Int. J. Human–Computer Studies 210 (2026) 103763

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

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

Under what influence: Measuring AI influence to fit user profiles in decision-making

Andrea Campagner[a,b], Caterina Fregosi[a], Chiara Natali[a], Federico Cabitza[a,c,*]

[a] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy
[b] IRCCS Ospedale Galeazzi Sant'Ambrogio, via Cristina Belgioioso 173, Milano, 20157, Italy
[c] Digital Health and Wellbeing Center, Fondazione Bruno Kessler (FBK), Via Sommarive 18, Povo, Trento, 38123, Italy

HIGHLIGHTS

- Developed novel metrics to evaluate AI's influence beyond user agreement with AI.
- Identified biases impacting AI influence, such as self-anchoring and automation bias.
- Applied framework to four medical studies with 330 clinicians and 15,000 decisions.
- Revealed up to 81% alignment but variances in appropriate reliance and influence.
- Highlighted need for adaptive AI systems to match user expertise.
- Demonstrated that influence metrics uncover dynamics missed by traditional reliance.

ARTICLE INFO

Keywords:
Calibrated trust
Appropriate reliance
Automation bias
Artificial intelligence
Decision support systems
Influence

ABSTRACT

Artificial Intelligence (AI) has become a pivotal tool in augmenting human decision-making across various domains, yet its influence on user decisions often lacks comprehensive evaluation. While technical performance metrics such as accuracy and efficiency dominate AI design, integrating human-centered approaches that consider trust and reliance remains underexplored. This study addresses the knowledge gap in understanding how AI systems influence decision-making quality, calibrated to user profiles, including their expertise, skills, professional role, confidence, and reliance tendencies.

We present a novel and comprehensive metric framework for evaluating AI influence, emphasizing behavioral patterns and measurable improvements in decision outcomes beyond simple alignment with AI recommendations. The framework is applied to four medical domain case studies—MRI, ECG, X-ray, and ENDO – with user groups spanning specialists, sub-specialists, and trainees. Results reveal that while human and AI systems achieve high agreement rates (up to 81%), AI influence on decision quality varies significantly. Notably, X-ray decision-making showed the highest influence index (0.27), while MRI decisions exhibited substantial self-anchoring bias (6.94), undermining the potential positive impact of AI. Influence metrics unveiled nuances missed by agreement scores, highlighting domain-specific biases and opportunities to optimize AI-human interaction.

This research underscores the necessity of adapting the type of AI system and affordances to user characteristics and attitudes of reliance to foster calibrated trust and improve decision outcomes. Our findings inform the design of AI systems that better support diverse user needs and align with human decisions, driving progress toward human-centered AI integration in high-stakes domains.

1. Introduction

One of the key tenets of the human-centered approach to developing Artificial Intelligence (AI) systems is ensuring that the system's characteristics (e.g., the type of output it generates and the manner in which it is presented) and modes of interaction (e.g., when and for what purpose it is used) *fit* the user (Carroll and Rosson, 1992), meaning they are aligned with the user's characteristics, objectives, and practices. This alignment is intended to enhance user performance, foster

* Corresponding author at: Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy.
Email address: federico.cabitza@unimib.it (F. Cabitza).

https://doi.org/10.1016/j.ijhcs.2026.103763
Received 17 December 2024; Received in revised form 5 February 2026; Accepted 10 February 2026
Available online 20 February 2026

the development of new skills, or maintain those essential to their work and sense of agency and control (Legaspi et al., 2024). Achieving this fit can be accomplished through *configuration*, *adaptation*, or *co-evolvement* (Yang et al., 2020), depending on the timing (respectively, pre-deployment, at runtime, or across successive deployment cycles) and the time scale of the process.

This work focuses on obtaining information derived from AI use and human-AI interaction outcomes, which can serve as criteria for achieving and enhancing fit during each stage mentioned above (i.e., configuration, adaptation, co-evolvement) based on user characteristics such as skills, perceptions, roles, and expertise (among others). This information pertains to how *appropriately* users *rely on* the system for decision-making and, crucially, how *appropriately* the system *influences* their decisions. The idea is that once we have operationalized the concepts of *reliance* and *influence* through appropriate empirical metrics, designers and deployers of AI systems can conduct studies within specific decision-making contexts to gather empirical evidence about which configuration or type of AI optimizes system *influence* and ensures its *appropriateness* (Cao et al., 2024). This approach is grounded in the assumption that an AI system that users neither rely on nor find influential in decision-making is ineffective and, consequently, both useless and a waste of resources, if not worse.

Our work aligns with recent literature on the use of AI systems in decision-making contexts, which has emphasized the importance of measuring and evaluating *reliance* (Kahr et al., 2024; Schemmer et al., 2023b) among decision-makers and assessing its *appropriateness* (a comprehensive survey of this body of research will be presented in the next section). Our work also builds on the proposal by Famiglini et al. (2024), who argued for basing technological adaptations (at any stage mentioned above) on empirical evidence, prioritizing stronger evidence as outlined in the hierarchy proposed in their work.

In the remainder of this paper, we will first motivate the focus on *reliance* and *influence* by characterizing their essential features and argue that our work aims to propose a metric framework that moves beyond the concept of simple agreement between users and AI systems. This framework is designed to provide more robust metrics for reliance and influence than those currently employed in the HCI and AI research community. We will then present the framework in detail, including its implementation in an online prototype. Following this, we will describe four user studies conducted in recent years in medical diagnostic settings, involving 330 practitioners with diverse characteristics and over 15,000 decisions. We will report the results of applying the metric framework to the decisions observed in these studies.

The subsequent discussion will focus on interpreting these results to guide the design of systems optimized for specific categories or types of users (in terms of role, expertise, perceptions, and skills) within a given context of use. Ideally, this context would be the same settings where the studies were conducted. However, depending on the strength of the evidence, the findings may also be applicable to other settings with comparable characteristics.

2. Motivations and background

In the current phase of technological development, AI systems are predominantly designed and developed with a focus on algorithmic optimization to excel in specific tasks, such as image recognition, natural language processing, or autonomous driving. In knowledge-intensive tasks, the prevailing approach prioritizes technical performance metrics - such as accuracy, efficiency, and speed – often without explicitly addressing human values or broader societal considerations.

In response to these practices, the concept of Human-Centered Artificial Intelligence (HCAI) has gained prominence, offering an alternative perspective on the design, development, and deployment of AI (Schmager et al., 2023). HCAI emphasizes the centrality of human needs and aspirations, acknowledging the profound effects AI systems have on individuals, societies, and the broader human experience (Bingley et al., 2023; Shneiderman, 2022).

One of the most widely recognized definitions of HCAI emphasizes its focus on "amplifying, augmenting, and enhancing *human performance* in ways that make systems reliable, safe, and *trust-worthy*" (Shneiderman, 2020) (our emphasis). To operationalize this vision, it is essential to define what constitutes human performance, moving beyond a simplistic conflation of performance with the mere use of the system, and where trust can be observed.

To this aim, the concept of *information value chain*, introduced by Coiera (2019) (see Fig. 1), offers a valuable framework to contextualize human performance from an HCAI perspective, particularly in decision-making contexts. This concept allows for a more nuanced understanding of how AI systems interact with and support human decision-making processes, and underscores that the generation of outputs by decision-support systems – often the primary focus of AI researchers – is just one step in a broader sequence that determines the real value and impact for users and deploying organizations (Herrmann and Pfeiffer, 2023). The first critical opportunity for AI systems to influence human decisions lies in their capacity to first gain human trust, at least for a specific case, and hence exert an effect in reducing human error rates (Bansal et al., 2019), thereby altering human actions and increasing the likelihood of better outcomes. Although this final step is pivotal in justifying the integration of AI systems into decision-making contexts, there is limited evidence demonstrating AI's capacity to significantly improve the final outcomes of human actions (Cresswell et al., 2020, 2024; Moja et al., 2019).

Nevertheless, meaningful effects can only emerge from the integration of humans and machines when the entire process beyond output generation is considered (Cabitza et al., 2023c). This necessitates a deeper examination of how AI systems influence the actual environments or cases of interest. In this contribution, we focus on the second step after output generation: the stage where the machine's influence on decision-making should manifest, as a consequence of human trust, creating opportunities to improve human action. This step is foundational for HCAI research, as it represents the point where the promise of AI integration as a trustworthy augmentation tool is realized. Furthermore, it is crucial to evaluate the machine's effects on human capabilities to ensure alignment with fundamental rights, the long-term sustainability of its impact, the development and maintenance of human skills, and the preservation of individuals' sense of agency and responsibility.

It is important to note that measuring the effect of machine outputs on human decision-making solely in terms of decision alignment, such as agreement and switch rate, that is the number of times humans change their minds after considering the AI output, is insufficient. Recent research has shifted towards a more nuanced construct: (appropriate) reliance (Lee and See, 2004). Assessing the degree and appropriateness of human reliance on AI systems is critical for achieving an optimal human-machine relationship, which is instrumental in fulfilling the ultimate goal of the information value chain: improving human action outcomes.

Fig. 1. The information value chain in a human-centered AI context, inspired by Coiera (Coiera, 2019).

Building on this foundation, this contribution advocates for moving further, and beyond the concept of reliance, to introduce and operationalize the notion of *influence*. This concept emphasizes the differential utility of a decision support system as a factor contributing to the achievement of specific goals at the end of the value chain. We argue that evaluating steps that follow the system output, shifting focus from abstract accuracy metrics to the tangible impact on human decisions and actions, is vital. This perspective is particularly relevant in knowledge-intensive domains such as medical diagnostics, where decision-making often relies on tacit forms of expertise. Empirical evidence shows that expert diagnostic reasoning frequently depends on implicit knowledge, visual inspection strategies (e.g., identification of hot spots, use of magnification), and structured interpretive routines (Plass et al., 2022). These cognitive and perceptual mechanisms, while difficult to quantify, play a critical role in shaping diagnostic judgments and actions. Therefore, evaluating AI systems solely on accuracy performance overlooks the extent to which they align with and support these human cognitive strategies. Similar observations have been made in applied domains such as radiology, where recent studies emphasize the importance of designing AI tools that complement domain-specific expertise and situated decision-making practices (Sorantin et al., 2022). Assessing influence on the basis of observable behaviors, and associating that construct with the rate of correct decisions made by the human being after consulting with the AI system (i.e., the final accuracy of the combined human-AI system) may be the preliminary step in figuring out which type of AI is best suited for which type of user, on the basis of their profile, including information such as role, expertise, seniority, and skill, or their degree of appropriate reliance, i.e., the degree of comprehension of the system's capacities at the level of the individual case and individual decision to be made, and thus of how well calibrated their trust (Buçinca et al., 2021) in the system appears to be.

In what follows, we will argue that to tailor an AI system to a specific user—whose characteristics, such as role, expertise, perceptions, or skills, are partially known – it is essential to adopt a grounded, evidence-based design approach (Famiglini et al., 2024). This approach ensures that the system can be configured during setup, adapted in real-time, or improved over time, with a focus on the decision-making step of the information value chain and guided by the findings of empirical research. Achieving this requires the development of novel metrics to better understand this critical step and to substantiate the value of AI in real-world human contexts.

3. Related work

In this section, we introduce key concepts and provide an overview of related work concerning appropriate reliance and trust calibration. Specifically, we explore how these notions have been conceptualized and measured in previous research, highlighting the diversity of approaches adopted in the literature (Benda et al., 2022).

Influence. So far, we have assumed a general understanding of the term *influence*. Broadly, influence refers to an agent's ability to affect another's behavior, perception, or cognition. This definition encompasses *interaction* between agents, including humans and AI systems, where mutual influence may occur. Strategies for exerting influence range from physical control, such as coercion, to communication-based approaches, including deception (Umbrello and Natale, 2024), persuasion (Burr et al., 2018), nudging (Luciano, 2024), and the provision of specific affordances (Famiglini et al., 2024).

In this work, we propose a more narrowly defined concept of AI influence. Specifically, we define *influence* as the system's capacity, through its output, to *actually* change the user's mind (and hence their decision), or modify their perception, feelings, or opinion about a decision. This change may also include increasing or decreasing the user's confidence in the decision's correctness. Machine influence occurs within a specific decision context and depends on various factors in addition to the system's output. While system output may include advice, explanations (Cabitza et al., 2023b), similar cases (Cabitza et al., 2024), or counterfactual examples (Amann et al., 2022), other contextual factors include the user's initial confidence - shaped by prior interactions with the system or what social learning theory and cultural anthropology call its reputational *prestige*, see Henrich et al. (2015) - their level of expertise, and their understanding of the system's likely accuracy in the specific case at hand. More broadly, this understanding of how the machine operates forms a mental model often referred to as the *theory of machine* (Logg, 2022). Machine influence shapes user behavior and decision-making, ultimately reinforcing or eroding user trust in the machine based on perceived system reliability and alignment with their goals (Hoffman et al., 2013).

It is crucial to clearly distinguish AI influence from the closely related concept of AI persuasion. Persuasive technologies are explicitly designed to change users' attitudes or behaviors through persuasion and social influence (De Vries et al., 2017). An effective persuader is defined as an entity that successfully convinces a recipient to adopt a specific viewpoint or behavior (Baek and Falk, 2018). This characterization positions AI persuasion as a form of technology-mediated behavioral control, one that is inherently teleological and intentional, designed to achieve predefined goals aligned with the persuader's objectives (Floridi, 2024). As Fogg (1998) succinctly states, "persuasion requires intentionality."

In contrast, the changes in behavior or attitudes resulting from AI influence are not driven by an intent to persuade users to adopt a particular, AI-backed stance. Rather, designing for an appropriate level of AI influence focuses on enhancing the collaborative performance of the human-AI dyad by tailoring the system's outputs to align with user capabilities and needs, even when this leads to decisions that diverge from the AI's recommendations.

Thus, summing things up, AI influence refers to the effect an AI system exerts on a decision-making process, shaped by a combination of technical, socio-technical, and cognitive factors. Key elements include the system's validity, its effectiveness in communicating this validity, and the user's ability to comprehend the system's output, evaluate its reliability, and integrate it into their decision-making process. These factors can be operationalized through metrics such as positive predictive values, calibrated confidence scores, and degrees of appropriate reliance.

Trust and reliance. The terms *trust* and *reliance* are often used interchangeably in the literature, although they represent different aspects of human-AI interaction, with trust serving as a primary contributing factor to the reliance (Klingbeil et al., 2024; Schaschek et al., 2024). Several studies further blur this distinction by defining behavioral trust through observable actions, effectively using them as proxies for user trust (Zhou et al., 2019; Papenmeier et al., 2022). Specifically, such studies propose that reliance can be measured through behavioral indicators of trust in an AI system, such as the frequency with which users adjust their decisions to align with the AI's predictions. However, trusting an entity and believing it to be correct in a specific instance are two fundamentally different concepts.

In this work, we align with the research perspective that opposes conflating trust and reliance, as suggested by (Lee and See, 2004), and focus on reliance as a measurable quantity, in contrast to trust. Trust refers to an emotional or attitudinal stance, representing a psychological belief in the system's competence and reliability. On the other hand, reliance is defined as observable behavior: the decision to either adhere to or override a system's recommendation (Lee and See, 2004).

In scenarios where an initial AI suggestion is provided, the human agent retains the discretionary power to evaluate the recommendation and decide whether to follow it or pursue an alternative course of action. While trust and reliance are closely related, they do not necessarily align; users might trust an AI system yet choose not to rely on it due to various

contextual factors, such as task complexity or personal self-confidence (Hoff and Bashir, 2015).

