\%$. As mentioned earlier, we generally consider $n \to \infty$, in which case this range becomes 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 $\mathscr{A}$, the range of attainable decision-making accuracy Acc_{final} is*

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

The maximum of this accuracy range will be attained whenever the human maximizes correct adherence and correct overrides given a degree of adherence $\mathscr{A}$, since $Acc_{final} = \mathscr{A}_{correct} + \mathscr{O}_{correct}$. Hence, in the ideal case, we would have $\mathscr{A}_{correct} + \mathscr{O}_{correct} = 100\%$; which immediately implies that $\mathscr{A}_{wrong} = \mathscr{O}_{wrong} = 0\%$. This would be case (a) in Figure 2. However, as we can see in Proposition 1, this is only possible when $\mathscr{A} = \mathscr{A}_{correct} = Acc_{AI}$, 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 things:

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

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

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 2, 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 Figures 3 ($Acc_{AI} = 70\%$) and 4 ($Acc_{AI} = 90\%$). On the horizontal axes we have the human adherence to AI recommendations, $\mathscr{A} \in [0, 100\%]$. The vertical axes show the decision-making accuracy, $Acc_{final} \in [0, 100\%]$. The filled rectangular area in red and green combined constitutes the attainable decision-making accuracy for any given $\mathscr{A}$. We distinguish red and green to highlight areas where the human in the loop complements the AI (green, $Acc_{final} > Acc_{AI}$) or impairs it (red, $Acc_{final} < Acc_{AI}$) regarding accuracy. The green dashed vertical line

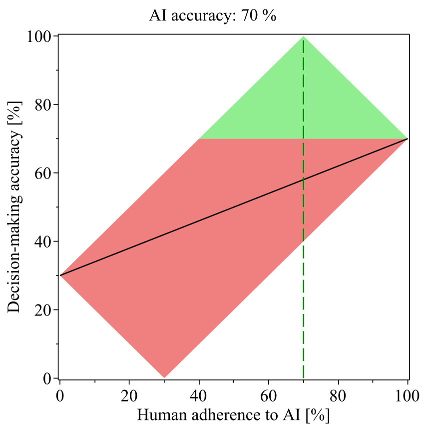


Figure 3. The area of attainable decision-making accuracy for a given AI accuracy of 70%. The red area indicates $Acc_{final} < Acc_{AI}$; green indicates $Acc_{final} > Acc_{AI}$; the green dashed line indicates the level of adherence where $Acc_{final} = 100\%$ is attainable; the black line indicates the expected Acc_{final} when humans cannot discern correct and wrong.

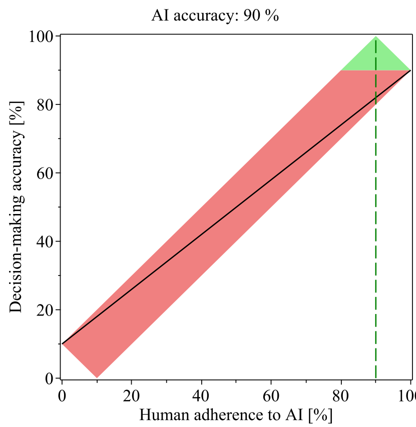


Figure 4. The area of attainable decision-making accuracy for a given AI accuracy of 90%. The red area indicates $Acc_{final} < Acc_{AI}$; green indicates $Acc_{final} > Acc_{AI}$; the green dashed line indicates the level of adherence where $Acc_{final} = 100\%$ is attainable; the black line indicates the expected Acc_{final} when humans cannot discern correct and wrong.

indicates the level of $\mathscr{A} = Acc_{AI}$, 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 3 $\rightarrow$ Figure 4), the colored area decreases; and for $Acc_{AI} = 100\%$ it becomes a line, in which case Proposition 1 collapses into $Acc_{final} = \mathscr{A}$.

