Rams, Hounds and White Boxes: Investigating Human-AI Collaboration Protocols in Medical Diagnosis

Federico Cabitza[a,b,*], Andrea Campagner[a,*], Luca Ronzio[c], Matteo Cameli[d], Giulia Elena Mandoli[d], Maria Concetta Pastore[e], Luca Sconfienza[b,f], Duarte Folgado[g], Marília Barandas[g], Hugo Gamboa[g,h]

[a] *Department of Computer Science, Systems and Communication, University of Milano-Bicocca, Milan, Italy*
[b] *IRCCS Istituto Ortopedico Galeazzi, Milan, Italy*
[c] *IRCCS Istituto San Raffaele, Milan, Italy*
[d] *University of Siena, Siena, Italy*
[e] *Azienda Ospedaliera Maggiore Della Carita' Di Novara, Novara, Italy*
[f] *Department of Biomedical Sciences for Health, Università degli Studi di Milano, Milan, Italy*
[g] *Associação Fraunhofer Portugal Research, Lisbon, Portugal*
[h] *Laboratório de Instrumentação, Engenharia Biomédica e Física da Radiação (LIBPhys-UNL), Departamento de Física, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa, Lisbon, Portugal*

Abstract

In this paper, we study human-AI collaboration protocols, a design-oriented construct aimed at establishing and evaluating how humans and AI can collaborate in cognitive tasks. We applied this construct in two user studies involving 12 specialist radiologists (the knee MRI study) and 44 ECG readers of varying expertise (the ECG study), who evaluated 240 and 20 cases, respectively, in different collaboration configurations. We confirm the utility of AI support but find that XAI can be associated with a "white-box paradox", producing a null or detrimental effect. We also find that the order of presentation matters: AI-first protocols are associated with higher diagnostic accuracy than human-first protocols, and with higher accuracy than both humans and AI alone. Our findings identify the best conditions for AI to augment human diagnostic skills, rather than trigger dysfunctional responses and cognitive biases that can undermine decision effectiveness.

Keywords: Human-AI collaboration protocols; Artificial Intelligence; Explainable AI; cognitive biases; automation bias

1. Introduction

In a recent editorial, Elmor and colleagues [25] noted that "there are complex interactions between a computer algorithm output and the interpreting physi-

[*] Authors contributed equally

cian... the extent to which physicians may be influenced by the many types and timings of computer cues remains unknown".

Despite increasing research into the application of AI systems in healthcare [62, 70], the potential impact of AI on clinical decision making is still poorly understood [29], and particularly so for systems featuring complex interaction between clinicians and AI, such as those endowed with eXplainable AI (XAI) [18, 38]. Several studies [10, 33, 39, 42, 44, 64] have demonstrated the benefits that AI systems offer, but others raised concerns about the emergence of biases, such as *automation bias* [46, 44], *algorithmic aversion* [11, 20] or deskilling [36, 63], which can lead to an increase in clinical errors.

As noted above and as highlighted in recent experimental studies [29], not only the quality of decision support, but also the way information is presented to the clinicians, and how the clinicians interact with the decision support, can have a significant impact on clinical decision-making [61, 16, 32]. When evaluating the impact of AI-based decision support systems in clinical practice, AI-based medical solutions should thus not only be tested as stand-alone apps, but also in real-world scenarios with real people to evaluate how human agents perceive AI-generated recommendations and explanations [21, 64].

In practice, several features of the interaction between clinicians and their computational decision aids could have a potential influence on trust, adoption, and clinical accuracy, and should therefore be precisely stipulated [6]. These include: the information exchanged between users and machines, including how the advice is conveyed; the order in which users and machines exchange such information, including whether users are required to produce a temporary decision output before obtaining advice from the machine [7]; whether or not explanations should be given and what kind (such as feature rankings, pixel attribution maps, textual justifications); what kind of support can be given (e.g., sensitive rather than specific [12], or calibrated rather than uncalibrated [67]); the doctors who are the intended users should be identified, along with the configurations that are associated with significant effects.

Various studies have recently aimed to evaluate more complex interactions between human and AI agents in the clinical setting. The study of Gaube et al. [29] involved 265 doctors with different levels of task expertise who were asked to assess chest radiology images for the presence of abnormalities, with the support of either an AI or a second human agent. The diagnostic accuracy was found to be similar in the two cases, even though the clinicians trusted more the support of a colleague rather than that of the AI. Tschandl et al. [64] asked 302 clinicians to analyze dermoscopic photographs of benign and malignant skin alterations, both with and without the use of AI. They compared three protocols, in which the AI support was presented in different formats: the confidence scores associated with all pertinent diagnoses; the confidence score of a malignant alteration; or a selection of similar pictures along with the respective diagnoses. The authors found that collaboration with AI improved the examiners' diagnosis accuracy only in the first case, and significantly so.

Thus, we draw on the above mentioned studies and introduce the concept of an human - artificial intelligence collaboration protocol (HAI-CP)[1]. We define an HAI-CP as "the instance of a process schema that stipulates the use of AI tools by competent practitioners to perform a certain task or do a certain job".

The expression "use of AI" is actually more complex than it looks, as it stands for any configuration of learning and interaction parameters that can be combined together to create a single protocol. These parameters regard at least 5 dimensions (acronymized as AFOOT):

- **Affordance**, that is what the AI system affords in terms of functionalities and task automation (e.g., case retrieval, case comparison, case classification, decision justification);
- **Fit**, that is how the AI system fits into the existing work practice, e.g., in terms of order of presentation and or degree of automation;
- **Optimization**, that is what the AI system is optimized for in the learning phase (e.g., accuracy, calibration, utility, complex/simple cases);
- **Output**, that is what the AI system produces as content, e.g., single classes, confidence scores, list of categories, textual or visual explanations);
- **Target**, that is who is the intended target user in terms of profiling data as e.g., expertise, experience, job title, role).

Adopting the concept of HAI-CP helps considering all of these dimensions (and possibly others) to make each and every AI intervention tailored for a specific work setting or, even, situation (e.g., if a situation of emergency is detected or the user is recognized as associated with a specific role or experience, the most suitable protocol could be adopted). In particular, we use this concept in two explorative user studies in which we focus on the effects of AI and XAI support on diagnostic decision making, and of other process options that have yet to be examined in the specialized literature. We assess solutions differing by two dimensions, namely Fit and Output: 1) the effect of AI support presentation order to examine whether to show the AI's recommendation before any human assessment or only after the human provided their initial assessment can make a difference; 2) the effect of presenting an explanation enriching the AI's advice. The first comparison is motivated by the relevant body of work about priming and framing effects in decision making [48, 59]; while the second comparison is

[1] A collaboration protocol is a specific version of the more general concept of interaction protocol. Although adopting the term collaboration is not a neutral choice (no terminological choice really is), we also believe that it is opportune to adopt a term that specifically concerns "work settings, that is, [work practice] under conditions of severe constraints" [57]. In so doing, we recognize that the concept of interaction is necessarily broader and capable to include any informal, entertainment or ludic settings, and, more generally, information and knowledge retrieval activities that are not necessarily associated with a formal task or with tasks mutually associated with other tasks in the context of more complex and articulated processes.

motivated by the conjecture that AI support should be tailored to the users' expertise or expectations [5, 49].

Our main aim is then to present the results of an explorative analysis by which we compare two second-opinion classes of HAI-CPs, the AI-first and the human-first configurations: we refer, in the title, to the AI systems involved in these configurations as *rams* and *hounds*, respectively (as a suggestive animal metaphor). These configurations have previously been studied in the literature concerned with the detection and study of cognitive biases in decision making arising due to interaction with AI support systems, especially in crowdsourcing scenarios [9, 32], as well as in studies investigating the cognitive effect of explanations in XAI [7, 66] . Compared with these previous study, and differing from a crowdsourcing setting, our study focuses on subject-matter experts (i.e., medical doctors), who also have familiarity with computerized support systems (e.g., in computer aided diagnosis). Furthermore, compared with previous studies, our study is the first to decouple the provisioning of AI classification and its explanation, in order to evaluate the separate impact of these forms of support. The above mentioned configurations and protocols will be evaluated in two diagnostic settings. These are, respectively, knee lesion MRI interpretation and ECG reading (see Figures 1 and 2, respectively, in which each activity sequence, or scenario, depicts a protocol represent in Business Process Modeling Notation (BPMN)). We assess any differences in overall effectiveness in the different HAI-CPs, which can then inform adoption policies of AI-based decision support systems in real practice, and thus an evidence-based design of this class of support [2].

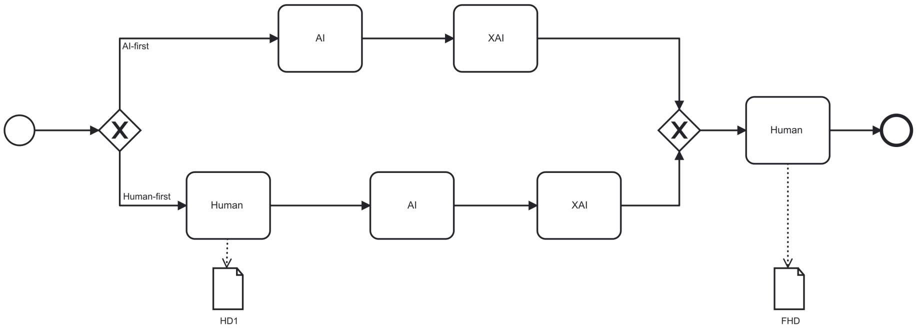


Figure 1: A BPMN diagram representing the H-AI collaboration protocols of the MRI reading study (each sequence is a protocol). HD1 denotes the first human decision and FHD the final human decision that closes the protocols, which is the one written in the final report. Four HAI-CPs can be instantiated from the BPMN diagram: AI-FHD, AI-XAI-FHD, HD1-AI-FHD, HD1-AI-XAI-FHD (all of which may involve either a Novice or an Expert reader).

