technologies

MDPI

Article

Process over Skill: Testing Kasparov's Law and Coordination Protocols in Hybrid Human–AI Decision-Making for Medical Diagnosis

Alessia Papale [1,*], Gloria Lopiano [1], Andrea Campagner [1] and Federico Cabitza [1,2]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Viale Sarca 336, 20126 Milan, Italy; g.lopiano1@campus.unimib.it (G.L.); andrea.campagner@unimib.it (A.C.); federico.cabitza@unimib.it (F.C.)
[2] Digital Health and Wellbeing Center, Fondazione Bruno Kessler (FBK), Via Sommarive 18, 38123 Trento, Italy
* Correspondence: alessia.papale@unimib.it

Abstract

Artificial intelligence (AI) is increasingly being integrated into Clinical Decision-Support Systems (CDSSs), shifting attention from algorithmic performance alone to the broader sociotechnical conditions that shape effective human–AI collaboration. In this study, we investigated whether nine displacement-based structured coordination protocols can improve the collective diagnostic decision-making of hybrid human–AI teams (16 board-certified radiologists and a simulated AI model) in a radiological double-reading task for vertebral fracture detection from X-ray images. Among the protocols tested, the Accuracy-Oriented, Confidence-Oriented, and Presumptuous strategies achieved the highest (balanced) accuracy overall, with up to 97% among strong clinicians and 92% among weak ones, significantly outperforming simpler methods like majority voting. Conversely, approaches optimized for a single metric (e.g., sensitivity or specificity) introduced performance trade-offs. Benefits were strongest among less proficient clinicians, which exhibited substantial and consistent improvements, while proficient clinicians showed limited gains and occasional declines. Critically, Kasparov's Law emerged as a comparative framework for empirically evaluating coordination quality relative to the diagnostic task, clinical objective, and clinician proficiency by identifying situations in which less proficient clinicians supported by superior coordination protocols outperformed more proficient clinicians operating under inferior ones. These findings demonstrate that coordination design is a critical determinant of hybrid human–AI decision-making, highlighting that a well-structured process can be more relevant than individual components' performance and support process-centered approaches to the development and evaluation of CDSSs.

Keywords: coordination protocols; interaction protocols; human–AI interaction; process design; Kasparov's Law; human–AI collaboration; human–AI synergy; human–AI teaming; hybrid intelligence; collective intelligence; double reading; Clinical Decision-Support Systems

Academic Editor: Giovanni Delnevo

Received: 10 April 2026
Revised: 10 June 2026
Accepted: 12 June 2026
Published: 17 June 2026

1. Introduction

1.1. Sociotechnical Challenges of Clinical AI

Artificial intelligence (AI) is rapidly reshaping clinical decision-making through AI-powered Clinical Decision-Support Systems (CDSSs), whose performance increasingly approaches—and, in some narrowly defined tasks, even surpasses—that of human experts [1]. Over the past few decades, CDSSs have evolved from early knowledge-based,

rule-driven systems such as MYCIN [2] and INTERNIST-I [3] to contemporary machine learning models capable of extracting latent patterns from large-scale clinical data. This progression reflects a broader shift toward data-driven healthcare, in which digital technologies function as cognitive extenders [4], supporting clinicians in managing growing informational complexity and time pressure. By enhancing clinical reasoning, streamlining workflows, and reducing human error, modern AI systems have demonstrated effectiveness in tasks such as the early detection of sepsis and diabetic retinopathy, thereby enabling timely intervention and improved patient outcomes [5,6].

Several researchers have explored the persistent challenges that hinder the successful adoption and integration of CDSSs in real-world settings [7,8]. A major barrier that undermines clinicians' trust and complicates ethical and legal accountability is the limited interpretability of many contemporary AI systems. As noted by Guidotti and colleagues [9], a substantial portion of high-performing models—particularly those based on deep learning—operate as "black boxes", generating accurate predictions without providing a transparent rationale. In clinical settings, such opacity is especially problematic because medical decision-making is inherently justificatory: clinicians are required to articulate and defend their decisions to patients, colleagues, and regulatory authorities, while ultimately retaining both professional accountability and legal responsibility. Moreover, limited interpretability complicates reliance calibration, that is, the ability to appropriately trust a system's recommendations by accepting them when correct and rejecting them when incorrect, detecting potential anomalies and inherent biases [10,11]. This often leads to automation bias, defined as the tendency of clinicians to over-rely on system recommendations, especially under conditions of cognitive load, fatigue, or time pressure [12]. Over time, such over-reliance may reduce clinician vigilance and independent judgment, potentially contributing to deskilling and raising concerns about the preservation of long-term professional competencies [13–16].

These findings indicate that CDSS deployment is not merely a technical undertaking, but a sociotechnical challenge that cannot be adequately understood if AI systems are treated as isolated technical artifacts evaluated solely in terms of predictive performance [17]. They operate within complex clinical environments and are actively incorporated by clinicians into their reasoning processes as they interpret and use system recommendations. Thus, the potential advantages associated with hybrid configurations do not automatically arise from merely pairing clinicians with AI systems. According to the National Academies of Sciences, Engineering, and Medicine [18], effective human–AI integration requires that human operators be able to understand and anticipate the behavior of intelligent agents, establish an appropriate level of trust, make accurate decisions using AI-generated outputs, and effectively control and manage the deployed systems.

From this perspective, clinical AI may increasingly be reframed through the lens of Human–AI teaming (HAIT), namely configurations involving "one or more people and one or more AI systems requiring collaboration and coordination to achieve successful task completion" [19]. This perspective shifts the analytical focus toward the interactional and organizational dynamics through which human and artificial agents jointly produce decisions. As also formalized in Friedman's Fundamental Theorem of Biomedical Informatics [20], the critical unit of analysis needs to be relocated from the technology alone to the human–technology ensemble. In such contexts, the quality of coordination protocols, reliance calibration, role allocation, communication structures, and aggregation mechanisms may become as important as predictive accuracy itself. Consequently, designing and evaluating AI as a "monadic" entity obscures the fundamentally relational and systemic nature of clinical decision-making environments.

1.2. Hybrid Intelligence: Synergy or Augmentation?

Taking these issues into consideration, a growing body of research has begun to explore the paradigm of hybrid intelligence, a framework that emphasizes the deliberate and synergistic integration of human and machine capabilities [21–23]. According to Hemmer and colleagues [24], this approach aims to orchestrate joint performance, whereby AI provides computational speed, scalability, and pattern recognition, while humans contribute intuition, context sensitivity, ethical judgment, and adaptability in uncertain or novel scenarios [25]. Wilson and Daugherty [26] argue that the value of hybrid intelligence lies in designing systems that enable a level of joint performance exceeding what either humans or machines could achieve independently. Without such integration, the system would merely default to the better-performing actor, rather than producing a genuinely cooperative advantage.

Noteworthy studies have attempted to assess the achievability of such configurations. For instance, a recent meta-analysis by Vaccaro, Almaatouq, and Malone [27] provides one of the most systematic examinations of this concept. The authors define human augmentation as scenarios *"where the human–AI group performs better than the human alone"* [27]. While this improvement is shown to be reliable and replicable, the results differ when considering human–AI synergy, namely situations *"where the human–AI group performs better than both the human alone and the AI alone"*. In this case, their findings indicate that such synergy is relatively rare. On average, combined systems often fail to outperform their individual components. Moreover, enhancements such as confidence scores or model rationales do not consistently lead to improved outcomes. Several factors may account for these underwhelming results. One explanation lies in miscalibrated reliance, where human users may exhibit either over-reliance by uncritically accepting AI suggestions [28,29] or under-reliance by dismissing valuable AI input [29,30]. Another explanation concerns the inefficient allocation of decision tasks across human and machine agents, where system design fails to leverage complementary strengths [25,27].

Considering the medical domain, a recent literature review by Liu and colleagues [31] highlights a nuanced set of patterns regarding the effectiveness of HAIT in clinical decision-making. Overall, hybrid configurations tend to outperform human-only performance on average, supporting the general promise of AI-assisted CDSS. However, such improvements are typically modest, rarely approach the idealized form of full complementarity in which system-level failures arise only from simultaneous errors by both human and AI components, and do not consistently exceed the performance of AI systems alone. A key determinant of effectiveness is the teaming mode, namely the temporal structure through which AI recommendations are presented to clinicians. Evidence suggests that simultaneous configurations, in which clinicians evaluate a case while concurrently observing the AI output, are generally more effective than sequential settings, in which clinicians first produce an independent judgment and only subsequently consult the AI. Another important factor is clinician expertise. Junior physicians tend to benefit more substantially from AI assistance, whereas senior clinicians exhibit smaller gains, with improvements that are often minimal or absent, particularly in sequential interaction settings.