Contrasting the red and green areas, we immediately see that up to a certain level of $\mathscr{A}$ there is no possibility to reach the green area, where $Acc_{final} > Acc_{AI}$. We also see that the minimum level of $\mathscr{A}$ for which the human in the loop may complement the AI increases as Acc_{AI} increases ($\mathscr{A} = 40\%$ in Figure 3 $\rightarrow$ $\mathscr{A} = 80\%$ in Figure 4). Finally, when $\mathscr{A} \geq Acc_{AI}$, attaining a decision-making accuracy in the green area is always possible. We characterize this in the following corollary:

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

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

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

However, fixing Acc_{final} at 70%, different levels of $\mathscr{A}$ correspond to different vertical positions within the rectangle: $\mathscr{A} = 40\%$ corresponds to a position at the very northern border of the rectangle, whereas any $\mathscr{A} \in [70\%, 100\%]$ corresponds to a position on the horizontal line that separates the red and green areas. This means that a given decision-making accuracy can be achieved through strikingly different reliance behaviors. We address this, as well as the role of the black separating lines in Figures 3 and 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 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 $\mathscr{A}$, we would expect the human to adhere to $\mathscr{A}\%$ of correct AI recommendations and $\mathscr{A}\%$ of wrong AI recommendations. At $Acc_{AI} = 70\%$, this implies that $\mathscr{A}\%$ of 70% are correct adherences, $\mathscr{A}\%$ of 30% are wrong adherences, $(100 - \mathscr{A})\%$ of 70% are wrong overrides, and $(100 - \mathscr{A})\%$ of 30% are correct overrides. When we have $\mathscr{A} = 70\%$, this would imply $\mathscr{A}_{correct} = 49\%$, $\mathscr{A}_{wrong} = 21\%$, $\mathscr{O}_{correct} = 9\%$, and $\mathscr{O}_{wrong} = 21\%$, with a decision-making accuracy of $\mathscr{A}_{correct} + \mathscr{O}_{correct} = 58\%$. This corresponds to the intersection of the black line with the dashed green vertical line in Figure 3. We generalize this in the following proposition.

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

$$\begin{aligned} Acc_{final}(\mathscr{A}) &= \mathscr{A} \cdot Acc_{AI} + (100\% - \mathscr{A}) \cdot (100\% - Acc_{AI}) \\ &= (100\% - Acc_{AI}) + \underbrace{(2 \cdot Acc_{AI} - 100\%)}_{>0} \cdot \mathscr{A}, \end{aligned}$$

for a given AI accuracy Acc_{AI}.

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

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

Having established the expected decision-making accuracy when humans are 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 humans adhere to and override 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 levels of adherence to AI recommendations, and we characterize the reliance behavior that is associated with the maximum and minimum decision-making accuracy for given $\mathscr{A}$.

Case: $0 \leq \mathscr{A} \leq 100\% - Acc_{AI}$ Since we assume that $Acc_{AI} > 50\%$, we have $\mathscr{A} < Acc_{AI}$ 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 overrides a correct AI recommendation, i.e., $\mathscr{O}_{wrong} > 0$. From Proposition 1 we also see that the **maximum** achievable decision-making accuracy in that case is $100\% - Acc_{AI} + \mathscr{A}$, which is achieved when $\mathscr{A}_{correct} = \mathscr{A}$. Using the definition of $\mathscr{A}$ and relationships from (1), this directly implies that $\mathscr{A}_{wrong} = 0$, $\mathscr{O}_{correct} = 100\% - Acc_{AI}$, and $\mathscr{O}_{wrong} = Acc_{AI} - \mathscr{A} > 0$. The **minimum** achievable decision-making accuracy, on the other hand, is attained when adherence only happens to wrong AI recommendations, hence, $\mathscr{A}_{wrong} = \mathscr{A}$. Similar to above, we this implies that $\mathscr{A}_{correct} = 0$, $\mathscr{O}_{wrong} = Acc_{AI}$, and $\mathscr{O}_{correct} = 100\% - Acc_{AI} - \mathscr{A}$.

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

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

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

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

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

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

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

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 $Q(\mathscr{A}) \in [0,1]$ for the quality of reliance given Acc_{AI}, such that a value of 1 corresponds to the maximum attainable decision-making accuracy, and 0 to the minimum. First, we derive the following corollary from Proposition 1:

Corollary 6 *The width W of the range of attainable values for Acc_{final} is:*

$$W = \begin{cases} 2 \cdot \mathscr{A} & \text{if } 0 \leq \mathscr{A} \leq 100\% - Acc_{AI} \\ 2 \cdot (100\% - Acc_{AI}) & \text{if } 100\% - Acc_{AI} < \mathscr{A} \leq Acc_{AI} \\ 2 \cdot (100\% - \mathscr{A}) & \text{if } Acc_{AI} < \mathscr{A} \leq 100\%. \end{cases}$$

