Five Degrees of Separation: Investigating the Unexpected Potential of Displaced Human-AI Collaboration Protocols for Apter AI Support

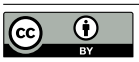
FEDERICO CABITZA*, University of Milano-Bicocca, Italy and IRCCS Ospedale Galeazzi - Sant'Ambrogio, Italy
ANDREA CAMPAGNER*, IRCCS Ospedale Galeazzi - Sant'Ambrogio, Italy
CATERINA FREGOSI, University of Milano-Bicocca, Italy
MATTEO CAMELI, Department of Medical Biotechnologies, Division of Cardiology, University of Siena, Italy
ENRICO GALLAZZI, U.O.C. Patologia Vertebrale e Scoliosi, ASST Gatano Pini - CTO, Italy
LUCA MARIA SCONFIENZA, Department of Biomedical Sciences for Health, University of Milan, Italy and IRCCS Ospedale Galeazzi - Sant'Ambrogio, Italy
GIAN EUGENIO TONTINI, Department of Pathophysiology and Transplantation, University of Milan, Italy and Gastroenterology and Endoscopy Unit, Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico, Italy

The integration of AI into decision-making processes offers substantial benefits, particularly in enhancing accuracy and efficiency. However, long-term consequences, such as over-reliance, skill erosion, and loss of human agency, present significant challenges. This study investigates various human-AI collaboration protocols – traditional, inhibition, displacement, and replacement – across multiple medical settings, including radiological imaging, ECG, and endoscopy. We introduce a novel framework that includes a choice nomogram and qualitative assessment tool, designed to optimize both decision accuracy and socio-technical impacts. Our findings reveal that the displacement protocol consistently outperformed others in several contexts, achieving 87% accuracy in MRI analysis, 89% in x-ray reading and 85% in endoscopy; conversely, the traditional protocol was most effective only in ECG analysis, with 82% accuracy. These results demonstrate that no single protocol is universally optimal, highlighting the need for context-specific selection to ensure effective and sustainable AI-supported decision-making, with a focus on balancing short-term performance with long-term human factors.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods**; **HCI theory, concepts and models**; • **Computing methodologies** → *Artificial intelligence*; • **Information systems** → *Decision support systems.*

*Both authors contributed equally to this research.

Authors' Contact Information: Federico Cabitza, University of Milano-Bicocca, Milan, Italy and IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milan, Italy, federico.cabitza@unimib.it; Andrea Campagner, IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milan, Italy, andrea.campagner@unimib.it; Caterina Fregosi, University of Milano-Bicocca, Milan, Italy, caterina.fregosi@unimib.it; Matteo Cameli, Department of Medical Biotechnologies, Division of Cardiology, University of Siena, Siena, Italy, matteo.cameli@unisi.it; Enrico Gallazzi, U.O.C. Patologia Vertebrale e Scoliosi, ASST Gatano Pini - CTO, Milan, Italy, enrico.gallazzi@asst-pini-cto.it; Luca Maria Sconfienza, Department of Biomedical Sciences for Health, University of Milan, Milan, Italy and IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milan, Italy, luca.sconfienza@unimi.it; Gian Eugenio Tontini, Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy and Gastroenterology and Endoscopy Unit, Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy, gianeugenio.tontini@unimi.it.



ACM 2573-0142/2025/11-ARTCSCW420
https://doi.org/10.1145/3757601

Additional Key Words and Phrases: Human-AI interaction, Decision making, Human-AI collaboration protocols, Evidence-based design

ACM Reference Format:
Federico Cabitza, Andrea Campagner, Caterina Fregosi, Matteo Cameli, Enrico Gallazzi, Luca Maria Sconfienza, and Gian Eugenio Tontini. 2025. Five Degrees of Separation: Investigating the Unexpected Potential of Displaced Human-AI Collaboration Protocols for Apter AI Support. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW420 (November 2025), 28 pages. https://doi.org/10.1145/3757601

1 Introduction

1.1 The Promise and Pitfalls of AI in Decision-Making

The potential of artificial intelligence (AI) across various decision-making domains is widely recognized, especially for its promises to improve decision accuracy, efficiency and user satisfaction. More recently, however, also other effects of AI in work settings have attracted interest in the specialized literature, regarding the so-called unintended consequences that are associated with the deployment of AI in knowledge-intensive contexts [52], including complex fields such as medicine [11, 18]. Among the most well-known of these negative effects are the development of over-reliance and complacency by usersy [47]. In the short term, these behaviors may lead to machine-induced errors (cf. automation bias [60]) or opportunistic behaviors such as over-dependence; in the long term, they may contribute to skill erosion [68], impaired learning [4], technostress [87] and loss of agency [28, 46, 91]. While it remains an open question whether these detrimental effects can ultimately nullify the advantages of AI or undermine decision-making, human control, and agency [39, 46, 58, 65], there is broad consensus that AI solutions significantly impact the performance of human teams that adopt them. Consequently, it is crucial to measure these effects. In this study, we examine how different approaches to integrating AI into decision-making processes influence the performance of human teams operating in complex, knowledge-intensive environments. Adopting a socio-technical perspective, we conceptualize AI systems as interventions embedded within organizational workflows. We present findings from four user studies conducted in the medical domain, with a particular focus on decision accuracy as the primary performance metric. Our results indicate that the most effective mode of human-AI integration is not always the most intuitive or conventional; in some cases, less collaborative configurations outperform more interactive ones. By systematically evaluating alternative integration protocols, this work contributes to the evidence-based design of AI-supported decision-making systems.

1.2 Human-AI Collaboration as a Socio-Technical Concern

Evidence increasingly shows that simply having humans and AI "collaborate" does not always result in improved outcomes, and the effectiveness of such collaboration depends on task type and complexity [88].

In fact, recent findings emphasize the need to evaluate human-AI performance comprehensively, to determine under which conditions collaborative human-machine configurations actually outperform humans or AI working independently [3, 88, 95]. Recent theoretical and design-oriented contributions have emphasized the need to conceptualize human-AI collaboration not merely as interaction, but as a socio-technical arrangement shaped by task structure, role configuration, and coordination mechanisms. For example, Wang et al. [90] advocate for importing principles from human-human collaboration into the design of AI systems, while Yang et al. [94] highlight the specific challenges posed by AI in sketching and prototyping collaborative workflows, particularly

due to the adaptive and opaque characteristics of many systems. Empirical research by Green and Chen [31] further illustrates how algorithmic decision aids may distort human judgment or introduce new forms of bias, even when technically accurate, highlighting the importance of understanding how humans interpret and integrate algorithmic outputs in practice.

Consequently, the success of AI interventions in complex clinical settings depends not only on technological factors but also on a comprehensive understanding of clinical workflows and interdependencies within healthcare teams [80]. These interdependencies, which span organizational, clinical, political, and behavioral domains, highlight the need for design approaches that are sensitive to the specific context of use. For instance, studies in clinical decision support emphasize the need to design AI systems that do not merely replicate human tasks but integrate with the tacit and experiential knowledge that healthcare practitioners employ in their daily work [2].

This aligns well with the concept, particularly prevalent in the field of medical informatics, that AI systems should be considered as *AI interventions* [48], meaning as essential components of broader organizational and socio-technical processes aimed at enhancing knowledge-intensive work processes and the associated practices of coordination and communication within specific work settings.

1.3 Designing Human-AI Collaboration Protocols

Conceiving and implementing an AI intervention involves much more than designing an AI system and training its embedded statistical model on large amounts of data. It rather requires an in-depth understanding of the surrounding organizational context, which includes not only the working methods and practices but also the expectations, attitudes, and competencies of prospective users. Even more crucially, it involves designing what has been termed in [9] a *human-AI collaboration protocol*. These protocols define not only the main features of the interactive AI system (such as its affordances and types of output) but also its proper usage (i.e., when users should engage with the system, for what tasks, and who can use it, based on expertise profiles and user roles). The aim to conceive and implement an AI intervention would then be to design and evaluate a given collaboration protocol, adjusting certain independent variables as if they were parameters to be configured, and then observing specific phenomena associated with indicators, e.g., the decision accuracy of the AI-supported team, their efficiency and satisfaction, in short the usability of the system. This approach helps identifying the configuration (i.e., the protocol instance) that optimizes the likelihood of achieving positive outcomes and, at the same time, mitigates the probability of adverse side effects of the intervention.

Taking this approach allows studies aimed at evaluating and comparing different AI solutions and decision aids to create a body of empirical knowledge that can inform the *evidence-based design* of these systems [26]. Specifically, the AI-as-intervention approach enables the accumulation of study results that, depending on their experimental design (such as qualitative vs quantitative, retrospective vs prospective, observational vs randomized), provide AI solution designers with evidence of varying strength and credibility, upon which they can then choose the most appropriate solution for a given setting. This approach parallels the design of interactive digital systems with other disciplines within the broad spectrum that Herbert Simon associated with the term "sciences of the artificial," such as the architectural design of specific working environments like hospitals[81].

1.4 Our Contribution and Research Focus

In the, admittedly still limited, body of studies aimed at comparing alternative human-AI collaboration protocols, most research focuses on solutions that vary in the type of output they provide, including the presence and nature of explanations [54, 89]. However, fewer studies address the

comparison of different processes, that is, the various ways in which humans and AI systems can collaborate or integrate their individual contributions toward the formulation of a final decision in classification tasks.

Our study contributes to the Computer-Supported Cooperative Work (CSCW) tradition by investigating how different human-AI collaboration protocols affect decision-making in complex work settings, with a focus on the medical domain. We selected four distinct areas—radiological imaging, electrocardiography, orthopedic radiography, and gastrointestinal endoscopy—based on their diversity in diagnostic complexity, types of clinical expertise, and current levels of AI integration. These settings exemplify knowledge-intensive work under uncertainty and time pressure, where AI systems are increasingly adopted yet their implications for professional practice remain contested. By applying and comparing protocols across these varied contexts, we show how our framework can support researchers, designers, and practitioners in empirically identifying the protocol that best fits the priorities of a given decision setting—balancing performance, human-AI coordination, and reliance dynamics [29]. In doing so, we extend CSCW's core concerns with socio-technical systems, emphasizing the need to consider not only the artifact (i.e., the AI system) but also the protocol [12, 75], through which it is embedded in coordination mechanisms, decision structures, and digitally mediated choice architectures.

To conceptualize the different collaboration configurations examined in this work, we draw on a framework introduced by James Pierce [63], which identifies distinct modalities of human-technology engagement—namely, inhibition, displacement, foreclosure, and replacement. Although not originally developed for decision-support systems, this framework offers a critical lens for examining how design can intentionally shape, limit, or negate interaction with technology. In our context, it enables a principled classification of human-AI collaboration protocols along both epistemic and socio-technical dimensions, including decision accuracy, user agency, and long-term skill retention. The vocabulary of the framework - drawn from design studies and infused with critical intent - helps foreground design choices that deliberately inhibit, delay, or bypass AI advice, rather than assuming direct collaboration as inherently beneficial. This makes it particularly well-suited to exploring responsible AI integration in professional work practices.