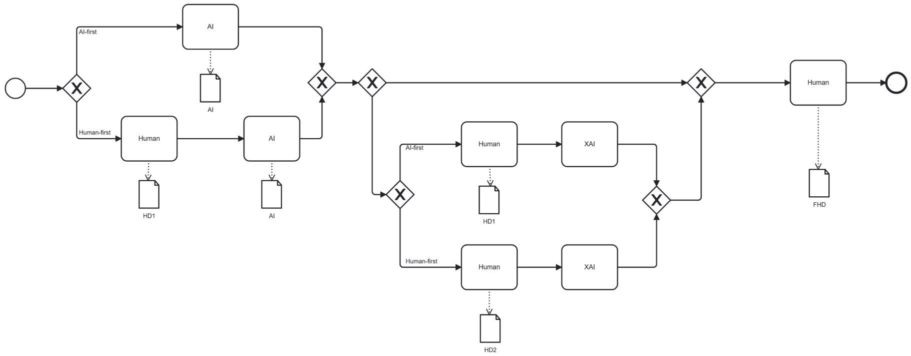


Figure 2: A BPMN diagram representing the H-AI collaboration protocols of the ECG reading study (each sequence is a protocol). HD1 denotes the first human decision , HD2 the second human decision and FHD the final human decision that closes the protocols, which is the one written in the final report. Eight HAI-CPs can be instantiated from the BPMN diagram: HD1-AI-FHD, HD1-AI-HD2-XAI-FHD, AI-FHD, AI-HD1-XAI-FHD (all of which may involve either a Novice or an Expert reader).

2. Methods

2.1. Experimental Design and Data Collection

2.1.1. Knee MRI study

In the MRI reading study, we involved 12 board-certified radiologists, from hospitals and healthcare centers throughout Italy, with different levels of expertise (8 with higher expertise, or subspecialists, and 4 with lower expertise, or specialists). The radiologists were asked to report their best diagnoses about 240 knee Magnetic Resonance Imaging (RMI) exams, which we previously extracted from the MRNet dataset[2], with the support of an AI system. In particular, for each of the 240 cases, the radiologists were invited to assess whether the case presented a ligament or knee abnormality, or neither. They received advice from a simulated AI system whose average accuracy on the cases was 80%.

The study was structured as a factorial design with two factors: presentation order (human-first vs AI-first) and availability of explanations (yes vs no). Both factors were within-subject. Comparisons based on raters' expertise were between-subjects. More in detail, for each case, the radiologists were asked to provide their diagnoses on the selected cases, and received the AI recommendations according to two different collaboration protocols: for one half of the cases (120 out of 240), the AI provided its diagnosis before the case was interpreted by the radiologist (AI-first protocols) on the same page as the radiologist could provide their diagnosis. For the other half, the radiologists had to first propose a tentative diagnosis, which was recorded, and only then they were shown the diagnostic advice proposed by the AI system. They could then either confirm their initial diagnosis or change it in light of the machine's advice (that is

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

human-first protocols). In addition, for each protocol, in half of the cases (i.e., 60 from each protocol) the diagnosis provided by the AI systems was accompanied by XAI decisional support in the form of an activation map (generated through the GradCAM method [58]). This highlighted the regions of the MRI that were considered to be most relevant by the AI system. The other half of the cases (i.e., 60 from each protocol) had no XAI support. The complete list of HAI-CPs for the knee MRI study is reported in the top half of Table 3. We also collected the initial diagnoses of each respondent in the human-first group (see HD1 in Table 3), which were used to evaluate their basal, unsupported accuracy compared to the MRNet ground truth.

The experiment was conducted using an online multi-page questionnaire, implemented on the LimeSurvey[3] platform (version 3.23). The respondents were shown the 240 cases in random order. For each case, the questionnaire showed three views of the MRI (i.e., images on the axial, sagittal and coronals planes), and the items indicating whether the imaging presented any clinically-detectable abnormalities.

2.1.2. ECG Study

The ECG reading study involved 44 cardiology residents and specialists (25 residents, 19 specialists), from the Medicine School of the University Hospital of Siena (Italy). They annotated 20 ECG cases, previously selected by a cardiologist from a random set of cases extracted from the ECG Wave-Maven repository[4] based on their complexity characteristics. The study participants had to provide their diagnoses, both supported and not supported by a simulated AI system. The accuracy of the simulated AI was 70% (in terms of the ECG Wave-Maven gold standard)[5].

The study was structured as a factorial design with two factors: presentation order (human-first vs AI-first) and availability of explanations (yes vs no). The first factor was between-subjects, while the second factor was within-subject. Comparisons based on raters' expertise were between-subjects. More in detail, the ECG readers were randomly divided in two different groups, which were equivalent for expertise, to evaluate human-first and AI-first HAI-CPs. For each ECG case, the readers in the human-first group, after being shown the trace of the ECG together with a brief case description, had to first provide an initial diagnosis (in free text format). These respondents were then shown the diagnosis proposed by the AI. After being shown the AI support, the respondents could revise their initial diagnosis before being shown the textual explanation and having to provide their final diagnosis. By contrast, the readers in the AI-first group were shown the AI-proposed diagnosis together with the ECG trace and case description, and only afterwards they were asked to provide

[3] https://www.limesurvey.org/

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

[5] This rate was considered appropriate because in a previous study [54] we observed a slightly lower average accuracy in a similar population of readers.

their own diagnosis. Finally, they were shown the textual explanations, and asked whether they wanted to revise their initial diagnoses. To avoid negative priming, the first five cases of the questionnaire shown to the ECG readers were all associated with a correct diagnosis and a correct explanation from the AI support. Although the participants had been told that the explanations were automatically generated by the AI system, like the diagnostic advice, these had been prepared by a cardiologist. 40% of the explanations were prepared to be incorrect or not completely pertinent to the cases. The complete list of HAI-CPS for the ECG study is reported in the bottom half of Table 3. We also collected the initial diagnoses of each respondent in the human-first group (see HD1 in Table 3), which were used to evaluate their basal, unsupported accuracy compared to the ECG Wave-Maven ground truth.

The experiment was delivered using a web-based questionnaire set up through the LimeSurvey platform (version 3.23), to which the readers had been individually invited by email.

2.2. Reliance patterns and Diagram

To analyze the effects of AI support on human decision-making, in Table 1 we define variables corresponding to AI-supported human judgment reliance patterns[6]. The names of the reliance patterns are inspired by the framework previously discussed in [41]. Intuitively, over-reliance refers to trusting the machine even when this is against one's judgment; self-reliance is not trusting the machine when this is against one's judgment; under-reliance is trusting the machine so little as to change one's mind if the machine agrees with one's judgment.

Table 1: Definition of all possible decision and reliance patterns between human decision makers and their AI. In the first three columns, 0 denotes an incorrect decision, and 1 a correct decision. We associate the attitude towards the AI in each possible decision pattern (in terms of trust [41]), which leads to either accepting or discarding the AI'advice, and the main related cognitive biases.

Human judgment	AI support	Final decision	Reliance pattern	Biases and Effects
0	0	0	detrimental reliance	automation complacency
0	0	1	beneficial under-reliance	extreme algorithmic aversion
0	1	0	detrimental self-reliance	conservatism bias
0	1	1	beneficial over-reliance	algorithm appreciation
1	0	0	detrimental over-reliance	automation bias
1	0	1	beneficial self-reliance	algorithmic aversion
1	1	0	detrimental under-reliance	extreme algorithmic aversion
1	1	1	beneficial reliance	confirmation bias (on later cases)

Based on the above mentioned variables we define four metrics aimed at evaluating the performance of HAI-CPs in terms of reliance patterns and related biases, as shown in Table 2. RBT and RBMT are equivalent to the relative positive self-reliance (RSR) and relative positive AI-reliance (RAIR), which are

[6] We acknowledge that, at the macro level, trust is a complex concept [60], associated with the trustee's reputation and the trustor's risk propensity. reliance patterns refer to expressions of trust at a micro level, that is, the attitude that leads the trustor to accept (or reject) the trustee's suggestion at the individual decision level.

introduced and discussed in [55] as metrics of automation reliance, i.e., the behavioral effect of trust.

Table 2: Metrics to evaluate the effects and impact of an AI decision support system on decision making.