Geometrically, W corresponds to the distance between the upper and lower vertical boundary of the rectangle (see, e.g., Figures 3 and 4) for a fixed $\mathscr{A}$. With that, we can define our metric $Q(\mathscr{A})$ as follows:

$$Q(\mathscr{A}) := \begin{cases} \dfrac{(\mathscr{A}_{correct} + \mathscr{O}_{correct}) - (100\% - Acc_{AI} - \mathscr{A})}{W} & \text{if } 0 \leq \mathscr{A} \leq 100\% - Acc_{AI} \\ \dfrac{(\mathscr{A}_{correct} + \mathscr{O}_{correct}) + (100\% - Acc_{AI} - \mathscr{A})}{W} & \text{if } 100\% - Acc_{AI} < \mathscr{A}. \end{cases} \quad (2)$$

Note that since Acc_{AI} and $\mathscr{A}$ are fixed, maximizing the quality of reliance, $Q(\mathscr{A})$, corresponds to maximizing $\mathscr{A}_{correct} + \mathscr{O}_{correct} = Acc_{final}$, and we have seen what this entails in terms of reliance behavior for any value of $\mathscr{A}$ in the previous subsection. Note that $Q(\mathscr{A})$ is not constant in cases where humans cannot discern correct and wrong AI recommendations. In this case, using Proposition 3, we obtain $Q(30\%) = 0.7$, whereas $Q(70\%) = 0.3$. We may think of this as follows: while $\mathscr{A} = 70\%$ leads to a higher expected decision-making accuracy of $Acc_{final} = 58\%$ (vs. $Acc_{final} = 42\%$ for $\mathscr{A} = 30\%$), the attainable accuracy in either case is $[40\%, 100\%]$ at $\mathscr{A} = 70\%$ and $[0, 60\%]$ in the case of $\mathscr{A} = 30\%$. Hence, the accuracy relative to the "potential" is much worse in the case of $\mathscr{A} = 70\%$. This will be relevant in the following section.

4. Understanding the Effects of Interventions

Our visual framework can be used to depict empirical results in AI-informed 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 adherence and overrides, or the ability of the human to complement or not the AI. Another key usage of the framework is its ability to understand and disentangle the effects of interventions, such as explanations or other means of decision support [3]. For that, let us consider the following hypothetical example: through a randomized experiment, we have collected data where humans are making decisions in the presence of two different types of explanations (● and ●) vs. a baseline without explanations (●). We can think of these interventions as movements in our visual framework, as depicted in Figure 5. The black point corresponds to a situation where a human cannot discern correct and wrong AI recommendations and adheres to $\mathscr{A} = 50\%$. Now, 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 ($\mathscr{A} = 50\% \rightarrow \mathscr{A} = 30\%$), 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—where this intervention leads to an *increase* in adherence to AI recommendations ($\mathscr{A} = 50 \rightarrow \mathscr{A} = 90\%$). At the same time, reliance quality drops from 0.5 to 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. Finally, note that since the purple point lies below the black line, this corresponds to reliance behavior that is of lower quality according to (2) than in cases where the human decides at random which AI recommendations to adhere to or override. This implies that different interventions can have seemingly similar effects on decision-making accuracy but drastically different effects on reliance behavior. Our framework enables us to disentangle these dimensions.

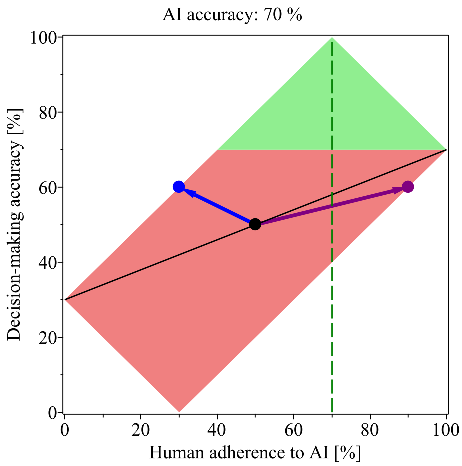


Figure 5. Visualizing the effects of different interventions (● and ●) on reliance behavior and decision-making accuracy.

5. Discussion and Conclusion

In this work, we propose a framework to understand and analyze the interdependence between reliance behavior and decision-making accuracy in AI-assisted decision-making.

We show that any given *quantity* of humans' adherence to AI recommendations is associated with a specific range of attainable decision-making accuracy, depending on the *quality* of reliance, i.e., humans' ability 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, both in terms of reliance quantity and quality. This has implications for assessing the effectiveness of interventions, such as explanations or other forms of decision support, in AI-assisted decision-making. For instance, our work highlights the importance of assessing and reporting *both* effects on accuracy *and* reliance behavior in order to derive meaningful implications on how such interventions affect decision-making. Specifically, we show an example of how assessing only effects on accuracy may lead to the wrong conclusion that an intervention was not effective when in reality it changed reliance behavior significantly. Even more worryingly, by not measuring or reporting effects on reliance behavior, we may conclude that an intervention led to an increase in decision-making accuracy, without understanding that this increase was driven solely by an increase in adherence *quantity* while the ability to discern correct and wrong AI recommendations dropped.