In the following sections, after recalling the Pierce framework, we illustrate its application through four user studies conducted by our research group in the medical domain. These studies focus on a critical dimension in medical contexts: decision accuracy (understood as the inverse of error rate). This concept aligns with an epistemological perspective on a construct that has recently garnered significant attention in the human-AI interaction community – appropriate reliance [14, 72]. In light of this concept, we propose a method for identifying the protocol associated with the highest (estimated) accuracy. We will discuss the surprising finding that, in some cases, the best protocol does not involve traditional modes of collaboration between experts and AI systems.

2 Collaboration protocols for evidence-based AI

As hinted above, a human-AI collaboration protocol is any integrated set of rules, stipulating the articulation of interdependent activities between humans and AI systems [75], that defines the modalities about how (and if) humans interact with AI supporting tools in decision tasks.

Ideally, different AI solutions and different ways of deploying them, that is what is described and stipulated by different collaboration protocols [92], should be compared for a given task and decision setting, so that their positive and negative effects assessed, and the associated evidence evaluated in order to make an informed design decision [26, 66].

In their more general form, different protocols can also differ for what in the specialist literature is often referred to as levels of automation (LOA) [62], a spectrum of solutions ranging from total

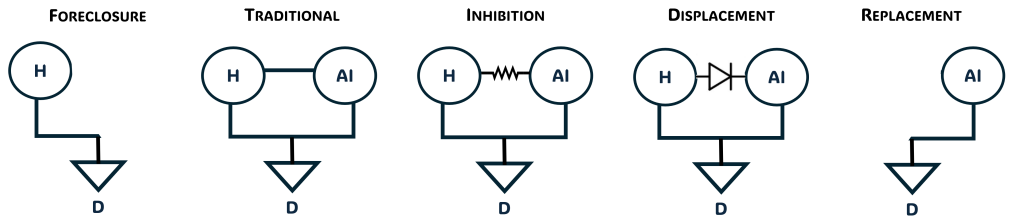


Fig. 1. A figure illustrating the primary human-AI collaboration protocols in decision-making contexts, using an electric circuit metaphor. Each protocol depicts the relationships between the human decision-maker (H) and the AI system, leading to a final decision (D). For the inhibition and displacement protocols, the metaphor indicates that in the former, designers introduce forms of resistance and constraints between humans and machines, while in the latter, there is no direct feedback from the machine, though it processes data from the human component of the socio-technical system.

absence of automated support to total delegation to automation. To navigate in this spectrum, we were inspired by the evocative framework proposed by Pierce [63], to pinpoint five main configurations or main protocols: protocols of AI foreclosure, traditional interaction with AI, inhibited interaction, AI displacement and replacement by AI (see Figure 1).

- *Foreclosure* refers to protocols where humans do not use AI supports at all (0 in the LOA scale [62]) for a given task or work objective, mainly because the use of AI in any form is associated with unacceptable levels of reliability, transparency, or risk of negative consequences and externalities.
- *Traditional support* refers to protocols (also called AI-first [9], or concurrent [85]) that stipulate that decision makers receive AI advice and any other piece of related information (e.g., confidence scores, textual/visual explanations) alongside the case to be considered, and are expected to make their decisions considering this output. This class of protocols is typically associated with the concept of individual augmentation [24, 41] and exhibits a wide range of possible solutions. These solutions can be characterized by various levels of automation (excluding levels 0 and 10) and differ in every collaborative aspect depending on the context, system-task characteristics, and individual differences [74], with the consistent requirement that all involve synchronous and co-located interaction. This interaction can be aligned with various configurations [1], such as human-in-command, human-in-the-loop, human-on-the-loop, or human-over-the-loop. The specific configuration depends on factors such as task allocation, the role of AI [20], the level of human oversight and machine autonomy [87], the initiative exhibited by the machine [35], the automation level of different tasks [62], and other features.
- *Inhibition support* refers to protocols where human-AI interaction is somehow inhibited or constrained through features, choice architectures [17], and interfaces designed to hinder or prevent mindless reliance on AI. Pierce [63] distinguishes between the design of technologies that actively hinder user use, misuse or abuse [61] through a wide range of attributes and physical affordances; and technologies intended to inhibit their own interactions (self-inhibition). An instance of this latter class are systems that abstain from [13] or reject [38] giving an advice, like in the *Intelligent Decision Assistance* setting [73], in the *evaluative XAI* framework [53] and in the *judicial AI* approach [10]. This class of protocols differs from conventional approaches that focus on bridging the gulfs between users and machines [57]

and improving the user-friendliness and immediacy of interactions between humans and machines [42]. In fact, inhibited or constrained protocols are rather aimed at increasing what has been termed *cognitive friction* between the user and the system [10, 19, 55] to avoid potential unintended consequences of AI while still reaping its benefits. In what follows, we will consider one of the simplest forms of inhibition: delayed or detained interaction, also referred to by various terms such as human-first [9], consequential [6], sequential [85], proactive [44], silent [43], or update [32]. All these protocols share the requirement that human decision-makers must report their initial decision before being exposed to the AI's 'second opinion.' After this initial decision, they are asked to finalize their decision, thus introducing deliberate friction in the use of AI by necessitating that AI use is inhibited (as a constraint) until humans make a preliminary and comprehensive decision.

- *Displacement support* refers to protocols where humans do *not* directly engage with AI support and this latter is removed even (but not necessarily) "literally, from its typical or currently occupied position.[...] Displacing goes beyond inhibiting by aiming to more substantially hindering routine or common interaction with technology, but without completely negating its existence and potential for use. " [63]. Convenient solutions to displace AI regard designing forms of indirect interaction, such as the settings where AI prompts a secondary decision from a different human in case of a human-machine disagreement. In these cases, the AI would then consolidate a final decision based on majority consensus, which may be weighted by factors such as confidence or expertise. Different methods of performing such an AI displacement are conceivable [37]. In our implementation of this protocol, as it will be clear in the next section, we considered a solution intentionally designed for simplicity in which the initial evaluator remains unaware of the AI's differing opinion and there is no direct communication between the two human interpreters. More complex variations of this protocol could involve communicating the second evaluator's findings back to the initial decision-maker and facilitating their synchronous interaction in cases of differing opinions, or also having the AI act as an orchestrator among the human evaluators [67].
- *Replacement* refers to the solution in which humans are no longer involved in the decision (level 10 of automation, out-of-the-loop configurations [22], or Automated Decision Making [76]), because considered a source of unacceptable bias, arbitrariness or noise [40].

Although this framework has been referenced multiple times within the HCI community [63, 64], such recognition has not yet led to widespread practical applications. Partly for this reason, we find Pierce's framework particularly well-suited and intellectually stimulating for the type of methodological reflection our proposal aims to advance, despite some arguably unconventional terminological choices. Terms such as inhibition and foreclosure, which we interpret in their original, evocative sense of constraint and removal, reflect a critical orientation that aligns with our interest in rethinking how humans and machines should (or should not) collaborate within specific work settings.

Given their abstract nature, these protocol classes are designed to encompass a broad range of human-AI collaboration patterns, particularly the three core classes illustrated in Figure 1. Nonetheless, for the sake of simplicity in the user studies reported in the following sections, all protocols have been operationalized in the form diagrammatically represented in Figure 2.

Given these distinct ways to enable the interaction between humans and AI (if any), it is of crucial importance to understand, in a given setting, which collaboration protocol would provide the best results, not only in terms of resulting accuracy and efficiency [84], but also in terms of possible effects on human behaviors in regard to over- (or under-) reliance on the AI support, and

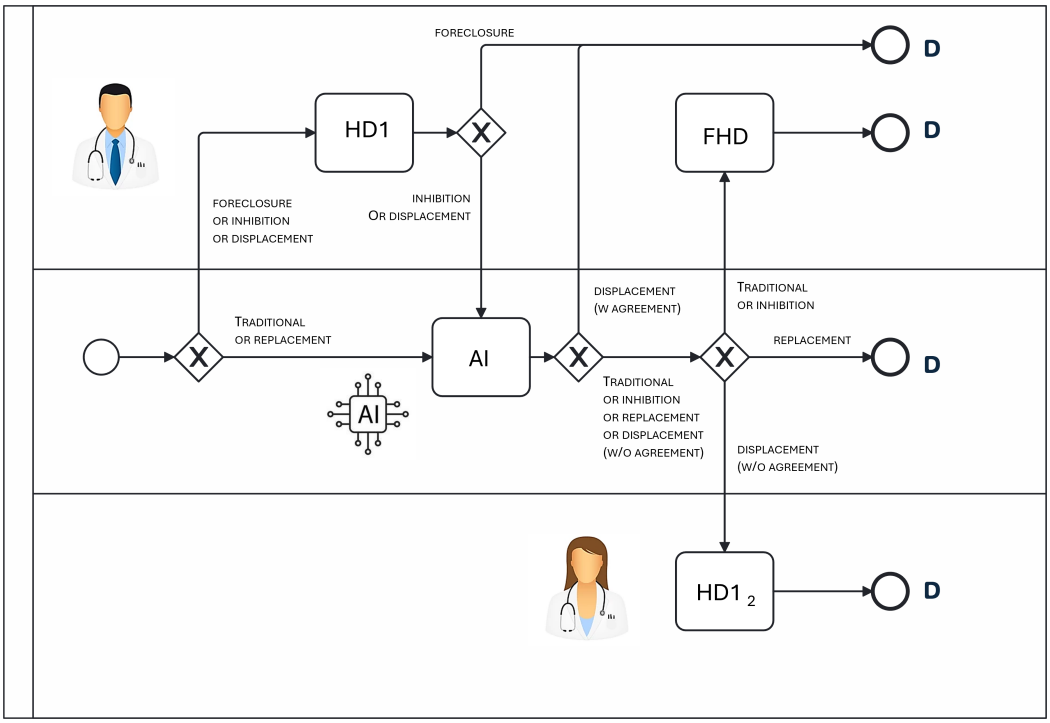


Fig. 2. BPMN (Business Process Model and Notation) diagram illustrating the collaboration protocols applied in the user studies. The diamond shapes with a cross represent exclusive gateways, directing the decision flow to only one outgoing path based on the evaluation of specified conditions. The unlabeled diamond shapes signify convergence points for multiple incoming sequences that serve as junctures for the merging of flows. The final states represent the decisions made in the diagnostic task. HD1 tasks refer to judgments expressed by humans without consulting AI advice, while FHD tasks represent decisions made after consulting AI advice.

risks of dependence on the technological support. In the following sections, we will describe our investigation to compare these protocols in four user studies of realistic complexity.

3 Collaboration protocols and associated decision dimensions

Each collaboration protocol presented above corresponds to a strategy to put humans and AI more or less close to each other and involve them in a more or less tight relationship, and is thus associated with specific characteristics and risks [11, 16, 78]. In the next section, we will study the effect of such protocols from an empirical point of view for four different medical case studies. In this section, we will present a methodological contribution that allows to evaluate these protocols both from a qualitative and quantitative point of view. To this aim we will consider five main quality dimensions: *accuracy*, that is the decision performance of the given collaboration protocol; *complexity*, intended as the reciprocal of efficiency, that is the number of human decisions to get a certain and acceptable level of accuracy [84]; as well as three additional dimensions, which are more concerned with the human-system relationship, namely *dependence*, *interaction*, and (risk of) *deskilling*. Dependence refers to the degree to which the overall socio-technical decision-making setting depends on the AI system for effective decision-making, or to the potential impact of a

sudden failure of the technological component within the socio-technical setting. Reliance pertains to the extent to which human decision-makers rely on technology for making accurate decisions by deferring to its judgments. This concept also encompasses risks associated with over-reliance, such as automation bias (i.e., making mistakes caused by incorrect AI advice). Deskilling represents a long-term consequence of continuous and routine reliance, encompassing phenomena such as skill erosion, impaired learning, technostress [87] and the oversimplification of expert work.