Abbr.	Metric	Formula
RBOD	Relative Beneficial Over-distrust	beneficial under-reliance / (beneficial under-reliance + detrimental reliance)
RBT	Relative Beneficial Trust	beneficial reliance / (beneficial reliance + detrimental under-reliance)
RBOT	Relative Beneficial Over-trust	beneficial over-reliance / (beneficial over-reliance + detrimental self-reliance)
RBD	Relative Beneficial Distrust	beneficial self-reliance/ (beneficial self-reliance + detrimental over-reliance)

The four metrics can be visualized in the Reliance Pattern Diagram. In Figure 3 we report an empty Reliance Pattern Diagram, through which we can evaluate different types of reliance patterns (according to Table 2) and the corresponding biases facilitated by a HAI-CP. Code to generate the Reliance Pattern Diagram can be accessed on GitHub at `https://github.com/AndreaCampagner/qualiMLpy/blob/master/viz/trust_diagram.py`, while a web service to generate the Reliance Pattern Diagram from data is available at `https://mudilab.github.io/evaluate-human-ai-interaction/`.

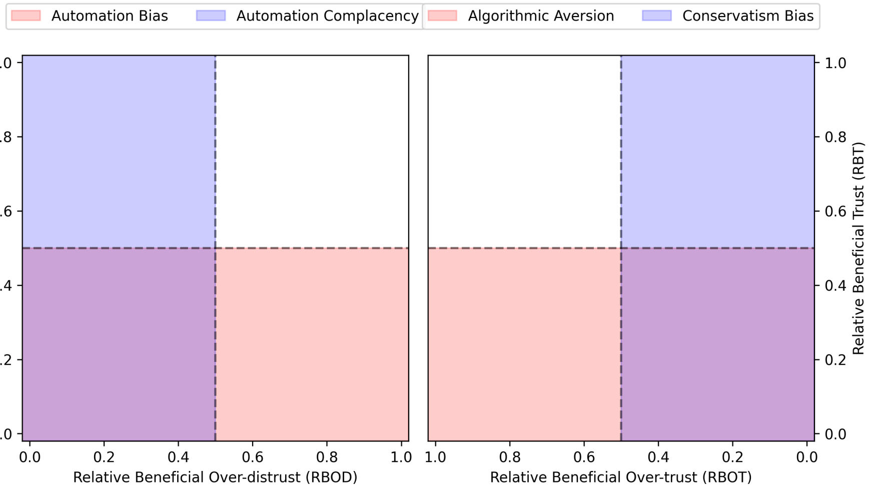


Figure 3: An empty Reliance Pattern Diagram. The diagram in the left panel depicts the distrust-related patterns, which are the relative beneficial over-distrust (RBOD) against the relative beneficial distrust (RBD). Low RBD values correspond to an increased risk of automation bias, while low RBOD values correspond to an increased risk of automation complacency. The diagram in the right panel depicts the trust-related patterns, which are the relative beneficial over-trust (RBOT) against the relative beneficial trust (RBT): low RBT values correspond to an increased risk of (extreme) algorithmic aversion, while low RBOT values correspond to an increased risk of conservatism bias.

2.3. Statistical Analysis

Based on the two previously described case studies, we were interested in investigating the following hypotheses:

1. Do the main factors under analysis (i.e., human-first vs AI-first protocol, provisioning of explanations, readers' expertise) have any effect on accuracy?
2. Do the individual HAI-CPs (i.e. the interactions between presentation order and provisioning or not of explanations) under analysis have any effect on accuracy?
3. Did the provisioning of AI support and/or explanations allow the readers to improve their accuracy, in comparison with their unsupported accuracy, or to surpass the accuracy of the AI support?
4. Were there any effects (in terms of increased risk of automation-related biases) of computer support on reliance patterns and behavioural correlates of trust?

In both the knee MRI and ECG studies, the statistical analysis of the above mentioned research questions was performed by means of a statistical hypothesis testing approach. In all cases, the units of analysis were the per-rater accuracies. We considered two levels of analysis: the main factors, aggregated level (i.e., human-first vs AI-first protocols, raters' expertise, availability of explanations) and the HAI-CPs, individual conditions level (i.e., the pairwise comparison of all considered HAI-CPs, which in turn consist of all the pairwise interactions between presentation order and availability of explanations). We decided to focus on these two levels of analyses to investigate both the effects of the main factors of variation in a HAI-CP as well as the effects of each specific HAI-CP. We separately also considered the effect of readers' expertise on the HAI-CPs.

In regard to the adopted statistical procedures, since data was not normally distributed, we only considered non-parametric tests. We applied the Mann-Whitney U test [47] for between-subjects comparisons, and Wilcoxon signed rank test [68] for within-subject comparisons. In both bases, the tests were selected as non-parametric alternatives to t-test procedures, since as mentioned before the collected data was not normally distributed. In all cases, to control the false discovery rate (i.e., Type 1 errors) due to multiple testing, p-values were adjusted by means of the Benjamini-Hochberg procedure. The statistical significance of the findings was taken at the 95% confidence level (that is, $\alpha = .05$). In all cases, effect sizes were computed using the Rank Biserial Correlation [17][7].

3. Results

3.1. Knee MRI Study

The distribution of accuracy for the considered HAI-CP and the baseline HD1, are reported in Table 3 and illustrated in Figure 4a.

We first considered the main factors of analysis (see Figure 4a). The difference in accuracy due to the presentation order (human-first vs AI-first) was

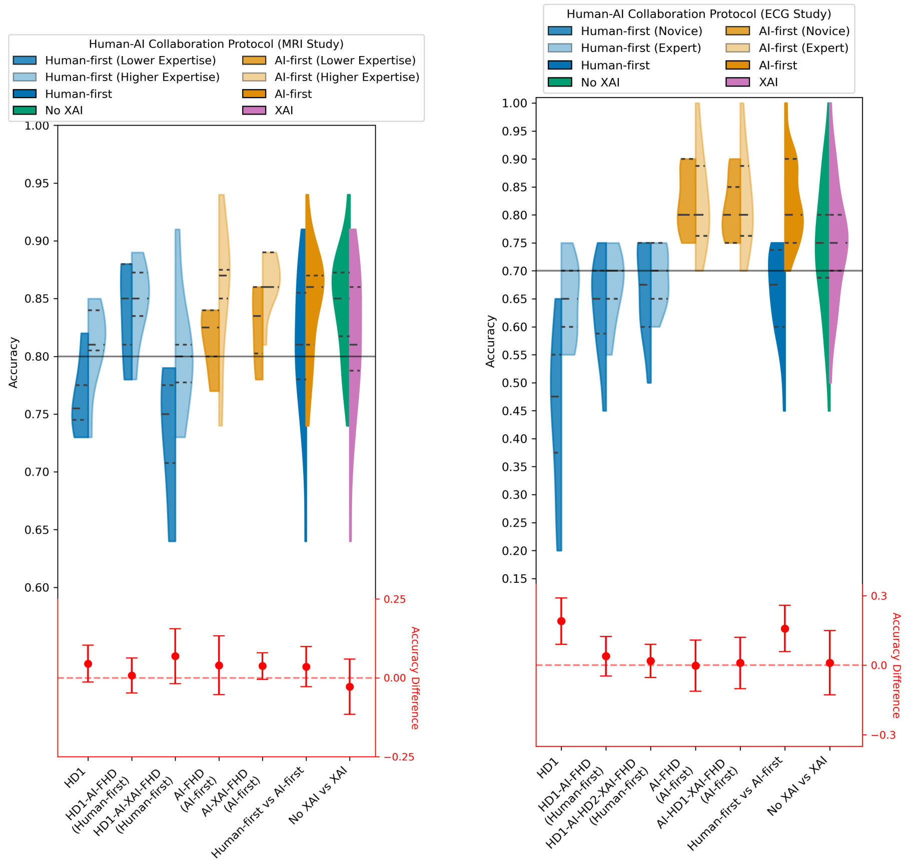
[7] Effect sizes were interpreted according to the following scale: RBC $<$ .05: negligible; .05 $\leq$ RBC $<$ .1: negligible-to-small; .1 $\leq$ RBC $<$ .2: small; .2 $\leq$ RBC $<$ .3: small-to-medium; .3 $\leq$ RBC $<$ .5: medium; .5 $\leq$ RBC $<$ .8: medium-to-large; RBC $\geq$ .8: large.

Table 3: The results for each protocol and for the baseline HD1, from the MRI and ECG studies. Accuracy is defined as the ratio between the number of correct diagnoses and the total number of cases (with a 95% confidence interval).