Reliance metrics. *Appropriate reliance* in human-AI decision-making has been defined and measured in multiple ways across the literature (Eckhardt et al., 2024), but it is commonly described as relying on AI predictions when they are accurate and rejecting them when they are not (Lee and See, 2004; Schemmer et al., 2023b; Schmitt et al., 2021; Guo et al., 2024b). Several metrics have been developed to measure appropriate reliance or its key components (e.g., under- or over-reliance)Schaschek et al. (2024). However, assessing reliance in a way that accurately reflects AI's real influence on decision-making remains challenging.

Common metrics, such as Agreement Fraction or Percentage (Yin et al., 2019), Reliance Rate (Yu et al., 2019), Failure Detection (Merritt et al., 2015), and the metrics of appropriate and inappropriate reliance proposed by Brachman et al. (2022), have been extensively used in studies of human-AI interaction (Wang and Yin, 2021; He et al., 2023b). While these metrics have often been employed to quantify people's trust and reliance on AI recommendations (Lu and Yin, 2021), they actually measure *agreement*, that is, the degree to which the decision of the humans and the AI recommendations match.

In particular, the **Agreement Fraction** measures the number of decision tasks where the user's final decision matches the AI's prediction, divided by the total number of tasks (expressed as a percentage in the Agreement Percentage).

Reliance Rate refers to the proportion of consistent decisions made with the system over a set number of consecutive trials.

Failure Detection is calculated as the percentage of times users' final decisions disagreed with incorrect automation advice (i.e., correct disagreements) divided by the percentage of incorrect disagreements. Similarly, the metrics defined by Brachman et al. (2022) measure appropriate reliance as the ratio of items where the user agreed with the correct advice to the total number of correct advice items, and inappropriate reliance as the ratio of items where the user agreed with an incorrect advice to the total number of incorrect advice items.

Although these metrics offer insights into the decision-making process, they have a key limitation in that they do not account for the user's initial decision (HD1, see Table 2). Therefore, they cannot determine whether the AI genuinely influenced the user's judgment, as the user may have already agreed with the AI before seeing its suggestion. As a result, these metrics are insufficient for assessing reliance and the true influence of the AI on the user's decision-making process, but should rather be interpreted as metrics of simple agreement.

The limitation of not considering users' initial decisions has been addressed by implementing additional metrics that do not solely rely on the agreement rate between human and AI decisions. Examples in this line of work include the Switch Fraction (Yin et al., 2019), Weight of Advice (WoA) (Harvey and Fischer, 1997), Weight of Own Estimate (WOE) (Yaniv and Kleinberger, 2000), Relative Positive AI Reliance (RAIR) (Schemmer et al., 2022), Relative Positive Self-Reliance (RSR) (Schemmer et al., 2022) and Deception of Reliance (DoR) (Morrison et al., 2024).

The **Switch Fraction** or Percentage, employed in several decision support system studies by He et al. (2023a); Lu and Yin (2021); Ma et al. (2024), is defined as the ratio of the number of tasks where the subject's initial prediction disagreed with the model's prediction, but the final prediction aligned with the model, to the total number of tasks where the initial prediction disagreed with the model. This metric measures how often users, after initially disagreeing with the AI, subsequently changed their final decision to match the AI's suggestion. As such, it provides a more stringent measure of reliance on AI, as it tracks explicit decision changes.

The **Weight of Advice (WoA)**, introduced by Harvey and Fischer (1997) and further defined by Yaniv (2004), is a a measure of advice utilization (Logg et al., 2019) and it is calculated as the ratio of the difference between the final decision and the initial decision, divided by the difference between the advice and the initial decision. This proportion quantifies the extent to which users modify their decisions after receiving advice. A positive WoA indicates that users adjusted their responses toward the AI's advice, while a negative WoA suggests they discarded it.

A related metric to WoA, the **Weight of Own Estimate (WOE)** (Yaniv and Kleinberger, 2000), measures the weight users place on their initial judgment relative to AI advice, with a value of 1.0 indicating complete reliance on their original decision (100% discounting of the advice). It is calculated as the ratio of the difference between the advice and the final decision, divided by the difference between the advice and the initial decision. A key limitation of the Switch Fraction, WoA and WOE is that they do not distinguish between appropriate and non-appropriate reliance, that is, they do not distinguish cases in which a decision switch resulted in an improvement from those cases in which instead it did not. Additionally, from a more statistically-oriented perspective, the Switch Rate takes the form of a conditional rate, which are harder to interpret than commonly used statistical indicators of effect (Schechtman, 2002) (i.e, unconditional rates, odd ratios or relative risks) and also have worse statistical properties (e.g., confidence intervals for conditional rates are usually bigger than for other indicators, as they use a reduced number of samples).

WoA and WOE, by contrast, only provide informative indicators in case of numerical decisions: for categorical ones, they can only assume values equal to 1 (when both the initial and final user decisions disagree with the AI decision), 0 (when the final user decision and the AI decision agree) or ∞ (when the initial user decision and AI decision agree, but then the user switches their decision).

A more advanced approach was proposed by Schemmer et al. (2022), which considered two different metrics to model reliance, namely the Relative Positive AI Reliance (RAIR) and Relative Self-Reliance (RSR), to provide deeper insight into decision-making behavior.

Relative AI Reliance (RAIR) reflects the proportion of instances in which users change their incorrect initial decisions to follow correct AI advice. It is calculated by dividing the number of cases where participants were initially wrong, received correct AI advice, and subsequently changed their decision to the correct one (reliance pattern 011, see Table 2), by the total number of instances where participants were initially wrong and the AI provided correct advice, regardless of the final decision (reliance patterns 011 + 010, see Table 2).

Relative Positive Self-Reliance (RSR), on the other hand, refers to instances of correct self-reliance when the AI advice is incorrect. It is calculated by dividing the number of cases where participants were initially correct, received incorrect AI advice, and maintained their correct decision (reliance pattern 101, see Table 2), by the total number of instances where participants were initially correct and the AI provided incorrect advice regardless of the final decision (reliance patterns 101 + 100, see Table 2). Both metrics range from 0 to 1, with higher values indicating more appropriate reliance on the AI. The authors refer to the theoretical goal of achieving an RSR (Relative Self Reliance) and RAIR (Relative AI Reliance) metric of "1" as the optimal **Appropriateness of Reliance (AoR)** (Schemmer et al., 2023a,b). Nonetheless, AoR does not provide a metric of appropriate reliance as it can only take two values (0 or 1), thus not allowing an easy comparison between different AI systems, and also as it focuses on over- and under-reliance rather than appropriate reliance properly defined. Additionally, and similarly to the Switch Rate, WoA and WOE, RSR and RAIR take the form of conditional rates.

Table 1
Summary of the metrics discussed in related work.

Metric	Definition	Limitation
Agreement Fraction (Yin et al., 2019)	$\frac{\text{Agreements between the subject's and model's prediction}}{\text{Total number of tasks}}$	Agreement measure, does not consider the user's initial decision (H1)
Reliance Rate (Yu et al., 2019)	$\frac{\text{Consistent decisions made with the system}}{\text{Set number of consecutive trials}}$	Agreement measure, does not consider the user's initial decision (H1)
Failure Detection (Merritt et al., 2015)	$\frac{\%\text{ Correct disagreements}}{\%\text{ Incorrect disagreements}}$	Agreement measure, does not consider the user's initial decision (H1)
Appropriate Reliance (Brachman et al., 2022)	$\frac{\text{Decisions where the user agreed with the correct advice}}{\text{Total number of correct advice}}$	Agreement measure, does not consider the user's initial decision (H1)
Accuracy Wid (Harvey and Fischer, 1997)	$\frac{\text{Correct final decisions following initial disagreement}}{\text{Total number of decisions with initial disagreement}}$	Does not consider cases with initial agreement
Switch Fraction (Yin et al., 2019)	$\frac{\text{Number of switch (from disagreement to agreement)}}{\text{Total number of initial disagreement}}$	Does not consider whether switches are for the better or for the worse
Weight of Advice (WoA) (Harvey and Fischer, 1997; Yaniv, 2004; Logg et al., 2019)	$\frac{\lvert\text{Final Decision} - \text{Initial Decision}\rvert}{\lvert\text{Advice} - \text{Initial Decision}\rvert}$	Informative only for numerical decisions
Weight of Own Estimate (WOE) (Yaniv and Kleinberger, 2000)	$\frac{\lvert\text{Advice} - \text{Final Decision}\rvert}{\lvert\text{Advice} - \text{Initial Decision}\rvert}$	Informative only for numerical decisions
Relative AI Reliance (RAIR) (Schemmer et al., 2022)	$\frac{\text{Positive AI Reliance}}{\text{Positive AI Reliance} + \text{Negative Self} - \text{Reliance}}$	–
Relative Self-Reliance (RSR) (Schemmer et al., 2022)	$\frac{\text{Positive Self} - \text{Reliance}}{\text{Positive Self} - \text{Reliance} + \text{Negative AI Reliance}}$	–
Appropriateness of Reliance (AoR) (Schemmer et al., 2023b)	$AoR = \begin{cases} 1 & RSR = 1 \text{ and } RAIR = 1 \\ 0 & \text{otherwise} \end{cases}$	Not informative (can only take values 0 and 1)
Appropriate Reliance (Guo et al., 2024a)	$\frac{\text{Number of correct switches (from disagreement to agreement)}}{\text{Total number of cases}}$	Does not consider the rejection of AI's recommendation when it is wrong

Table 2
Reliance pattern: definition of all possible decisional configurations between human decision makers and their AI system. In the first three columns, 0 denotes an incorrect decision, and 1 denotes a correct decision. We associate the attitude towards the AI in each possible reliance pattern, which leads to either accepting or discarding the AI'advice, and the main related cognitive biases. AI support can be further decoupled into advice and explanation.

Human judgment (HD1)	AI support (AI)	Final decision (FHD)	Reliance pattern	Other formulations	Facilitating Biases
wrong (0)	wrong (0)	wrong (0)	detrimental reliance	/	automation complacency
wrong (0)	wrong (0)	right (1)	beneficial under-reliance	/	extreme algorithmic aversion
wrong (0)	right (1)	wrong (0)	detrimental self-reliance	Detrimental overriding	self-anchoring bias
wrong (0)	right (1)	right (1)	beneficial over-reliance	Correct adherence	algorithm appreciation
right (1)	wrong (0)	wrong (0)	detrimental over-reliance	Detrimental adherence	automation bias
right (1)	wrong (0)	right (1)	beneficial self-reliance	Corrective overriding	algorithmic aversion
right (1)	right (1)	wrong (0)	detrimental under-reliance	/	extreme algorithmic aversion
right (1)	right (1)	right (1)	beneficial reliance	/	confirmation bias (in later cases)

Guo et al. (2024b) define appropriate reliance as the probability that a rational agent will choose to follow the AI's recommendation when there is a disagreement between the human and the AI. While this metric correctly identifies one component of appropriate reliance (adopting the AI recommendation when it is correct), it forgoes the other key component of rejecting the AI's incorrect recommendation.

In this section, we present a summary of the metrics used in the literature to measure (appropriate) reliance and their limitations (see Table 1). In the following sections, we introduce our proposed framework for assessing AI influence, using established metrics along with some of our own additions.

4. The metric framework

In this section, we present and discuss the various metrics that constitute our framework, including those previously proposed in the literature as well as our original contributions. The foundation of our framework is the notion of a *reliance pattern*, which is a decisional configuration in which both the human user and the AI DSS have formulated a decision. After the introduction of an AI DSS into a decision-making context, eight distinct behavioral patterns can be identified based on two factors: the extent of agreement between human and AI decisions and the resulting error rate in the final human decision. These reliance patterns, assuming a binary decisional outcome, are detailed in Table 2.

By considering the frequency of occurrence of each reliance pattern and their mutual relationships, we can define several metrics, as summarized in Table 3. In the following, we provide an intuitive explanation of the rationale and significance of each of the metrics. In the definition of the metrics, we adopt the following notational conventions: the number of occurrences of a given reliance pattern is denoted using a three-character label representation of the pattern[1] derived from Table 2 (e.g., 000 represents the number of occurrences where both the human and the AI system are initially wrong, and the final decision is also incorrect). Additionally, we use the term $AIER$ to denote the sample average of the AI-supported error (FHD) individual rates, and UER to denote the sample average of the unsupported (HD1) error rates.

[1] We note that in such expressions the symbols 0/1 are used solely as indicator values for incorrect/correct judgments with respect to ground truth.

Table 3
Metrics of the proposed framework of AI influence. For each metric we report either its analytical formulation (when easily given) or a description of it in natural language.

Metrics of agreement	Description
Percent of Agreement	$\frac{\text{number of agreements between AI and FHD}}{N}$
Error rate in Agreement	$\frac{\text{number of negative agreements between AI and FHD}}{\text{number of agreements between AI and FHD}}$
Metrics of reliance	**Description**
Dominance Strength	$\frac{number\ of\ induced\ changes}{number\ of\ recommendations\ given}$
Dominance orientation	the difference between the number of changes for the right (1) and the number of changes for the wrong (0), wrt the total number of changes
Deference Strength	$\frac{number\ of\ changes\ leading\ to\ agreement}{number\ of\ recommendations\ given}$
Deference Orientation	the difference between the number of changes leading to agreement with the correct decision (1) and the number of changes leading to agreement with the incorrect decision (0), relative to the total number of changes
AI decision Impact	$\frac{1-AIER}{AIER}\frac{UER}{1-UER}$, represents the odds ratio comparing the likelihood of making errors when aided by AI to the likelihood of making errors when unaided.
Team AI Effect on Decision	$\frac{\bar{UER}-\bar{AIER}}{USD}$, the difference between the AI-supported and the unsupported average accuracy scores wrt the unsupported standard deviation (cf. Glass' Delta)
Number Needed of Decisions	Number of AI-supported decisions that must be made before avoiding a mistake that would have been made without the AI support
Automation Bias	$\frac{detrimental\ over-reliance}{N-detrimental\ over-reliance}\frac{N-beneficial\ self-reliance}{beneficial\ self-reliance}$
Self-anchoring Bias	$\frac{detrimental\ self-reliance}{N-detrimental\ self-reliance}\frac{N-beneficial\ over-reliance}{beneficial\ over-reliance}$
Appropriate reliance level	$\frac{number\ of\ positive\ trust\ and\ distrust\ pattern\ occurrences}{number\ of\ recommendations-given}$
Metrics of influence	**Description**
Appropriate influence	$1-\frac{1-\text{Appropriate reliance}}{1-\mathbb{E}[\text{Appropriate reliance}]}$
Influence Index	IRR(FHD, AI) - IRR(HD1, AI)

Percent of agreement. The percent of agreement (also referred to as the Agreement Fraction or Reliance Rate in the literature) measures the extent to which the human decision (FHD) and AI advice align. This is defined as:

$$\frac{\overbrace{100+011+000+111}^{\text{Sum of cases where the final decision aligns with system advice}}}{\underbrace{N}_{\text{Total number of advice provided by the AI}}} \tag{1}$$

Error rate in agreement. The error rate in agreement is defined as the proportion of agreements that were associated with a wrong decision: that is, the cases in which AI and the human decision (FHD) coincided, but both were wrong. It is expressed as:

$$\frac{\overbrace{100+000}^{\text{Sum of cases where agreement with system advice led to incorrect decisions}}}{\underbrace{100+011+000+111}_{\text{Sum of cases where the final decision aligns with system advice}}} \tag{2}$$

Dominance strength. Dominance is essentially the extent to which users change their minds after consulting the AI and whether they do so for better or worse (positive or negative orientation). Dominance Strength (also called Switch Rate in the literature) measures the ratio of changes made by users induced by AI suggestions, relative to the total number of AI suggestions provided, regardless of whether the final decision aligns with the AI's advice. That is, in formula:

$$\frac{\overbrace{011+100+110+001}^{\text{Sum of cases where users made changes based on AI suggestions}}}{\underbrace{N}_{\text{Total number of advice provided by the AI}}} \tag{3}$$

It indicates how frequently a user modifies their decision based on the AI's advice. The metric ranges from 0 to 1, where a value closer to 1 indicates that the AI frequently affects users to change their decisions, while a value closer to 0 suggests that the AI has minimal or no impact on user behavior.