3.1 The choice nomogram for evaluating the accuracy of collaboration protocols

In regard to accuracy, our main contribution consists of a *choice nomogram*, that is a diagram to help practitioners select, based on prior information about a given decision setting, the collaboration protocol that will result in the highest decision accuracy. In particular, our methodological contribution stems from the observation that the accuracy of the collaboration protocols considered above, from a theoretical point of view, can be estimated on the basis of only three parameters: the baseline accuracy of the human decision-makers and the AI, and the so-called *appropriate reliance* [34, 45, 72], that is, the ability of humans to accept the AI right advice (either confirming their initial interpretation, if available, or changing their mind) and discard wrong advice[1].

We first focus on the case of the foreclosure, replacement and displacement protocols, as in these protocols there is no direct dependence among the different decisions that concur in the formulation of the final decision. Indeed, in both foreclosure and replacement no such dependence actually exists (because only a human or AI, respectively, participates in making the decision), while in displacement the dependence only relates to the probability with which the first human decision-maker and the AI will agree on an instance's classification (also known as percentage of agreement, or P_o, in the content analysis literature [86]). Thus, we assume that the decisions involved in these three protocols are independent of each other: then, clearly, only two parameters are needed to compare these three protocols, namely the average human accuracy and the accuracy of their decision aid (obviously, under the assumption that these scores generalize to new cases). Thus, let H be the average accuracy of the human decision makers in the considered scenario, and let AI be the accuracy of the AI. These correspond to the (average) accuracy rates of the foreclosure and replacement protocols, respectively, that is:

Estimated decision accuracy under Foreclosure — Average human accuracy

$$Foreclosure(H) = H\ ; \tag{1}$$

$$Replacement(AI) = AI\ . \tag{2}$$

Estimated decision accuracy under Replacement — AI accuracy

As for the case of displacement, assuming independence of human decisions, we note that the final decision FHD recommended by this protocol will be correct in exactly three cases: 1) HD1 and AI agree and both are correct, with probability $H * AI$; 2) HD1 and AI disagree, AI is correct and HD2 (the second human rater) agrees with the AI, with probability $(1 - H) * AI * H$; 3) HD1 and AI disagree, HD1 is correct and HD2 agrees with HD1, with probability $H * (1 - AI) * H$.

[1]As hinted above, this discussion pertains solely to the *epistemological* perspective on decision appropriateness, which complements the *ethical* perspective. The former addresses questions of right and wrong in terms of ontology and truth content, while the latter focuses on moral or axiological considerations. In this context, a parallel can be drawn with what Parasuraman [61] termed *misuse* and *disuse*- reliance patterns that undermine accuracy —and *abuse*, which involves avoidable reliance that nonetheless affects other socio-technical dimensions, such as sense of agency, responsibility, skill degradation, and social learning.

Estimated decision accuracy under Displacement

Probability that HD1 and AI disagree, AI is correct, FHD agrees with AI

$$Displacement(H, AI) = H * AI + (1 - H) * AI * H + H * (1 - AI) * H. \quad (3)$$

Probability that HD1 and AI agree and are both correct

Probability that HD1 and AI disagree, HD1 is correct, FHD agrees with HD1

As for the inhibition and traditional protocols, we note that the average human accuracy and AI accuracy are not sufficient, by themselves, in determining (even on average) their accuracy. Indeed, in both of these protocols, there is a complex dependence between the AI support and the final decision [33, 36], for instance because the AI advice corroborates an initial human interpretation, or because it induces changes (or switches [30]) either for the worse (automation bias) or for the better (appropriate reliance [15]).

Such dependence grounds on the ability of the human decision maker to accurately discern when the AI recommendation is correct or wrong, i.e., on appropriate reliance. As mentioned above, we thus consider a third additional parameter, *AR* which denotes the appropriate reliance level. Following the definition of *AR* in [33], this parameter can be expressed as the probability of a correct adherence or correct override of the AI decision [72, 77]), with the caveat that the configurations in which all decisions agree, that is the probability that HD1, AI and FHD are all correct or all wrong, must be treated distinctly from others, so as to discount cases in which there is no actual reliance on the AI support (i.e., the human and AI judgments coincide, but the human simply ignored the AI support and confirmed their initial opinion). In particular, for cases in which all decisions are correct (resp., wrong) one should consider only cases that led to an increase (resp., decrease) in confidence between HD1 and FHD [72], as in these cases the AI support had an impact on the final decision (though this impact is not visible on the final decision itself, it affects the degree of confidence attached to it). Thus, *AR* can be quantified as the the (weighted) number of times humans accept the AI right advice (either confirming their initial interpretation, if available, or changing their mind by either switching their decision or changing their confidence) and discard wrong advice. In formulas:

Appropriate reliance level

Probability that the human correctly relies on the AI

Probability that the human correctly discards AI advice

$$\begin{aligned} AR = {} & P(\text{FHD correct, AI correct, HD1 wrong}) + P(\text{FHD correct, AI wrong, HD1 correct}) \\ & + P(\text{FHD correct \& confidence} \uparrow, \text{AI correct, HD1 correct}) \\ & + P(\text{FHD wrong \& confidence} \downarrow, \text{AI wrong, HD1 wrong}) \end{aligned} \quad (4)$$

Probability that human and AI agree on the correct answer and final confidence increases

Probability that human and AI agree on the wrong answer and final confidence decreases

Finally, we introduce an additional assumption, namely that the probability of extreme algorithmic aversion [7], i.e. cases where the human and AI initially agree on their decision but this leads to a decision switch in the final decision ($FHD \neq AI$ whereas $AI = HD1$.) is negligible[2]. Thus, based on the above assumptions, we estimate the average expected accuracy for the traditional collaboration protocol according to the following formula:

Estimated decision accuracy under Traditional

$$Traditional(H, AI, AR) = max(AR, AI). \quad (5)$$

Accuracy under Traditional is at least as large as the maximum between appropriate reliance and the AI accuracy

[2]Such an observation has indeed been confirmed in previous empirical investigations [9]

Table 1. Quanti-qualitative summarization of the associations between dimensions of human-system interaction and the collaboration protocols evaluated in our study. The darker the circle, the stronger the association. The strength assessments are deliberately qualitative and intended as general indicators to aid in choosing the most appropriate protocol for the specific decision-making task in a complementary way with respect to accuracy comparisons. Besides the association between the protocols and their costs and side effects, we report the related decision complexity (which is inversely proportional with efficiency), in terms of the number of human decisions for the same accuracy on n cases. Here P_o stands for percentage of agreement between the first human decision maker and the AI.

Protocol	Complexity	Efficiency	Dependence	Interaction	Deskilling
Replacement	0	●●●	●●●	○	●●●
Traditional	$\propto n$	●●	●●	●●●	●●
Inhibition	$\propto 2 * n$	●	●	●	●●
Displacement	$\propto n + (1 - P_o) * n$	●	●	●	●
Foreclosure	$\propto n$	●●	○	○	○

Indeed, if the human simply blindly follows the AI recommendation, the resulting accuracy would be equal to at least AI; on the other hand, if the human decision-maker has an appropriate reliance level equal to AR this also represents a lower bound on the value of $Traditional(H, AI, AR)$ [33].

By contrast, the final decision obtained by following the inhibition protocol will be correct in exactly two cases: 1) human and AI agree and are correct, with probability $H * AI$; 2) the human and AI disagree and human appropriately relies (or does not rely) on the AI advice depending on the correctness of this latter, with probability $(H * (1 - AI) + (1 - H) * AI) * AR$. Hence, we obtain that the estimated average accuracy of the inhibition protocol can be computed as

Estimated decision accuracy under Inhibition

Probability that Human and AI disagree and human appropriately relies (or does not) on AI

$$Inhibition(H, AI, AR) = H * AI + (H * (1 - AI) + (1 - H) * AI) * AR\,. \quad (6)$$

Probability that Human and AI agree and are both correct

Grounding on the estimates of accuracy for the five considered collaboration protocols, we propose a 3D choice nomogram, illustrated in Figure 3 in terms of some of its relevant cross-sections. By means of this nomogram, one can graphically select the optimal (in terms of accuracy) collaboration protocol based on the average human accuracy, the AI accuracy and the appropriate reliance level: by representing case studies in terms of single points in the diagram whose coordinates are given by these pieces of information, the nomogram shows which protocol to select based on the region in which the point ends up belonging to, which in turn depends on which among H, AI, $Displacement(H, AI)$, $Traditional(H, AI, AR)$ and $Inhibition(H, AI, AR)$ is largest, as this latter identifies the protocol which would provide the highest decision accuracy.

From a more general point of view, this tool allows designers and practitioners to estimate the overall accuracy of a hybrid human-AI system by directly calculating the above formulas and considering only three parameters: the average accuracy of human decision-makers (by assuming their decision independent), the accuracy of the AI, and an estimate of appropriate reliance [45] (i.e., the degree to which decision-makers' trust in AI is calibrated).

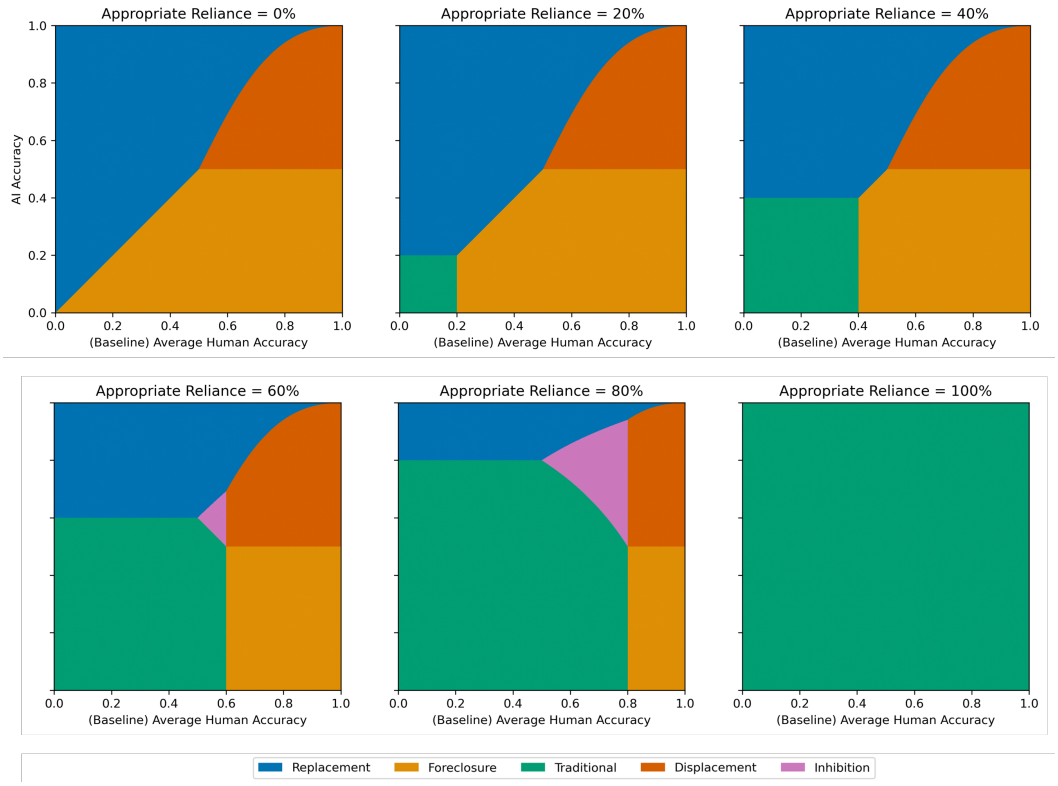


Fig. 3. The 3D nomogram for selecting the most suitable collaboration protocol based on the average accuracy of human decision-makers, the accuracy of the AI system, and the appropriate reliance level. The figure shows six cross-sections of this cube at specific levels of appropriate reliance: 0%, 20%, 40%, 60%, 80%, and 100%.