1.3. Human–AI Coordination Protocols in Hybrid Decision-Making

Human–AI coordination protocols—in this paper, the expression coordination protocols and interaction protocols will be used interchangeably as synonyms—comprise both the topological ensemble of human and computational actors and the explicit set of rules articulating their activities in order to perform a common task (e.g., how human and AI contributions are sequenced and integrated into a joint decision process, and how human and AI components exchange information and manage disagreement). In this per-

spective, AI systems are conceptualized as sociotechnical interventions embedded within organizational workflows and professional practices.

Several classes of these coordination protocols have been proposed. Building on Pierce's framework of human–technology relations, Cabitza and colleagues [32] distinguish five major configurations for AI-supported decision-making: foreclosure, traditional interaction, inhibition, displacement, and replacement. Among others, we focus in particular on this taxonomy, as it provides a useful theoretical lens for the present study.

Foreclosure refers to settings in which AI systems are intentionally excluded from the decision process [32]. This configuration corresponds to a complete absence of automation support and is generally adopted when AI systems are considered insufficiently reliable, transparent, or trustworthy for a given task. At the opposite end of the spectrum lies replacement, in which decision authority is fully delegated to the AI system and humans are removed from the loop. While replacement may maximize efficiency, it also introduces substantial risks related to dependency, loss of human agency, and long-term deskilling.

Between these extremes lie protocols that preserve varying degrees of human involvement. The most widespread configuration is the traditional protocol, also referred to as AI-first, concurrent, or human-in-the-loop interaction [32]. In this setting, humans receive AI recommendations concurrently with the case information and are expected to integrate algorithmic advice into their decision-making process. However, empirical research has shown that this configuration may also induce automation bias [33]. As a consequence, the quality of outcomes in traditional interaction settings depends heavily on appropriate reliance on AI advice. To mitigate these risks, several authors have proposed the inhibition protocol, that deliberately introduces cognitive friction in order to prevent passive reliance on the machine and to preserve critical reflection and professional autonomy [32].

An extreme form of cognitive friction is represented by the displacement protocol, which consists in literally displacing the AI system from the physical or abstract location it would normally occupy in the decision-making process. In displacement settings, humans and AI systems independently and asynchronously produce complete judgments within a parallel decision-making scheme, and these judgments are subsequently reconciled through aggregation procedures [32]. In other words, they could be understood as systematic ways of combining distinct individual decisions into a single collective decision. Several aggregation rules commonly adopted in collective decision-making contexts—including majority voting, plurality voting, and weighted aggregation methods—have been extensively developed within social choice theory [34] and reviewed by Kameda and colleagues in the context of information aggregation and collective intelligence [35]. A key feature of displacement protocols is that humans and machines are integrated into teams without direct interaction. This seemingly counterintuitive design choice is motivated by a concern that has only recently become an object of systematic inquiry, namely the need to reduce their mutually detrimental influence [36,37]. This concern applies both to the influence of humans on machines in online and interactive learning systems, which use human feedback to better align with human judgment [38], and to the influence of machines on humans, when the latter over-rely on AI systems and are thereby led into error [33], when they acquire the biases embedded in machine recommendations [39], and when prolonged patterns of improper reliance progressively contribute to deskilling and erosion of expertise [16]. Moreover, recent empirical evidence suggests that this indirect form of hybrid decision-making can outperform both traditional interaction and AI-only configurations in a series of medical user studies involving radiology, orthopedics, and gastrointestinal endoscopy [32].

In this work, we specifically focus on displacement protocols.

1.4. Kasparov's Law: A Process-Oriented Evaluation Framework

In sociotechnical settings, performance does not reside in the artifact per se but in the coordinated—either interactive or displaced—system formed by humans and machines. Despite this important conceptual reorientation, evaluation frameworks rarely consider human–AI coordination protocols as an independent variable. Conversely, the relevance of this oversight is vividly illustrated in the domain of chess by insights attributed to Garry Kasparov [40] and echoed in recent studies [41]. His view was later synthesized by Brynjolfsson and McAfee into what has been called "Kasparov's Law" [42], expressed in the form of a comparative formula:

> Weak Human + Machine + Better Process > Strong Human + Machine + Inferior Process

The formula conveys a crucial insight: the quality of the coordination protocol can outweigh the individual strength of either the human or the machine. Even a moderately skilled human, when paired with an AI system through a well-designed interaction protocol, can outperform a stronger expert supported by the same technology but embedded in a poorly structured process. In other words, hybrid intelligence is process-sensitive, and coordination protocols can be more consequential than either raw computational power or individual expertise alone.

In line with this perspective, our objective is to empirically investigate whether structured coordination protocols—in this case displacement-based schemes—could:

- (RQ1)systematically influence the performance of human–AI ensembles relative to the standalone performance of their individual components;
- (RQ2) enable human–AI ensembles composed of less accurate clinicians, but better coordinated among themselves and with the AI system, to outperform human–AI ensembles composed of more accurate clinicians operating under inferior coordination protocols in a clinically relevant diagnostic task.

Consistent with the research questions, we formulate two general hypotheses. With respect to RQ1, we hypothesize that at least one structured coordination protocol systematically affects the performance of human–AI ensembles, leading to performance differences relative to the standalone performance of their individual components (H1). With respect to RQ2, we hypothesize that there exists at least one pair of coordination protocols such that less proficient raters operating under a more effective coordination process can achieve higher diagnostic performance than more proficient raters operating under a less effective coordination protocol (H2).

In this context, we define process quality as the overall quality of the human–AI coordination process with respect to the goals and constraints of a given use case or operational setting. Rather than being reducible to a single metric, process quality is conceptualized as a multidimensional construct that manifests across several performance-, behavioral-, and cognition-related dimensions. These include short-term performance outcomes, such as efficacy with respect to the task objective (i.e., accuracy, sensitivity, specificity), and efficiency (i.e., achieving such performance with minimal expenditure of time and resources), as well as the behavioral and cognitive consequences of the interaction (e.g., appropriate reliance, understanding, perceived utility, satisfaction, trust, automation bias, cognitive load, and potential deskilling) [43]. However, because displacement protocols do not involve direct interaction with the machine at the decision level, these behavioral and cognitive consequences are not directly applicable in our case.

Empirical studies systematically evaluating structured displacement protocols in hybrid human–AI decision-making remain relatively rare. Cabitza and colleagues [41] introduced and compared multiple coordination protocols in a radiological double-reading

setting for the detection of knee lesions by means of Magnetic Resonance Imaging (MRI), providing early evidence that structured coordination frameworks can outperform individual performances, although the displacement paradigm had not yet been explicitly theorized as a distinct design approach. More recently, Cabitza and colleagues [32] formally introduced displacement protocols in human–AI interaction and showed that majority voting-based schemes can outperform both traditional protocols and AI-only configurations. However, this study focused primarily on majority voting approaches and did not investigate how process quality varies as a function of clinician proficiency.

Building on this line of research, we test whether Kasparov's Law generalizes to a further diagnostic task in a simulated double-reading radiology setting, in which two or more clinicians (often referred to as observers or readers) independently assess the same medical images for diagnostic interpretation [44]. Specifically, we focus on the detection of vertebral fractures from X-ray images. Compared with prior work, we examine multiple human–AI configurations that vary in clinician expertise and interaction protocol structure, thereby expanding the coordination space under investigation through the evaluation of nine protocols and explicitly modeling clinician proficiency as a moderating variable.

Our findings demonstrate that interaction protocols play a significant role in overall human–AI system performance. Consequently, the evaluation of CDSSs must critically examine not only the capabilities of the AI system itself, but also how humans and machines work together. In this regard, Kasparov's Law provides a valuable conceptual anchor for advocating process-centered evaluation frameworks, in which synergy and augmentation arise not merely from the sum of individual competencies, but from the orchestration of contributions across the system. The effectiveness of AI in practice depends as much on effective process design as on algorithmic innovation [45]. Without this dual focus, even highly accurate AI models may fail to produce meaningful improvements in real-world clinical outcomes.

The remainder of this paper is structured as follows. In Section 2, we present our methodological approach, providing a detailed account of the radiological task under investigation, the coordination protocols we designed, and the quantitative techniques employed to compare human–AI system performance across multiple scenarios. Section 3 reports the results of these experiments and discusses their implications. Finally, in Section 4, we summarize the paper, acknowledge its limitations, and suggest potential directions for future research aimed at optimizing human–AI collaboration in clinical decision-making.

2. Materials and Methods

2.1. Dataset Preparation

We recruited 16 board-certified radiologists from Gaetano Pini Hospital in Milan (Italy) to participate in the data collection phase. The experiment was conducted using a structured questionnaire developed with LimeSurvey (6.5.17), an open-source platform for online survey administration. All ethical standards were rigorously observed throughout the study, with particular attention to participant confidentiality and responsible collection, storage, and management of data.