Study	Collaboration Protocol	Description	Accuracy
MRI	HD1 (Lower Expertise)	No support	77% ± 3%
	HD1 (Higher Expertise)	No support	81% ± 3%
	HD1-AI-FHD (Lower Expertise)	AI support (Human-first)	84% ± 4%
	HD1-AI-FHD (Higher Expertise)	AI support (Human-first)	85% ± 3%
	HD1-AI-XAI-FHD (Lower Expertise)	AI+XAI support (Human-first)	73% ± 5%
	HD1-AI-XAI-FHD (Higher Expertise)	AI+XAI supoprt (Human-first)	80% ± 5%
	AI-FHD (Lower Expertise)	AI support (AI-first)	81% ± 3%
	AI-FHD (Higher Expertise)	AI support (AI-first)	85% ± 5%
	AI-XAI-FHD (Lower Expertise)	AI+XAI support (AI-first)	83% ± 3%
	AI-XAI-FHD (Higher Expertise)	AI+XAI support (AI-first)	86% ± 2%
ECG	HD1 (Novice)	No support	45% ± 8%
	HD1 (Expert)	No support	66% ± 5%
	HD1-AI-FHD (Novice)	AI support (Human-first)	63% ± 5%
	HD1-AI-FHD (Expert)	AI support (Human-first)	68% ± 5%
	HD1-AI-HD2-XAI-FHD (Novice)	AI+XAI support (Human-first)	67% ± 5%
	HD1-AI-HD2-XAI-FHD (Expert)	AI+XAI support (Human-first)	69% ± 4%
	AI-FHD (Novice)	AI support (AI-first)	83% ± 3%
	AI-FHD (Expert)	AI support (AI-first)	82% ± 5%
	AI-HD2-XAI-FHD (Novice)	AI+XAI support (AI-first)	82% ± 3%
	AI-HD2-XAI-FHD (Expert)	AI+XAI support (AI-first)	82% ± 5%

significant and associated with a medium effect size (p-value: .033, RBC: .46). On average, AI-first protocols reported an higher accuracy than human-first ones. By contrast, the difference in accuracy due to the availability of explanations was not significant but was nonetheless associated with a medium-to-large effect size (p-value: .095, RBC: .62).

We then considered the individual conditions (i.e., the individual HAI-CPs). First, we observe that there was no significant difference between lower expertise and higher expertise readers for any of the considered protocols (HD1-AI-FHD: .931, HD1-AI-XAI-FHD: .071, AI-FHD: .141, AI-XAI-FHD: .100) or the baseline accuracies (HD1: .170). Nonetheless, the associated effect sizes were all medium-to-large (HD1: .53, HD1-AI-XAI-FHD: .69, AI-FHD: .56, AI-XAI-FHD: .59) except for the one associated with the HD1-AI-FHD HAI-CP (RBC: .06). For this reason, and also due to low test power associated with comparisons based on expertise, we did not consider readers' expertise in subsequent analyses. The p-values and effect sizes for the pairwise comparisons among HAI-CPs are reported in Figure 5. Due to low sample sizes, none of the pairwise comparisons was associated with a significant difference. Nonetheless, the effect sizes for the comparison between the unsupported baseline and the AI-first protocols, as well as that for the AI-supported human-first protocol (i.e. HD1-AI-FHD), were medium-to-large or large. Similarly, the XAI-supported human-first protocol (i.e. HD1-AI-XAI-FHD) reported lower performance than all other considered HAI-CPs, and all such comparisons, though not significant, were associated with a medium or large effect size. By contrast, all the other differences were not statistically significant, and the associated effect sizes were



(a)

(b)

Figure 4: Distribution of the readers' accuracy for the considered HAI-CPs in the knee MRI study (left, 4a) and in the ECG study (right, 4b). Top: Violinplots of the accuracy distributions, with dashed lines denoting the median and quartiles of the distributions, while the solid line denotes the accuracy of the AI. Bottom: Pointplot of the differences in accuracy between the 2 levels of readers' expertise (computed for each protocol p as the average accuracy for protocol p for higher expertise readers, minus the average accuracy for protocol p for lower experise readers), as well as for the comparison between human-first and AI-first and XAI vs no XAI protocols; dots represented the average difference while bars represent the 95% confidence interval of the difference.

negligible-to-small.

XAI support provisioning had a small positive effect for AI-first protocols. Indeed, the XAI-supported AI-first protocol reported a higher accuracy, as depicted in Figure 6a. However, the difference was not significant (see Figure 5). In contrast, XAI support had a negative effect for the human-first protocols, as shown in Figure 6b.

As shown in Figures 4a and Appendix A.1, the AI support (see protocols

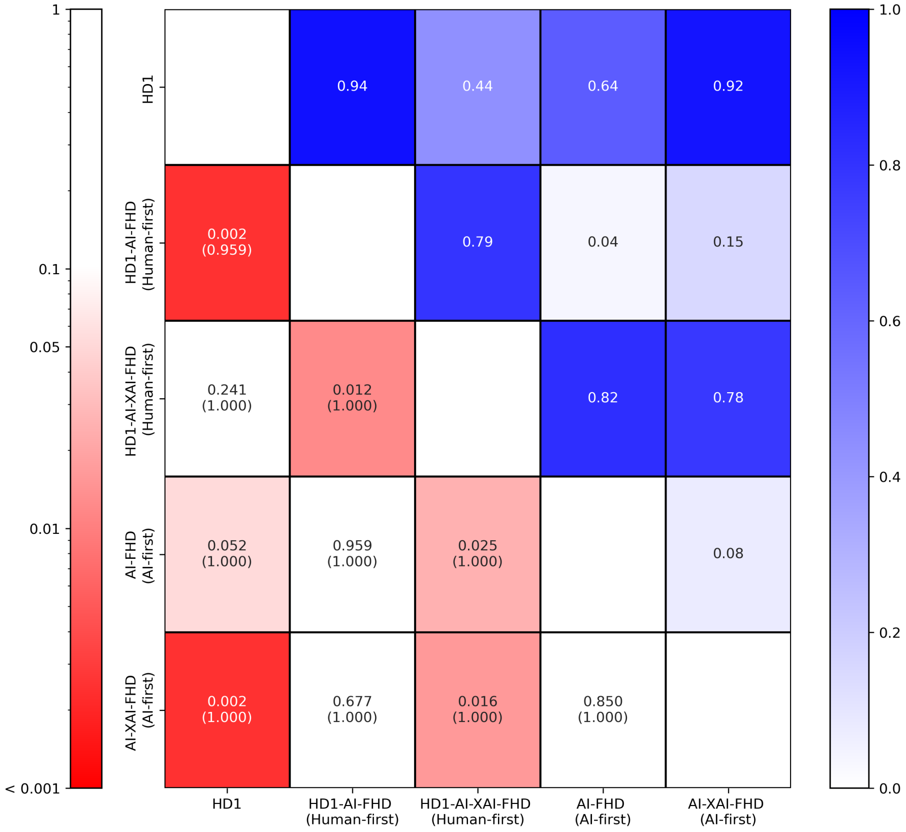


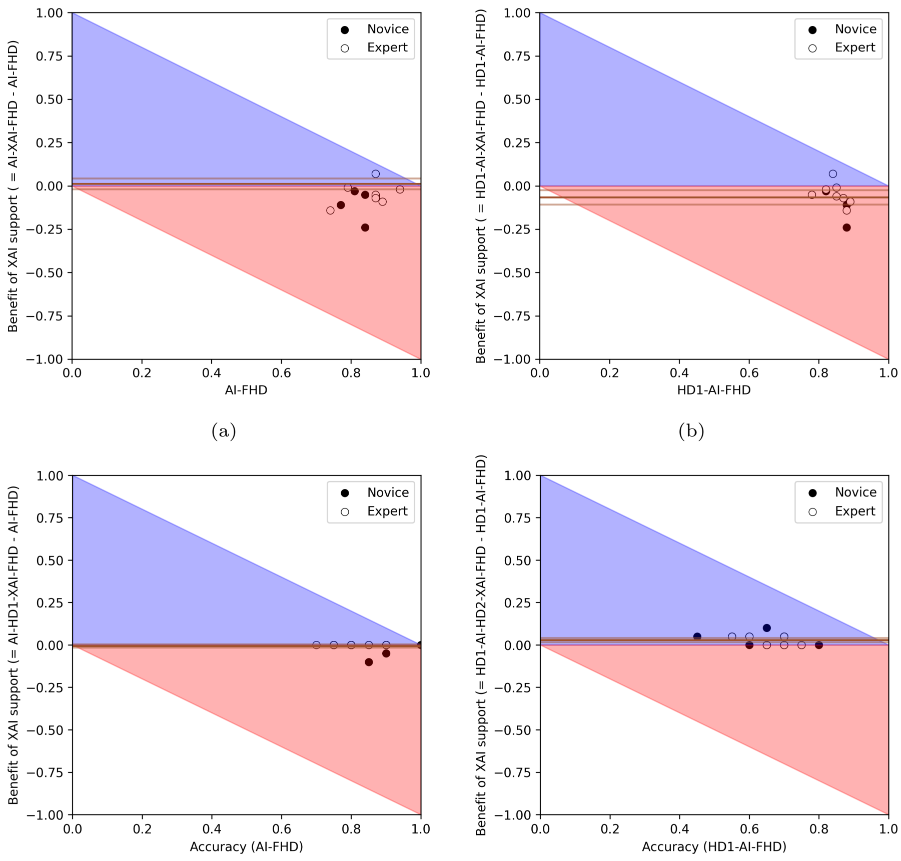
Figure 5: Comparison of the accuracy of the HAI-CPs (with no stratification for type of AI support), in the knee MRI study. P-values and adjusted p-values (in brackets) for the comparisons are depicted below the diagonal, with p-values lower than .1 colored red: the darker the hue, the lower the p-value. Effect sizes for the comparisons are depicted above the diagonal, colored blue: the darker the hue, the higher the effect size.

HD1-AI-FHD and AI-FHD) had a beneficial effect compared to the baseline HD1. Remarkably, the AI support had a significant beneficial effect also compared with the performance of the AI alone (p-value: .001), see Figure 4a.