3.2 A qualitative tool for evaluating the dimensions of human-system interaction

In Table 1, by contrast, we present a qualitative tool that illustrates the associations between each collaboration protocol and some key dimensions related to the *ethical*[3] component of appropriate reliance: complexity (and, correspondingly, efficiency), dependence, reliance and deskilling. This tool is designed to be indicative, using circular glyphs to represent what each protocol optimizes, guarantees to a certain extent or entails in terms of risk.

Thus, the higher the number of full circles (●) the higher the association between a dimension and the protocol: for instance, the replacement protocol is associated with the highest efficiency, since no human decision is necessary, but also with the highest risk of dependence – since failing the AI, no decision is made – and deskilling –since humans are not supposed to make decision any longer and can lose their capability to do so. A decreasing number of circles corresponds to lower levels of, respectively, efficiency, dependence (or loss of agency), interaction (or reliance), and skill degradation (deskilling). For example, in the inhibition protocol, humans are routinely required to make an initial decision before receiving AI input, resulting in lower dependence compared to the traditional approach. The lowest level in each dimension is represented by an empty circle (○), indicating a complete absence of that dimension. For instance, in the foreclosure protocol, there

[3]from a consequentialist perspective [21].

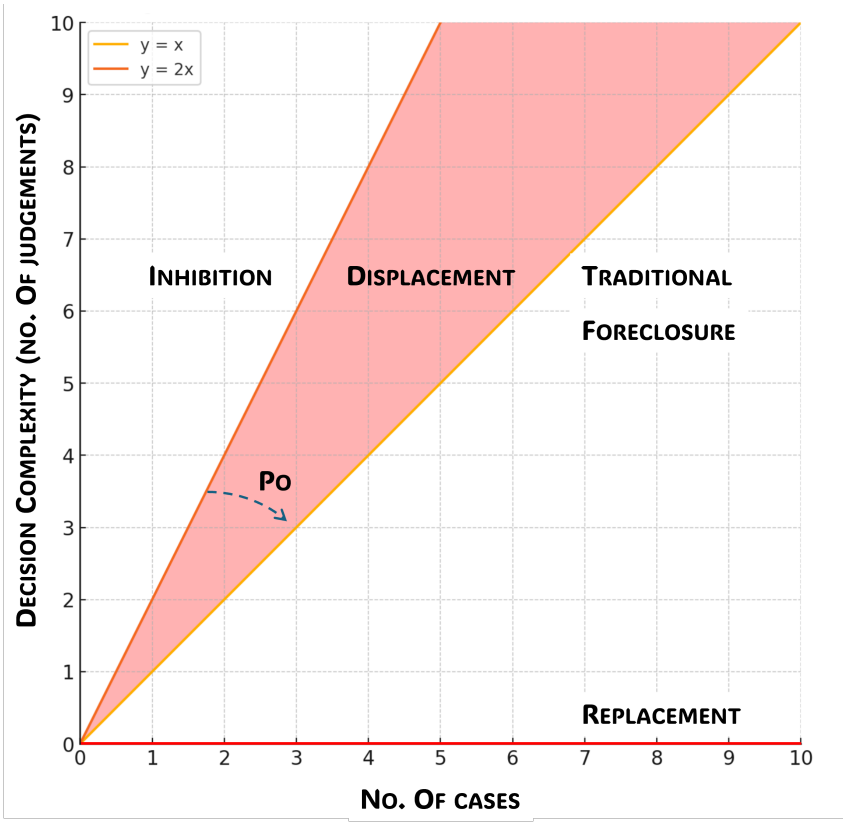


Fig. 4. A diagram illustrating the relationship between the number of cases and decision complexity, defined as the number of human judgments required to reach a final decision. Decision complexity is inversely related to efficiency, assuming constant accuracy. Among the protocols, inhibition protocols demonstrate the lowest efficiency, characterized by the highest decision complexity. In contrast, traditional and foreclosure protocols exhibit the highest efficiency with the lowest decision complexity, despite differing in other aspects. Displacement protocols show variability, aligning with the complexity of inhibition protocols when human-AI agreement (Po) is absent, and with the complexity of traditional and foreclosure protocols when agreement is at its maximum.

is no interaction between users and the AI system, as users must make decisions independently, without any system-provided support. The dimension of complexity, which is complementary to efficiency, merits some further discussion. Indeed, compared with the other dimensions analyzed in Table 1, complexity can also be associated with a quantitative assessment, that is an estimate of the number of human judgments required to formulate a final decision. That is, being n the number of cases or instances to be assessed, Table 1 provides an asymptotic indication of the decision complexity of each protocol. This numerical assessment can be further analyzed in terms of the classification diagram provided in Figure 4. This diagram associates with any given number of cases and collaboration protocol, the corresponding number of human judgments required to come to a final decision: in particular, for the case of displacement, as mentioned also when discussing accuracy, its complexity depends on the rate of agreement between the human decision makers and AI support, which in the diagram is depicted in terms of the P_o, the percentage of agreement.

When considered in its generality, though obviously the actual impacts associated with each protocol will depend on the task at hand and its other multiple contextual factors, Table 1 (along with the diagram in Figure 4) aims to simply illustrate general and intuitive associations, which nevertheless must be thoroughly considered when choosing the aptest collaboration protocols for a given decision setting. The nomogram and qualitative table outlined above, then, are intended to provide complementary perspectives on the considered collaboration protocols, with the aim of informing potential practitioners about their advantages, limitations and best operating conditions, as well as how they balance among different dimensions (including, but not limited to, accuracy). Therefore, practitioners should use these tools together, as a guide for selecting the most appropriate protocol, for any given setting.

4 User studies for collaboration protocol investigation

We applied different collaboration protocols in four user studies in the medical domain. These studies involved real practitioners completing diagnostic tasks based on retrospective real-world cases of varying complexity, presented in a controlled laboratory setting. According to the ten-level Hierarchy of Evidence for AI and XAI empirical studies recently introduced by Famiglini et al. [26], this design corresponds to Level 5 evidence. This hierarchy was inspired by analogous models in evidence-based medicine and health informatics, and ranks empirical studies based on factors such as experimental design, realism of cases, involvement of real practitioners, and setting. Level 5 denotes single quasi-experimental studies involving retrospective real or simulated cases evaluated by real practitioners in a simulated environment. While not the strongest level, this still represents user-centered, empirical research that contributes meaningful evidence for the evaluation and design of human–AI collaboration protocols. In each of the studies, participants were involved through a a combination of convenience-basis and snowball sampling: in particular, in each of the studies, we directly contacted the responsible of the units or school, who then selected the other participants by, either, including members of their research group (in the MRI and X-RAYS studies), or involving students and professionals at the respective hospital (in the ECG and ENDO studies). Below, we outline the main characteristics of these user studies.

The Magnetic Resonance Imaging (MRI) case study. In this radiological study, we engaged 14 board-certified radiologists from IRCCS Ospedale Galeazzi Sant'Ambrogio (Milan, Italy) and other hospitals across Italy. They were tasked with analyzing 240 Magnetic Resonance Images (MRI) for knee lesion classification, selected from the MRNet dataset[4]. The experiment was conducted via an online multi-page questionnaire on the LimeSurvey platform. Each respondent viewed the 240 cases in a randomized order, with the questionnaire displaying three views of each MRI (axial, sagittal, and coronal planes). The radiologists provided their diagnoses (abnormal vs normal) on each case, supported by a simulated AI system with an accuracy of 80%. This study adopted a one-factor, within-subject design, in which we varied the collaboration protocol (traditional vs. inhibition), with one half of the cases (120 out of 240) supported by the traditional collaboration protocol, and the remaining half by the inhibition one.

The Electrocardiogram reading (ECG) case study. In this cardiological study, we involved 44 medical professionals from the Medicine School of the University Hospital of Siena (Italy). Their task was to classify heartbeat patterns from 20 electrocardiograms (ECG), selected by a cardiologist from the ECG Wave-Maven repository [56], using a web-based questionnaire developed on the LimeSurvey platform (version 3.23). The participants provided their diagnoses on each case, interacting with a simulated AI system having an accuracy of 70%. The diagnoses were then compared with the ground

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

Table 2. Summary of quantitative characteristics from the reported medical user studies. Human Accuracy refers to the accuracy of all participants prior to interacting with the AI system. For the MRI and ECG studies, Human Accuracy was calculated based only on the HD1 decisions for cases supported with the inhibition protocol. Percent of Agreement (Po) represents the proportion of times where humans and AI independently proposed the same diagnosis. Reliance measures the proportion of times where humans and AI agreed on the diagnosis after participants received AI support. Appropriate reliance measures the proportion of times users made the right decision regarding whether to trust, or distrust, the AI advice. †: This figure represents the number of participants who completed at least one case; due to the anonymous nature of the study, this is the only data available. ‡: This figure reflects the total number of cases prepared, though only 95 participants provided a final diagnosis for all cases. Additionally, 128 participants fully diagnosed at least 60 cases, and 200 participants diagnosed at least 30 cases.

Study	Num. Participants	Num. Cases	Num. Judgments	Human Accuracy	AI Accuracy	Po	Reliance	Appropriate Reliance
MRI	14	240	5040	0.76	0.80	0.68	0.68	0.27
ECG	44	20	1300	0.54	0.70	0.61	0.78	0.84
XRAYS	16	18	576	0.79	0.78	0.61	0.75	0.45
ENDO	256†	90‡	27930	0.76	0.84	0.74	0.81	0.51

truth from the ECG Wave-Maven repository. This study adopted a one-factor, between-subject design, in which we varied the collaboration protocol (traditional vs. inhibition): participants were randomly divided into two groups that interacted with the AI according to one of the two different collaboration protocols.

The radiography (X-RAYS) case study. In this orthopedic study, we focused on the identification of traumatic thoraco-lumbar fractures using X-ray imaging. The study engaged 16 orthopedists of varied experience levels, who classified 18 previously selected X-ray images, using a multi-page online questionnaire hosted on the LimeSurvey platform. The participants interacted with an AI support system, having an accuracy of 78%, following an inhibition collaboration protocol.