Dominance orientation. Dominance Orientation compares the number of changes in human decision that results in correct judgments with those that lead to incorrect ones, relative to the total number of changes. It is defined as:

Number of changes that led to correct decision — Number of changes that led to incorrect decision

$$\frac{011 + 001 \quad - 100 - 110}{011 + 100 + 110 + 001}$$

Cases where users made changes based on AI suggestions

(4)

It is expressed as a number between -1 and $+1$, where a positive value indicates that changes induced by AI advice more often lead to correct decisions, while a negative value suggests that AI advice more often results in incorrect decisions.

Deference strength. Deference strength (also referred to as Switch Fraction in the literature) represents the portion of dominance leading to a human-AI agreement. It is defined as:

Cases where users changed their decisions to align with the AI's advices

$$\frac{011 + 100}{010 + 011 + 100 + 101}$$

Cases where users made changes based on AI suggestions

(5)

It measures how often users change their minds to align with the AI's suggestion. It's a number between 0 and 1, a higher value indicates that users frequently defer to the AI's recommendations by adjusting their decisions in agreement with the AI.

Deference orientation. Deference Orientation is the difference between the number of decision changes that result in agreement with a correct AI decision (1) and those that result in agreement with an incorrect AI decision (0), expressed as a proportion of the total number of changes, that is:

Difference between correct and incorrect alignments with AI advice

$$\frac{011 - 100}{011 + 100}$$

Total sum of alignments

(6)

This metric ranges from -1 to $+1$, where positive values indicate that deference was predominantly positive (meaning users tended to accept correct recommendations provided by the AI, while rejecting incorrect ones).

AI decision impact. It is measured as the odds ratio comparing two scenarios, namely: the odds of the human user making an error with AI support (denoted as AIER, AI-supported error rate) and without AI support (denoted as UER, unsupported error rate). It is defined as:

Odd of the user making an error with AI support

$$\frac{1 - AIER}{AIER} \quad \frac{UER}{1 - UER}$$

Odd of the user making an error without AI support

(7)

Thus, values of AI Decision Impact greater than 1 denote a positive impact of the AI support on the final decision, as they correspond to cases where UER (unsupported error rate) is higher than AIER (AI-supported error rate). In other words, introducing AI support decreases the error rate. Conversely, values lower than 1 denote a negative impact, as they are associated with UER being lower than AIER, meaning that the introduction of AI support has increased the error rate. When interpreted as the ratio of $(1 - AIER)/(1 - UER)$ to $(AIER)/(UER)$, the AI decision impact can be understood as the ratio between relative benefit and relative risk. These are two widely used measures in the statistical analysis of ecological, cohort, medical, and intervention studies (Cook and Sackett, 1995; Malenka et al., 1993). Unlike other metrics, this ratio of ratios accounts for both positive and negative outcomes, integrating these effects into a single comprehensive measure. This characteristic makes AI Decision Impact particularly valuable for trade-off analyses, where understanding an intervention's overall impact may clarify trade-offs, and in goal prioritization, where decision-making should aim to maximize benefits while also minimizing harms.

Team AI effect on decision. The difference between the average accuracy scores of the AI-supported and unsupported samples, in relation to the standard deviation of the unsupported sample, that is:

$$\frac{UER - AIER}{USD} \tag{8}$$

Average error rate of the unsupported users

Average error rate of the AI-supported users

Standard deviation in the error rates of the unsupported users

where USD is the standard deviation of the user individual unsupported error rates. Intuitively, the Team AI Effect on Decision quantifies the effect of the AI support in terms of a standardized effect similar to Glass' Delta (Hedges, 1981): the metric takes a positive value if and only if the accuracy improves after being exposed to the AI support, that is if the AI had a positive effect. Furthermore, it takes values larger than 1 if and only if the improvement described above is at least as large as the average deviation in the baseline accuracies (hence, the observed improvement in performance due to being exposed to the AI is larger than what could be expected by chance due to purely random variations in the users' population).

Number needed of decisions (NND). This metric calculates the number of AI-supported decisions required to prevent a single error that would have occurred without the support. This is defined as the minimum sample size based on the Team AI Effect on Decision A (interpreted as an effect size) through the formula, originally proposed by Furukawa and Leucht (2011) to compute the number needed to treat in the context of epidemiological studies (Cook and Sackett, 1995):

$$\frac{1}{\Phi(A - \mathcal{N}(UER)) - (1 - UER)} \tag{9}$$

The expected accuracy when supported by AI

where Φ is the cumulative distribution function of a standard Gaussian random variable (that is, a Gaussian random variable with mean equal to 0 and standard deviation equal to 1), $\mathcal{N}$ is the corresponding density function. This formula expresses that, given a certain effect size representing the effect the AI support had on the final decision, the number needed of aided decisions will differ according to the response threshold one expects by random chance (that is, $\mathcal{N}(UER)$) and the accuracy obtained when unaided associated with that threshold (that is, $1 - UER$).

Automation bias. Refers to the negative aspect of reliance on AI, characterized by AI-induced commission errors (Cabitza et al., 2023a). It occurs when users are misled by incorrect AI advice, resulting in mistakes they would not have made independently, reflecting an over-reliance on faulty AI recommendations: thus, intuitively, automation bias can be understood as the lack of a component of appropriate reliance, namely the ability to appropriately discard wrong AI recommendations. In terms of the reliance patterns, automation bias corresponds to the detrimental over-reliance pattern: the more frequently this pattern is observed, the stronger the automation bias. In the proposed metric of automation bias, the detrimental over-reliance pattern is compared with the beneficial self-reliance in terms of an odds ratio:

$$\frac{100}{N - 100} \Big/ \frac{N - 101}{101} \tag{10}$$

Inverse of odds of users rejecting incorrect AI advice compared to all other cases

Odds of users following incorrect AI advice

If the automation bias metric is lower than, or close to, 1, then there is no significant indication of automation bias, since the users are less or equally likely to be misled by the AI as they are to rightfully ignore its wrong advice. By contrast, a value of the metric much larger than 1 provides a significant evidence of automation bias since the user is much more likely to retain rather than discard a wrong advice. This metric is related to the Relative Positive Self-Reliance (RSR) proposed in (Schemmer et al., 2022): automation bias is equal to 0 if and only if RSR is equal to 1, and it is equal to 1 if and only if RSR is equal to 0. Compared with the RSR, however, it offers a more natural interpretation, as it is expressed in terms of an odds ratio (which represents the relative strength of automation bias compared with the opposite behavior of discarding wrong AI support) rather than a conditional rate; furthermore, it has a larger effective sample size (equal to N, the whole sample size, compared with $100 + 101 \leq N$ for the RSR), which can in general result in smaller estimation uncertainty (e.g., when computing confidence intervals).

Self-anchoring bias. Self-anchoring Bias reflects a preference for human over algorithmic advice, resulting in reduced reliance on AI interventions (Dietvorst et al., 2015). It occurs when users are not convinced to follow correct AI advice, but rather reject it reflecting an over-reliance on their own faulty judgments. Similar to automation bias, self-anchoring bias can be understood as the lack of a component of appropriate reliance, namely the ability to identify and appropriately accept correct AI recommendations. In terms of reliance patterns, self-anchoring bias corresponds to the detrimental self-reliance pattern. The considered metric for self-anchoring bias is defined similarly to the corresponding one for automation bias, as it compares the detrimental self-reliance pattern with the beneficial over-reliance pattern using an odds ratio:

$$\frac{010}{N - 010} \Big/ \frac{N - 011}{011} \tag{11}$$

Inverse of odds of users following correct AI advice

Odds of users rejecting correct AI advice

If the self-anchoring bias metric is smaller than or close to 1, then there is no significant indication of the presence of self-anchoring bias,

since the users are more or equally likely to accept the correct recommendations of the AI as they are to reject them. By contrast, a value of the metric much larger than 1 provide a significant evidence of self-anchoring bias, as users are much more likely to discard rather than retain a correct advice. Similarly to the relationship between automation bias and RSR, this metric is related to the Relative Positive AI Reliance (RAIR) proposed by (Schemmer et al., 2022).

Appropriate reliance. Intuitively, AR is an indication of how well users determine, from any clue, whether they can trust or should distrust the machine's advice. We operationalize this concept as the ratio of the sum of beneficial trust and distrust pattern occurrences to the total number of recommendations provided. The beneficial reliance patterns are 011 and 111, while beneficial distrust patterns are 101, 001 and 000. Notably, the metric considers not only reliance patterns associated with a decision switch, but also patterns in which the AI recommendation and human user decision agree (i.e., the patterns 000 and 111): to assess whether the AI influenced the decision-making in these cases, we consider the change in the confidence the human users attaches to its decision between HD1 and FHD. In particular, 111 is considered a beneficial trust pattern if it is associated with an increase in confidence, while 000 is considered a beneficial distrust pattern if it is associated with a decrease in confidence: the rationale behind this approach lies in the observation that, even though the AI recommendation was not sufficient to make the human user change its mind, it still had an impact. Consequently, the formula for the appropriate reliance level can be defined as:

Beneficial trust and distrust pattern

$$\frac{011 + 101 + 001 + 111\uparrow + 000\downarrow}{N} \quad (12)$$

Confidence Decrease

Confidence Increase

Total number of advice provided by the AI

where $\uparrow$ (resp., $\downarrow$) denote an increase (resp., decrease) in confidence between HD1 and FHD, and N is the total number of recommendations given. Appropriate reliance combines, and goes beyond, the proposals by Schemmer et al. (2023b) and Guo et al. (2024b): compared to the proposal of Schemmer et al., our metric provides a single indicator combining the elements that appear in the RAIR and RSR components; while compared to the proposal of Guo et al. it considers an additional component of appropriate reliance (that is, the probability of discarding a wrong AI recommendation). In particular, appropriate reliance is related to automation bias and self-anchoring bias (see below), as well as RAIR and RSR. Indeed, it holds that if the appropriate reliance level is equal to 1, then both automation bias and self-anchoring bias are equal to 0 (if and only if both RAIR and RSR are equal to 1).[2] The converse implication does not, however, automatically hold, due to the terms $111\uparrow + 000\downarrow$ in the definition of the appropriate reliance.

Appropriate influence. Appropriate influence aims to address the shift in focus from reliance to influence, as highlighted in the previous section. It can be defined as the extent to which a system's advice makes user reliance appropriate beyond what would occur by random chance. As such, it allows researchers to distinguish between informed and uninformed reliance, in virtue of a model of chance through which to calculate the likelihood that users may appear to appropriately rely (that is, accept correct advice, or discard incorrect advice) on the system's advice by chance, and then to discount this factor from the appropriate reliance.

Analytically, the formula for the appropriate influence stems from calculating the expected value of the appropriate reliance level metric under the assumption of no influence, that is, all observed cases of appropriate reliance are due to chance. This is equivalent to assuming that the decisions of the human user and AI are entirely independent. Hence, let AI be the accuracy of the AI. Under the uniform prior on the confidence increasing or decreasing (that is, the confidence has the same probability, equal to 0.5, of increasing or decreasing between HD1 and FHD), the expected value of appropriate reliance can be computed as:

The expected frequency of a beneficial trust pattern

$$UER * AI * (1 - UER) + \quad (13)$$

[2] If Appropriate reliance level is equal to 1, then $011 + 101 + 011 + 000\downarrow + 111\uparrow = N$, which implies $100 = 0$, as well as $010 = 0$, hence, automation bias = self-anchoring bias = 0 and RAIR = RSR = 1.

$$+ \underbrace{\frac{(1-UER)*AI*(1-UER)}{2}}_{\text{The expected frequency of the 111 pattern with a confidence increase}} + \underbrace{\frac{UER*(1-AI)*UER}{2}}_{\text{The expected frequency of the 000 pattern with a confidence decrease}} + \tag{14}$$

$$+ \underbrace{(1-UER)*(1-AI)*(1-UER) + UER*(1-AI)*(1-UER)}_{\text{The expected frequency of a beneficial distrust pattern}} . \tag{15}$$

The formula for the appropriate influence is then defined as:

$$1 - \frac{\overbrace{1-\text{Appropriate reliance}}^{\text{Portion of appropriate reliance explained by mere chance}}}{1-\mathbb{E}[\text{Appropriate reliance}]} . \tag{16}$$

Intuitively, the appropriate influence is larger than 0 if and only if the observed appropriate reliance is larger than what could be expected by chance. In particular, under the convention that $\frac{0}{0} = 0$, the appropriate influence is equal to 1 if and only if the appropriate reliance is equal to 1. By contrast, the appropriate influence can also take negative values, whenever the observed appropriate reliance is smaller than what could be expected by chance: in particular, if the appropriate reliance is equal to 0 or $\mathbb{E}[\text{Appropriate reliance}] = 1$, then the appropriate influence takes value tending to $-\infty$.

Influence index. The Influence Index measures how persuasive the AI support is. It is the difference in chance-adjusted agreement scores between the final decision (FHD) and the AI and the first judgment (HD1) and the AI, that is:

$$\underbrace{IRR(FHD, AI)}_{\text{Chance-adjusted agreeement between supported users and AI}} - \underbrace{IRR(HD1, AI)}_{\text{Chance-adjusted agreeement between unsupported users and AI}} . \tag{17}$$

To gauge the Influence Index, we measure how often users change their decisions to align with the AI's advice, regardless of whether the advice is correct or incorrect. IRR denotes a chance-discounted inter-rater agreement metrics, such as Krippendorff α. Thus, the Influence Index quantifies the change in chance-discounted agreement between a human user and the AI after the user has been exposed to AI support. An Influence Index greater than 0 indicates that the AI influenced the user, as the exposure resulted in an increased agreement rate. Conversely, a negative value signifies that the AI support reduced the agreement rate.

5. Methods of the user studies

To demonstrate our framework, we applied it to decision performance data from four user studies conducted in the medical domain. These studies involved simulated AI systems tested under controlled conditions using real-world retrospective cases (see Fig. 2). Below, we summarize the key features of the user studies, which provided evidence classified as level 5 Famiglini et al. (2024) regarding the AI systems under evaluation.

5.1. The MRI user study

This radiology-based study involved 12 certified radiologists from IRCCS Ospedale Galeazzi Sant'Ambrogio (Milan, Italy) and other hospitals throughout the country. The task was to classify knee lesions using 120 MRI scans selected from the MRNet dataset.[3] The study was carried out using an online, multi-page LimeSurvey questionnaire where participants viewed the MRI cases in random order, with each case displaying three standard views (axial, sagittal, and coronal), and had to rate the scans for presence of lesions (ligament or knee abnormality). The radiologists provided diagnoses with the aid of a simulated AI system operating at 80% accuracy. A human-first design was employed, in which, for every case, the users were first asked to provide their initial decision (HD1), then they were shown the support of the AI (in the form of a clear-cut diagnosis), subsequently they had to formulate a final decision (FHD). The baseline accuracy (unsupported) of the users (evaluated based on the HD1 decisions) was 79%. Users were also asked to evaluate the confidence in their decision, both for their initial (C1) and final (FC) decisions, through a 6-item ordinal scale ranging from value 1 (no confidence at all) to 6 (absolute confidence in the decision). In half of the cases (60), users were also presented with explainable support (in the form of activation maps, generated through the GradCAM method, highlighting the regions of the MRI image with color denoting their importance for the AI decision) together with the support of the AI. Further details can be found in (Cabitza et al., 2023c).