Each radiologist was presented with a set of 18 X-ray images carefully selected to represent a diverse range of cases commonly encountered in specialized orthopedic clinical practice. Participants were asked to independently assess the presence or absence of a fracture in each case, providing a binary response (1 for "fracture present" and 0 for "fracture absent"). In addition to the binary classification, participants were required to indicate their level of confidence in each response using a four-point scale, which was subsequently normalized during preprocessing to ensure consistency across analyses.

In two instances, radiologists omitted the confidence rating. These missing values were imputed using the mean confidence score assigned by other participants for the same image to preserve dataset completeness. Although imputing these two missing confidence values preserves the overall distributional properties of the confidence scores, it reduces within-case variance for the affected observations and may marginally influence protocols that use confidence values to resolve disagreement. However, because only 2 of the 288 total ratings were affected, the expected impact on aggregate protocol performance is negligible.

Based on the radiologists' responses and the established diagnostic ground truth, we computed accuracy, balanced accuracy, sensitivity, and specificity for each participant (see Table 1). Accuracy (acc) was calculated as the proportion of correctly classified cases among all cases, $(TP + TN)/(TP + TN + FP + FN)$. Balanced accuracy (bal acc) was computed as the arithmetic mean of sensitivity and specificity, $(sens + spec)/2$, thereby providing a performance measure that is robust to potential class imbalance. Sensitivity (sens), corresponding to the true positive rate, was defined as the proportion of correctly identified positive cases, computed as $TP/(TP + FN)$. Specificity (spec), corresponding to the true negative rate, was calculated as the proportion of correctly identified negative cases, computed as $TN/(TN + FP)$.

In addition, the dataset included simulated recommendations generated by a well-calibrated AI-based CDSS, whose performance metrics are reported in Table 1. To reproduce real-world variability in confidence reporting, a random confidence value between 0.50 and 1.00 was generated for each AI prediction. Consistent with the assumption of good calibration, correct AI predictions were assigned confidence scores ranging from 0.76 to 1.00, whereas incorrect predictions were assigned confidence scores ranging from 0.50 to 0.74.

Table 1. Baseline accuracy (acc), balanced accuracy (bal acc), sensitivity (sens), and specificity (spec) of all participants involved in the study, including the AI.

	AI	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
acc	0.89	0.89	1	0.83	1	0.78	1	0.89	0.89	0.78	0.83	0.89	0.78	0.94	0.89	0.83	0.89
bal acc	0.89	0.89	1	0.84	1	0.78	1	0.89	0.89	0.78	0.84	0.90	0.78	0.95	0.89	0.84	0.89
sens	0.89	0.89	1	0.89	1	0.78	1	0.78	0.89	0.89	0.78	0.79	0.89	1	1	0.78	0.78
spec	0.89	0.89	1	0.78	1	0.78	1	1	0.89	0.67	0.89	1	0.67	0.89	0.78	0.89	1

These different data were later used to test the human–AI coordination protocols described in the following section.

2.2. Human–AI Interaction Protocols Simulations

After collecting the responses and performance metrics for both the human raters and the simulated AI-based CDSS, we implemented a double-reading procedure within a pseudo-arbitration framework [44]. This approach, commonly used in diagnostic workflows, involves two observers independently reviewing the same case in sequence without influencing each other's judgments. When the two initial interpretations disagree—that is, when one classifies the case as normal and the other as pathological, or vice versa—a third observer is introduced. This third observer, often referred to as an arbiter, also evaluates the case independently and remains blind to the prior disagreement to minimize confirmation bias and groupthink. The final decision is then determined by majority vote among the three observers, simulating a consensus mechanism often employed in clinical practice.

Building on this structure, we implemented nine protocol variants to model different forms of human–AI coordination. These protocols were not intended to exhaust the space of possible displacement strategies, but rather to represent a set of interpretable and clinically plausible coordination archetypes reflecting different operational priorities and decision-making practices. Each protocol operationalized a distinct approach to

integrating decisions within the diagnostic process, including consensus-based aggregation, competence- and confidence-weighted disagreement resolution, risk-sensitive strategies prioritizing either sensitivity or specificity, and different decision styles. This framework allowed us to examine how different coordination protocols influence diagnostic accuracy, balanced accuracy, sensitivity, and specificity, as well as how protocol design impacts the performance of clinicians with varying levels of proficiency:

- *Simple-Majority*: The first and second observer provide their judgments. The third observer is involved if and only if the first two observers disagree. The final decision is the majority choice of the three observers in the team.
- *Accuracy-Oriented*: The three observers provide their judgments. If they agree, their decision is taken as final. If they disagree, the decision is made by comparing the average accuracy of the two agreeing observers with the accuracy of the dissenting one. If the average of the pair is higher, their shared decision is chosen; otherwise, the dissenting opinion prevails.
- *Confidence-Oriented*: The three observers provide their judgments. If they agree, their decision is taken as final. If they disagree, the decision is made by comparing the average confidence of the two agreeing observers with the confidence of the dissenting one. If the average of the pair is higher, their shared decision is chosen; otherwise, the dissenting opinion prevails.
- *Specificity-Oriented*: The first observer provides their judgment and the second observer is involved if and only if the first observer deemed the case abnormal. If the first two observers disagree, then the third observer is also involved. The final decision is made by comparing the average $spec_i \times conf_i$, where specificity also accounts for how confident the observer is in this particular case, of the two agreeing observers with the same product computed for the dissenting one. If the average of the pair is higher, their shared decision is chosen; otherwise, the dissenting opinion prevails.
- *Sensitivity-Oriented*: The first observer provides their judgment and the second observer is involved if and only if the first observer deemed the case normal. If the first two observers disagree, then the third observer is also involved. The final decision is made by comparing the average $sens_i \times conf_i$, where sensitivity also accounts for how confident the observer is in this particular case, of the two agreeing observers with the same product computed for the dissenting one. If the average of the pair is higher, their shared decision is chosen; otherwise, the dissenting opinion prevails.
- *Cautious*: The first two observers provide their judgments, accuracy, and subjective confidences. If they agree in their judgment, the team accepts the decision only if both have accuracy and confidence above a defined threshold α_{acc} and α_{conf} ($acc_i \geq \alpha_{acc}$ and $conf_i \geq \alpha_{conf}$), both equal to the quantile 0.25 of human's accuracy and confidence respectively in the considered case. Otherwise, if the two observers disagree or if at least one of them has accuracy or confidence below the threshold α, the third observer is involved. The final decision is made by comparing the average $acc_i \times conf_i$, where accuracy also accounts for how confident the observer is in this particular case, of the two agreeing observers with the same product computed for the dissenting one. If the average of the pair is higher, their shared decision is chosen; otherwise, the dissenting opinion prevails.
- *Presumptuous*: The first two observers provide their judgments, accuracy, and subjective confidences. If both of them have accuracy above a defined threshold α_{acc} ($acc_i \geq \alpha_{acc}$), equal to the quantile 0.25 of the human accuracies in the considered case, the team's decision is the same as the one provided by the observer with greater confidence. If at least one observer has accuracy below the threshold α_{acc}, the team follows the observer with the highest accuracy.

- *Positive-Gated Review*: The first observer provides their judgment. The second observer is involved if and only if the interpretation of the first observer is abnormal and, in that case, the decision of the team is the same as the second observer's.
- *Negative-Gated Review*: The first observer provides their judgment. The second observer is involved if and only if the interpretation of the first observer is normal and, in that case, the decision of the team is the same as the second observer's.

Specifically, for each X-ray and each permutation of observers, we determined whether the final classification (fracture vs. non-fracture) was correct and then computed accuracy, balanced accuracy, sensitivity, and specificity for each protocol across all evaluated cases. Then, 95% confidence intervals (CIs) were subsequently estimated using a binomial approximation. Throughout Section 3, performance differences are considered significant when the 95% CIs of the compared metrics do not overlap, a conservative criterion that approximately corresponds to a two-sided test with $\alpha < 0.01$ for proportions [46].

To assess the reliability of both individual-level performance and group-level decisions, we calculated the Intraclass Correlation Coefficient (ICC), a statistical measure commonly used to evaluate inter-rater, test–retest, and intra-rater reliability [47]. We employed ICC(2,1), which is appropriate when raters are treated as a random sample from a larger population and the objective is to assess agreement among individuals rating the same set of cases. According to the guidelines proposed by Cicchetti [48], the resulting ICC value of 0.63 [0.47, 0.80] indicates good reliability, thereby supporting the validity of the comparative analyses across protocols and participants.

2.3. Testing Kasparov's Law

To empirically investigate the quality of coordination protocols in human–AI decision-making, we adopted Kasparov's Law as a comparative methodological framework:

> Weak Human + Machine + Better Process > Strong Human + Machine + Inferior Process

Importantly, we did not assume *a priori* that any protocol was intrinsically "superior" or "inferior". Instead, we operationalized Kasparov's Law as a relational and comparative framework for empirically identifying coordination processes capable of producing compensatory or augmentative effects. Specifically, we examined whether clinicians with lower baseline performance, when paired with a given protocol, could outperform stronger clinicians operating under an alternative protocol. When such differences were observed and found to be significant, they provided evidence that coordination processes can play a decisive role in shaping team performance, particularly when the clinician is less proficient.