The gastrointestinal endoscopy (ENDO) case study. In this gastroenterological study, we focused on the assessment of gastrointestinal bleeding and of inflammatory lesions through small bowel capsule endoscopy, and on the assessmenf of ulcerative colitis activity through proctosigmoidoscopy. The study engaged 256 participants, among specialists (79%) and trainees (21%) of varied experience levels, who diagnosed up to 90 high-quality endoscopic short videos (15-20 seconds each), using a multi-page online questionnaire hosted on the LimeSurvey platform. The participants interacted with an AI support system, having an accuracy of 80%, following an inhibition collaboration protocol.

4.1 Methods for the analysis of the user studies

In all of the considered four case studies we compared the resulting final accuracy applying the five different collaboration protocols with varying degrees of AI integration in the diagnostic task (replacement, traditional, inhibited/constrained, displaced, and foreclosed). While the general idea associated with the protocols is metaphorically represented in Figure 1, the actual implementation of the protocols in the user studies is described as follows and depicted in Figure 5:

- *Replacement*: we simply considered the actual AI accuracy;
- *Foreclosure*: we considered the HD1 accuracy of the participants in the studies. In particular, for the MRI study we focused on the cases (120 out of 240) in which participants interacted with the AI following the inhibition protocol; for the ECG study we focused only on participants

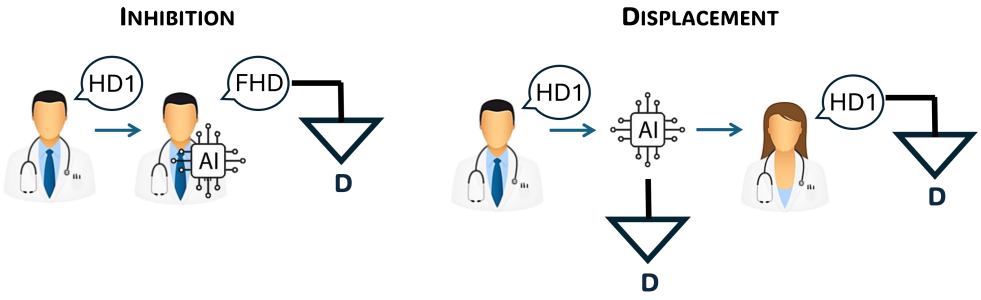


Fig. 5. Graphical representation of the inhibition and displacement collaboration protocols implemented in the user studies. Temporal sequences are indicated by the blue arrow. HD1 represents the initial judgment made by the human decision maker when evaluating a single case. FHD denotes the final decision, made by the human decision maker after receiving support from the AI system. In the implemented displacement protocol (other implementations are possible), the AI system autonomously decides whether to confirm the HD1 of the initial decision maker (if it concurs with this judgment) or to allow the HD1 of another decision maker to become the final decision (based on a majority vote that considers both the AI and the initial human judgment).

in the inhibition group. In both cases, the HD1 decision for scenarios involving the traditional collaboration protocol were excluded from this analysis;

- *Traditional*: we considered this protocol only for the ECG and MRI case studies. In regard to the MRI study, we considered the FHD decisions corresponding to cases in which participants interacted following the traditional collaboration protocol; in the case of the ECG study we focused instead only on the FHD decisions of participants in the traditional group;
- *Inhibition*: we considered the FHD decision of the respondents (see Figure 5). For the XRAYS and ENDO case studies we considered the total number of cases, as these case studies were originally conducted following only the inhibition collaboration protocol; for the MRI case study we only considered the FHD decisions for the 120 cases (for all the respondents) in which participants interacted with the AI system through the inhibition collaboration protocol; while for the ECG case study, we only focused on the FHD decisions for participants in the inhibition collaboration protocol group;
- *Displacement*: we adopted a simulation strategy, described in Algorithm 1 and depicted in Figure 5. First, a randomly selected participant made an initial decision (HD1) for a given case. Then, the AI's judgment for the same case was considered. If the AI agreed with HD1, that decision was finalized. If not, a second randomly selected participant was involved, and the final decision (FHD) was determined by the majority vote between the AI and the two human judgments (which always coincided with the second participant's judgment, obviously).

To compare the accuracy associated with the different collaboration protocols, we conducted an interval-based analysis, comparing the respective 95% confidence intervals: two protocols were considered significantly different if the respective 95% C.I.s did not overlap . The 95% confidence intervals were computed by means of the normal approximation approach. This approach to assess statistical significance was selected as it was similar to, but more robust to violations of parametric

Algorithm 1 Simulation Procedure for the Displacement collaboration protocol

```
procedure DISPLACEMENT(iters: number of simulation iterations)
    acc ← 0
    for t = 1, ..., iters do
        Select a participant p_i uniformly from the set of participants
        Select a case r uniformly for the set of cases
        if p_i(r) == AI(r) then
            acc += p_i(r) == GT(r)          ▷ GT(r) is the ground truth label for case r
        else
            Select a participant p_j ≠ p_i uniformly from the set of participants
            acc += p_j(r) == GT(r)
        end if
    end for
    return acc/iters
end procedure
```

assumptions, than a two-sample t-test at significance level 0.05[5], while at the same time being easily interpretable. In regard to the computation of the C.I. intervals, for the traditional, inhibition and foreclosure protocols, first the accuracy of each participant was computed: these were used, in turn, to compute the average accuracy p associated to that collaboration protocol, as well as the corresponding standard deviation σ. Then, the corresponding 95% confidence intervals were calculated as $p \pm z_{\alpha=0.025}\sigma$, where $z_{\alpha=0.025} \sim 1.96$. For the replacement protocol, we computed the accuracy rate p of the AI: then, the corresponding 95% C.I. was computed as $p \pm \sqrt{\frac{p(1-p)}{n}}$, where n was the total number of judgments. Finally, for the displacement protocol, the accuracy p of the protocol was computed as described in Algorithm 1: then, the corresponding 95% C.I. was computed as $p \pm \sqrt{\frac{p(1-p)}{iters}}$.

Subsequently, we used the results of the case studies to illustrate the application of, as well as validate, our methodological contribution (specifically, the proposed nomograms). To this end, in each of the study we computed the three required parameters H, AI and AR and compared the results of the case studies with the recommendations provided by our nomogram.

4.2 Results

In Table 3 we report the results of the four case studies. Displacement was the best collaboration protocol (in terms of accuracy) in three case studies out of four, namely the MRI, XRAYS and ENDO case studies. For the MRI and ENDO case studies, displacement was significantly better than all other protocols; by contrast, in the XRAYS case study displacement was not significantly different from the second best protocol (i.e., inhibition). Instead, for the ECG case study, the best collaboration protocol was the traditional one, which was significantly better than all other protocols. In all considered cases we found that including both humans and AI in the decisional process (as in the traditional, inhibition and displacement protocols) offered an advantage compared with both the replacement and foreclosure scenarios: in particular, the difference was always statistically significant.

We then illustrate the application of our nomogram, as shown in Figure 6: we illustrate, for each case study, the nomogram with the corresponding value of observed appropriate reliance. Our nomogram was able to correctly predict the most accurate protocol for all of the considered studies.

[5]That is, if the t-test would claim statistical significance, also the comparison of confidence intervals would. By contrast, if the comparison of confidence intervals claims non-significance the t-test could still claim statistical significance. In particular, the adopted approach is still valid for non-normally distributed samples.

Table 3. The results from the user studies (columns) for each protocol (rows). 95% confidence intervals are indicated. In each user study, the best-performing protocol is highlighted in bold. It is important to note that the Foreclosure figure represents baseline human accuracy, which at least one human-AI collaboration protocol has consistently surpassed in all studies.

	MRI	ECG	XRAYS	ENDO
Traditional	.84 ± .02	**.82 ± .04**	n/a	n/a
Inhibition	.77 ± .01	.66 ± .01	.88 ± .01	.80 ± .00
Displacement	**.87 ± .01**	.63 ± .01	**.89 ± .01**	**.85 ± .00**
Foreclosure	.76 ± .00	.54 ± .01	.79 ± .01	.77 ± .00
Replacement	.80 ± .00	.70 ± .02	.78 ± .01	.82 ± .00

5 Discussion

In this section, we discuss the two main findings of our analysis. The first is a descriptive observation concerning the four studies as a whole, which could have significant implications for the design of human-AI configurations and AI affordances. The second finding involves the technical prediction of human performance in hybrid settings and the identification of the optimal human-AI collaboration protocol.

5.1 'No risk, no gain', or the importance of trying different ways

The first observation, as indicated in Table 3, is that no single protocol is suitable for all settings. This finding challenges the mainstream assumption that the best outcomes from integrating AI into work practices arise simply from direct interaction between humans and AI systems, which we refer to as the traditional approach. In fact, the traditional approach outperformed the other protocols only once, in the ECG study—particularly the one involving more users with relatively low expertise. Conversely, the protocol that surpassed all others in the remaining studies, the displacement protocol, explicitly stipulates that humans and AI should *not* interact directly, but rather collaborate through a distributed and asynchronous team effort. The traditional approach, regardless of its implementation, reflects an individualistic stance—often referred to as the human-in-the-loop approach—within the augmentation narrative of human-AI collaboration initiated by Engelbart [23] in the 1960s. The displacement protocol, by contrast, offers a more realistic interpretation of this narrative, recognizing the collaborative nature of professional work and the potential for AI to orchestrate or integrate individual contributions. This approach aligns with the computer-in-the-group concept proposed by Malone [51] and further developed by Shneiderman [79].

At this point, however, we are not claiming that any one protocol is consistently superior to the others. On the contrary, we are making two other points. First and foremost, evaluating or simulating the effects of collaboration protocols in specific decision contexts is essential for the informed and responsible deployment of AI-based systems, contributing to an evidence-based approach to AI decision support [26]. In our analysis, we confirmed an intuitive conjecture: different protocols influence performance in varying ways, leading to different levels of final accuracy. However, while accuracy is critical in decision-making, it is not the only performance dimension that should be considered. In fact, when considering other socio-technical dimensions, some of which are outlined in Table 1, the selection of the most suitable collaboration protocol is far from trivial and should not be taken lightly. This is particularly important given the need to ensure human capital sustainability [70], a sense of agency [28, 91], and satisfaction, in addition to efficacy and efficiency. We will revisit this point in the conclusion of this discussion.