5.2. The ECG user study

In this cardiology-oriented study, 21 medical professionals from the Medicine School of the University Hospital of Siena (Italy) participated in classifying heartbeat patterns from 20 ECGs. These ECGs were curated by a cardiologist from the ECG Wave-Maven repository.[4] Using a web-based LimeSurvey questionnaire, participants provided diagnoses while interacting with a simulated AI system with 70% accuracy. The study followed a human-first protocol: for each case, users were initially asked to provide an initial decision (HD1), they were then shown the AI support (in the form of a diagnosis), after which they had to formulate

[3] https://stanfordmlgroup.github.io/competitions/mrnet/

[4] https://ecg.bidmc.harvard.edu/maven/mavenmain.asp

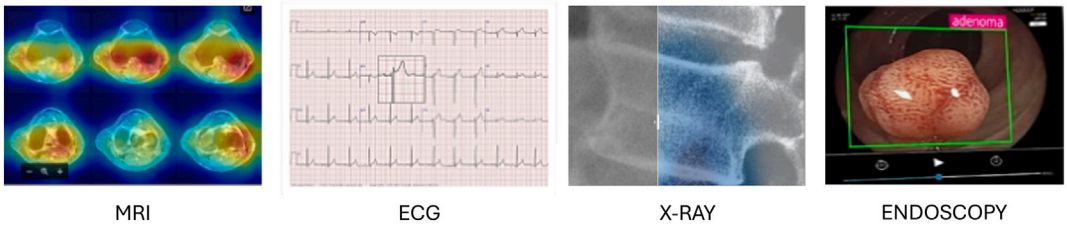

Fig. 2. Screenshots from the simulated AI systems employed in the user stdies.

a second decision (HD2), subsequently they were shown an explainable AI support (in the form of a textual explanation describing the observed ECG trace), and they were finally asked to formulate a final decision (FHD). Although the participants had been told that the explanations were automatically generated by the AI system, like the diagnostic advice, these had been prepared by a cardiologist: 40% of the explanations were prepared to be incorrect or not completely pertinent to the cases. The baseline accuracy (unsupported) of the users (evaluated based on the HD1 decisions) was 56%. For further details, see (Cabitza et al., 2023c).

5.3. The x-ray user study

This orthopedic study focused on detecting traumatic thoraco-lumbar fractures in X-ray images. Seven orthopedists with varying levels of expertise participated in classifying 12 X-ray images, using a multi-page LimeSurvey questionnaire. The participants worked with an AI support system with an accuracy of 78%. The AI system, a ResNeXt-50 model, was trained to detect thoraco-lumbar lesions based on a dataset consisting of 630 vertebral cropped X-rays, which had been collected at the Spine Surgery Centre of the Niguarda General Hospital of Milan (Italy) from 2010 to 2020, from 151 trauma patients over 18 years old, split into 328 no-fracture images (52%) and 302 fracture ones. The experiment utilized a human-first interaction protocol: for each case, users were initially asked to provide an initial decision (HD1), they were then shown the AI support (in the form of a diagnosis) together with an explainable AI support (in the form of a saliency map), and they were finally asked to formulate a final decision (FHD). Users were also asked to rate their confidence, for both their initial (C1) and final (FC) decisions, through a 6-item Likert scale ranging from value 1 (no confidence at all) to 6 (absolute confidence in the decision). Further information is available in (Cabitza et al., 2022).

5.4. The endoscopy user study

The gastroenterological study involved the assessment of gastrointestinal bleeding and inflammatory lesions through small bowel capsule endoscopy, as well as to evaluate ulcerative colitis activity through proctosigmoidoscopy. A total of 274 participants, comprising specialists (79%) and trainees (21%) of different experience levels, were tasked with diagnosing up to 90 short endoscopic video clips (15-20 s each) with and without AI assistance via a LimeSurvey-based questionnaire, amounting to 12,000 paired video assessments. The task involved indicating the presence or absence of small bowel lesions with high bleeding potential (30 paired videos), the presence or absence of small bowel lesions with significant inflammatory potential (30 paired videos), and providing a score on the Ulcerative Colitis Index of Severity (UCEIS, 30 paired videos). The support was provided in form of a simulated AI system with a set diagnostic accuracy of 80%, a sensitivity of 86%, and a specificity of 75%. To indicate the presence of lesions, the system displayed geometric cues (arrows, circles) and, in the UCEIS setting, also reported a suggested severity score at the end of the video. The interaction protocol involved a human-first approach: after watching the video without AI assistance, participants registered their initial decision (HD1) as well as their confidence in their own response (C1, on a scale from 1 to 4). Then, the video was shown again, but with an AI overlay highlighting target lesions. Participants then provided their final decision (FHD), their final confidence (FC), and an assessment of the system's utility (on a scale from 1 to 4, including an *I don't know* option). For more details, refer to (Tontini et al., 2025).

5.5. Statistical analysis

We analyzed the effect of AI on decision-making based on the metric framework described in Section 4. For each study, we considered both the general results, calculated based on the entire sample, as well as a stratified analysis, in which we calculated the metrics separately for the different user profiles and types of AI considered. Additionally, for the MRI and X-ray studies, where confidence was also measured, we considered an additional stratification based on the confidence. Specifically, we stratified the population in two strata: low-confidence (LC) and high-confidence (HC) users; a user was considered an LC (resp. HC) user, depending on whether their average confidence was lower (resp., higher) than the median of the sample's average confidence. Finally, for all studies, we also considered a stratification based on skill, in which we distinguished the samples' population based on their HD1 accuracy. Specifically, for each study, we stratified users into low-performers (LP) and high-performers (HP) based on whether their HD1 decision accuracy was lower or higher than the median HD1 accuracy for that study.

Statistical analysis of the results was based on interval analysis (Altman et al., 2013): for each metric, we computed the corresponding 95% confidence interval through a bootstrap-based approach using Efron (1987) Bias-corrected and Accelerated (BCa) bootstrap algorithm. Statistical significance was assessed based on interval comparison: two metrics' values were deemed significantly different (at the 95% confidence level) if, considering only that comparison/isolating that research question from the others,[5] the corresponding 95% confidence intervals did not overlap.

6. Results and discussion

The results obtained by applying the metric framework proposed in this work to the aforementioned user studies are presented in Table 4 and are discussed in the following sections. All results were generated using an online tool that implements the metrics described in Section 4. This tool is accessible at https://haiiassessment-metimeter.eu.pythonanywhere.com/ or its proxy https://www.entechne.com/metimeter/haiassessment/. The associated Python libraries are available on GitHub and can be found via the Software section of our laboratory's website: https://mudilab.disco.unimib.it/software/. In the following sections we illustrate how the proposed metrics could be applied in a practical context, highlight which kinds of discussion they can stimulate, as well as illustrate the existence of macro-level differences between settings and scenarios and how our framework could be used in such a sense to guide design decisions (Table 5).

[5] The risk of false discovery rate depends on which and how many other comparisons/research questions are considered.

Table 4
The results from the user studies (columns) for each metric (rows) presented in this work.

Metric	Endoscopy	X-ray	ECG	MRI	All
Percent of Agreement	0.81	0.72	0.79	0.71	0.79
Error rate in Agreement	0.11	0.04	0.27	0.09	0.12
Dominance Strength	0.08	0.24	0.10	0.06	0.08
Dominance orientation	0.38	0.39	0.71	0.18	0.38
Deference Strength	0.07	0.19	0.10	0.03	0.07
Deference Orientation	0.44	0.64	0.71	0.82	0.46
AI decision Impact	1.21	2.00	1.61	1.06	–
Team AI Effect on Decision	0.34	1.23	0.66	0.18	0.36
Number Needed of Decisions	12	6	5	19	11
Appropriate reliance	0.51	0.46	0.52	0.28	0.51
Automation Bias	0.21	0.19	0.30	0.02	0.19
self-anchoring bias	1.90	0.20	0.96	6.94	2.00
Appropriate influence	−0.05	−0.22	0.04	−0.59	−0.05
Influence Index	0.13	0.27	0.18	−0.00	0.14

Across all studies, the results consistently demonstrate that influence-based metrics such as appropriate influence and the influence index provide a more nuanced and fine-grained understanding of how AI impacts human decision-making than agreement, dominance, or reliance scores alone. Percent of agreement remains high and relatively stable across all studies, as seen with scores ranging from 0.71 (MRI) to 0.81 (Endo). While agreement scores reveal alignment between human and AI decisions, they fail to indicate whether this alignment translates into improved decision quality or meaningful shifts in decision-making behavior.

Appropriate reliance scores highlight varying levels of trust and distrust in AI, with Endo (0.51) and ECG (0.52) showing more appropriate reliance compared to X-ray (0.46) and MRI (0.28). While in the x-ray study no bias seems to be prominent in negatively impacting reliance, in the MRI study, conversely, the self-anchoring bias is not only much greater than the automation bias (6.94 vs 0.02), which is virtually absent, but it is by far the highest detected in all studies. This is confirmed by the results regarding influence, which in that study was negligible as expressed by the influence index (0.00) and the least appropriate (−0.59), whereas a negative value indicates that influence has been observed to be about half of what would be expected if it were due entirely to chance. All these measurements indicate that the sub-specialists involved in the MRI study overlooked the correct AI advice too often and suggest an information campaign to make them more aware of the potential of the technology for the diagnostic task under consideration: this is still relevant, since, as evidenced by the value of NND it only takes 19 images to be diagnosed with AI support to avoid an error that would have been made without the use of AI.

On the other hand, since the highest value in appropriate reliance is just half of the theoretical maximum (0.52), and so it is the mean value across all studies (0.51), we can argue that there is a great deal of room for improvement regarding the optimal way to integrate machine advice into human decision-making in all studies. However, reliance scores lack the depth to account for whether such reliance improves decision outcomes beyond chance.

Influence metrics uncover critical dynamics hidden by agreement and reliance scores. The influence index reveals that X-ray (0.27) and ECG (0.18) benefit the most from AI advice, while MRI (0.00) shows negligible improvement. These findings contrast with reliance metrics, which suggest relatively similar trust patterns for X-ray and Endo, yet only X-ray demonstrates substantial decision improvement. Similarly, appropriate influence is negative across most cases, especially MRI (−0.59), showing that reliance in these contexts often fails to produce beneficial decision-making beyond random chance. Notably, ECG stands out with a positive appropriate influence (0.04), suggesting that decisions in this domain are more effectively shaped by AI interactions.

Bias measures further contextualize these patterns. Automation bias is particularly high for ECG (0.30), suggesting over-reliance on AI recommendations, whereas MRI shows minimal automation bias (0.02),

Table 5
Summary of study insights and AI support strategies aimed at fostering Appropriate Influence and achieving a positive impact across various user types and AI support modalities.

Study	User Profile	Recommended AI Type	Rationale	Suggested Intervention
MRI	Specialist	XAI (significant) or None (suggestive)	Less negative Appropriate Influence, AI Decision Impact not significantly positive	Reduce CB
	Sub-Specialist	XAI (significant) or None (suggestive)	Less negative Appropriate Influence, AI Decision Impact not significantly positive	Reduce CB
	Low-Confidence	AI (significant)	Much higher and positive Appropriate Influence	Reduce CB
	High-Confidence	XAI (significant) or None (suggestive)	Less negative Appropriate Influence, AI Decision Impact not significantly positive	Reduce CB
	Low-Performers	AI (suggestive) or None (suggestive)	AB and CB are lower for AI, AI Decision Impact not significantly positive	Reduce CB
	High-Performers	XAI (significant) or None (suggestive)	Less negative Appropriate Influence, AI Decision Impact not significantly positive	Reduce CB
ECG	Expert	AI (significant) or None (suggestive)	Less negative Appropriate Influence, AI Decision Impact not significantly positive	Reduce CB
	Novice	AI (significant)	Much higher and positive Appropriate Influence	Slightly reduce AB
	Low-Performers	AI (significant)	Much higher and positive Appropriate Influence	Reduce CB and AB
	High-Performers	AI (significant)	Much higher and positive Appropriate Influence	Reduce CB
X-Ray	Specialist	None (suggestive)	AI Decision Impact not significantly positive	Slightly reduce AB and CB
	Sub-Specialist	None (suggestive)	AI Decision Impact not significantly positive	–
	Resident	XAI (significant)	AI Decision Impact significantly positive	–
	High-Confident	XAI (significant)	AI Decision Impact significantly positive	–
	Low-Confident	None (significant)	Dominance Direction and Team AI Effect on Decision significantly negative	Reduce extreme algorithm aversion, slightly reduce AB and CB
	High-Performers	None (suggestive)	AI Decision Impact not significantly positive	Slightly reduce CB
	Low-Performers	XAI (significant)	AI Decision Impact significantly positive	–
Endoscopy	Specialist	AI (significant)	AI Decision Impact significantly positive	Reduce CB
	Trainee	AI (significant)	AI Decision Impact significantly positive	Reduce CB
	High-Performers	AI (significant)	AI Decision Impact significantly positive	Reduce CB
	Low-Performers	AI (significant)	AI Decision Impact significantly positive	Reduce CB

potentially reflecting underutilization of AI. Self-anchoring bias paints an inverse picture, with MRI users (6.94) heavily resistant to AI advice, in stark contrast to X-ray users (0.20), who exhibit almost no such reluctance. These biases explain much of the variance in influence metrics: high self-anchoring bias in MRI corresponds with negligible influence index and highly negative appropriate influence, highlighting how resistance to AI undermines its potential benefits.

Taken together, these results confirm that influence metrics offer a richer and more detailed assessment of AI's impact, revealing the interplay of bias, decision shifts, and alignment that agreement and reliance scores cannot capture. For instance, reliance scores suggest broad trust in AI across domains like Endo, X-ray, and ECG, but only X-ray demonstrates substantial positive influence on decision quality. Conversely, MRI reliance is low, and influence metrics confirm this is due to high self-anchoring bias rather than AI ineffectiveness. These insights reinforce the value of influence-based measures in evaluating human-AI interaction and underscore the importance of domain-specific interventions, such as bias training or adjustments to AI system design, to optimize decision-making outcomes.

6.1. Interpretation of the MRI results

The results of the MRI user study are shown in Table 6 and Fig. 3.

At first glance, agreement metrics suggest a broadly positive effect of AI support. Indeed, all configurations of AI and user expertise (as well as confidence) showed a moderate to large agreement, as shown by the Percent of Agreement metric being generally larger than 70%, except for specialists and high confidence users supported by XAI. In all cases, irrespective of the user expertise or initial confidence level, agreement between FHD and AI was significantly higher when the support was not accompanied by explanations. Similar observations can be made in terms of the accuracy of the final decisions: when the AI and FHD decisions agreed, users erred on approximately one case out of 10. Furthermore, as could be expected, it can be observed that sub-specialists were significantly more accurate than specialists irrespective of the type of the AI support. Interestingly, as in the case of agreement, Error Rate in Agreement was significantly lower when the provided support was not explainable, except for users with low initial confidence, where no significant difference between the two types of AI support could be detected.