The results of the reliance patterns-based analysis are reported in Figure 7. The XAI-supported human-first protocol (i.e. HD1-AI-XAI-FHD) reported significantly lower relative beneficial distrust and relative beneficial over-distrust (i.e., higher automation bias and automation complacency) than the human-first protocol without explanation support (i.e. HD1-AI-FHD).

3.2. ECG Study

We collected a total of 1352 responses from the 44 ECG readers, of which 21 considered the human-first protocols and the remaining 23 the AI-first protocols.



(a)

(b)

(c)

(d)

Figure 6: Benefit diagrams for the HAIICPs in the knee MRI study (top, 6a and 6b) and the ECG study (bottom, 6c and 6d), showing the benefit of providing XAI support versus not providing it for both AI-first (left, 6a and 6c) and human-first protocols (right, 6b and 6d) . The dots represent the accuracies of the readers, and the brown lines the average difference in accuracy between the two protocols, along with the corresponding 95% confidence interval. The blue region denotes an improvement in error rates, while the red region denotes a worsening.

The distribution of accuracy of the HAI-CPs and the baseline HD1, are reported in Table 3, and illustrated in Figure 4b.

We first considered the main factors of analysis (see Figure 4b). The difference in accuracy due to the presentation order (human-first vs AI-first) was statistically significant and was associated with a large effect size (p-value: $<$ .001, RBC: .90). AI-first protocols reported a significantly higher accuracy than human-first ones. By contrast, the difference in accuracy due to the availability of explanations was not significant and was associated with a medium-to-large effect size (p-value: .081, RBC: .55).

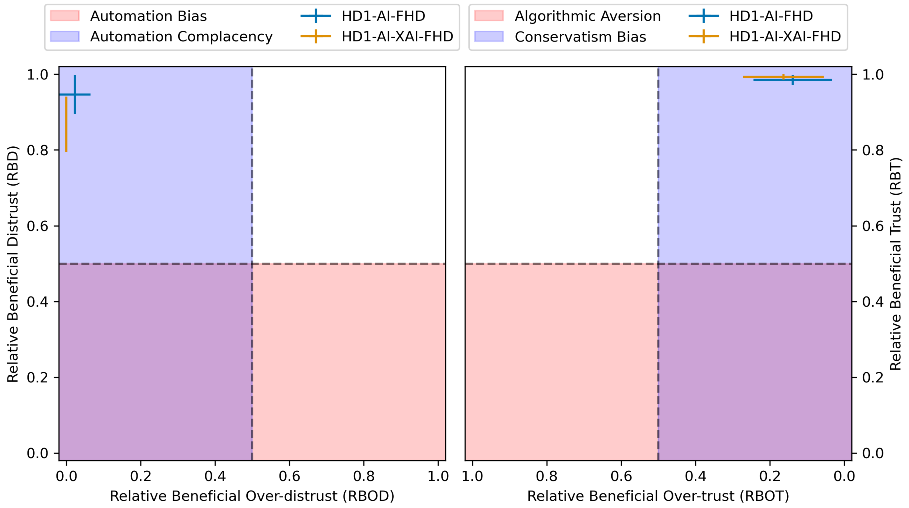


Figure 7: Reliance Pattern Diagram, showing the reliance patterns and potential effect of AI and XAI support for hound protocols in the knee MRI study.

We then considered the individual conditions (i.e., the individual HAI-CPs). First, we observe that there was no significant difference between novice and expert readers for the considered protocols (HD1-AI-FHD, pvalue: .665; HD1-AI-HD2-XAI-FHD, p-value: .965; AI-FHD, p-value: .490; AI-HD1-XAI-FHD, p-value: .782), which were also associated with small or negligible-to-small effect sizes (HD1-AI-FHD, RBC: .13; HD1-AI-HD2-XAI-FHD, RBC: .02; AI-FHD, RBC: .18; AI-HD1-XAI-FHD, RBC: .08). By contrast, the unsupported accuracy levels (HD1) were significantly different (p-value: .009) and associated with a medium-to-large effect size (RBC: .74). For this reason, and also due to low test power associated with comparisons based on expertise, we considered raters' expertise only for the unsupported HD1 baseline in subsequent analyses. The p-values and effect sizes for the pairwise comparison among HAI-CPs are reported in Figure 8. Novice readers, as well as expert ones (supported by any protocol) reported a significantly higher accuracy than that reported by the unsupported novice readers. By contrast, the expert readers were able to significantly improve their baseline performance only when support by AI-first protocols. These latter protocols, i.e. the two AI-first protocols, were significantly better than all other protocols as well as better than the unsupported readers, with no distinction for novice and expert ones. By contrast, the two human-first protocols were not associated with a significantly higher accuracy than that reported by the unsupported expert readers.

XAI support had a negligible effect for AI-first protocols, as depicted in Figures 6c and 8. Similarly, XAI support had a small positive, but not significant, effect for human-first protocols, though in this case the effect size was small-to-medium, see Figures 6d and 8.

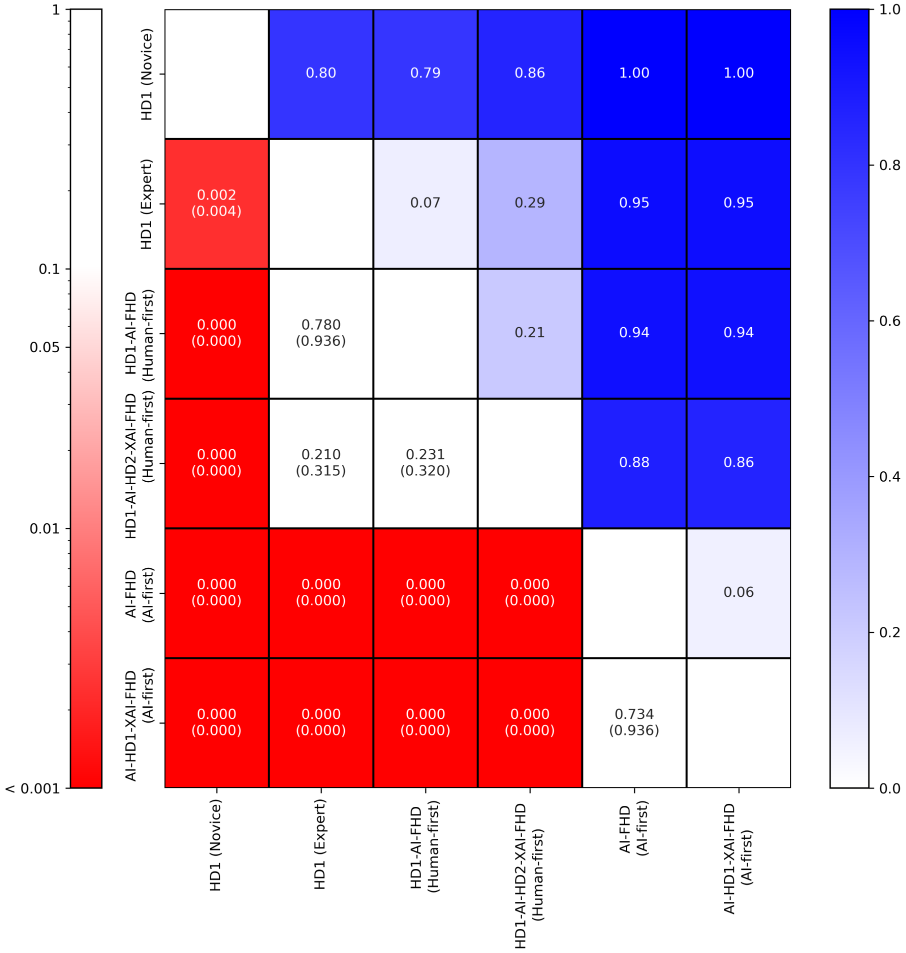


Figure 8: Comparison of the accuracy of the HAI-CPs (with no stratification for type of AI support), in the ECG study. P-values and adjusted p-values (in brackets) for the comparisons are depicted below the diagonal, with p-values lower than .1 colored red: the darker the hue, the lower the p-value. Effect sizes for the comparisons are depicted above the diagonal, colored blue: the darker the hue, the higher the effect size.

As shown in Figures 4b, 8 and Appendix A.2, the provisioning of AI support had a significant beneficial effect compared to the baseline HD1. Remarkably, the provisioning of AI support had a significant beneficial effect also compared with the performance of the AI alone (p-value: $< .001$), and especially so for AI-first protocols, see Figure 4b.

The results of the reliance patterns-based analysis are reported in Figure 9. The XAI-support human-first protocol (i.e. HD1-AI-XAI-FHD) reported significantly lower relative beneficial distrust and higher relative beneficial over-trust (i.e., higher automation bias and lower conservatism bias) than the human-first protocol without explanation support (i.e. HD1-AI-FHD HAI-CP).