We also infer interesting properties when the human in the loop cannot discern correct and wrong AI recommendations, i.e., when the probability of adhering to or overriding a given AI recommendation is independent of the correctness of AI recommendations. In practice, this may occur when a task is too difficult for the human to solve. In such cases, we show that the human may never be expected to complement the AI, meaning that the decision-making accuracy will be strictly lower than the initial AI accuracy—except when the human adheres to *all* AI recommendations, in which case the decision-making accuracy will be equal to the AI accuracy. Another interesting implication of this analysis is that expected decision-making accuracy is linearly increasing in the quantity of adherence to AI recommendations, i.e., decision-making accuracy may be increased by solely adhering to more AI recommendations. This must be considered when interpreting empirical findings.

Finally, we infer that under- and over-reliance[4] is not symmetrical regarding their implications for decision-making accuracy. While the human may complement the AI when over-relying by systematically adhering to correct recommendations and overriding wrong ones, there is no hope for improvements in accuracy over the AI baseline when the human under-relies past a threshold of $\mathscr{A} < 2 \cdot Acc_{AI} - 100\%$. Notably, this threshold may be very high when the AI performs well—for instance, at an AI accuracy of 90%, any adherence $\mathscr{A} < 80\%$ can *never* lead to a decision-making accuracy that is better than the AI. Especially when the human in the loop is not aware of such high AI performance, it might be unrealistic to expect them to complement the AI.

Our framework is currently applicable to binary decision-making tasks with an AI system in place that performs better than chance. A natural extension would be to include cases with more than two decision alternatives. In such cases, our reliance taxonomy would have to be altered to account for situations where overriding a wrong AI recommendation may still lead to a wrong decision. Our visual framework is also limited in its use for situations where we want to compare empirical findings across studies with *different* AI accuracy. Extending it to account for varying AI accuracy would involve a 3-dimensional visual with a third axis on Acc_{AI}. Finally, we might think of cases where the metric of decision-making performance is not accuracy but, for instance, fairness [5].

[4]Recall that we define *under-reliance* globally as $\mathscr{A} < Acc_{AI}$, and vice versa for over-reliance.

References

[1] Samir Passi and Mihaela Vorvoreanu. Overreliance on AI: Literature review. *Technical Report MSR-TR-2022-12. Microsoft*, 2022.

[2] Heliodoro Tejeda, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, pages 1–18, 2022.

[3] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-AI decision making: A survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

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

[5] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. On explanations, fairness, and appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2209.11812*, 2022.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*, pages 1–16, 2021.

[7] Patrick Hemmer, Max Schemmer, Michael Voessing, and Niklas Kuehl. Human-AI complementarity in hybrid intelligence systems: A structured literature review. *PACIS*, page 78, 2021.

[8] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

[9] Max Schemmer, Patrick Hemmer, Niklas Kuehl, Carina Benz, and Gerhard Satzger. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916*, 2022.

[10] Sebastian W Schuetz, Zachary R Steelman, and Rhonda A Syler. It's not just about accuracy: An investigation of the human factors in users' reliance on anti-phishing tools. *Decision Support Systems*, 163:113846, 2022.

[11] Zhuoran Lu and Ming Yin. 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*, pages 1–16, 2021.

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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, 2021.

[13] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on AI systems during decision-making. *arXiv preprint arXiv:2212.06823*, 2022.

[14] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. AI-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–27, 2021.

[15] Antino Kim, Mochen Yang, and Jingjing Zhang. 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, 2023.

[16] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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*, pages 295–305, 2020.

[17] Vivian Lai, Han Liu, and Chenhao Tan. "Why is 'Chicago' deceptive?" Towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[18] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. Improving human-AI collaboration with descriptions of AI behavior. *arXiv preprint arXiv:2301.06937*, 2023.

[19] Johannes Jakubik, Jakob Schoeffer, Vincent Hoge, Michael Voessing, and Niklas Kuehl. An empirical evaluation of predicted outcomes as explanations in human-AI decision-making. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part I*, pages 353–368. Springer, 2023.