Based on individual accuracy scores, participants were divided into two groups: a "strong" rater group and a "weak" rater group. Specifically, participants whose accuracy was equal to or above the median value of all accuracy scores (0.89) were classified as "strong", while those scoring below this threshold were classified as "weak". To further ensure that this grouping reflects a meaningful difference in diagnostic performance, we computed the Intraclass Correlation Coefficient (ICC(2,1)) [47]. The results showed that the "strong" group had an ICC of 0.75 [0.61, 0.88], while the "weak" group had an ICC of 0.47 [0.28, 0.70]. According to the interpretive guidelines proposed by Cicchetti [48], these values correspond to excellent (ICC $\geq$ 0.75) and fair (0.40 $\leq$ ICC < 0.60), respectively. Overall, these findings provide statistical support for the grouping strategy, indicating that the two sets of raters differ meaningfully in diagnostic consistency and overall performance.

Following this stratification, we implemented a full factorial comparison across expertise levels and coordination protocols. In practice, we systematically constructed and evaluated all possible human–AI configurations obtained by pairing clinicians classified as "strong" and "weak" with each available coordination protocol. The resulting experimental

space was explored through direct pairwise comparisons between two contrasting configurations: (1) "weak" clinicians operating under a given protocol, and (2) "strong" clinicians operating under a different protocol. For each configuration, diagnostic performance was computed separately in terms of accuracy, sensitivity, and specificity, and these metrics were analyzed independently within each pairwise comparison.

3. Results

3.1. Overall Performance

Table 2 and Figure 1 present the aggregate results of the simulation, considering all participants irrespective of individual diagnostic proficiency. Among the evaluated strategies, the Accuracy-Oriented, Confidence-Oriented, Presumptuous protocols consistently achieved the significantly highest overall (balanced) accuracy scores, reaching 0.92 [0.91, 0.93], 0.91 [0.90, 0.91], and 0.91 [0.91, 0.92] respectively. Conversely, protocols designed to prioritize either sensitivity (Sensitivity-Oriented) or specificity (Specificity-Oriented) exhibited a pronounced trade-off, excelling in their targeted metric but significantly underperforming in the complementary dimension compared to all other protocols. Notably, the Positive-Gated Review and Negative-Gated Review protocols also demonstrated strong overall (balanced) accuracy (0.89 [0.89, 0.90] and 0.88 [0.87, 0.88]), with Positive-Gated Review achieving the significantly highest specificity (0.97 [0.97, 0.97]) and Negative-Gated Review the significantly highest sensitivity (0.96 [0.96, 0.96]). When focusing specifically on sensitivity under Sensitivity-Oriented and Negative-Gated Review protocols and on specificity under Positive-Gated Review, team performance exceeded the standalone performance of both the average human rater and the AI system, thus constituting clear instances of human–AI synergy [27]. From a practical standpoint, these findings suggest that coordination protocols can be strategically selected based on clinical priorities. In screening contexts where minimizing false negatives is critical to ensure that no true cases are missed, sensitivity-oriented schemes may be preferable, whereas in confirmatory or resource-constrained settings where it is better to minimize false positives to avoid overdiagnosis or unnecessary interventions, specificity-oriented protocols may be more appropriate. More broadly, the results indicate that optimizing the structure of decision aggregation, rather than focusing exclusively on improving individual human or algorithmic performance, can yield measurable gains in diagnostic effectiveness.

Overall, protocols that weigh decisions based on rater accuracy or confidence, or that allow for less conservative decision-making, appear to offer greater diagnostic benefit. In contrast, both the Simple-Majority and Cautious protocols yielded mediocre performance, consistently failing to improve upon baseline human or AI results and, in many cases, even leading to performance declines. This suggests that, in a general radiological double-reading context, these strategies may be suboptimal.

The fact that some coordination schemes produce measurable synergy while others systematically undermine it underscores that hybrid intelligence is not guaranteed by mere combination, but is critically shaped by the structure of interaction and decision aggregation. Even when the same human and machine components are involved, differences in coordination protocols can determine whether the ensemble outperforms its constituents or falls below them.

Table 2. Accuracy (acc), balanced accuracy (bal acc), sensitivity (sens), and specificity (spec) scores for all participants across the different human–AI interaction protocols, along with 95% confidence intervals (CI). Differences between protocols can be assessed by examining non-overlapping CIs within each performance dimension, which provide an indication of superior or inferior diagnostic performance in terms of accuracy, balanced accuracy, sensitivity, and specificity.

Protocol	spec [95% CI]	sens [95% CI]	acc [95% CI]	bal acc [95% CI]
Simple-Majority	0.86 [0.84, 0.88]	0.77 [0.74, 0.80]	0.82 [0.80, 0.84]	0.82 [0.79, 0.84]
Accuracy-Oriented	0.93 [0.92, 0.93]	0.92 [0.91, 0.92]	0.92 [0.91, 0.93]	0.92 [0.91, 0.93]
Confidence-Oriented	0.91 [0.91, 0.92]	0.91 [0.90, 0.91]	0.91 [0.90, 0.91]	0.91 [0.90, 0.91]
Specificity-Oriented	0.90 [0.89, 0.90]	0.62 [0.61, 0.63]	0.77 [0.76, 0.78]	0.76 [0.75, 0.77]
Sensitivity-Oriented	0.58 [0.57, 0.59]	0.93 [0.93, 0.94]	0.75 [0.74, 0.76]	0.76 [0.75, 0.77]
Cautious	0.77 [0.77, 0.78]	0.84 [0.83, 0.85]	0.81 [0.80, 0.82]	0.81 [0.80, 0.81]
Presumptuous	0.91 [0.90, 0.91]	0.92 [0.91, 0.92]	0.91 [0.91, 0.92]	0.91 [0.91, 0.92]
Positive-Gated Review	0.97 [0.97, 0.97]	0.81 [0.81, 0.82]	0.89 [0.89, 0.90]	0.89 [0.89, 0.90]
Negative-Gated Review	0.80 [0.79, 0.81]	0.96 [0.96, 0.96]	0.88 [0.87, 0.88]	0.88 [0.87, 0.88]

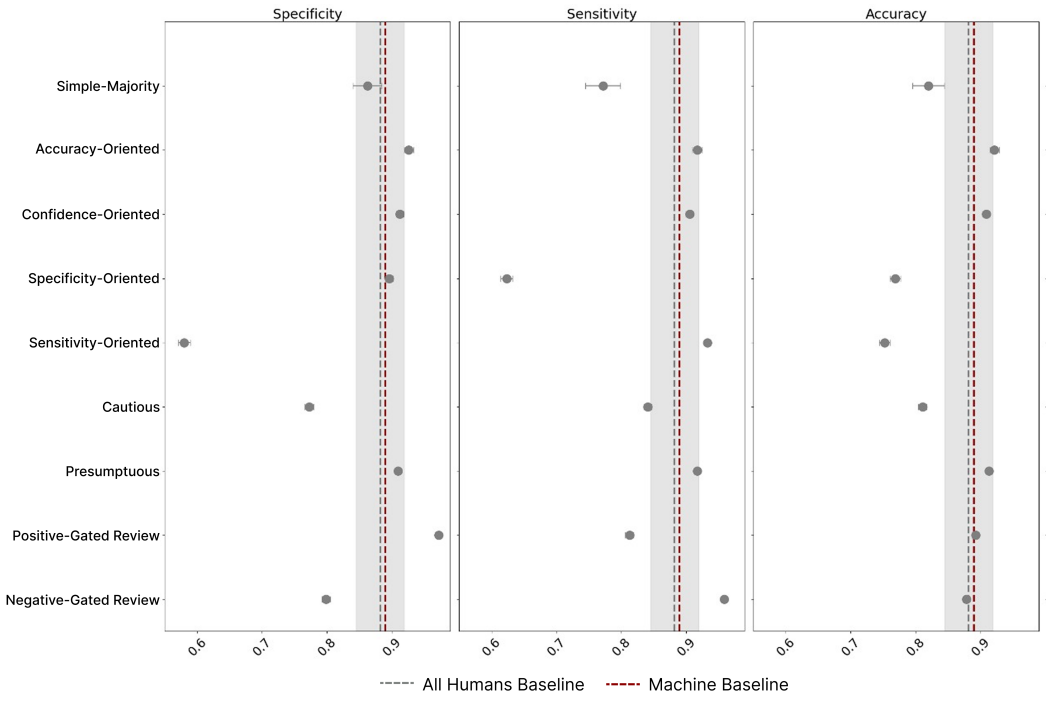

Figure 1. Accuracy, sensitivity, and specificity scores for all participants across the different human–AI interaction protocols, along with their 95% confidence intervals represented as error bars. Dashed lines represent the baseline performance of the AI and the mean baseline performance of human clinicians when operating independently. The lighter-colored bands surrounding the lines indicate the corresponding confidence intervals (95%). Differences between protocols can be assessed by examining non-overlapping CIs within each performance dimension, which provide an indication of superior or inferior diagnostic performance in terms of accuracy, balanced accuracy, sensitivity, and specificity.