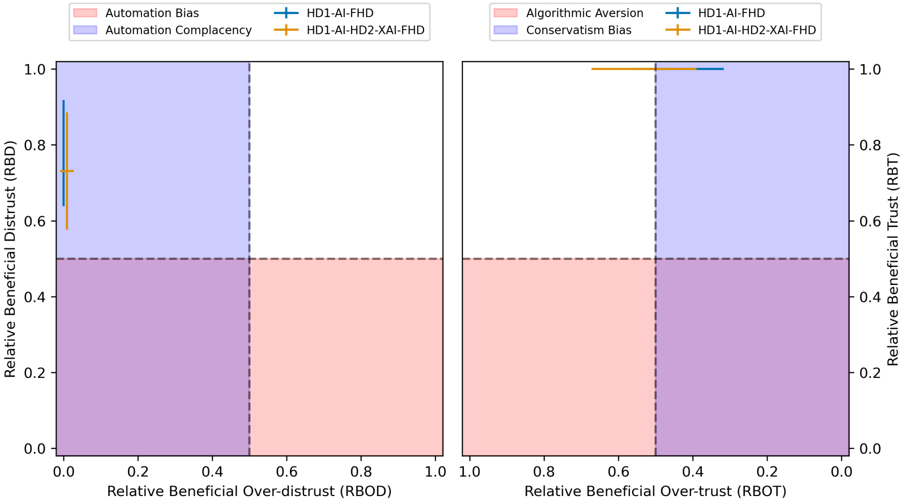


Figure 9: Reliance Pattern Diagram, showing the reliance patterns and potential effect of AI and XAI support for hound protocols in the ECG study.

4. Discussion

In this article, the concept of human-AI collaboration protocols (HAI-CP) is proposed to design and evaluate different ways in which users and their AI tools can interact to have their work done and to make better decisions.

The studies presented in this article are aimed at comparing the effectiveness of HAI-CPs that differ in terms of specific options chosen for their originality or relevance. First, we contribute to the emerging body of works (e.g., [8, 53]) which examine bias in AI and XAI support, and focus on the priming effect [11] of AI advice. To this aim, we compared protocols differing in terms of whether physicians must express a diagnostic judgment before being influenced (for better or worse) by the machine's advice, or not (human-first and AI-first protocols, respectively). To our knowledge, our studies are the first ones that aim to contribute on this issue in regard to clinical tasks in which the involved users were subject-matter experts: indeed, even if other studies compared AI-first and human-first protocols (as mentioned in the introduction), these latter studies focused on crowdsourcing settings that only involved laypersons [9, 32]. We also compared protocols that embed some form of XAI support (i.e., activation maps or written explanations) with those that provide categorical support without any explanation. Some other studies in the literature focus on this latter issue [7]. For example, Alufaisan et al. [1] compared human decision accuracy without AI, with an AI decision support but no XAI, and with both AI and XAI support. They involved a cohort of 300 lay volunteers and found significant evidence that AI-based support improves user decision accuracy, but provide no conclusive evidence that XAI has any meaningful impact on accuracy. Bansal et al. [4] observed that, while AI support improves user's accuracy, this was not increased by explanations. In contrast, explanations increased the chance that humans accept the AI's recommendation, regardless of its correctness. Similarly, Paleja et al. [50] studied human-AI teaming in the context of games and evaluated the impact of XAI support on users with different levels of expertise. The authors showed that while XAI had a positive impact on beginner-level users, it had a largely negative effect on expert-level users. As mentioned in the introduction, an interesting design aspect of our studies regards decoupling the effect due to the AI advice and the effect due to the related explanation: to our knowledge, this decoupling has never been considered before.

In regard to our results, we found a significant effect of AI support in improving human decisions, which was observed in both studies and for different HAI-CPs and was particularly evident for the lower-expertise users (see Figures 4b and Appendix A.2). These results are not particularly surprising, as they provide additional confirmation to the growing body of literature which focuses on the benefits of AI support in clinical decision making [1, 29, 33, 39, 42, 64]. However, we observe two particularly significant points.

First, by focusing on human-first protocols in the ECG study, we note that AI support helped the less-expert users align their diagnostic performance with those of the experts [52], even though the basal performance of the two groups was significantly different (see Figure 4b). This effect can be attributed to lower algorithmic aversion (or higher trust) in physicians with less expertise compared to the experts, which was also observed in previous studies [20, 29], and especially in [11] where novices expressed a lower 'prejudice against the machine' in their diagnostic choices. Indeed, the improvement due to AI support was significantly lower for the expert readers and not strong enough to enable them to surpass or even equal the performance of AI support alone (see Figure 4b).

Second, the performance improvement observed in both studies occurred despite the relatively low accuracy of the AI support (70% accuracy in the ECG study, 80% accuracy in the knee MRI study). In the knee MRI study, AI-supported protocols had a diagnostic accuracy that was significantly higher than the accuracy of both the unaided readers and AI support alone, irrespective of the order of presentation of its advice (i.e., human-first vs AI-first). A similar effect was also highlighted in the ECG study for the AI-first protocols. We believe this last result to be particularly remarkable in light of the recent interest about the so-called *complementarity effect* in the human-AI interaction literature. Complementarity refers to the alleged phenomenon by which human-AI teams could achieve better accuracy than both humans and AI alone. While this effect has been widely investigated [3, 35, 40, 71], few studies have found convincing evidence for its existence. In this sense, we believe that our studies represent an important first positive step in this direction since, as mentioned previously, we highlighted a synergistic interaction between human readers and AI in both of the considered user studies. Furthermore, our results, and especially so those for the MRI case study, also support previous findings reported in [10], according to which good hybrid-team performance can also be achieved with relatively low-accuracy decision support, i.e., with systems exhibiting lower accuracy than the average accuracy of the physicians. We conjecture this to be due to humans' tendency to improve their performance as a consequence of a trigger to reflection when leveraging other interpretations, especially when they consider themselves to be the only responsible for the decision. We discuss this in more detail when we consider the AI-first vs human-first comparison.

By contrast, our findings about the effect of XAI support were more controversial. While XAI support had a small positive effect on diagnostic accuracy in the AI-first protocols and for expert readers, in the human-first protocols we observed a small beneficial effect in the ECG study but a relevant detrimental effect in the knee MRI study, and especially so for the less expert readers. A possible explanation for this finding stems from the effect previously investigated in [56], indicating that explanations increase trust in AI and may then result in an increase of over-reliance and automation bias. These controversial findings are aligned with the previous work mentioned above [1, 4, 50] and confirm arguments that, despite its intuitive appeal [15], explainability for patient-level decision making is unlikely to maximise decision accuracy [30, 51]. Indeed, explanations, by increasing the persuasiveness of AI support [22], may even have negative consequences by inducing a false sense of confidence (confirmation bias, fixation), trust misplacement or automation bias [23, 24, 26, 56]. As an example of this issue, we note that, as emerged from the reliance patterns-based analysis presented in Figure 7, while the radiologists involved in the MRI study under-relied on the support independently of whether XAI support was available or not, the availability of explanations increased the chance that the advice of the AI was trusted by the clinicians. Furthermore, this increase in trust due to the XAI support was observed irrespective of whether the AI advice was right or wrong. Consequently, the XAI-supported protocol reported an increased risk of automation bias and automation complacency than the protocol without XAI support. A similar finding was observed also for the ECG study, as presented in Figure 9: indeed, also in this latter case the availability of explanations significantly increased the risk of automation bias. In this sense, the reliance patterns-based analysis allows to highlight a potentially harmful effect of explanations in terms of emergence of biases due to increased trust towards AI suggestions. We believe that this effect should be further investigated to better understand the value of explanations, as well as the best ways to provide explanatory support to users. Indeed, even though our results suggest that explanations could not be useful for improving accuracy, their effect on users' confidence and trust should be further investigated.

The most interesting and partly counter-intuitive finding that we derive from our two studies is that the order of presentation of the AI support has a significant effect on the diagnostic accuracy, as conjectured in previous research [16, 32]. Remarkably, AI-first protocols were found to be significantly more effective (that is more accurate) than the human-first ones. That is, by recalling the metaphor presented in the title, using AI as a ram is better than using it as a hound. This result was not expected, because some sort of framing effect (such as priming) from presenting the AI advice before an idea of the case is obtained independent of it could be conjectured to be likely, as was also shown in previous studies [9, 32]. The opposite effect was instead observed: allowing the readers to initially see the AI advice helped them form a better understanding of the cases, thus achieving a diagnostic accuracy that was not only higher than the baseline human accuracy (that is, the accuracy of the unassisted HD1 protocol) but also higher than the AI alone. A similar effect was not generally observed for human-first protocols: in both studies these latter protocols reported higher accuracy than those of the human readers alone, but only in the knee MRI study the AI supported protocol (with no XAI support) was more accurate than the AI alone. We conjecture two possible explanations for this effect. First, this could be traced back to some form of conservatism bias in the physicians, which is activated in human-first protocols when the AI opposes their initial assessment, or to some type of fixation [45] or automation complacency, when the AI corroborates their initial interpretation (see also Figure 7) . Both effects could then induce a stronger anchoring bias, and specifically a form of belief perseverance [65], even when the AI is correct and differs from the human interpretation. Similarly, a competitive response, or the need to exploit the suggestion, could be activated when readers are initially given the case and another interpretation (as in the AI-first protocols), which may contribute to the better performance of what is perceived to be a team decision. Second, this could be associated with the System 1-System 2 metaphor popularized by Kahneman [43]. Human-first protocols could be associated with a System 1 (i.e. gut feelings-based [31]) response from the users, whereas the subsequent interaction with the AI is not sufficient for the users to provide a better rationalization of the cases at hand. On the other hand, we conjecture that AI-first protocols could be associated with a sort of *replacement effect* by which the humans' gut feeling is replaced by the AI support which is implicitly interpreted as a sort of a mediated System 1 response. This then allows the human users to provide a more rational (i.e. System 2-like) interpretation. Finally, concerning the difference with the results obtained in previous studies: we notice, as we did previously, that our study differs from the previous ones in the specialist literature in that our experiments involved subject-matter experts, i.e., clinicians, rather than laypersons. In both of the considered settings (ECG reading and MRI interpretation), clinicians have some familiarity with automated or computerized support systems: for example, in MRI interpretation, computer-aided diagnostic (CAD) systems have been in use since the early '90s [69], and in ECG reading the main vendors of ECG equipment also deliver simple CAD features even with the most economic devices. We believe that this difference in decision support familiarity, as well as re the expertise of the involved users for the considered tasks, could explain the results that we observed, and motivate further research in similar settings. In any case, we believe that the implementation of human-first (hound) protocols should still be encouraged, at least at the beginning of a digitization project, and regularly over small intervals of time, especially in light of the results observed above and derive from the reliance patterns-based analysis. The potential for any deskilling due to practices that rely on digital support