By contrast, reliance metrics offer a rather different perspective. Appropriate reliance, while relatively stable across configurations, indicates that users appropriately relied on the AI in only about one quarter of the cases. This value was the smallest among the considered case studies, and significantly so. Only one of the groups, users with initial low confidence supported by AI, had a significantly higher appropriate reliance, approximately close to one case out of two (which was more aligned with the other case studies). This suggests that users who were initially unsure about their decision may be more willing to reflect upon and accept the value of the AI suggestions. Such a difference was not however observed in terms of NND, for which there were no significant differences among the different configurations.

To explain the discrepancy between the pictures drawn by the agreement and reliance metrics, we can observe that AI support exerted only a moderate to small dominance and deference, consistently below 10%, and in some cases even significantly smaller than 5%. As a result, a relevant portion of the observed agreement between FHD and AI was simply due to the HD1 decision already aligning with the AI support, rather than reflecting the influence of the AI. Dominance was notably stronger for specialists than for sub-specialists as well as for low confidence users compared to high-confidence users. By contrast, except for specialists, there were no differences in dominance strength between AI with and without explainability, while there were significant, but pragmatically small, differences in terms of deference strength. By contrast, there were significant differences both in terms of dominance direction and deference orientation: in all cases, except for sub-specialists, the number of decision switches for the better was significantly higher when the AI was not accompanied by explanations. In this case, that is for non-explainable AI support, notably specialists had a significantly larger fraction of decision switches for the better than sub-specialists (approximately one case out of three against one out of ten), and the same held for low confidence users over high confidence ones (with all decision switches being for the better against one out of two). The picture reversed when considering explainable support where, notably, low confidence users seem to have been negatively impacted by it, and significantly so.

A deeper understanding of the difference between agreement and reliance can be found by examining the Automation Bias and Self-anchoring Bias metrics. In all cases, except for the low confidence users supported by AI, users exhibited a markedly high self-anchoring bias, as the null hypothesis of no self-anchoring bias was rejected (with the confidence intervals not containing the value 1). Consequently, in most cases, users rejected AI suggestions even when they were correct, and accepting them would have led to improvement. Furthermore, the same configurations showed no significant evidence of automation bias, as the null hypothesis of no automation bias could not be rejected. These results further explain why appropriate reliance was low compared to agreement: in all of the mentioned groups, users simply disregarded AI advice, regardless of whether it was correct (and following it would have led to improvement) or incorrect (and following it would have led to a worsening). As a result, the patterns 000 and 111 were relatively common, explaining the relatively high Percent of Agreement, while the pattern 011 occurred much less frequently than the pattern 010. In contrast, the group of AI-supported low confidence users showed no evidence of either self-anchoring or automation bias or the lack of them, as the confidence interval for the two metrics straddled both sides of 1.

The limited impact of AI support, which might be overlooked when considering agreement alone, becomes more evident when looking at the influence metrics. In terms of the Influence Index, when we account for the possibility that decision switches could also be due to chance, AI support showed only modest influence on users. While the chance-adjusted agreement between FHD and AI was significantly higher than that between HD1 and AI, this difference was relatively small, amounting to less than one (chance-adjusted) case out of 5. Furthermore, for all groups except the AI-supported low confidence users, the observed appropriate reliance was significantly lower than would be expected by chance, as indicated by the consistently negative Appropriate Influence metric. At the same time, similarly to how the difference in appropriate reliance between the AI-supported low confidence users and the other groups was large and significant (approximately 25%), accounting for chance reveals an even larger difference, with AI-supported low confidence users being the only group where Appropriate Influence was positive and significantly so, and also much larger than for the other groups.

6.2. Interpretation of the ECG study

The results of the ECG user study are shown in Table 7 and Fig. 4.

Similarly to the MRI case study, the results of the ECG study also reveal markedly different insights depending on whether one examines agreement, reliance, or influence metrics, underscoring the limitations of relying solely on agreement scores to evaluate the impact of AI support.

At first glance, the Percent of Agreement suggests a broadly positive effect of AI support, with consistently high values across all configurations (exceeding 70%) and notably higher scores for novices compared to experts. Additionally, explainable AI (XAI) appears to improve agreement for both novices and low performers, while also reducing the Error Rate in Agreement (hence, improving the accuracy in the agreement cases), and significantly so.

However, reliance metrics offer a more nuanced view. Appropriate reliance, while relatively stable across configurations, indicates that

Table 6
The results from the MRI user study, stratified by level of user expertise and type of AI support.

Metric	Specialist - XAI	Specialist - AI	Sub-specialist - XAI	Sub-specialist - AI	HC - XAI	HC - AI	LC - XAI	LC - AI	HP - XAI	HP - AI	LP - XAI	LP - AI
Percent of Agreement	0.70 [0.69,0.71]	0.73 [0.73,0.74]	0.67 [0.66,0.67]	0.74 [0.74,0.75]	0.68 [0.68,0.68]	0.74 [0.74,0.75]	0.71 [0.69,0.73]	0.75 [0.73,0.77]	0.7 [0.69,0.7]	0.75 [0.74,0.75]	0.67 [0.66,0.67]	0.73 [0.73,0.74]
Error Rate in Agreement	0.15 [0.14,0.15]	0.07 [0.06,0.07]	0.1 [0.1,0.11]	0.07 [0.07,0.07]	0.12 [0.12,0.12]	0.07 [0.07,0.07]	0.11 [0.1,0.12]	0.13 [0.11,0.14]	0.13 [0.12,0.13]	0.06 [0.06,0.06]	0.11 [0.11,0.12]	0.08 [0.08,0.08]
Dominance Strength	0.08 [0.08,0.08]	0.06 [0.06,0.07]	0.04 [0.04,0.04]	0.04 [0.04,0.04]	0.05 [0.05,0.06]	0.05 [0.04,0.05]	0.1 [0.09,0.11]	0.1 [0.09,0.11]	0.07 [0.07,0.08]	0.07 [0.07,0.07]	0.04 [0.03,0.04]	0.03 [0.02,0.03]
Dominance Orientation	0.14 [0.1,0.19]	0.38 [0.33,0.42]	0.18 [0.14,0.22]	0.1 [0.05,0.15]	0.12 [0.09,0.15]	0.27 [0.24,0.31]	−0.33 [−0.45,–0.22]	1.0 [1.0,1.0]	0.17 [0.14,0.21]	0.08 [0.04,0.12]	0.14 [0.09,0.2]	0.6 [0.55,0.65]
Deference Strength	0.07 [0.07,0.08]	0.06 [0.05,0.06]	0.04 [0.04,0.04]	0.03 [0.03,0.03]	0.05 [0.05,0.05]	0.04 [0.03,0.04]	0.07 [0.06,0.08]	0.1 [0.09,0.11]	0.07 [0.07,0.07]	0.05 [0.05,0.05]	0.03 [0.03,0.03]	0.02 [0.02,0.03]
Deference Orientation	0.26 [0.22,0.31]	0.57 [0.53,0.62]	0.3 [0.26,0.34]	0.43 [0.38,0.48]	0.24 [0.21,0.28]	0.54 [0.5,0.57]	0.0 [−0.15,0.15]	1.0 [1.0,1.0]	0.26 [0.22,0.29]	0.37 [0.33,0.41]	0.33 [0.28,0.39]	0.78 [0.74,0.82]
AI Decision Impact	1.06 [0.72,1.54]	1.15 [0.75,1.76]	1.04 [0.79,1.37]	1.03 [0.75,1.41]	1.03 [0.82,1.30]	1.08 [0.83,1.41]	0.86 [0.30,2.49]	1.58 [0.42,6.01]	1.07 [0.78, 1.48]	1.04 [0.72,1.50]	1.03 [0.75,1.40]	1.10 [0.78,1.56]
Team AI Effect on Decision	0.19 [0.12,0.25]	0.34 [0.3,0.39]	0.12 [0.1,0.14]	0.08 [0.05,0.12]	0.09 [0.08,0.11]	0.22 [0.2,0.24]	0.2 [0.15,0.26]	0.18 [0.16,0.21]	0.21 [0.1,0.32]	0.1 [0.05,0.14]	0.08 [0.05,0.1]	0.32 [0.28,0.36]
NND	17 [0,462]	12 [0,153]	26 [0,115]	47 [0,116]	33 [1,66]	19 [4,34]	14 [9,20]	15 [13,18]	16 [0,136]	41 [17,66]	39 [0,166]	13 [5,22]
Automation Bias	0.22 [0.10,0.49]	0.10 [0.03,0.31]	0.09 [0.04,0.18]	0.06 [0.02,0.16]	0.12 [0.07,0.21]	0.06 [0.03,0.13]	0.58 [0.07,4.68]	0.18 [0.01,4.00]	0.15 [0.08,0.30]	0.13 [0.06,0.29]	0.08 [0.03,0.22]	0.03 [0.00,0.13]
Self-anchoring Bias	4.19 [2.19,8.05]	3.54 [1.78,7.04]	8.66 [4.83,15.52]	6.40 [3.28,12.48]	7.04 [4.49,11.04]	4.83 [2.96,7.90]	6.22 [1.00,38.88]	1.48 [0.26,8.60]	3.98 [2.29,6.90]	3.51 [1.87,6.58]	11.88 [5.77,24.46]	7.14 [3.41,14.95]
Appropriate Reliance	0.26 [0.25,0.26]	0.26 [0.26,0.27]	0.28 [0.28,0.29]	0.27 [0.26,0.27]	0.27 [0.26,0.27]	0.27 [0.26,0.27]	0.27 [0.25,0.28]	0.5 [0.48,0.52]	0.27 [0.26,0.27]	0.21 [0.21,0.22]	0.28 [0.28,0.29]	0.32 [0.32,0.32]
Appropriate Influence	−0.55 [−0.56, −0.54]	−0.57 [−0.58, −0.56]	−0.53 [−0.53, −0.52]	−0.62 [−0.63, −0.61]	−0.55 [−0.55, −0.54]	−0.61 [−0.61, −0.6]	−0.39 [−0.42, −0.35]	0.07 [0.03,0.11]	−0.59 [−0.6, −0.58]	−0.73 [−0.74, −0.72]	−0.48 [−0.49, −0.47]	−0.48 [−0.49, −0.47]
Influence Index	0.15 [0.14,0.16]	0.12 [0.12,0.13]	0.08 [0.07,0.08]	0.04 [0.04,0.05]	0.1 [0.1,0.11]	0.07 [0.06,0.07]	0.17 [0.14,0.19]	0.18 [0.15,0.21]	0.15 [0.14,0.15]	0.11 [0.1,0.11]	0.06 [0.05,0.06]	0.03 [0.03,0.04]

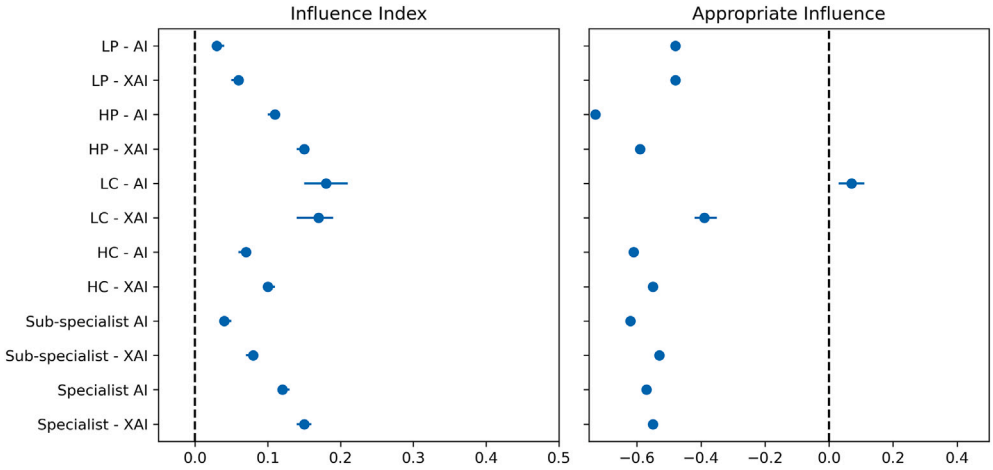

Fig. 3. Results from the MRI user study, showing the Influence Index (left) and Appropriate Influence (right) metrics, stratified by user skills (LP stands for Low Performers, HP for High Performers), user confidence ('LC' stands for Low Confidence, 'HC' for High Confidence), level of expertise, and type of AI support ('AI' refers to advice-only support, while 'XAI' represents explainable support). Ninety-five percent confidence intervals were generated using bootstrapping.

users appropriately relied on the AI in only about half of the cases. This uniformity obscures the stark differences in how AI impacts decision-making. For instance, novices using non-explainable AI achieved the highest appropriate reliance, suggesting they were not only more willing to follow AI advice than other groups (as highlighted by the higher percent of agreement) but also had better trust calibration. Also in terms of NND, a significantly smaller number of decisions would be needed to bring about an improvement in decision accuracy for the AI-supported novices and low-performers than for other configurations.

To explain the discrepancy between the pictures drawn by the agreement and reliance metrics, we can observe that AI support exerted only a moderate to small dominance and deference, consistently below 25%. As a result, a significant portion of the observed agreement between FHD and AI was simply due to the HD1 decision already aligning with the AI support, rather than reflecting the influence of the AI. Dominance was notably stronger for novices than for experts, as well as for low performers than high performers. Furthermore, AI without explainability resulted in significantly higher dominance than when explanations were provided. In fact, novices and low performers with AI support changed their opinion in one case out of four, and in all of these cases, the revised user decision (FHD) was in agreement with the AI. In contrast, for all other configurations, the number of decision switches was significantly smaller than 10%, indicating that AI support influenced human decisions in fewer than one case out of ten.

A deeper understanding of the difference between agreement and reliance can be found by examining the Automation Bias and Self-anchoring Bias metrics. In all cases, except for the AI-supported novices and low-performers, users exhibited high self-anchoring bias, as the null hypothesis of no self-anchoring bias was rejected (with the confidence intervals not containing the value 1). Consequently, in most cases, users rejected AI suggestions even when they were correct, and accepting them would have led to improvement. The same configurations showed no significant evidence of automation bias. These results further explain why appropriate reliance was low compared to agreement: in the mentioned groups, users simply disregarded AI advice, regardless of whether it was correct (and following it would have led to improvement) or incorrect (and following it would have led to a worsening). As a result, the patterns 000 and 111 were relatively common, explaining the relatively high Percent of Agreement, while the pattern 011 occurred much less frequently than the pattern 010. In contrast, the AI-supported novices and low-performers showed no evidence of either self-anchoring or automation bias, although the presence of automation bias could not be ruled out, as the confidence interval for the Automation Bias metric in both groups straddled both sides of 1.

The limited impact of AI support, which might be overlooked when considering agreement alone, becomes more evident when looking at the influence metrics. In terms of the Influence Index, when we account for the possibility that decision switches could also be due to chance, AI support showed only modest influence on users. While the chance-adjusted agreement between FHD and AI was significantly higher than that between HD1 and AI, this difference was relatively small, except for the AI-supported novices and low performers. Furthermore, for all groups except the AI-supported novices, low-performers and high-performers, the observed appropriate reliance was significantly lower than would be expected by chance, as indicated by the consistently negative Appropriate Influence metric. At the same time, while the differences in appropriate reliance between the AI-supported novices, low-performers and high-performers and the other groups were small but significant (approximately 5%), accounting for chance reveals a much larger difference, with AI-supported novices and low performers being the only groups where Appropriate Influence was positive and significantly so, and also much larger than for the other groups. In particular, novices reported the highest value of appropriate influence.