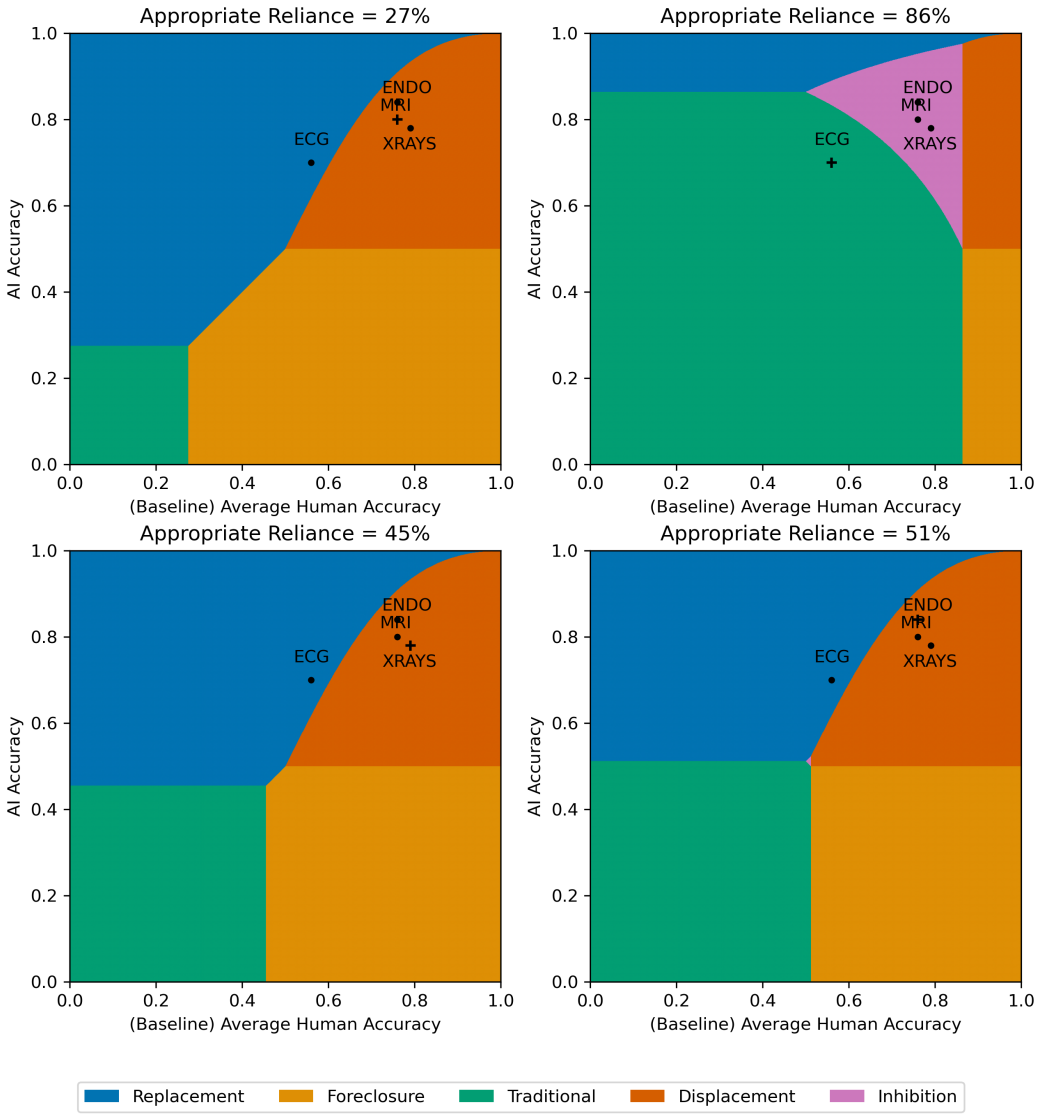


Fig. 6. Nomograms populated with results from four case studies. Each study is represented by a dedicated nomogram, indicating the level of appropriate reliance observed. Within each nomogram, the specific case study is marked with a '+' symbol, identifying the protocol suggested for that case study. The glyphs associated with the other studies are included for comparison, enabling a 'what-if' analysis to determine the most suitable protocol at different levels of appropriate reliance.

Moreover, different protocols exert varying effects due to multiple, intertwined factors, such as the nature of the decision task, the attitudes of the decision-makers, their expertise, and their prior experience with decision support systems. These aspects, although notoriously elusive and difficult to operationalize, correlate with observable behavioral characteristics, such as AI accuracy (related to previous interactions), average human accuracy (related to expertise), and the users'

ability to discern whether the machine will be able to assist them. This ability, which is beginning to attract scholarly interest [49, 50], is often described with tentative terms such as "mentalizing about AI", "theory of machine" or " machine reading[6]." It is typically operationalized in terms of *appropriate reliance*, which refers to the extent to which users understand when to rely (or not) on the machine's advice, even when they are unsure of the correct answer. This leads us to the second main contribution of our analysis.

The medical case studies were deliberately chosen to reflect the heterogeneity of professional decision-making contexts. MRI and endoscopy involve highly specialized image-based diagnostics performed by experienced clinicians, while ECG and X-ray tasks entail more structured classifications by professionals with varied expertise. This heterogeneity enabled us to test the protocols in contrasting settings and demonstrate that traditional interaction models are not necessarily optimal—even when focusing solely on accuracy. It also illustrated how our nomogram and evaluation framework can guide context-sensitive protocol selection. While further validation in non-medical domains is needed, this diversity enhances the applicability of our contribution by showing how the framework adapts to distinct decision environments.

5.2 'Starting well is half the game', or how to set up the nomogram

The decision effectiveness of human-AI hybrid work or collaboration can be accurately estimated by knowing the average human accuracy, the AI accuracy, and a reliable measure of users' ability to determine whether AI is correct, that is appropriate reliance. If a protocol is defined by its impact on these three performance dimensions, and if a specific decision setting can be characterized by these measures, then the optimal protocol for that setting can be identified.

This statement has been proved by the validation of the nomogram shown in Figure 3. Indeed, the values of accuracy predicted by the nomogram for the best protocols in the case studies (resp., MRI - displacement: 87%; XRAYS - displacement: 88%; ECG - traditional: 86%; ENDO - displacement: 89%) were close to the actual observed values, with a maximum error of less than 5%. For this reason, the proposed nomogram can be used not only to provide a qualitative indication of the best protocol to adopt, but also to estimate quantitative information on the expected accuracy of different protocols. This makes it a valuable tool for stakeholders to make informed decisions regarding the adoption of various collaboration protocols, while considering local feasibility and the prioritization of needs.

To this end, further explanations are necessary to understand how this tool can be applied and used in practice, particularly in terms of how the aforementioned parameters can be accurately obtained. The baseline accuracy of AI and human participants can be obtained in various ways: for instance, by conducting a pilot study to measure these performance scores (noting that in such a study, AI and human participants can operate independently without requiring interaction), or by using estimates from previous, related studies. Estimating appropriate reliance is less simple. In fact, the idea of measuring this latter concept has only recently received attention in the specialized literature [71, 72, 77]. We outline three strategies to achieve this: 1) A pilot study to estimate appropriate reliance in the desired context (this is what we also do in our case studies). Unlike baseline accuracy assessments, this requires the adoption of the inhibition collaboration protocol to assess the actual reliance of humans on AI. Our method can then be applied to estimate accuracy for all protocols, despite only applying the inhibition protocol in the pilot study. 2) Using data from

[6]These terms are clearly inspired by their psychological counterparts, which describe the human ability to discern, from clues and behavioral cues, whether others are benevolent (intentions), sincere (willingness), and competent (capabilities). In the context of machine-based systems, particularly decision aids, and assuming benevolence and sincerity, these terms refer to the human ability to determine whether the machine is correct in its proposals or recommendations.

previous, comparable studies. 3) Establishing a baseline for desired appropriate reliance analytically, as outlined in [77].

The three approaches mentioned differ significantly in their data requirements, assumptions about accuracy, and generalizability. Intuitively, the second and third methods bypass the need for empirical data from the specific context, as they estimate appropriate reliance either by reusing existing data or through purely analytical methods. In contrast, the first method requires a pilot study and the collection of relevant usage data within the targeted setting. Specifically, the third method allows for generalization to other contexts, provided that baseline accuracy and appropriate reliance values remain relatively stable. However, this method introduces additional assumptions beyond those required by our approach. For example, employing the formulation in [77], one can estimate the minimum level of reliance necessary to achieve the desired final decision accuracy (FHD) as:

$$R = FHD + AI - 1, \tag{7}$$

where R: Minimum level of reliance necessary; AI: AI accuracy; FHD: Desired final decision accuracy

which is valid if AI is $> .5$, as it is often the case. Though this estimation formula is conceptually simple and could in principle be used to determine the reliance from only the desired level of accuracy, its validity rests, critically, on the assumption that reliance is always appropriate[7]. This overlooks the reality where reliance on AI can sometimes be detrimental, such as when AI recommendations are wrong. Thus, the effectiveness of the proposed nomogram using the third method depends on the validity of the assumptions used to calculate an analytical estimate of appropriate reliance. In contrast, both the first and second methods assume that the level of appropriate reliance can be applied to varied, yet related, contexts. For the second method, this implies that findings are generalizable across different studies. Conversely, the first method makes a less stringent assumption, suggesting generalizability across multiple observations within the same setting. These assumptions are much less stringent than those required for the third method, as they are not only inherent to our approach but are also reasonable within the context of using AI-based decision support systems. The assumption of in-domain generalizability, necessary for the first method, aligns with the expectations of current machine learning methodologies. Conversely, out-of-domain generalizability, which is crucial for the second method, presents a greater challenge but is still a common assumption for the transferability of AI systems. Notably, implementing the first method enables the subsequent application of the second method by other researchers in similar settings, provided the same assumptions hold, making the first method the most effective strategy when feasible.

5.3 'Unity is strength', or the importance of collective hybrid intelligence

Returning to the empirical results from our four use cases, it is noteworthy that displacement protocols resulted in the best performance in three out of four cases, and this difference was statistically significant. This highlights an important insight: achieving super-human performance does not require a super-human AI, but rather the effective leveraging of the collective intelligence of a team augmented by AI. Notably, the AI can be less accurate than the human team members. To illustrate this finding, we refer to the diagram depicted in Figure 7.

The simplifying assumptions behind the diagram depicted in Figure 7 are as follows: a) the average accuracy of the human decision makers is known; b) human variance around this estimate

[7]Indeed, the formulation in [77] considers the reliance, which is simply the probability with which the human follows the AI recommendation, rather than the appropriate reliance.

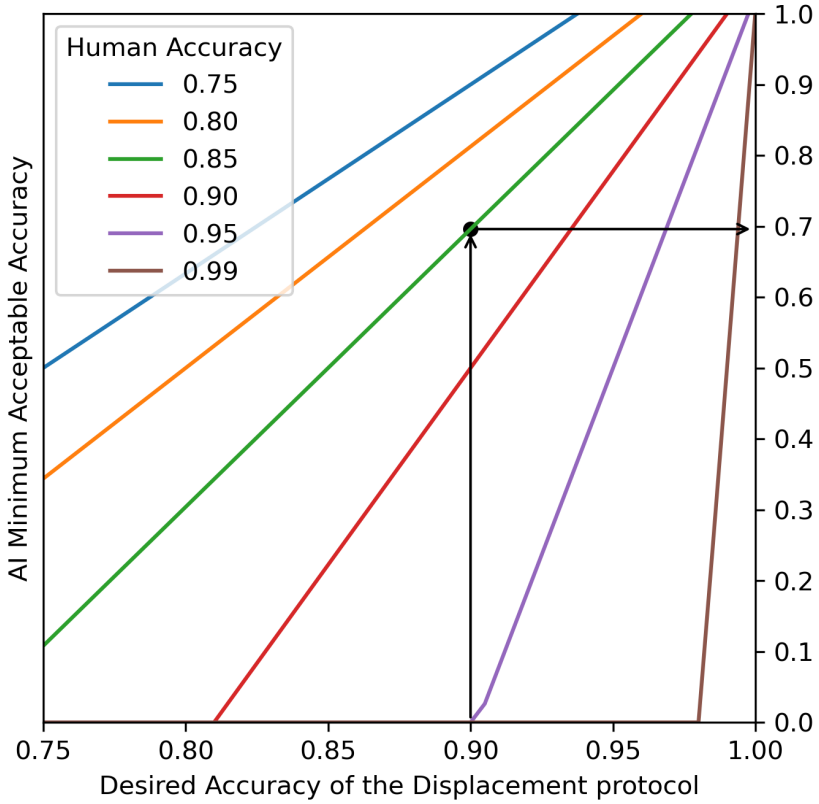


Fig. 7. A diagram for determining the minimum acceptable accuracy (MAA) of an AI in a binary decision setting, for the displacement protocol. The arrows demonstrate how to use the diagram to read the MAA on the vertical axis, once the desired accuracy from applying the displacement protocol has been set according to the average accuracy of the human, as indicated by the colored segments where the dashed line intersects.

is practically negligible (i.e., experts exhibit similar performance); and c) the judgement of experts (including the AI) are independent. If these three preconditions are met, then the formula in Equation 3 represents the probability that the majority decision of the group of three raters (thus, including the AI) is indeed correct. For instance, if we consider a common average accuracy, such as 85% (which was observed among super-specialists radiologists in [9]), and a reasonable increase in accuracy of 5%, we can see that the displacement protocol used in our studies would require an AI with a significantly lower accuracy, approximately 70%. This lower accuracy AI would be less expensive or exclusive, yet still effective.