[19, 28], as well as the possible emergence of automation-related biases, can then be regularly assessed[8], and our findings could then be replicated in real-world conditions. The collection of the initial unsupported judgment also enables an evaluation of the *value* of the information provided by the AI, in terms of the ability to change human decisions and, for those decisions, to change outcomes for the better [14]. If we assume that outcomes can be improved by simply identifying the correct diagnosis, the AI systems used in the MRI and ECG studies induced a decision change in approximately 1 case out of 20 and 3 cases out of 20, respectively, with the increased number of decision changes in the ECG study mainly due to the novice readers. In both studies, a mistake occurred in just under half of the cases due to these changes, which would likely be avoided without the AI suggestion. This finding opposes Friedman's "Fundamental Theorem" of Informatics [27] (usually denoted as simply as '$H + C > H$'), which is an usually unstated assumption that the use of any computational technology (C) should leave humans (H) better off than not using it. Thus, although AI-first protocols yield better results, human-first protocols, supported by an analysis of reliance patterns, could enable long-term *technovigilance* [13] of the effects of automation on human decision performance.

Despite the relevance of the reported results, our study has some limitations, which stem from its exploratory nature. First, between-subjects comparisons in regard to readers' expertise (in both the MRI and ECG studies) were underpowered, due to the limited number of involved users. Despite this limitation, we believe that our results could provide some insight about the expected effect sizes for such comparisons and could then inform the power analysis of future studies pursuing statistical significance. Furthermore, even though the considered user studies were realistic, they were conducted in a serious game settings and not in a real-world scenario. Thus, future studies should evaluate our findings in clinical practice, to avoid potential laboratory effects [34]. As a final limitation, we note that even though one of the main aims of our work was to investigate some dimensions related to trust (of a human expert towards an AI system), we focused mainly on issues related to accuracy and explainability. Nonetheless, other relevant dimensions could be related to trust and reliance patterns, such as *robustness* [37]: indeed, robustness, reliability, replicability and contestability [51] (even more so than explainability) could be seen as essential components for trustworthiness and trust-building. Thus, we believe that future research should be devoted at exploring the effects of AI robustness (or lack of thereof) on human-AI interaction.

[8] Obtaining the exam readers' perceptions about their confidence in their final decisions and about the complexity of the case can inform comparisons of the confidence levels and error rates.

5. Conclusion

In summary, in this paper we compared various methods to include AI and XAI aids into diagnostic decision making, which we refer to as *human-AI collaboration protocols* and proposed the adoption of this concept in future evaluations of AI-based decision support systems. We investigated whether XAI support, in terms of both visual aids and textual explanations, has a significant effect on diagnostic accuracy. Our findings confirm the utility of AI support, however we found that XAI aids can be associated with what has been referred to as the "white-box paradox" [11], which has recently been observed in other settings [8, 56], i.e., a null or detrimental effect. Furthermore, and most notably, we compared protocols that differed in terms of when the machine's advice is given to human readers, i.e., either simultaneously with the case or after an initial diagnosis was formally obtained, which we refer to as ram and hound protocols, respectively. Even when the AI was less (or equally) accurate than the average human reader, we found that the order of presentation matters: AI-first protocols are associated with higher diagnostic accuracy than human-first protocols (and higher than both human basal accuracy and AI accuracy). This finding suggests the best conditions in which AI can actually augment human diagnostic skills (ram protocols), rather than trigger dysfunctional responses and cognitive biases (such as algorithmic aversion and conservatism bias) that can undermine decision effectiveness (human-first protocols).

Author Contribution Statement

F.C. and A.C. contribute equally to the research and manuscript drafting. In particular, they both conceived and planned the experiments and interpreted the results; the knee MRI experiment design was revised by L.S.; the ECG experiment design was revised by L.R.. AA.VV. annotated the MRI images; A.C. produced the activation maps for the knee MRI study. L.R. chose the ECG images and produced the related explanations. M.C. coordinated the ECG study. L.S. coordinated the MRI study. F.C. designed and developed the online questionnaire. A.C. analyzed the collected responses and produced the data visualizations. F.C. and A.C. interpreted the results. F.C. and A.C. wrote the manuscript. D.F. and M.B. surveyed the literature, produced the related work section and managed the quality of the references. H.G. supervised their work. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Code and Data Availability

Code for the data analysis and for generating the data visualizations, along with the raw and aggregated data required to replicate our experiments, is publicly available from `http://www.entechne.com/shared/Nature_Data_Code.zip`.

Competing Interests

The authors declare no competing interests. All participants of the study agreed to use of the collected data for academic research and have continuous ability to withdraw that consent.

List of Acronyms

- AI: artificial intelligence
- XAI: explainable artificial intelligence
- HAI-CP: human-artificial intelligence collaboration protcol
- AFOOT: affordance-fit-optimization-output-target
- BPMN: Business Process Modeling Notation
- MRI: Magnetic Resonance Imaging
- ECG: Electro-cardiogram
- HD1: first human decision
- HD2: second human decision
- FHD: final human decision

References

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

[2] Ammenwerth, E. (2015). Evidence-based health informatics: how do we know what we know? *Methods of Information in Medicine*, *54*, 298–307.

[3] Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (pp. 2–11). volume 7.

[4] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).

[5] Bental, D. S., Cawsey, A., & Jones, R. (1999). Patient information systems that tailor to the individual. *Patient education and counseling*, *36*, 171–180.

[6] van Berkel, N., Skov, M. B., & Kjeldskov, J. (2021). Human-ai interaction: intermittent, continuous, and proactive. *Interactions*, *28*, 67–71.

[7] Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society* (pp. 78–91).

[8] Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proc. of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* AIES. New York, NY: ACM. To appear.

[9] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1–21.

[10] Cabitza, F., Campagner, A., & Sconfienza, L. M. (2021). Studying human-ai collaboration protocols: the case of the kasparov's law in radiological double reading. *Health Information Science and Systems*, *9*, 1–20.

[11] Cabitza, F., Campagner, A., & Simone, C. (2021). The need to move away from agential-ai: Empirical investigations, useful concepts and open issues. *International Journal of Human-Computer Studies*, *155*, 102696.

[12] Cabitza, F., Campagner, A., Zotti, F., Ravizza, A., & Sternini, F. (2020). All you need is higher accuracy? on the quest for minimum acceptable accuracy for medical artificial intelligence. (pp. 159–166).

[13] Cabitza, F., & Zeitoun, J.-D. (2019). The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Annals of translational medicine*, *7*.

[14] Coiera, E. (2016). A new informatics geography. *Yearbook of Medical Informatics*, *25*, 251–255.

[15] Combi, C., Amico, B., Bellazzi, R., Holzinger, A., Moore, J. H., Zitnik, M., & Holmes, J. H. (2022). A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine*, (p. 102423).

[16] Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *AIAA 3rd Intelligent Systems Conference* (pp. 2004–6313).

[17] Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, *21*, 287–290.

[18] Cutillo, C. M., Sharma, K. R., Foschini, L., Kundu, S., Mackintosh, M., & Mandl, K. D. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ digital medicine*, *3*, 1–5.

[19] Dey, S., Karahalios, K., & Fu, W.-T. (2018). Getting there and beyond: Incidental learning of spatial knowledge with turn-by-turn directions and location updates in navigation interfaces. In *Proceedings of the symposium on spatial user interaction* (pp. 100–110).

[20] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*, 114.

[21] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, .

[22] Dragoni, M., Donadello, I., & Eccher, C. (2020). Explainable ai meets persuasiveness: Translating reasoning results into behavioral change advice. *Artificial Intelligence in Medicine*, *105*, 101840.

[23] Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M., Riedl, M. O. et al. (2021). The who in explainable ai: how ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*, .

[24] Ehsan, U., & Riedl, M. O. (2021). Explainability pitfalls: Beyond dark patterns in explainable ai. *arXiv preprint arXiv:2109.12480*, .

[25] Elmore, J. G., & Lee, C. I. (2022). Artificial intelligence in medical imaging—learning from past mistakes in mammography. In *JAMA Health Forum* (pp. e215207–e215207). American Medical Association volume 3.

[26] Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., & Holzinger, A. (2022). The explainability paradox: Challenges for xai in digital pathology. *Future Generation Computer Systems*, .