6.3. Interpretation of the x-ray results

The results of the X-ray user study are shown in Table 8 and Fig. 5.

Similarly to the MRI and ECG studies, also in the X-ray study the agreement metrics suggest a broadly positive effect of AI support. Indeed, all configurations of AI and user expertise (as well as confidence) showed a rather large agreement, as shown by the Percent of Agreement metric being generally larger than 70%. Agreement was significantly higher for specialists than for either residents or sub-specialists, as it was higher for low-confidence users than for high-confidence ones. The picture reverses, however, when looking at the error rate in agreement. Specialists erred significantly more than both specialists and sub-specialists when they agreed with the AI support (one case out of 10 against one case out of 20, for both residents and sub-specialists), similarly as low-confidence users erred significantly more than high-confidence ones (approximately one case out of 10 against less than one case out of 20). By contrast, low- and high-performers reported the same value of Percent of Agreement, while low-performers erred significantly less than high-performers in cases of agreement with the AI support.

Reliance metrics offer a picture which is different from the one that could be obtained by looking at agreement alone. Appropriate reliance, while relatively stable across configurations, indicates that users appropriately relied on the AI in only about one half of the cases. Residents appropriately relied on the AI support significantly more than both

Table 7
The results from the ECG user study, stratified by level of user expertise and type of AI support.

Metric	Novice - AI	Novice - XAI	Expert - AI	Expert - XAI	LP - AI	LP - XAI	HP - AI	HP - XAI
Percent of Agreement	0.82 [0.82,0.83]	0.86 [0.86,0.87]	0.72 [0.71,0.73]	0.73 [0.72,0.74]	0.78 [0.77,0.78]	0.82 [0.82,0.83]	0.78 [0.78,0.79]	0.79 [0.79,0.8]
Error Rate in Agreement	0.30 [0.29,0.31]	0.28 [0.28,0.29]	0.23 [0.23,0.24]	0.23 [0.23,0.24]	0.3 [0.29,0.3]	0.28 [0.27,0.28]	0.25 [0.24,0.26]	0.24 [0.23,0.25]
Dominance Strength	0.25 [0.24,0.25]	0.05 [0.04,0.05]	0.06 [0.06,0.06]	0.02 [0.01,0.02]	0.23 [0.23,0.24]	0.05 [0.05,0.05]	0.08 [0.07,0.08]	0.01 [0.01,0.01]
Dominance Orientation	0.73 [0.71,0.74]	0.82 [0.78,0.85]	0.45 [0.41,0.5]	1.0 [1.0,1.0]	0.68 [0.66,0.7]	0.83 [0.8,0.86]	0.71 [0.68,0.75]	1.0 [1.0,1.0]
Deference Strength	0.25 [0.24,0.25]	0.04 [0.04,0.04]	0.06 [0.06,0.06]	0.02 [0.01,0.02]	0.23 [0.23,0.24]	0.05 [0.04,0.05]	0.08 [0.07,0.08]	0.01 [0.01,0.01]
Deference Orientation	0.73 [0.71,0.74]	0.80 [0.76,0.84]	0.45 [0.41,0.5]	1.0 [1.0,1.0]	0.68 [0.66,0.7]	0.82 [0.79,0.85]	0.71 [0.68,0.75]	1.0 [1.0,1.0]
AI Decision Impact	2.07 [1.44,2.99]	1.18 [0.81,1.72]	1.13 [0.73,1.75]	1.08 [0.69,1.68]	1.90 [1.32,2.74]	1.20 [0.83,1.74]	1.28 [0.83,1.98]	1.05 [0.67,1.65]
Team AI Effect on Decision	1.23 [1.19,1.27]	0.44 [0.41,0.46]	0.38 [0.35,0.41]	0.25 [0.23,0.27]	1.02 [0.98,1.06]	0.51 [0.49,0.54]	0.74 [0.71,0.78]	0.18 [0.17,0.2]
NND	3 [3,4]	7 [6,9]	8 [0,24]	12 [9,16]	3 [3,4]	6 [5,8]	5 [5,6]	17 [3,22]
Automation Bias	0.62 [0.26,1.48]	0.11 [0.02,0.63]	0.13 [0.04,0.41]	0.02 [0.00, 0.31]	0.52 [0.23,1.18]	0.09 [0.02,0.46]	0.12 [0.03,0.44]	0.02 [0.00,0.38]
Self-anchoring Bias	0.51 [0.31,0.84]	2.27 [1.03,5.00]	3.80 [1.71,8.42]	8.32 [2.67,25.97]	0.73 [0.45,1.17]	2.71 [1.30,5.68]	1.82 [0.88,3.77]	8.63 [2.27,32.74]
Appropriate Reliance	0.56 [0.55,0.56]	0.50 [0.50,0.50]	0.50 [0.50,0.50]	0.50 [0.50,0.51]	0.54 [0.54,0.54]	0.5 [0.5,0.5]	0.52 [0.52,0.53]	0.51 [0.5,0.51]
Appropriate Influence	0.23 [0.22,0.23]	−0.02 [−0.03,-0.02]	−0.04 [−0.04,-0.03]	−0.06 [−0.06,-0.05]	0.19 [0.18,0.2]	−0.02 [−0.02,-0.01]	0.02 [0.01,0.03]	−0.07 [−0.07,-0.06]
Influence Index	0.41 [0.40,0.42]	0.08 [0.07,0.08]	0.13 [0.12,0.13]	0.03 [0.03,0.03]	0.4 [0.39,0.41]	0.08 [0.07,0.09]	0.15 [0.14,0.16]	0.02 [0.02,0.03]

specialists and, even more so, sub-specialists: indeed, the differences between all of the three groups were statistically significant. This suggests that appropriate reliance negatively correlates with professional expertise. Similarly, low-confidence users, who were initially unsure about their decision, had a significantly larger appropriate reliance than high-confidence ones. Also, low-performers relied appropriately on the AI significantly more often than high-performers, confirming the effect observed on residents, specialists and sub-specialists that in the X-ray case studies appropriate reliance seemed to be negatively correlated with skill or expertise.

To explain the discrepancy between the pictures drawn by the agreement and reliance metrics, we can observe the dominance and deference metrics. Similarly to the previous two studies, also in this case AI support exerted only a moderate to small dominance and deference. As a result, a relevant portion (more than two cases out of three) of the observed agreement between FHD and AI was simply due to the HD1 decision already aligning with the AI support, rather than reflecting the influence of the AI. Dominance, as well as deference, was notably stronger for residents than for both specialists and sub-specialists. Also in terms of dominance orientation and deference orientation, residents made a significantly larger number of decision switches for the better than all other groups. Similarly, specialists switched their decisions significantly more often than sub-specialists and significantly more often did so for the better. The same pattern was also observed between low-performers and high-performers, in which the AI exerted a significantly larger dominance on the former group, which also reported a significantly larger deference as well number of switches for the better. By contrast, high-confidence users switched their decisions significantly more often than low-confidence users (although the difference disappeared when we consider switches that resulted in an agreement with the AI). The difference was even more relevant when considering the orientation of these decision switches, where not only did the high-confidence users significantly more often switch their decisions for the better, but also low-confidence users more often than not changed their decisions for the worse.

The Automation Bias and Self-anchoring Bias metrics provide a similar analysis, although less clearly evident than for the MRI and ECG studies due to the reduced small size. In all cases, there was no evidence of either automation bias or self-anchoring bias. However, the presence of either automation bias and self-anchoring bias could be excluded only for residents, sub-specialists, low-performers and high-confidence users: by contrast, for both specialists, low-confidence users, and high-performers the presence of self-anchoring bias could not be disproved.

The limited impact of AI support, which might be overlooked when considering agreement alone, becomes more evident when looking at the influence metrics. In terms of the Influence Index, when we account for the possibility that decision switches could also be due to chance, AI support showed moderate influence on users, with significant differences across levels of user expertise, skill and confidence. Specifically, residents and specialists were both significantly more influenced than sub-specialists, low-confidence users were significantly more influenced than high-confidence ones, and low-performers were significantly more influenced than high-performers. Interestingly, even though low-confidence users were the group which was more strongly influenced among the considered ones, it was the one for which more decision switches were for the worse. Furthermore, for all groups except the residents, the observed appropriate reliance was significantly lower than would be expected by chance, as indicated by the consistently negative Appropriate Influence metric. While for residents the appropriate influence was positive, it was not significantly so. Interestingly, even though low-confidence users made more decision switches for the worse than for the better (as the dominance orientation for this group was negative), their appropriate influence was higher than for all other groups except the residents. This finding may be explained by seeing that the deference orientation for the same group was significantly positive: thus,

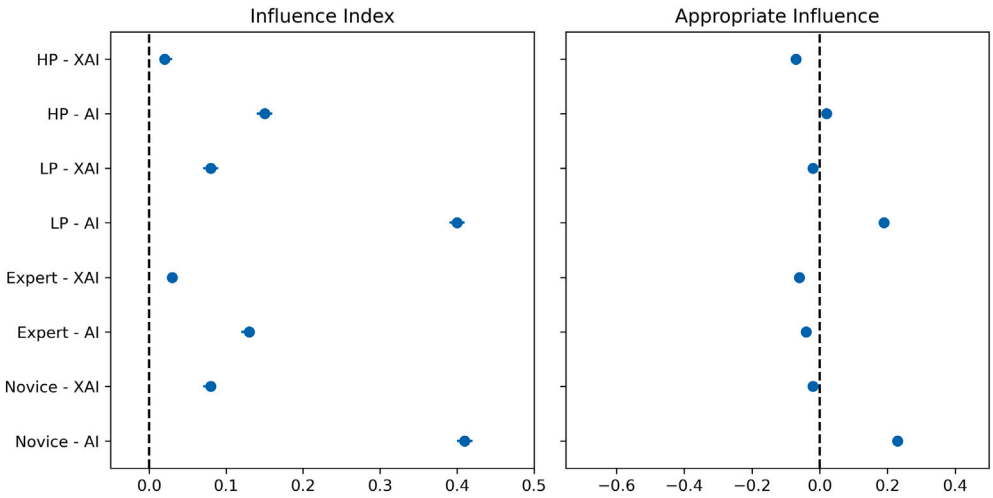

Fig. 4. Results from the ECG user study, showing the Influence Index (left) and Appropriate Influence (right) metrics, stratified by user skills ('LP' stands for Low Performers, 'HP' for High Performers), expertise level, and type of AI support ('AI' refers to advice-only support, while 'XAI' represents explainable support). Ninety-five percent confidence intervals were generated using bootstrapping.

the observed negative dominance is mostly due to the 110 pattern, associated with extreme algorithm aversion. As this pattern is not involved in the computation of appropriate reliance (and, consequently, of appropriate influence), this explains why the low-confidence users had higher appropriate influence (and reliance) than all other groups except the residents from a metric point of view. Intuitively, this behavior could be explained by conjecturing that a user who is not very confident in its own decision could end up doubting the AI support being correct when this latter agrees with their own decision.

The interpretation and insights that could be drawn from the results of the X-ray case study are more nuanced and less straightforward than for the MRI and ECG case studies. Indeed, except for the low-confidence users, neither the presence nor absence of automation and self-anchoring bias could be proven, thus not making it immediately clear how to intervene on the AI support (see also Section 6.5). Nonetheless, we note that for sub-specialists and high-performers, the self-anchoring bias metric was much larger than the automation bias one, while for specialists and low-performers the opposite was true: thus, a potential direction for improvement in sub-specialists and high-performers (resp., specialists and low-performers) would be to increase the likelihood that the AI support is accepted (resp. rejected). By contrast, for low-confidence users, as also described above, a potential way to improve the appropriate influence would be to reduce the risk of extreme algorithmic aversion: this could be done by informing the user about the relatively small likelihood that both his and the AI's interpretations are incorrect when they agree with each other (as an example, such information could be provided as a disclaimer to the user).

6.4. Interpretation of the endoscopy results

The results of the Endoscopy user study are shown in Table 9 and Fig. 6.

As in the other studies, also for Endoscopy case the Percent of Agreement is uniformly high across roles. This indicates a general alignment between human decisions and AI advice but does not distinguish how roles interact with or benefit from AI in terms of decision quality and reliance. A slightly different picture is captured by the Error Rate in Agreement, where specialists reported a significantly higher accuracy than trainees on the cases on which they agreed with the AI support, as well as low-performers reported a significantly higher accuracy than high-performers on the same cases.

In terms of reliance, however, we see that similarly to the ECG and X-ray studies, users appropriately relied on the AI support in only half of the cases, with no difference between specialists and trainees. The discrepancy between agreement and reliance can be understood through the lens of the Dominance Strength: users, irrespective of expertise or skill level, changed their opinion in less than one case out of 10 after receiving the support of the AI: in most cases, these switches led to agreement with the AI. The dominance, as well as the deference, was significantly higher for trainees than for specialists: however, by contrast, trainees switched their decision for the worse significantly more often than specialists. A similar analysis is also suggested by the Automation Bias and Self-anchoring bias. Irrespective of expertise or skill level, there was no significant evidence of the presence of automation bias, though the automation bias was slightly higher (but not significantly so) for the trainees than for all other groups. By contrast, for all of the groups there was significant evidence of the presence of self-anchoring bias: this confirms the analysis based on the dominance strength, and suggests that in most cases the users simply ignored the AI advice. Also, while the self-anchoring bias was higher for specialists than for trainees, there was no significant difference between the two user groups.

Influence metrics reinforce the observations made based on the reliance metrics, and also underscore the nuanced relationship between roles and AI. While trainees were influenced significantly more than specialists (with a difference of approximately one chance-adjusted case out of 20) and high-performers were similarly influenced significantly more than low-performers (albeit to a much smaller degree, with a difference of approximately one chance-adjusted case out of 50), all groups were only weakly influenced by the AI. Despite this difference in influence, appropriate influence was the same across all groups: users appropriately relied on the AI slightly less than expected by chance, as highlighted by the observed negative values of Appropriate Influence. Combined with the findings above, these results suggest that there are only minor differences between users of different expertise or skill levels in how they use and rely upon the AI: in all cases, the metrics suggest that the most likely cause for the low appropriate reliance is the presence of self-anchoring bias.

6.5. Further remarks

We explored appropriate reliance on AI-based decision support systems across four user studies in medical contexts. Our findings provide several insights into decision-makers' interaction with AI and highlight factors influencing trust calibration and reliance on these systems. Below, we discuss the main outcomes and their implications.

Table 8
Results of the X-ray study, in terms of Influence Index (left) and Appropriate Influence (right), stratified by level of user expertise and user confidence.