3.2. Performance of Strong and Weak Participants

Stratifying participants based on their baseline diagnostic accuracy revealed the moderating effects of the interaction protocols. Figures 2 and 3 present a comparative analysis of the performance of "strong" and "weak" clinicians across all human–AI interaction protocols, showing protocol-specific scores in the former and aggregated scores in the latter. As expected, clinicians who exhibited higher baseline performance continued to generally outperform their "weaker" counterparts—although not consistently nor always to a statistically significant extent—when collaborating with the AI regardless of the protocol used and across all metrics.

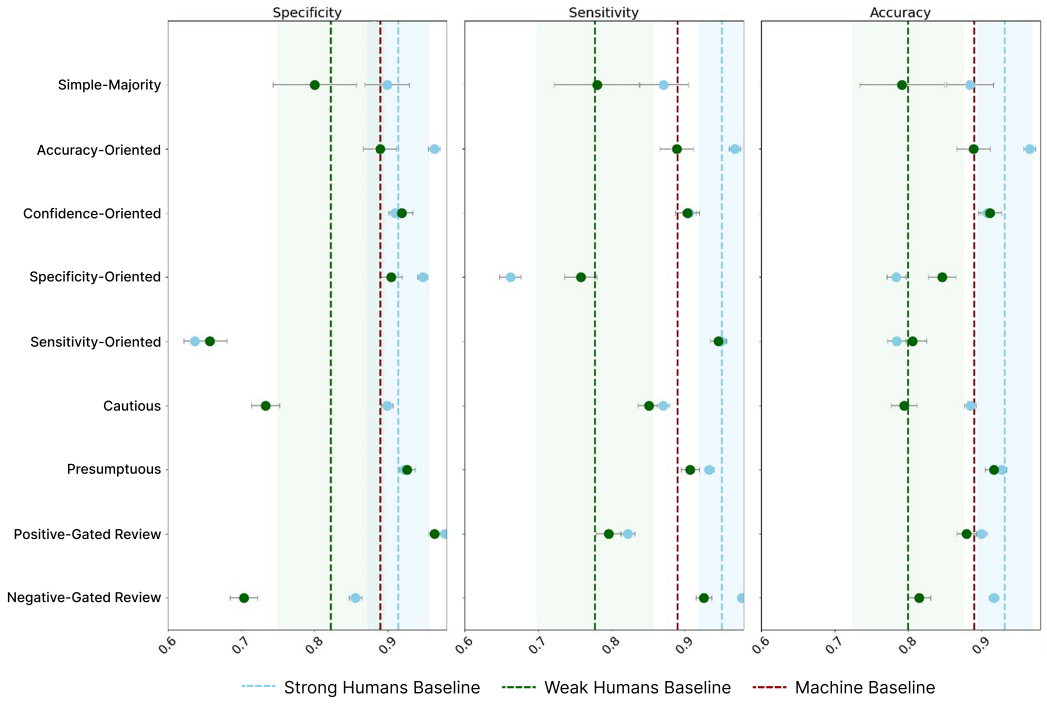

Figure 2. Accuracy, sensitivity, and specificity scores for all participants, grouped by strong (skyblue) and weak (green) clinicians, across the different human–AI interaction protocols, along with their 95% confidence intervals represented as error bars. Dashed lines represent the baseline performance of the AI and the mean baseline performance of human clinicians when operating independently. The lighter-colored bands surrounding the lines indicate the corresponding confidence intervals (95%). Differences between protocols can be assessed by examining non-overlapping CIs within each performance dimension, which provide an indication of superior or inferior diagnostic performance in terms of accuracy, balanced accuracy, sensitivity, and specificity.

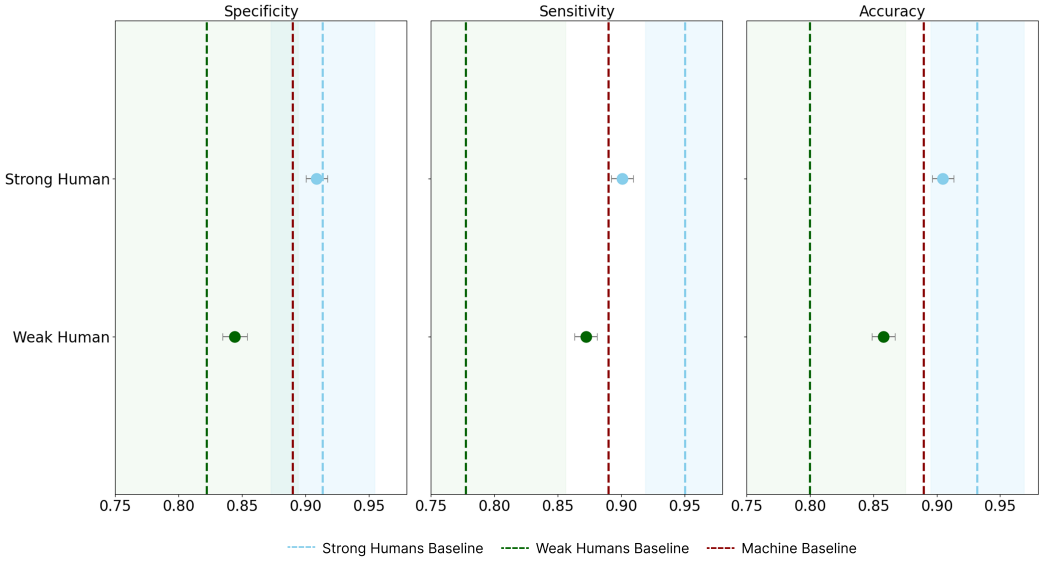

Figure 3. Comparison of accuracy, sensitivity, and specificity scores between strong (skyblue) and weak (green) clinicians on all protocols, along with their 95% confidence intervals represented as error bars. Dashed lines represent the baseline performance of the AI and the mean baseline performance of human clinicians when operating independently. The lighter-colored bands surrounding the lines indicate the corresponding confidence intervals (95%). Differences between protocols can be assessed by examining non-overlapping CIs within each performance dimension, which provide an indication of superior or inferior diagnostic performance in terms of accuracy, balanced accuracy, sensitivity, and specificity.

However, a more nuanced pattern emerges when we consider the impact of AI collaboration on performance change relative to baseline. "Strong" performers generally showed little to no improvement, and in some cases even a decline in performance, particularly and consistently under the Specificity-Oriented and Sensitivity-Oriented protocols. This decline is especially evident in sensitivity, where more than half of the protocols yielded lower scores compared to their human-only baselines. This suggests a potential ceiling effect [49], whereby participants who already perform at a very high level leave little room for further measurable improvement, thus limiting the observable benefits of AI assistance. In this case, senior clinicians may already operate close to the upper bound of diagnostic accuracy, so the marginal gain achievable through AI support is substantially reduced compared to less experienced clinicians, whose lower baseline performance allows for greater room for improvement. A ceiling effect has been observed in a large number of human–AI clinical teaming studies [31].

In contrast, "weaker" clinicians demonstrated substantial and consistent improvements across all metrics when using the Confidence-Oriented and Presumptuous protocols. When metrics are disaggregated, additional protocol-specific benefits emerged. In terms of specificity, improvements were also observed under the Positive-Gated Review protocol, while for sensitivity, gains were evident under the Accuracy-Oriented, Sensitivity-Oriented, and Negative-Gated Review protocols.

Once again, the findings suggest that human–AI coordination protocols should not be treated as universally optimal. The same protocol may benefit "weaker" clinicians while providing little advantage or even producing performance decrements for "stronger" ones, which might not even need an AI support. This implies that a "one-size-fits-all" deployment strategy in clinical settings may be suboptimal. However, certain protocols (notably, the Confidence-Oriented and Presumptuous) can function as performance equalizers, substantially raising the diagnostic floor by enabling "weaker" clinicians to approach higher performance levels, without eroding "strong" clinicians performance. From an organizational perspective, this may improve consistency of care and reduce variability across practitioners.

Tables 3 and 4 provide a detailed breakdown of the accuracy, balanced accuracy, specificity, and sensitivity scores achieved by "strong" and "weak" participants, respectively, across all human–AI interaction protocols.