The exploration of the displacement protocol in our study contributes to the broader CSCW discourse on asynchronous collaboration between humans and AI systems, offering a novel perspective on collective hybrid intelligence as a means to mitigate over-reliance and foster more sustainable work practices [27]. This approach underscores the value of distributed, indirect AI-human interaction in achieving contextually appropriate decisions while maintaining human agency.

5.4 'Nobody's perfect', or some methodological limitations

Given that we acknowledged certain aspects of decision-making as elusive, we must further discuss the limitations of our methodological contribution—the nomogram for identifying the collaboration protocol associated with the highest estimated final accuracy. These limitations primarily concern the simplifying assumptions we had to adopt.

First, the accuracy scores of both humans and the AI system, as well as the level of appropriate reliance, must be representative of future performance—that is, they must be generalizable to some extent, at least within the same decision setting. Second, the judgments of unaided humans and the AI system (and hence their baseline accuracies) must be independent. Third, the impact of extreme algorithmic aversion must be negligible, meaning that the likelihood of humans changing their minds simply because they realize they agree with the machine is very low. If any of these assumptions do not hold, our nomogram may fail to correctly identify the optimal collaboration protocol. However, it is important to note that these assumptions are relatively minimal and can reasonably be expected to apply in most general settings.

The first assumption is common in almost any sample-based investigation and is supported by proper randomization, a sufficiently large sample size, and limited inter-rater variability, which is typical in expert performance. The second assumption is also reasonable, as humans express their first decision (HD1) without any interaction with the AI system. However, this depends on the adaptiveness of the machine and the attitude of the humans, particularly novices. This leads to the third assumption: extreme algorithmic aversion is actually quite rare, as it contradicts other strong human heuristics, such as confirmation bias and fixation. Such aversion is usually motivated by strong prejudices against the machine [8] or extremely poor previous interactions with the system, which are more likely to result in neglect and disuse [61] rather than outright opposition.

In regard to our case studies, we must also mention some potential limitations. First and foremost, the four case studies only focused on the medical domain and ,furthermore, the considered settings all had different requirements and decision criteria, as well as different sample sizes. In particular, in two of the studies (the MRI study and the XRAYS study) the number of involved participants was relatively small, with a consequent reduction in statistical power. Obviously, this may limit the generalizability of our results, especially to other domains different from the medical one. At the same time, it is important to emphasize that we are not interested in horizontal comparisons among different studies but rather in showing how to apply our methodological framework, making the comparability of settings, requirements and sample sizes less relevant. Similarly, we do not make specific claims about the generalizability of our results, which are specific to each of the considered case studies. In fact, our primary conclusion is that there is no universally optimal collaboration protocol, as the effectiveness of a protocol is highly dependent on the specific characteristics of the setting and the trade-offs among the dimensions reported in Table 1 (and potentially others) that designers and practitioners consider feasible and acceptable. For this reason, we believe that our results are valuable in motivating further replication in diverse and more powered user studies to validate the usefulness of the nomogram and related methods, as well as applications in other work settings and domains.

5.5 'There Are More Things In Heaven And Earth', or the importance of going beyond the nomogram

This final point leads us to revisit what we aimed to address with Table 1 (and, similarly, Figure 4 for the dimension of decision complexity), which associates each collaboration protocol with relevant quality dimensions beyond accuracy.These dimensions include, but are not limited to,

efficiency (and its counterpart, complexity), the degree of direct interaction, and potential risks such as loss of agency [28, 59] and loss of skill [82].

While reducing errors remains a critical objective, from a CSCW perspective we recognize that any holistic evaluation of human-AI interaction must consider socio-technical factors that affect the sustainability of work practices and human agency in collaborative settings [5, 25, 93]. In line with this, while considering dimensions beyond accuracy is more challenging and often a neglected practice, our study underscores the importance of evaluating the socio-material consequences of AI deployment, particularly how the design of collaboration protocols influences long-term cognitive, social, and organizational outcomes.

Table 1 outlines the main dimensions that can be considered at the general level of collaboration protocols. However, more context-specific quality dimensions, intrinsic to the concept of usability (cf. ISO 9241), such as user satisfaction, must also be factored into the design of human-AI collaboration. This includes determining what information the AI system must provide to users and how this information is displayed (i.e., its affordances). Considering the long-term effects of implementing AI-based solutions in socio-technical environments is crucial. While short-term benefits may include increased decision-making accuracy, there is a risk that such systems could eventually lead to cognitive biases like automation bias, algorithmic aversion, or to risks of deskilling and user de-responsibilization [11]. Therefore, evaluating additional dimensions beyond accuracy is essential for a comprehensive assessment of different collaboration protocols. For example, if our nomogram indicates that two protocols have similar accuracy, as seen with inhibition and displacement in the XRAYS case study, or displacement and replacement in the ENDO case study, the qualitative assessment should guide the choice toward the protocol with a potentially lower risk of deskilling, such as inhibition [69], even if it is associated with a slightly lower estimate of final accuracy.

The findings of our study offer practical design implications for developing computing technologies that better support collaborative work. By identifying that no single AI collaboration protocol is universally optimal, our research underscores the importance of context-sensitive designs in AI systems intended to assist in cooperative environments. The displacement protocol, which consistently demonstrated superior performance in several use cases, suggests that asynchronous, distributed human-AI collaboration might be a promising direction for future design approaches. This protocol, which reduces direct reliance on AI by leveraging human judgment in decision-making, mitigates issues such as over-dependence and skill erosion – concerns frequently highlighted in CSCW literature.

Moreover, the empirical evaluation of protocols provides a basis for developing human-centered AI systems that account for both short-term performance and long-term human factors. This insight supports the need for design frameworks that not only optimize decision accuracy but also balance the cognitive load and agency of human workers. Thus, our work advocates for designs that maintain human expertise while ensuring AI systems function as collaborative tools rather than decision replacements.

These insights can inform the design of future AI-enhanced systems aimed at maintaining a sustainable balance between human input and machine intelligence in cooperative settings, fostering more resilient, efficient, and context-appropriate solutions in industries such as healthcare and beyond.

6 Conclusion

The argument that there is no single best way to deploy AI systems in organizational settings is largely uncontested. However, a knowledge gap persists regarding the broad categories that encompass the various human-machine configurations [83]. Additionally, there is limited understanding of the criteria for selecting one configuration over another in a specific work context.

To address this gap and illustrate how different modes of interaction can influence outcomes, we proposed a methodological framework that includes both a nomogram (with corresponding quantitative formulas) and a qualitative assessment table and diagram. This framework is intended to assist practitioners and users in comprehensively evaluating interaction modalities across multiple dimensions.Assessing these varying effects is essential for determining whether to adhere to conventional AI applications or adopt more frictional protocols [10, 55], aimed at mitigate the risks associated with over-reliance, automation bias, and the potential long-term consequences of AI abuse [61].

Our application of different collaboration protocols across four realistic user studies in a medical setting led us to a significant observation: while integrating AI into decision-making processes can indeed enhance organizational performance, this enhancement does not always occur through traditional protocols. Each collaboration protocol through which human decision-makers utilize decision support systems exerts unique effects on key usability dimensions, including effectiveness, efficiency, and (short- and long-term) impacts on users. As such, the findings from our studies demonstrate that no single mode of interaction is universally beneficial across all decision-making tasks and settings. Different levels of AI integration, particularly when users do not interact directly with AI (AI displacement), can sometimes prove to be the most effective. Thus, the appropriateness of various collaboration protocols depends on the specific decision task at hand, with certain protocols potentially offering superior outcomes.

This study represents a small but meaningful step toward establishing best practices for evaluating collaboration protocols, including unconventional ones, such as those envisioned by Pierce in his call for a progressive *undesign* of automation in critical decision-making tasks [63]. The goal is to implement protocols that maximize potential accuracy, while minimizing the socio-technical impact on decision-makers, that is their work on a more pragmatic and ethical level. The aim is to facilitate the deployment of *positive-sum automation*, where the net outcome of gains and losses is positive, sustainable, and achieved through solutions that leverage a strictly evidence-based design approach [26].

References

[1] Marc Anderson and Karën Fort. 2022. Human where? A new scale defining human involvement in technology communities from an ethical standpoint. *International Review of Information Ethics* 31, 1 (2022).

[2] Giulia Anichini, Chiara Natali, and Federico Cabitza. 2024. Invisible to Machines: Designing AI that Supports Vision Work in Radiology. *Computer Supported Cooperative Work (CSCW)* (2024), 1–44.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[4] Matt Beane. 2019. Learning to work with intelligent machines. *Harvard Business Review* 97, 5 (2019), 140–148.

[5] Glen Berman, Nitesh Goyal, and Michael Madaio. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[7] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making* 33, 2 (2020), 220–239.

[8] Federico Cabitza. 2019. Biases affecting human decision making in AI-supported second opinion settings. In *Modeling Decisions for Artificial Intelligence: 16th International Conference, MDAI 2019, Milan, Italy, September 4–6, 2019, Proceedings 16*. Springer, 283–294.

[9] Federico Cabitza, Andrea Campagner, Luca Ronzio, Matteo Cameli, Giulia Elena Mandoli, Maria Concetta Pastore, Luca Maria Sconfienza, Duarte Folgado, Marília Barandas, and Hugo Gamboa. 2023. Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine* 138 (2023),

102506.

[10] Federico Cabitza, Chiara Natali, Lorenzo Famiglini, Andrea Campagner, Valerio Caccavella, and Enrico Gallazzi. 2024. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial Intelligence in Medicine* 150 (2024), 102819.

[11] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. 2017. Unintended consequences of machine learning in medicine. *Jama* 318, 6 (2017), 517–518.

[12] Federico Cabitza and Carla Simone. 2013. Computational Coordination Mechanisms: A tale of a struggle for flexibility. *Computer Supported Cooperative Work (CSCW)* 22 (2013), 475–529.

[13] Andrea Campagner, Federico Cabitza, and Davide Ciucci. 2019. Three–way classification: Ambiguity and abstention in machine learning. In *Rough Sets: International Joint Conference, IJCRS 2019, Debrecen, Hungary, June 17–21, 2019, Proceedings*. Springer, 280–294.