Metrics	Resident	Sub-specialist	Specialist	HC	LC	LP	HP
Percent of Agreement	0.74 [0.74,0.75]	0.74 [0.74,0.75]	0.80 [0.79,0.81]	0.73 [0.73,0.74]	0.8 [0.79,0.81]	0.75 [0.74,0.76]	0.75 [0.74,0.76]
Error Rate in Agreement	0.05 [0.05,0.06]	0.05 [0.05,0.06]	0.10 [0.09,0.11]	0.03 [0.03,0.04]	0.13 [0.12,0.14]	0.05 [0.04,0.05]	0.07 [0.07,0.08]
Dominance Strength	0.31 [0.31,0.32]	0.18 [0.18,0.19]	0.25 [0.23,0.27]	0.25 [0.24,0.25]	0.22 [0.21,0.22]	0.3 [0.29,0.3]	0.17 [0.16,0.17]
Dominance Orientation	0.59 [0.56,0.61]	0.15 [0.12,0.19]	0.33 [0.27,0.39]	0.54 [0.52,0.56]	−0.06 [−0.11,-0.01]	0.46 [0.43,0.49]	0.24 [0.19,0.28]
Deference Strength	0.27 [0.26,0.28]	0.13 [0.13,0.14]	0.19 [0.18,0.21]	0.19 [0.19,0.2]	0.19 [0.18,0.2]	0.23 [0.23,0.24]	0.13 [0.13,0.14]
Deference Orientation	0.79 [0.77,0.82]	0.47 [0.44,0.51]	0.43 [0.36,0.50]	0.85 [0.83,0.87]	0.07 [0.02,0.12]	0.74 [0.72,0.76]	0.41 [0.37,0.45]
AI Decision Impact	3.29 [1.60,6.75]	1.27 [0.64,2.50]	1.71 [0.52,5.63]	2.86 [1.62,5.04]	0.92 [0.42,2.04]	2.75 [1.47,5.13]	1.33 [0.68,2.59]
Team AI Effect on Decision	6.45 [6.32,6.59]	0.97 [0.88,1.05]	–	1.3 [1.25,1.34]	−0.29 [−0.37,-0.22]	2.24 [2.12,2.36]	0.45 [0.39,0.5]
NND	4 [4,5]	9 [0,41]	–	5 [5,6]	Non-Positive Effect	5 [5,6]	11 [0,54]
Automation Bias	0.15 [0.05,0.49]	0.18 [0.07,0.48]	0.52 [0.10,2.64]	0.07 [0.02,0.22]	0.77 [0.28,2.11]	0.16 [0.06,0.42]	0.23 [0.09,0.62]
Self-anchoring Bias	0.13 [0.05,0.38]	0.35 [0.13,0.97]	0.24 [0.04,1.56]	0.12 [0.05,0.31]	0.62 [0.20,1.90]	0.06 [0.02,0.22]	0.66 [0.26,1.63]
Appropriate Reliance	0.53 [0.52,0.54]	0.40 [0.39,0.4]	0.47 [0.46,0.49]	0.43 [0.42,0.43]	0.53 [0.52,0.54]	0.49 [0.48,0.5]	0.41 [0.4,0.42]
Appropriate Influence	0.01 [−0.01,0.04]	−0.41 [−0.43,-0.4]	−0.17 [−0.21,-0.13]	−0.27 [−0.28,-0.25]	−0.07 [−0.1,-0.04]	−0.12 [−0.14,-0.1]	−0.34 [−0.36,-0.32]
Influence Index	0.33 [0.30,0.35]	0.22 [0.20,0.24]	0.33 [0.28,0.37]	0.16 [0.15,0.17]	0.46 [0.44,0.48]	0.29 [0.28,0.31]	0.25 [0.23,0.27]

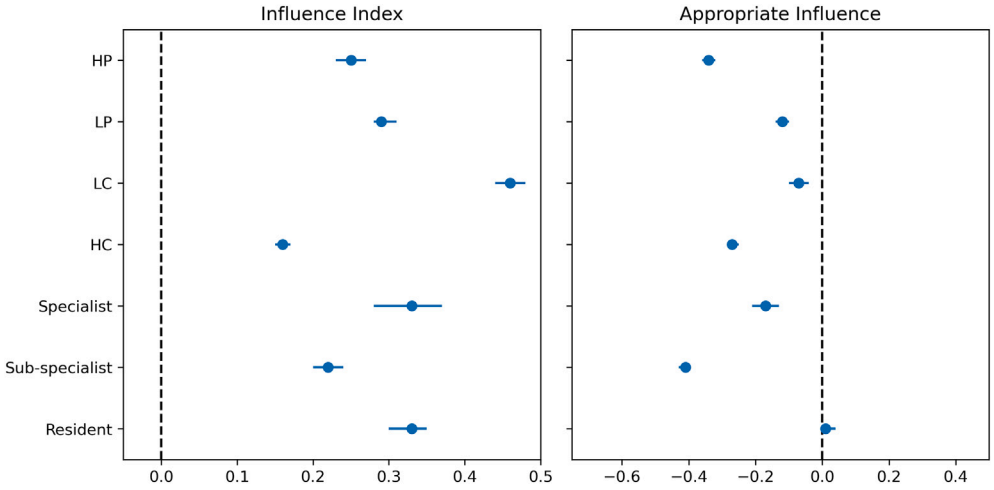

Fig. 5. Results from the X-ray user study, showing the Influence Index (left) and Appropriate Influence (right) metrics, stratified by user skills ('HP' stands for High Performers, 'LP' for Low Performers), user confidence ('LC' stands for Low Confidence, 'HC' for High Confidence) and professional role (level of expertise). Ninety-five percent confidence intervals were generated using bootstrapping.

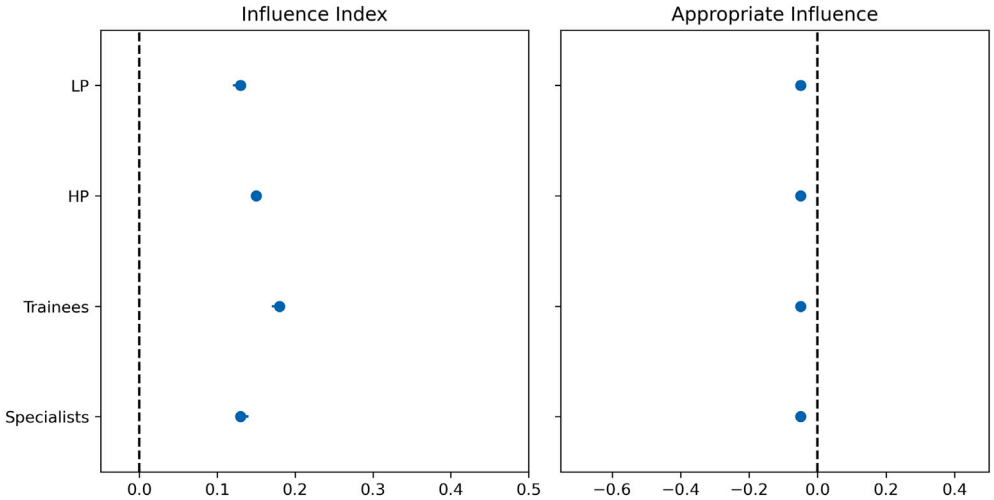

Fig. 6. Results from the Endoscopy user study, showing the Influence Index (left) and Appropriate Influence (right) metrics, stratified by user skills ('LP' stands for Low Performers, 'HP' for High Performers), and professional role (expertise). Ninety-five percent confidence intervals were generated using bootstrapping.

Table 9
The results from the Endoscopy user study, stratified by level of user expertise (HP, High Performers, LP Low Performers).

Metric	Specialists	Trainees	HP	LP
Percent of Agreement	0.81 [0.81,0.81]	0.81 [0.80,0.81]	0.81 [0.81,0.81]	0.81 [0.81,0.81]
Error Rate in Agreement	0.11 [0.11,0.11]	0.13 [0.13,0.13]	0.12 [0.12,0.12]	0.11 [0.11,0.11]
Dominance Strength	0.08 [0.08,0.08]	0.09 [0.09,0.10]	0.08 [0.08,0.08]	0.08 [0.08,0.08]
Dominance Orientation	0.37 [0.36,0.37]	0.35 [0.34,0.36]	0.35 [0.35,0.36]	0.4 [0.39,0.41]
Deference Strength	0.07 [0.07,0.07]	0.09 [0.08,0.09]	0.07 [0.07,0.07]	0.07 [0.07,0.07]
Deference Orientation	0.43 [0.42,0.44]	0.40 [0.39,0.41]	0.41 [0.41,0.42]	0.47 [0.46,0.49]
AI Decision Impact	1.20 [1.12,1.28]	1.21 [1.07,1.38]	1.19 [1.11,1.28]	1.22 [1.08,1.39]
Team AI Effect on Decision	0.33 [0.33,0.34]	0.35 [0.34,0.36]	0.34 [0.33,0.34]	0.35 [0.34,0.36]
NND	12 [12,13]	11 [11,12]	12 [12,13]	12 [12,13]
Automation Bias	0.20 [0.17,0.23]	0.28 [0.21,0.36]	0.21 [0.18,0.25]	0.20 [0.15,0.27]
Self-anchoring Bias	1.97 [1.76,2.20]	1.67 [1.36,2.06]	1.91 [1.70,2.14]	1.90 [1.55,2.33]
Appropriate Reliance	0.51 [0.51,0.51]	0.51 [0.51,0.51]	0.51 [0.51,0.51]	0.51 [0.51,0.51]
Appropriate Influence	−0.05 [−0.05,-0.05]	−0.05 [−0.05,-0.05]	−0.05 [−0.05,-0.05]	−0.05 [−0.05,-0.04]
Influence Index	0.13 [0.13,0.14]	0.18 [0.17,0.18]	0.15 [0.15,0.15]	0.13 [0.12,0.13]

The primary objective was to investigate how AI's decision support impacts decision-makers' reliance behaviors and to propose metrics for evaluation. We examined four medical fields: radiology, cardiology, orthopedics, and gastroenterology, where participants made diagnostic decisions using AI systems of varying accuracy. Reliance patterns differed across contexts, suggesting that reliance on AI is domain-specific and heavily influenced by interaction protocols and AI accuracy. Our key findings emphasize the need to focus on influence, rather than simple reliance, let alone agreement, to assess AI's effectiveness in decision-making environments.

Our results align with prior studies on AI reliance, particularly the notion that AI systems can induce both appropriate reliance and automation biases (e.g., Lee and See, 2004). However, our findings extend the current understanding by introducing metrics that capture not only reliance but also the influence of AI on cognitive processes. Metrics such as the *Influence Index* and *Appropriate Influence* offer more granular insights into how AI systems shape decisions, providing a more comprehensive framework for evaluating the interplay between human cognition and machine intelligence.

Additionally, our metric framework not only provides a detailed, static overview of a case study but also offers insights into addressing observed discrepancies among agreement, reliance, and influence. Specifically, in three out of four studies (MRI, ECG, and Endoscopy), we found significant evidence of self-anchoring bias across almost all groups. This bias explains why users relied on the AI less than expected, as the appropriate influence was consistently below zero for most groups. Despite the AI exerting a small but significant influence, users largely ignored its support and struggled to accurately identify cases where following the AI's correct suggestions would have improved their performance.

These findings suggest that corrective measures to enhance appropriate reliance should focus on increasing the likelihood that users correctly detect and accept accurate AI recommendations. While explanations are often proposed in the literature as a means to improve understanding and appropriate reliance, our results indicate that this approach is either insufficient or counterproductive in isolation. In the MRI and ECG studies, the provision of explanations not only exacerbated self-anchoring bias but also, particularly for less experienced or low-confidence users, worsened appropriate reliance and influence.

Thus, explanations alone are insufficient for improving reliance on AI in this context. Additional strategies should be explored to mitigate self-anchoring bias. Potential approaches include increasing users' awareness of the AI's superior performance, training users in the effective use of AI-based diagnostic systems, or adopting advanced interaction protocols that better arbitrate between AI and human decisions, thereby reducing the likelihood of users outright rejecting AI advice.

The observed differences in reliance between less and more experienced users, as well as between low- and high-confidence users, warrant further discussion. While the general conclusions mentioned above hold across the four case studies and most user groups, certain groups— AI-supported novices and low-performers in the ECG study, low-confidence users in the MRI study, and residents in the X-ray study—exhibited significantly different patterns. Specifically, for these groups, there was no significant evidence of self-anchoring bias, and their appropriate reliance significantly exceeded expectations.

In two of these groups (novices in the ECG study and residents in the X-ray study), automation bias was slightly more pronounced than self-anchoring bias, although neither bias was statistically significant. These findings suggest that less experienced users, who may exhibit a higher tendency to trust AI due to a greater perceived utility or familiarity with technological support, might benefit from strategies aimed at reducing the risk of automation bias.

One potential approach could involve providing less experienced users with uncertainty-aware AI systems that present a calibrated estimate of uncertainty in their recommendations. By increasing users' awareness that AI suggestions may sometimes be incorrect, such systems could help mitigate over-reliance and foster more appropriate trust in AI.

Thus, the main general insight that could be drawn from our studies pertains to the role of selecting most fit interaction protocols so as to optimize trust calibration and appropriate reliance in a *given setting* and for a *given user profile*. This suggests that AI systems and human-AI interaction protocols should be designed and adjusted based on evidence about the reliance behavior exhibited by their users, through a process-oriented, technovigilance-inspired approach that periodically updates system's design based on increasing and dynamically changing evidence about AI use (Cabitza and Zeitoun, 2019).

Obviously, our study is not without limitations. In particular, as we also mentioned at the beginning of Section 6, we want to emphasize that the main aim of the case studies was to illustrate the use of our main methodological contribution, that is, the metric framework for evaluating the phenomenon of reliance. In this sense, even though we adopted an approach grounding on statistical hypothesis testing, we note that our results should not be blindly generalized, as our analysis does not include correction measures for multiple testing, nor sophisticated approaches to model associations between user characteristics and reliance dimensions beyond stratification. In practical use, we would not suggest the use of all metrics or the application of all stratifications used in the paper, to avoid alpha error multiplication. Indeed, selecting the relevant strata and the most appropriate metrics, as well as evaluating claims of significance, must be carefully done by the researcher, based on the specifics of their setting: practitioners and researchers should carefully decide which metrics they prefer to adopt for their own studies, and they are therefore responsible for addressing issues of statistical significance in relation to the number and dependency of their research questions. Therefore, while our results highlight some relevant and suggestive conclusions that

can be helpful for informing the design of AI and XAI systems, we believe that such findings must be further assessed, also using more robust statistical standards. In this sense, we believe that the more generalizable finding of our results consists in highlighting the need to consider metrics that go beyond simple agreement and also account for reliance and chance effects (as in influence metrics). Hence, beyond our methodological proposal, our results also provide a guiding principle to select relevant metrics to measure the impact of AI systems in work practice.

More in general, in regards to the proposed framework, while we conceived it as general-purpose and not domain-specialized, our applicative exemplifications were all in the medical domain. Thus, we emphasize the need for larger-N, cross-context validation of the proposed indices, especially for establishing the observed results and, possibly, generalizing them beyond the clinical domain, as well as validating the proposed behavioral indices. That said, an important limitation of this study lies in its static, cross-sectional design. While this approach was appropriate for demonstrating the framework's applicability across diverse contexts, it does not account for how user reliance and AI influence may evolve over time. Nevertheless, the proposed framework is inherently compatible with longitudinal research designs. By tracking metric trajectories across repeated interactions or temporal phases, future studies could examine how patterns of reliance shift and how AI influence adapts in response to user experience, increasing familiarity, or varying task complexity.

Finally, we acknowledge that the current framework prioritizes observable behaviors over subjective constructs such as user satisfaction, perceived transparency, and perceived trust. This choice was deliberate: by focusing on measurable influence, we aim to provide objective and reproducible metrics that are less susceptible to context-specific biases and more generalizable across domains. Nonetheless, we recognize the importance of attitudinal and experiential dimensions in evaluating human–AI interaction. Future research may complement our behavioral metrics with longitudinal and self-reported data to support a more comprehensive understanding of trust calibration and reliance over time.

7. Conclusion

The findings of this research emphasize the complexity and nuance required to evaluate the influence of AI systems on human decision-making, particularly in high-stakes domains like healthcare. This study explored reliance and influence metrics beyond traditional agreement scores, identifying how these measures provide a deeper understanding of human-AI interaction dynamics. The studies we reported in this work broadly suggest that the effectiveness of AI systems in clinical decision-making depends not only on the algorithm's accuracy but also on its impact on user behavior.