Table 3. Accuracy (acc), balanced accuracy (bal acc), sensitivity (sens), and specificity (spec) scores for "weak" participants across all the different human–AI interaction protocols and the superior protocols specifically, along with 95% confidence intervals (CI). Differences between protocols can be assessed by examining non-overlapping CIs within each performance dimension, which provide an indication of superior or inferior diagnostic performance in terms of accuracy, balanced accuracy, sensitivity, and specificity.

Protocol	spec [95% CI]	sens [95% CI]	acc [95% CI]	bal acc [95% CI]
Simple-Majority	0.80 [0.74, 0.86]	0.78 [0.72, 0.84]	0.79 [0.73, 0.85]	0.79 [0.73, 0.85]
Accuracy-Oriented	0.89 [0.87, 0.91]	0.89 [0.87, 0.91]	0.89 [0.87, 0.91]	0.89 [0.87, 0.91]
Confidence-Oriented	0.92 [0.90, 0.93]	0.90 [0.89, 0.92]	0.91 [0.90, 0.93]	0.91 [0.90, 0.93]
Specificity-Oriented	0.90 [0.89, 0.92]	0.76 [0.74, 0.78]	0.85 [0.83, 0.86]	0.83 [0.81, 0.85]
Sensitivity-Oriented	0.66 [0.63, 0.68]	0.95 [0.93, 0.96]	0.81 [0.79, 0.83]	0.80 [0.78, 0.82]
Cautious	0.73 [0.71, 0.75]	0.85 [0.84, 0.87]	0.79 [0.78, 0.81]	0.79 [0.77, 0.81]
Presumptuous	0.93 [0.91, 0.94]	0.91 [0.90, 0.92]	0.92 [0.90, 0.93]	0.92 [0.90, 0.93]
Positive-Gated Review	0.96 [0.96, 0.97]	0.80 [0.78, 0.81]	0.88 [0.87, 0.89]	0.88 [0.87,0.89]
Negative-Gated Review	0.70 [0.69, 0.72]	0.93 [0.92, 0.94]	0.81 [0.80, 0.83]	0.81 [0.80, 0.83]
All Protocols	0.82 [0.75, 0.89]	0.77 [0.70, 0.86]	0.80 [0.72, 0.87]	0.80 [0.72, 0.87]

Table 4. Accuracy (acc), balanced accuracy (bal acc), sensitivity (sens), and specificity (spec) scores for "strong" participants across all the different human–AI interaction protocols and the inferior protocols specifically, along with 95% confidence intervals (CI). Differences between protocols can be assessed by examining non-overlapping CIs within each performance dimension, which provide an indication of superior or inferior diagnostic performance in terms of accuracy, balanced accuracy, sensitivity, and specificity.

Protocol	spec [95% CI]	sens [95% CI]	acc [95% CI]	bal acc [95% CI]
Simple-Majority	0.90 [0.87, 0.93]	0.87 [0.84, 0.90]	0.88 [0.85, 0.92]	0.89 [0.85, 0.92]
Accuracy-Oriented	0.96 [0.95, 0.97]	0.97 [0.96, 0.98]	0.97 [0.96, 0.97]	0.97 [0.96, 0.97]
Confidence-Oriented	0.91 [0.90, 0.92]	0.91 [0.90, 0.92]	0.91 [0.90, 0.92]	0.91 [0.90, 0.92]
Specificity-Oriented	0.95 [0.94, 0.95]	0.66 [0.65, 0.68]	0.78 [0.77, 0.80]	0.81 [0.79, 0.82]
Sensitivity-Oriented	0.64 [0.62, 0.65]	0.95 [0.94, 0.96]	0.78 [0.77, 0.80]	0.79 [0.78, 0.81]
Cautious	0.90 [0.89, 0.91]	0.87 [0.86, 0.88]	0.88 [0.88, 0.89]	0.88 [0.88, 0.89]
Presumptuous	0.92 [0.92, 0.93]	0.93 [0.93, 0.94]	0.93 [0.92, 0.93]	0.93 [0.92, 0.93]
Positive-Gated Review	0.98 [0.97, 0.98]	0.82 [0.81, 0.83]	0.90 [0.89, 0.91]	0.90 [0.89, 0.91]
Negative-Gated Review	0.86 [0.85, 0.86]	0.98 [0.97, 0.98]	0.92 [0.91, 0.92]	0.92 [0.91, 0.92]
All Protocols	0.91 [0.87, 0.95]	0.95 [0.92, 0.98]	0.93 [0.90, 0.97]	0.93 [0.90, 0.97]

For participants classified as "strong", the Accuracy-Oriented protocol once again yielded the highest (balanced) accuracy (0.97 [0.96, 0.97]), followed by the Presumptuous protocol (0.93 [0.92, 0.93]). As in previous analyses, the Positive-Gated Review and Negative-Gated Review protocols retained their respective advantages in specificity (0.98 [0.97, 0.98]) and sensitivity (0.98 [0.97, 0.98]).

On the other hand, among "weak" participants, the Presumptuous protocol emerged as the top performer in terms of (balanced) accuracy (0.92 [0.90, 0.93]), followed closely by the Confidence-Oriented protocol (0.91 [0.90, 0.93]). Although the Accuracy-Oriented protocol remained effective (0.89 [0.87, 0.91]), its advantage was less pronounced than among the "strong" subgroup. Once again, the Positive-Gated Review protocol maintained its lead in specificity (0.96 [0.96, 0.97]), while the Sensitivity-Oriented protocol achieved the best sensitivity score (0.95 [0.93, 0.96]), albeit at the cost of lower specificity (0.66 [0.63, 0.68]).

Interestingly, the Confidence-Oriented protocol presented a more stable performance across all metrics for both "weak" and "strong" clinicians.

3.3. Kasparov's Law

To evaluate whether effective coordination can compensate for differences in individual proficiency, we performed an exhaustive comparison of diagnostic performance across all human–AI configurations obtained by combining clinicians classified as as "strong" or "weak" with each available coordination protocol. For each performance metric—accuracy, sensitivity, and specificity—we constructed a pairwise comparison matrix reporting whether a given configuration involving less proficient clinicians achieved higher performance than a configuration involving more proficient clinicians under a different coordination protocol. The resulting matrices are visualized in Figures 4–9.

With respect to accuracy (see Figures 4 and 5), Accuracy-Oriented and Positive-Gated Review protocols emerged as superior to both the Specificity-Oriented and Sensitivity-Oriented protocols according to the Kasparov's Law comparison framework. The Confidence-Oriented and Presumptuous protocols extended this pattern by additionally outperforming the Cautious protocol. Furthermore, the Specificity-Oriented protocol was found to be superior to the Sensitivity-Oriented protocol, whereas the reverse relationship was never observed. However, a closer examination of these results suggests that much of this apparent superiority is driven by performance degradation among "strong" operating under less effective protocols rather than by substantial improvements among "weak" clinicians operating under more effective ones. A notable exception is the comparison between the Confidence-Oriented and Cautious protocols and the comparison between the Presump-

tuous and Cautious protocols. In both these cases, the superiority of the former cannot be attributed solely to a deterioration in the performance of proficient clinicians operating under the latter, but also to a significant gain in the performance of "weak" clinicians.