[27] Friedman, C. P. (2009). A "fundamental theorem" of biomedical informatics. *Journal of the American Medical Informatics Association*, *16*, 169–170.

[28] Gajos, K. Z., & Mamykina, L. (2022). Do people engage cognitively with ai? impact of ai assistance on incidental learning. In *27th International Conference on Intelligent User Interfaces* (pp. 794–806).

[29] Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, *4*, 1–8.

[30] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*, e745–e750.

[31] Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Penguin.

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

[33] Guermazi, A., Tannoury, C., Kompel, A. J., Murakami, A. M., Ducarouge, A., Gillibert, A., Li, X., Tournier, A., Lahoud, Y., Jarraya, M. et al. (2022). Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*, *302*, 627–636.

[34] Gur, D., Bandos, A. I., Cohen, C. S., Hakim, C. M., Hardesty, L. A., Ganott, M. A., Perrin, R. L., Poller, W. R., Shah, R., Sumkin, J. H. et al. (2008). The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology*, *249*, 47.

[35] Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-ai complementarity in hybrid intelligence systems: A structured literature review. *PACIS*, (p. 78).

[36] Hoff, T. (2011). Deskilling and adaptation among primary care physicians using two work innovations. *Health Care Management Review*, *36*, 338–348.

[37] Holzinger, A. (2021). The next frontier: AI we can really trust. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 427–440). Springer.

[38] Holzinger, A. T., & Muller, H. (2021). Toward human–ai interfaces to support explainability and causability in medical ai. *Computer*, *54*, 78–86.

[39] Hwang, E. J., Park, S., Jin, K.-N., Im Kim, J., Choi, S. Y., Lee, J. H., Goo, J. M., Aum, J., Yim, J.-J., Cohen, J. G. et al. (2019). Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA network open*, *2*, e191095–e191095.

[40] Inkpen, K. (2020). Does my ai help or hurt? exploring human-ai complementarity. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 2–2).

[41] Itoh, M., & Tanaka, K. (2000). Mathematical modeling of trust in automation: Trust, distrust, and mistrust. In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 9–12). SAGE Publications Sage CA: Los Angeles, CA volume 44.

[42] Jain, A., Way, D., Gupta, V., Gao, Y., de Oliveira Marinho, G., Hartford, J., Sayres, R., Kanada, K., Eng, C., Nagpal, K. et al. (2021). Development and assessment of an artificial intelligence–based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA network open*, *4*, e217249–e217249.

[43] Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

[44] Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C. P., Ball, R. L., Montine, T. J. et al. (2020). Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine*, *3*, 1–8.

[45] Klein, G. (2022). *Snapshots of the Mind*. The MIT Press.

[46] Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, *24*, 423–431.

[47] Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, (pp. 50–60).

[48] Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and brain sciences*, *37*, 1–19.

[49] Ooge, J., & Verbert, K. (2022). Explaining artificial intelligence with tailored interactive visualisations. In *27th International Conference on Intelligent User Interfaces* (pp. 120–123).

[50] Paleja, R., Ghuy, M., Ranawaka Arachchige, N., Jensen, R., & Gombolay, M. (2021). The utility of explainable ai in ad hoc human-machine teaming. *Advances in Neural Information Processing Systems*, *34*.

[51] Ploug, T., & Holm, S. (2020). The four dimensions of contestable ai diagnostics-a patient-centric approach to explainable ai. *Artificial Intelligence in Medicine*, *107*, 101901.

[52] Rafner, J., Dellermann, D., Hjorth, A., Verasztó, D., Kampf, C., Mackay, W., & Sherson, J. (2022). Deskilling, upskilling, and reskilling: a case for hybrid intelligence. *Morals & Machines*, *1*, 24–39.

[53] Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2020). Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *arXiv preprint arXiv:2010.07938*, .

[54] Ronzio, L., Campagner, A., Cabitza, F., & Gensini, G. F. (2021). Unity is intelligence: A collective intelligence experiment on ecg reading to improve diagnostic performance in cardiology. *Journal of Intelligence*, *9*, 17.

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

[56] Schemmer, M., Kühl, N., Benz, C., & Satzger, G. (2022). On the influence of explainable ai on automation bias. In *Thirtieth European Conference on Information Systems (ECIS 2022)*. Preprint at https://arxiv.org/abs/2204.08859.

[57] Schmidt, K., & Simonee, C. (1996). Coordination mechanisms: Towards a conceptual foundation of cscw systems design. *Computer Supported Cooperative Work (CSCW)*, *5*, 155–200.

[58] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

[59] Shin, D. (2020). How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in Human Behavior*, *109*, 106344.

[60] Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, *31*, 47–53.

[61] Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. (2000). Automation bias and errors: are crews better than individuals? *The International journal of aviation psychology*, *10*, 85–97.

[62] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, *25*, 44–56.

[63] Troya, J., Fitting, D., Brand, M., Sudarevic, B., Kather, J. N., Meining, A., & Hann, A. (2022). The influence of computer-aided polyp detection systems on reaction time for polyp detection and eye gaze. *Endoscopy*, .

[64] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J. et al. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, *26*, 1229–1234.

[65] Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*, 297–323.

[66] Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., & Krishna, R. (2022). Explanations can reduce overreliance on ai systems during decision-making. *arXiv preprint arXiv:2212.06823*, .

[67] Vodrahalli, K., Gerstenberg, T., & Zou, J. (2022). Uncalibrated models can improve human-ai collaboration. *arXiv preprint arXiv:2202.05983*, .

[68] Woolson, R. F. (2007). Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, (pp. 1–3).

[69] Yanase, J., & Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, *138*, 112821.

[70] Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, *2*, 719–731.

[71] Zhang, Q., Lee, M. L., & Carter, S. (2022). You complete me: Human-ai teams and complementary expertise. In *CHI Conference on Human Factors in Computing Systems* (pp. 1–28).

Appendix A. Appendix: Additional Results and Figures

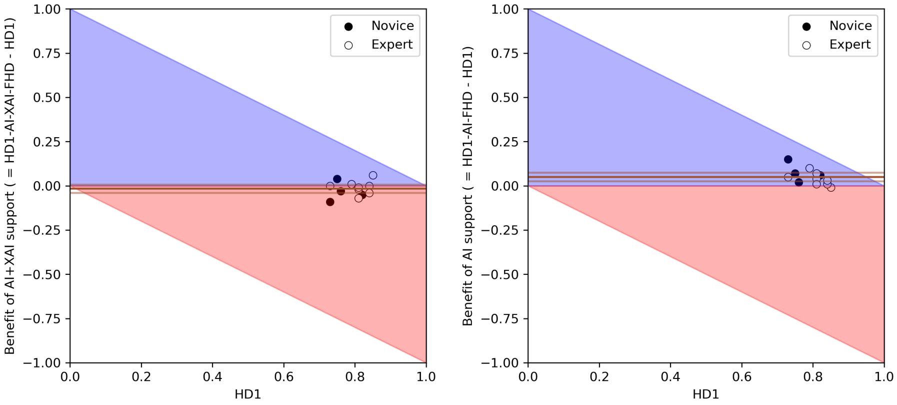


Figure Appendix A.1: Benefit diagrams for the hound protocols in the knee MRI study, showing the effect of providing XAI support (protocol HD1-AI-XAI-FHD, on the left) and AI support (protocol HD1-AI-FHD, on the right) versus providing no support (HD1). The dots represent the accuracies of the radiologists, and the brown lines the average difference in accuracy between the two protocols, along with the corresponding 95% confidence interval. The blue region denotes an improvement in error rates while the red region denotes a worsening

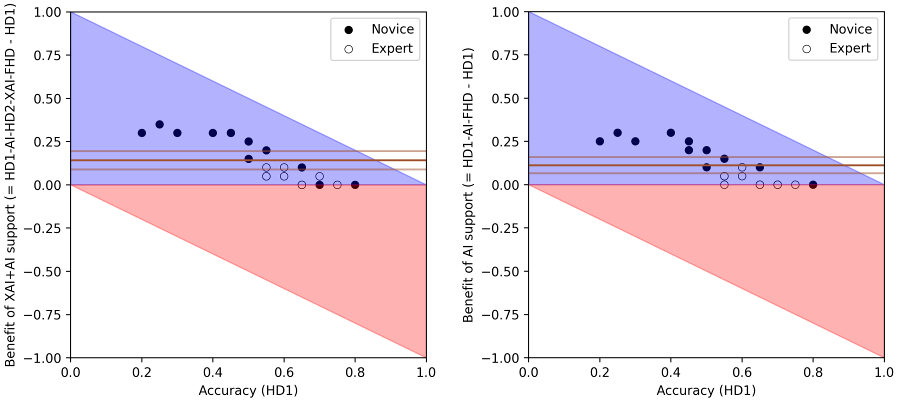


Figure Appendix A.2: Benefit diagrams for the hound protocols in the ECG study, showing the effect of providing XAI support (protocol HD1-AI-XAI-FHD, on the left) and AI support (protocol HD1-AI-FHD, on the right) versus providing no support (HD1). The dots represent the accuracies of the cardiologists, and the brown lines the average difference in accuracy between the two protocols, along with the corresponding 95% confidence interval. The blue region denotes an improvement in error rates while the red region denotes a worsening.