Our findings indicate that inappropriate reliance - whether excessive or insufficient - poses a significant risk, particularly when AI accuracy is suboptimal or interaction protocols do not foster well-calibrated trust. In particular, we found that self-anchoring bias and automation bias play a significant role in shaping user reliance and influence, with stark differences observed across modalities like MRI and X-ray diagnostics.

The implications of these findings are twofold. First, they underscore the need for adaptive AI systems capable of addressing specific user biases, such as self-anchoring or over-reliance, through tailored interaction protocols and decision support mechanisms. Second, the study highlights the necessity of fostering calibrated trust, ensuring that users appropriately rely on AI outputs in contexts where they enhance decision accuracy.

However, this work is not without limitations. The findings are based on studies conducted within specific medical domains and may not generalize to other fields without further validation. Nevertheless, the medical domain represents a critical and high-stakes environment where decision-making outcomes have immediate and measurable consequences, ensuring the findings' relevance to similarly high-stakes contexts. Additionally, the reliance on simulated AI outputs limits the exploration of real-world complexities, such as dynamic user feedback loops and evolving system accuracy. Despite this, the controlled nature of the simulations allows for isolating key variables and deriving robust metrics, which can serve as a foundation for more complex real-world applications. Addressing these limitations in future research could provide a more comprehensive understanding of AI influence.

Future research should focus on further refining the metrics for evaluating appropriate reliance and AI influence, particularly in real-world settings. This includes investigating how interaction protocols, user expertise, and domain-specific factors shape reliance behaviors. Expanding research into diverse application contexts beyond healthcare—such as autonomous systems, aviation, financial services, and legal decision-making—could provide deeper insights into tailoring AI systems to different professional environments. Additionally, developing training interventions to mitigate biases and studying their effectiveness in conjunction with advanced AI explainability tools represent promising directions for enhancing human-AI collaboration.

CRediT authorship contribution statement

Andrea Campagner: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Caterina Fregosi:** Writing – review & editing, Writing – original draft, Investigation. **Chiara Natali:** Writing – review & editing, Writing – original draft, Investigation. **Federico Cabitza:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Disclosure of interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve the fluency. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

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

Acknowledgements

C. Fregosi and F. Cabitza acknowledge funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision making. CUP: H53D23008090001 funded by the European Union - Next Generation EU.

C. Natali acknowledges the PhD grant awarded by the Fondazione Fratelli Confalonieri, which has been instrumental in facilitating her research pursuits. C. Natali also acknowledges the financial support provided by the Federal Commission for Scholarships for Foreign Students in the form of the Swiss Government Excellence Scholarship (ESKAS No. 2024.0002) for the academic year 2024–25.

Part of this research was supported by the Italian Ministry of Health – "Ricerca Corrente".

Data availability

Data will be made available on request.

References

Altman, D., Machin, D., Bryant, T., Gardner, M., 2013. Statistics with Confidence: Confidence Intervals and Statistical Guidelines. John Wiley & Sons.

Amann, J., Vetter, D., Blomberg, S.N., Christensen, H.C., Coffee, M., Gerke, S., Gilbert, T.K., Hagendorff, T., Holm, S., Livne, M., et al., 2022. To explain or not to explain?—artificial intelligence explainability in clinical decision support systems. PLOS Digit. Health 1, e0000016.

Baek, E.C., Falk, E.B., 2018. Persuasion and influence: what makes a successful persuader? Curr. Opin. Psychol. 24, 53–57.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W.S., Weld, D.S., Horvitz, E., 2019. Beyond accuracy: the role of mental models in human-ai team performance. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, pp. 2–11.

Benda, N.C., Novak, L.L., Reale, C., Ancker, J.S., 2022. Trust in AI: why we should be designing for appropriate reliance. J. Am. Med. Inform. Assoc. 29, 207–212.

Bingley, W.J., Curtis, C., Lockey, S., Bialkowski, A., Gillespie, N., Haslam, S.A., Ko, R.K.L., Steffens, N., Wiles, J., Worthy, P., 2023. Where is the human in human-centered AI? Insights from developer priorities and user experiences. Comput. Hum. Behav. 141, 107617.

Brachman, M., Ashktorab, Z., Desmond, M., Duesterwald, E., Dugan, C., Joshi, N.N., Pan, Q., Sharma, A., 2022. Reliance and Automation for human-ai collaborative data labeling conflict resolution. Proc. ACM Hum.-Comput. Interact. 6, 1–27.

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. Proc. ACM Hum.-Comput. Interact. 5, 1–21.

Burr, C., Cristianini, N., Ladyman, J., 2018. An analysis of the interaction between intelligent software agents and human users. Minds Mach. 28, 735–774.

Cabitza, F., Campagner, A., Angius, R., Natali, C., Reverberi, C., 2023a. 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.

Cabitza, F., Campagner, A., Famiglini, L., Gallazzi, E., La Maida, G.A., 2022. Color shadows (part i): exploratory usability evaluation of activation maps in radiological machine learning. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Springer, pp. 31–50.

Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., Holzinger, A., 2023b. Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable AI. Expert Syst. Appl. 213, 118888.

Cabitza, F., Campagner, A., Ronzio, L., Cameli, M., Mandoli, G.E., Pastore, M.C., Sconfienza, L.M., Folgado, D., Barandas, M., Gamboa, H., 2023c. Rams, hounds and white boxes: investigating human–ai collaboration protocols in medical diagnosis. Artif. Intell. Med. 138, 102506.

Cabitza, F., Natali, C., Famiglini, L., Campagner, A., Caccavella, V., Gallazzi, E., 2024. Never tell me the odds: investigating pro-hoc explanations in medical decision making. Artif. Intell. Med. 150, 102819.

Cabitza, F., Zeitoun, J.-D., 2019. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Ann. Transl. Med. 7.

Cao, S., Liu, A., Huang, C.-M., 2024. Designing for appropriate reliance: the roles of AI uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact. 8, 1–32.

Carroll, J.M., Rosson, M.B., 1992. Getting around the task-artifact cycle: how to make claims and design by scenario. ACM Trans. Inf. Syst. 10, 181–212.

Coiera, E., 2019. Assessing technology success and failure using information value chain theory. In: Applied Interdisciplinary Theory in Health Informatics. IOS Press, pp. 35–48.

Cook, R.J., Sackett, D.L., 1995. The number needed to treat: a clinically useful measure of treatment effect. Bmj 310, 452–454.

Cresswell, K., Callaghan, M., Khan, S., Sheikh, Z., Mozaffar, H., Sheikh, A., 2020. Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care: a systematic review. Health Inform. J. 26, 2138–2147.

Cresswell, K., de Keizer, N., Magrabi, F., Williams, R., Rigby, M., Prgomet, M., Kukhareva, P., Wong, Z.S.-Y., Scott, P., Craven, C.K., et al., 2024. Evaluating artificial intelligence in clinical settings—let us not reinvent the wheel. J. Med. Internet Res. 26, e46407.

De Vries, P.W., Oinas-Kukkonen, H., Siemons, L., Beerlage-de Jong, N., van Gemert-Pijnen, L., 2017. Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors: 12th International Conference, PERSUASIVE 2017, Amsterdam, the Netherlands, April 4–6, 2017, Proceedings, vol. 10171. Springer.

Dietvorst, B.J., Simmons, J.P., Massey, C., 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. 144, 114.

Eckhardt, S., Kühl, N., Dolata, M., Schwabe, G., A survey of ai reliance, arXiv preprint arXiv:2408.03948, 2024.

Efron, B., 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82, 171–185.

Famiglini, L., Campagner, A., Barandas, M., La Maida, G.A., Gallazzi, E., Cabitza, F., 2024. Evidence-based XAI: an empirical approach to design more effective and explainable decision support systems. Comput. Biol. Med. 108042.

Floridi, L., 2024. Hypersuasion–on ai's persuasive power and how to deal with it. Philos. Technol. 37, 1–10.

Fogg, B.J., 1998. Persuasive computers: perspectives and research directions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 225–232.

Furukawa, T.A., Leucht, S., 2011. How to obtain nnt from cohen's d: comparison of two methods. PLoS One 6, e19070.

Guo, Z., Wu, Y., Hartline, J., Hullman, J., A statistical framework for measuring ai reliance, arXiv preprint arXiv:2401.15356, 2024a.

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

Harvey, N., Fischer, I., 1997. Taking advice: accepting help, improving judgment, and sharing responsibility. Organ. Behav. Hum. Decis. Process. 70, 117–133.

He, G., Buijsman, S., Gadiraju, U., 2023a. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. Proc. ACM Hum.-Comput. Interact. 7, 1–29.

He, G., Kuiper, L., Gadiraju, U., 2023b. Knowing about knowing: an illusion of human competence can hinder appropriate reliance on AI systems. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–18.

Hedges, L.V., 1981. Distribution theory for glass's estimator of effect size and related estimators. J. Educ. Stat. 6, 107–128.

Henrich, J., Chudek, M., Boyd, R., 2015. The big man mechanism: how prestige fosters cooperation and creates prosocial leaders. Philos. Trans. R. Soc. B Biol. Sci. 370, 20150013.

Herrmann, T., Pfeiffer, S., 2023. Keeping the organization in the loop: a socio-technical extension of human-centered artificial intelligence. Ai Soc. 38, 1523–1542.

Hoff, K.A., Bashir, M., 2015. Trust in automation: integrating empirical evidence on factors that influence trust. Hum. Factors. 57, 407–434.

Hoffman, R.R., Johnson, M., Bradshaw, J.M., Underbrink, A., 2013. Trust in Automation. IEEE Intell. Syst. 28, 84–88.

Kahr, P.K., Rooks, G., Willemsen, M.C., Snijders, C.C.P., 2024. Understanding trust and reliance development in AI advice: assessing model accuracy, model explanations, and experiences from previous interactions. ACM Trans. Interact. Intell. Syst. 14, 1–30.

Klingbeil, A., Grützner, C., Schreck, P., 2024. Trust and reliance on ai—an experimental study on the extent and costs of overreliance on AI. Comput. Hum. Behav. 160, 108352.

Lee, J.D., See, K.A., 2004. Trust in automation: designing for appropriate reliance. Hum. Factors 46, 50–80.

Legaspi, R., Xu, W., Konishi, T., Wada, S., Kobayashi, N., Naruse, Y., Ishikawa, Y., 2024. The sense of agency in human–ai interactions. Knowl.-Based Syst. 286, 111298.

Logg, J.M., 2022. The psychology of big data: developing a "theory of machine" to examine perceptions of algorithms. In: Matz, S.C. (Ed.), The Psychology of Technology: Social Science Research in the Age of Big Data. American Psychological Association, pp. 349–378.

Logg, J.M., Minson, J.A., Moore, D.A., 2019. Algorithm appreciation: people prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. 151, 90–103.

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.

Luciano, F., 2024. Hypersuasion–on ai's persuasive power and how to deal with it. Philos. Technol. 37, 1–10.

Ma, S., Zhang, C., Wang, X., Ma, X., Yin, M., Beyond recommender: An exploratory study of the effects of different ai roles in ai-assisted decision making, arXiv preprint arXiv:2403.01791, 2024.

Malenka, D.J., Baron, J.A., Johansen, S., Wahrenberger, J.W., Ross, J.M., 1993. The framing effect of relative and absolute risk. J. Gen. Intern. Med. 8, 543–548.

Merritt, S.M., Lee, D., Unnerstall, J.L., Huber, K., 2015. Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. Hum. Factors 57, 34–47.

Moja, L., Polo Friz, H., Capobussi, M., Kwag, K., Banzi, R., Ruggiero, F., González-Lorenzo, M., Liberati, E.G., Mangia, M., Nyberg, P., et al., 2019. Effectiveness of a hospital-based computerized decision support system on clinician recommendations and patient outcomes: a randomized clinical trial. JAMA Netw. Open 2, e1917094.

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., Perer, A., 2024. The impact of imperfect XAI on human-ai decision-making. Proc. ACM Hum.-Comput. Interact. 8, 1–39.

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

Plass, M., Kargl, M., Nitsche, P., Jungwirth, E., Holzinger, A., Müller, H., 2022. Understanding and explaining diagnostic paths: toward augmented decision making. IEEE Comput. Graph. Appl. 42, 47–57.

Schaschek, M., Spatscheck, N., Winkelmann, A., 2024. For those about to rely-a taxonomy of experimental studies on AI reliance AI reliance. In: 19th International Conference on Wirtschaftsinformatik.

Schechtman, E., 2002. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat—which of these should we use? Value Health 5, 431–436.

Schemmer, M., Bartos, A., Spitzer, P., Hemmer, P., Kühl, N., Liebschner, J., Satzger, G., Towards effective human-ai decision-making: The role of human learning in appropriate reliance on ai advice, arXiv preprint arXiv:2310.02108, 2023a.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., Satzger, G., Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making, arXiv preprint arXiv:2204.06916, 2022.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., Satzger, G., 2023b. 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.

Schmager, S., Pappas, I., Vassilakopoulou, P., 2023. Defining human-centered AI: a comprehensive review of Hcai literature. In: MCIS'23: Proceedings of the Mediterranean Conference on Information Systems (MCIS).

Schmitt, A., Wambsganß, T., Söllner, M., Janson, A., 2021. Towards a trust reliance paradox? Exploring the gap between perceived trust in and reliance on algorithmic

advice. In: Proceedings of the International Conference on Information Systems (ICIS) 2021.
Shneiderman, B., 2020. Human-centered artificial intelligence: reliable, safe & trustworthy. Int. J. Hum.-Comput. Interact. 36, 495–504.
Shneiderman, B., 2022. Human-Centered AI. Oxford University Press.
Sorantin, E., Grasser, M.G., Hemmelmayr, A., Tschauner, S., Hrzic, F., Weiss, V., Lacekova, J., Holzinger, A., 2022. The augmented radiologist: artificial intelligence in the practice of radiology. Pediatr. Radiol. 52, 2074–2086.
Tontini, G.E., Pessarelli, T., Bertolotti, M., Aldinio, G., Barbaro, F., Calabrese, C., Caprioli, F., Elli, L., Hassan, C., Kopylov, U., Murino, A., Rondonotti, E., Toth, E., Natali, C., Cabitza, F., 2025. The use-AI study: a multicenter, single blind proof-of-concept study on the reliance of AI in gastrointestinal endoscopy. Underreview.
Umbrello, S., Natale, S., 2024. Reframing deception for human-centered AI. Int. J. Soc. Robot. 1–19.
Wang, X., Yin, M., 2021. Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. In: Proceedings of the 26th International Conference on Intelligent User Interfaces, pp. 318–328.
Yang, Q., Steinfeld, A., Rosé, C., Zimmerman, J., 2020. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In: Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems, pp. 1–13.
Yaniv, I., 2004. Receiving other people's advice: influence and benefit. Organ. Behav. Hum. Decis. Process. 93, 1–13.
Yaniv, I., Kleinberger, E., 2000. Advice taking in decision making: egocentric discounting and reputation formation. Organ. Behav. Hum. Decis. Process. 83, 260–281.
Yin, M., Wortman Vaughan, J., Wallach, H., 2019. Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems, pp. 1–12.
Yu, K., Berkovsky, S., Taib, R., Zhou, J., Chen, F., 2019. Do i trust my machine teammate? An investigation from perception to decision. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 460–468.
Zhou, J., Li, Z., Hu, H., Yu, K., Chen, F., Li, Z., Wang, Y., 2019. Effects of influence on user trust in predictive decision making. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–6.