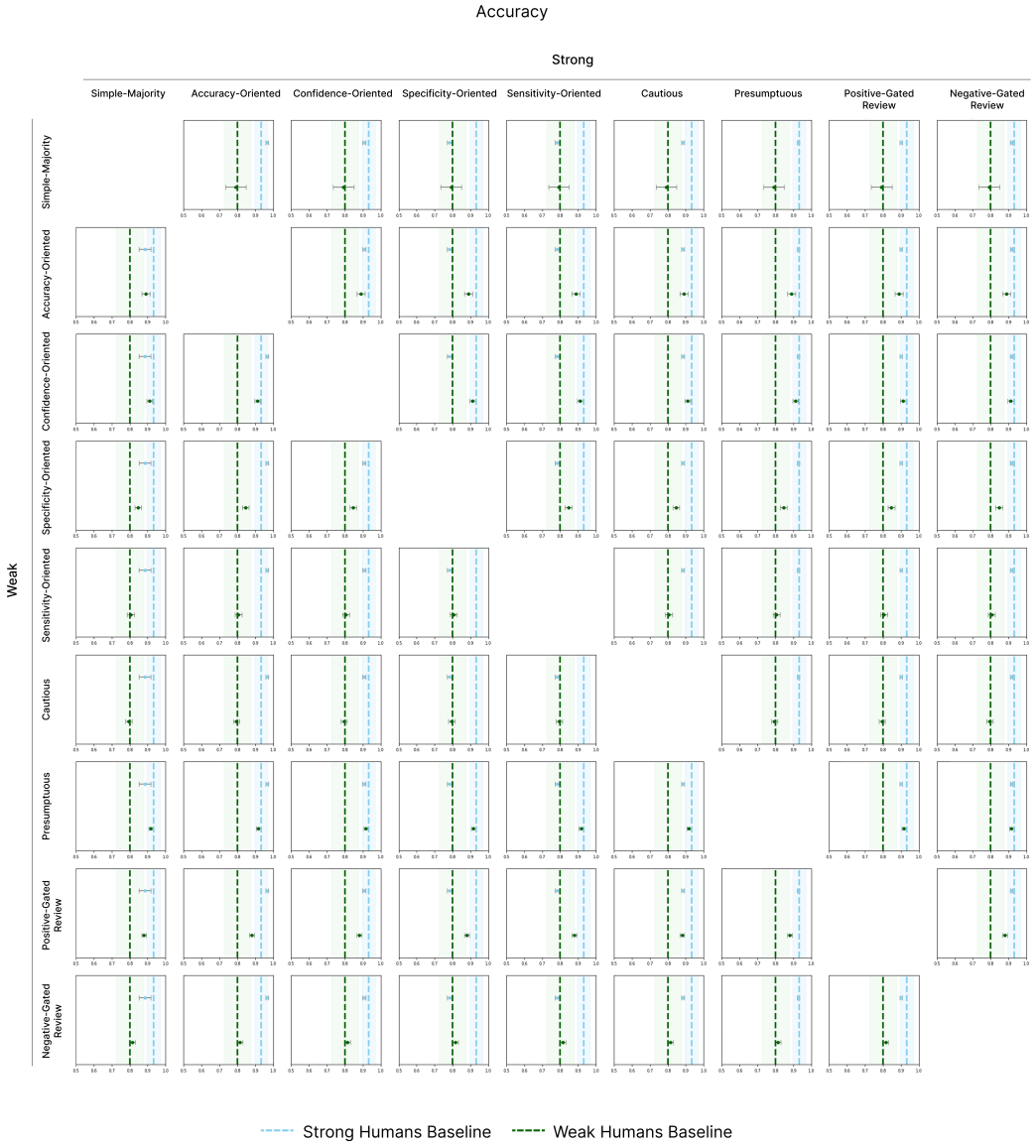

Figure 4. Pairwise comparison matrix for the Kasparov's Law analysis based on accuracy performance, along with 95% confidence intervals represented as error bars. Rows represent human–AI ensembles composed of weak (green) clinicians operating under a given coordination protocol, whereas columns represent human–AI ensembles composed of strong (skyblue) clinicians operating under a different coordination protocol. Each comparison indicates whether the accuracy achieved by weak clinicians exceeds that achieved by strong clinicians operating under an alternative coordination protocol. Dashed lines represent the mean baseline performance of human clinicians when operating independently. The lighter-colored bands surrounding the lines indicate the corresponding confidence intervals (95%). Differences between groups can be assessed by examining non-overlapping CIs within each comparison.

Figure 5. Pairwise comparison matrix for the Kasparov's Law analysis based on accuracy performance, along with 95% confidence intervals. Rows represent human–AI ensembles composed of weak (green) clinicians operating under a given coordination protocol, whereas columns represent human–AI ensembles composed of strong (skyblue) clinicians operating under a different coordination protocol. Each comparison indicates whether the accuracy achieved by weak clinicians exceeds that achieved by strong clinicians operating under an alternative coordination protocol. Differences between groups can be assessed by examining non-overlapping CIs within each comparison. Light blue cells indicate superior performance of the strong clinicians coordinated ensemble, green cells indicate superior performance of the weak clinicians coordinated ensemble, whereas gray cells indicate no significant difference between the two coordinated groups.

With respect to sensitivity (see Figures 6 and 7), the Specificity-Oriented protocol exhibited a pronounced degradation in performance among "strong" clinicians, to the extent that their sensitivity falls below the unaided performance of "weak" clinicians. Consequently, this protocol is consistently outperformed by all alternative coordination strategies, although improvements among "weak" clinicians also contribute to the observed effect. A similar, though less pronounced, pattern is observed for the Positive-Gated Review. Notably, the Sensitivity-Oriented protocol outperformed both the Simple-Majority and Confidence-Oriented, Negative-Gated Review protocol outperformed the Simple-

Majority, and Confidence-Oriented, Sensitivity-Oriented, Presumptuous, and Negative-Gated Review protocols all outperformed the Cautious protocol in terms of sensitivity.

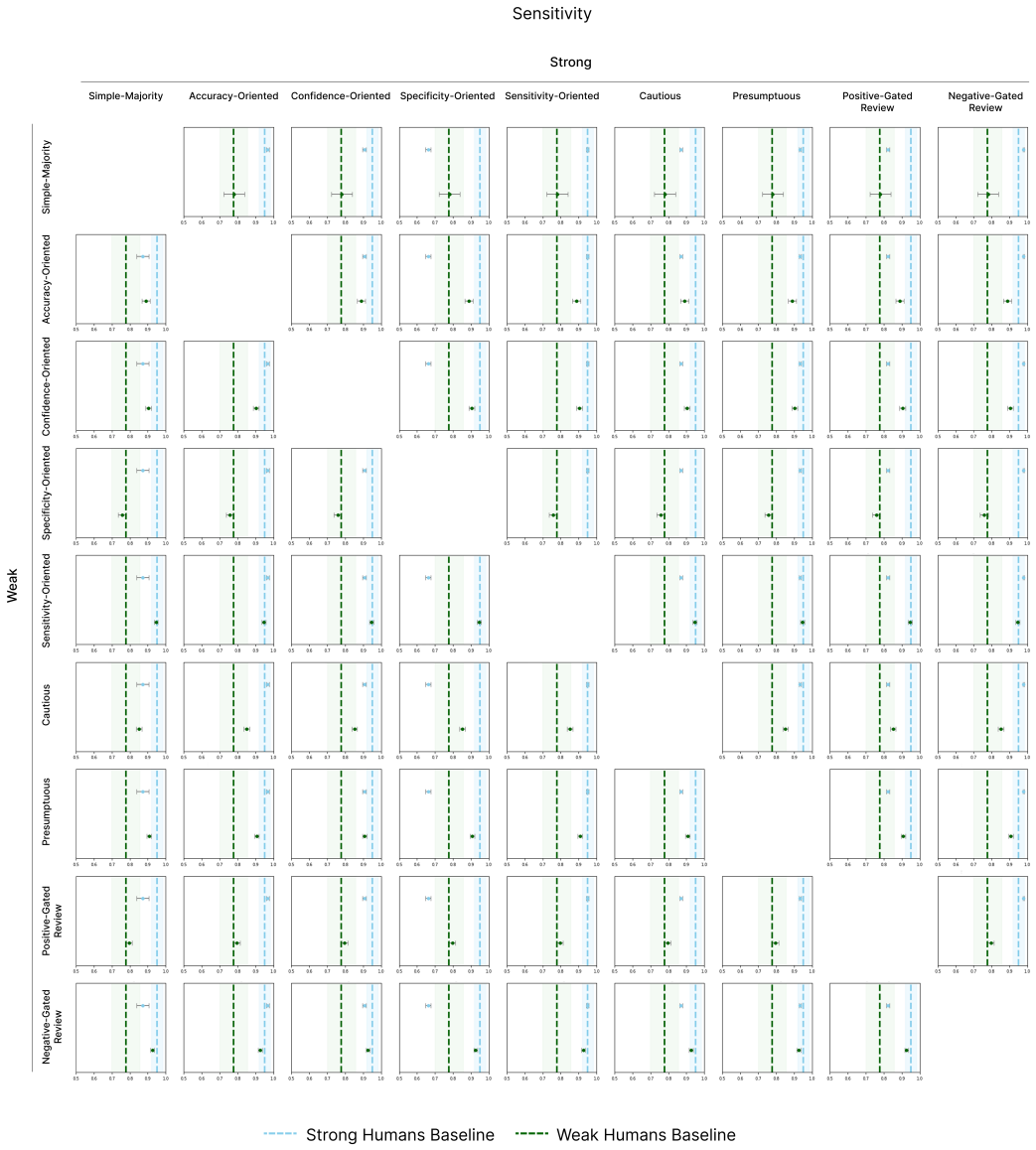

Figure 6. Pairwise comparison matrix for the Kasparov's Law analysis based on sensitivity performance, along with 95% confidence intervals represented as error bars. Rows represent human–AI ensembles composed of weak (green) clinicians operating under a given coordination protocol, whereas columns represent human–AI ensembles composed of strong (skyblue) clinicians operating under a different coordination protocol. Each comparison indicates whether the sensitivity achieved by weak clinicians exceeds that achieved by strong clinicians operating under an alternative coordination protocol. Dashed lines represent the mean baseline performance of human clinicians when operating independently. The lighter-colored bands surrounding the lines indicate the corresponding confidence intervals (95%). Differences between groups can be assessed by examining non-overlapping CIs within each comparison.