[14] Shiye Cao and Chien-Ming Huang. 2022. Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.

[15] Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. Designing for Appropriate Reliance: Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *arXiv preprint arXiv:2401.05612* (2024).

[16] Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, et al. 2023. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *arXiv preprint arXiv:2401.05377* (2023).

[17] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[18] Jonathan Chen and Andrew Beam. 2019. Potential trade-offs and unintended consequences of AI. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril* (2019), 89.

[19] Zeya Chen and Ruth Schmidt. 2024. Exploring a Behavioral Model of" Positive Friction" in Human-AI Interaction. *arXiv preprint arXiv:2402.09683* (2024).

[20] Gabriele Cimolino and TC Nicholas Graham. 2022. Two heads are better than one: A dimension space for unifying human and artificial intelligence in shared control. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[21] Virginia Dignum and Virginia Dignum. 2019. Ethical decision-making. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (2019), 35–46.

[22] Therese Enarsson, Lena Enqvist, and Markus Naarttijärvi. 2022. Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law* 31, 1 (2022), 123–153.

[23] Douglas C Engelbart. (1962) 2023. Augmenting human intellect: A conceptual framework. In *Augmented Education in the Global Age*. Routledge, 13–29.

[24] Douglas C Engelbart and William K English. 1968. A research center for augmenting human intellect. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*. 395–410.

[25] Johanne Mose Entwistle, Mia Kruse Rasmussen, Nervo Verdezoto, Robert S Brewer, and Mads Schaarup Andersen. 2015. Beyond the individual: The contextual wheel of practice as a research framework for sustainable HCI. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1125–1134.

[26] Lorenzo Famiglini, Andrea Campagner, Marilia Barandas, Giovanni Andrea La Maida, Enrico Gallazzi, and Federico Cabitza. 2024. Evidence-based XAI: An empirical approach to design more effective and explainable decision support systems. *Computers in Biology and Medicine* 170 (2024), 108042.

[27] Mingming Fan, Xianyou Yang, TszTung Yu, Q Vera Liao, and Jian Zhao. 2022. Human-ai collaboration for UX evaluation: effects of explanation and synchronization. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–32.

[28] Astrid Galsgaard, Tom Doorschodt, Ann-Louise Holten, Felix Christoph Müller, Mikael Ploug Boesen, and Mario Maas. 2022. Artificial intelligence and multidisciplinary team meetings; a communication challenge for radiologists' sense of agency and position as spider in a web? *European Journal of Radiology* 155 (2022), 110231.

[29] Ozlem Ozmen Garibay, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco, Stephen Fiore, Ivan Garibay, Keri Grieman, John Havens, Marina Jirotka, Hernisa Kacorri, Waldemar Karwoski, Joe Kider, Joseph Konstan, Sean Koon, Monica Lopez-Gonzales, Iliana Maifeld-Carucci, Sean McGregor, Gavriel Salvendy, Ben Shneiderman, Constantine Stephanidis, Christina Strobel, Carolyn Ter Holter, and Wei Xus. 2023. Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction* 39, 3 (2023), 391–437.

[30] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics* 83, 5 (2014), 368–375.

[31] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[32] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[33] Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. 2024. A Statistical Framework for Measuring AI Reliance. *arXiv preprint arXiv:2401.15356* (2024).

[34] Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*.

[35] Jessica He, David Piorkowski, Michael Muller, Kristina Brimijoin, Stephanie Houde, and Justin Weisz. 2023. Rebalancing worker initiative and AI initiative in future work: Four task dimensions. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–16.

[36] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2024. Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence. *arXiv preprint arXiv:2404.00029* (2024).

[37] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 453–463.

[38] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2024. Machine learning with a reject option: A survey. *Machine Learning* 113, 5 (2024), 3073–3110.

[39] W David Holford. 2022. 'Design-for-responsible'algorithmic decision-making systems: a question of ethical judgement and human meaningful control. *AI and Ethics* 2, 4 (2022), 827–836.

[40] Daniel Kahneman, Olivier Sibony, and Cass R Sunstein. [n. d.]. *Noise*.

[41] Hideki Koike, Jun Rekimoto, Junichi Ushiba, Shinichi Furuya, and Asa Ito. 2021. Human augmentation for skill acquisition and skill transfer. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–3.

[42] Steve Krug et al. 2014. Don't make me think, Revisited. *A Common Sense Approach to Web and Mobile Usability* (2014).

[43] Jethro CC Kwong, David-Dan Nguyen, Adree Khondker, Jin Kyu Kim, Alistair EW Johnson, Melissa M McCradden, Girish S Kulkarni, Armando Lorenzo, Lauren Erdman, and Mandy Rickard. 2024. When the model trains you: induced belief revision and its implications on artificial intelligence research and patient care—a case study on predicting obstructive hydronephrosis in children. *NEJM AI* 1, 2 (2024), AIcs2300004.

[44] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[45] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[46] Roberto Legaspi, Wenzhen Xu, Tatsuya Konishi, Shinya Wada, Nao Kobayashi, Yasushi Naruse, and Yuichi Ishikawa. 2024. The sense of agency in human–AI interactions. *Knowledge-Based Systems* 286 (2024), 111298.

[47] Peng Liu. 2023. Reflections on automation complacency. *International Journal of Human–Computer Interaction* (2023), 1–17.

[48] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara DarziJonathan J Deeks, et al. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health* 2, 10 (2020), e537–e548.

[49] Jennifer M Logg. 2022. The psychology of Big Data: Developing a "theory of machine" to examine perceptions of algorithms. (2022).

[50] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[51] Thomas W Malone. 2018. How human-computer 'superminds' are redefining the future of work. *MIT Sloan management review* 59, 4 (2018), 34–41.

[52] Anne-Sophie Mayer, Franz Strich, and Marina Fiedler. 2020. Unintended Consequences of Introducing AI Systems for Decision Making. *MIS Quarterly Executive* 19, 4 (2020).

[53] Tim Miller. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 333–342.

[54] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1

(2024), 1–39.

[55] Mohammad Naiseh, Reem S Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through friction: An approach for calibrating trust in explainable ai. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*. IEEE, 1–5.

[56] Larry A Nathanson, Charles Safran, Seth McClennen, and Ary L Goldberger. 2001. ECG Wave-Maven: a self-assessment program for students and clinicians.. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 488.

[57] Donald A Norman. 1986. Cognitive engineering. *User centered system design* 31, 61 (1986), 2.

[58] Osonde A Osoba, William Welser IV, and William Welser. 2017. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

[59] Marine Pagliari, Valérian Chambon, and Bruno Berberian. 2022. What is new with Artificial Intelligence? Human–agent interactions through the lens of social agency. *Frontiers in Psychology* 13 (2022), 954444.

[60] Raja Parasuraman and Dietrich H Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors* 52, 3 (2010), 381–410.

[61] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[62] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.

[63] James Pierce. 2012. Undesigning technology: considering the negation of design by design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 957–966.

[64] James Pierce. 2014. Undesigning interaction. *Interactions* 21, 4 (2014), 36–39.

[65] Jérémie Poiroux, Nolwenn Maudet, Karl Pineau, Emeline Brulé, and Aurélien Tabard. 2024. Design indirections: how designers find their ways in shaping algorithmic systems. *Computer Supported Cooperative Work (CSCW)* 33, 2 (2024), 173–204.

[66] Jan Pries-Heje and Richard Baskervillle. 2022. Evidence-based Information Systems Design: How Research Can Inform Sociotechnical AI Design. In *Pre-ICIS OASIS Workshop 2022: Value Positions and Criticality in Digital Transformation Research*.

[67] Clara Punzi, Roberto Pellungrini, Mattia Setzu, Fosca Giannotti, and Dino Pedreschi. 2024. AI, Meet Human: Learning paradigms for hybrid decision making systems. *arXiv preprint arXiv:2402.06287* (2024).

[68] Tapani Rinta-Kahila, Esko Penttinen, Antti Salovaara, Wael Soliman, and Joona Ruissalo. 2023. The vicious circles of skill erosion: a case study of cognitive automation. *Journal of the Association for Information Systems* 24, 5 (2023), 1378–1412.

[69] Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. 2024. AI and XAI second opinion: the danger of false confirmation in human–AI collaboration. *Journal of Medical Ethics* (2024).

[70] Lea Samek and Mariagrazia Squicciarini. 2023. *Impact of Artificial Intelligence in Business and Society*. Routledge, Chapter AI human capital, jobs and skills, 171–191.

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

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

[73] Max Schemmer, Niklas Kühl, and Gerhard Satzger. 2022. Intelligent Decision Assistance Versus Automated Decision-Making: Enhancing Knowledge Workers Through Explainable Artificial Intelligence. In *HICSS'22: Proceedings of the Hawaii International Conference on System Sciences*.

[74] Thomas Schilling, Rebecca Müller, Thomas Ellwart, and Conny H Antoni. 2024. Context-dependent preferences for a decision support system's level of automation. *Computers in Human Behavior Reports* 13 (2024), 100350.

[75] Kjeld Schmidt and Carla Simonee. 1996. Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work (CSCW)* 5 (1996), 155–200.

[76] Jakob Schoeffer. 2022. A Human-Centric Perspective on Fairness and Transparency in Algorithmic Decision-Making. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–6.

[77] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making. *arXiv preprint arXiv:2304.08804* (2023).

[78] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[79] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.
[80] Line Silsand and Gunnar Ellingsen. 2016. Complex decision-making in clinical practice. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 993–1004.
[81] Herbert A Simon. 2019. *The Sciences of the Artificial, reissue of the third edition with a new introduction by John Laird*. MIT press.
[82] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. *The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence.* Technical Report. Massachusetts Institute of Technology.
[83] Lucille Alice Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.
[84] Siddharth Swaroop, Zana Buçinca, Krzysztof Z Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *29th International Conference on Intelligent User Interfaces (IUI'24). ACM*.
[85] Heliodoro Tejeda, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2022. AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior* 5, 4 (2022), 491–508.
[86] Olga Towstopiat. 1984. A review of reliability procedures for measuring observer agreement. *Contemporary educational psychology* 9, 4 (1984), 333–352.
[87] Anna-Sophie Ulfert, Conny H Antoni, and Thomas Ellwart. 2022. The role of agent autonomy in using decision support systems at work. *Computers in Human Behavior* 126 (2022), 106987.
[88] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When Are Combinations of Humans and AI Useful? *arXiv preprint arXiv:2405.06087* (2024).
[89] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.
[90] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–6.
[91] Kilian Wenker. 2023. Who wrote this? How smart replies impact language and agency in the workplace. *Telematics and Informatics Reports* 10 (2023), 100062.
[92] Ben Wilson, Chiara Natali, Matt Roach, Darren Scott, Alma Rahat, David Rawlinson, and Federico Cabitza. 2025. Dimensions of Human-Machine Combination: Prompting the Development of Deployable Intelligent Decision Systems for Situated Clinical Contexts. In *Computer Supported Cooperative Work, Vol. 34*. Springer.
[93] Christine Wolf and Jeanette Blomberg. 2019. Evaluating the promise of human-algorithm collaborations in everyday work practices. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[94] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 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*. 1–13.
[95] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

Received October 2024; revised April 2025; accepted August 2025