Figure 7. Pairwise comparison matrix for the Kasparov's Law analysis based on sensitivity performance, along with 95% confidence intervals. Rows represent human–AI ensembles composed of weak (green) clinicians operating under a given coordination protocol, whereas columns represent human–AI ensembles composed of strong (skyblue) clinicians operating under a different coordination protocol. Each comparison indicates whether the sensitivity achieved by weak clinicians exceeds that achieved by strong clinicians operating under an alternative coordination protocol. Differences between groups can be assessed by examining non-overlapping CIs within each comparison. Light blue cells indicate superior performance of the strong clinicians coordinated ensemble, green cells indicate superior performance of the weak clinicians coordinated ensemble, whereas gray cells indicate no significant difference between the two coordinated groups.

A nearly symmetric pattern emerged when considering specificity (see Figures 8 and 9). In this case, the Sensitivity-Oriented and Negative-Gated Review protocols are consistently inferior under the Kasparov's Law comparison framework. This effect is again driven by a pronounced reduction in performance among "strong" clinicians—particularly severe for the Sensitivity-Oriented protocol and more moderate for the Negative-Gated Review protocol—together with a concurrent improvement in "weak" clinicians under alternative coordination schemes. Within this dimension, the Positive-Gated Review protocol stood out as the strongest overall performer, surpassing not only the Simple-Majority, Confidence-

Oriented, Specificity-Oriented, Cautious, and Presumptuous strategies, but crucially doing so without inducing a significant reduction in the performance of "strong" clinicians.

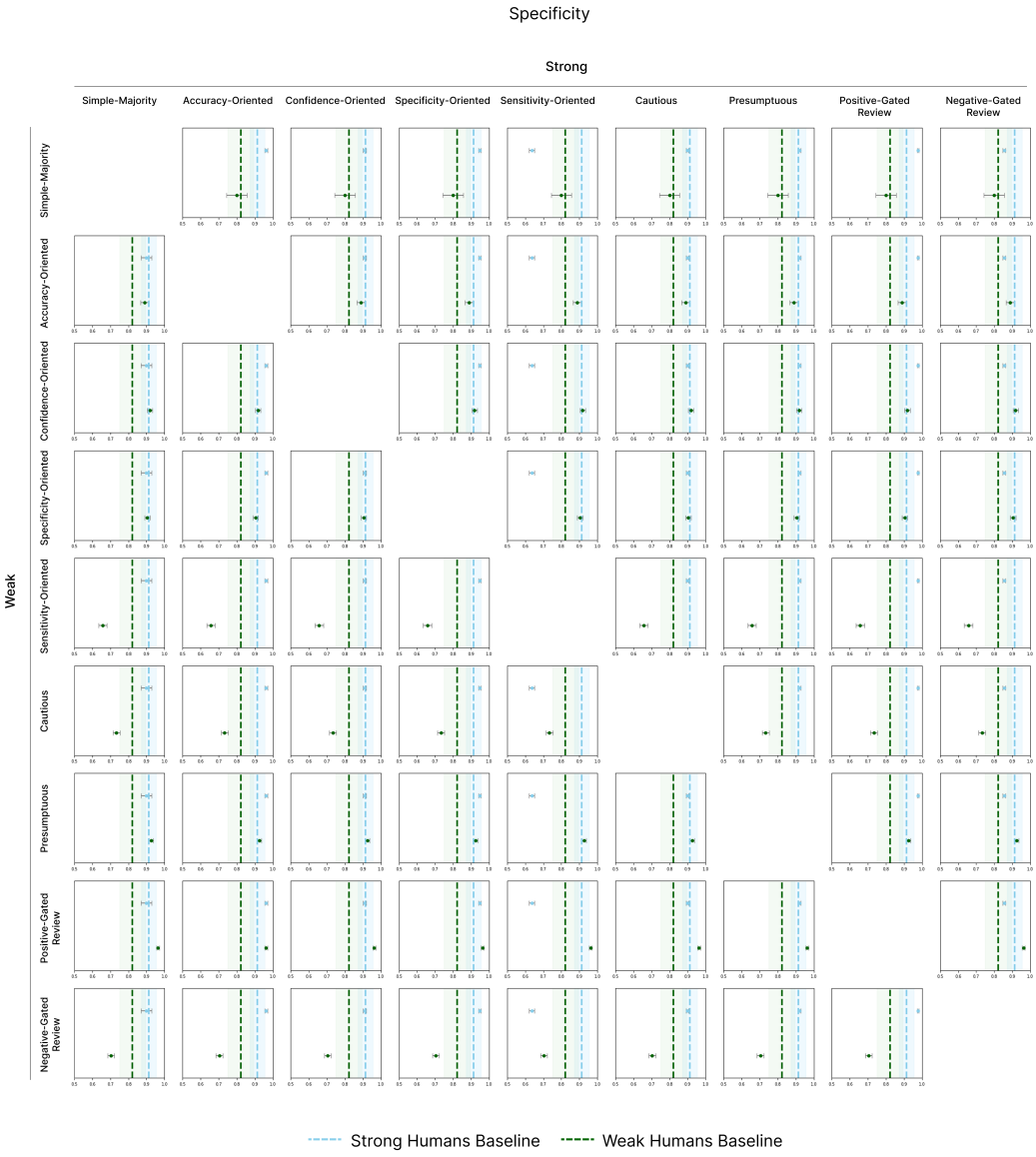

Figure 8. Pairwise comparison matrix for the Kasparov's Law analysis based on specificity performance, along with 95% confidence intervals represented as error bars. Rows represent human–AI ensembles composed of weak (green) clinicians operating under a given coordination protocol, whereas columns represent human–AI ensembles composed of strong (skyblue) clinicians operating under a different coordination protocol. Each comparison indicates whether the specificity achieved by weak clinicians exceeds that achieved by strong clinicians operating under an alternative coordination protocol. Dashed lines represent the mean baseline performance of human clinicians when operating independently. The lighter-colored bands surrounding the lines indicate the corresponding confidence intervals (95%). Differences between groups can be assessed by examining non-overlapping CIs within each comparison.

Figure 9. Pairwise comparison matrix for the Kasparov's Law analysis based on specificity performance, along with 95% confidence intervals. Rows represent human–AI ensembles composed of weak (green) clinicians operating under a given coordination protocol, whereas columns represent human–AI ensembles composed of strong (skyblue) clinicians operating under a different coordination protocol. Each comparison indicates whether the specificity achieved by weak clinicians exceeds that achieved by strong clinicians operating under an alternative coordination protocol. Differences between groups can be assessed by examining non-overlapping CIs within each comparison. Light blue cells indicate superior performance of the strong clinicians coordinated ensemble, green cells indicate superior performance of the weak clinicians coordinated ensemble, whereas gray cells indicate no significant difference between the two coordinated groups.

4. Conclusions

This study empirically investigated whether structured interaction protocols can affect the performance of human–AI ensembles in a simulated double-reading radiology setting for vertebral fracture detection from X-ray images, and whether better-coordinated teams with "weaker" human components can outperform worse-coordinated teams with "stronger" human components. Specifically, we simulated nine distinct human–AI interaction protocols involving 16 radiologists with varying diagnostic proficiency and a fixed-performance AI system.

Overall, the results show that human–AI performance is strongly shaped by the coordination protocol rather than solely by the capabilities of the human and artificial components of the system. Accuracy-Oriented, Confidence-Oriented, and Presumptuous protocols consistently achieved the highest overall (balanced) accuracy, while sensitivity-oriented and specificity-oriented protocols produced predictable trade-offs. Importantly, some protocols (e.g., Positive-Gated Review, Negative-Gated Review, and Sensitivity-Oriented for their target metrics) generated clear human–AI synergy, surpassing both standalone human and AI performance. In contrast, other protocols, such as Simple-Majority and Cautious, failed to improve performance and in some cases degraded it, confirming that hybrid intelligence is not guaranteed by aggregation alone. Stratified analyses further revealed that protocol effects are moderated by clinician baseline proficiency. "Strong" clinicians showed limited gains and occasional declines, suggesting ceiling effects, whereas "weak" clinicians exhibited substantial and consistent improvements, particularly under Confidence-Oriented and Presumptuous protocols. Moreover, these two schemes effectively raised the diagnostic floor without penalizing top performers.

Finally, when evaluated under the Kasparov's Law comparison framework, all coordination protocols satisfy the Law by enabling less proficient clinicians to outperform more proficient clinicians operating under alternative decision-making processes in at least one comparison and on at least one diagnostic metric. However, in many of these cases, the observed superiority is primarily explained by performance degradation among "strong" clinicians rather than by meaningful improvements among "weak" clinicians. The most notable exception is the Positive-Gated Review protocol, which, when compared with most other protocols, consistently enables less proficient clinicians to surpass their more proficient counterparts in terms of specificity while simultaneously preserving the performance of the latter. Consequently, although Kasparov's Law is satisfied whenever a less proficient but better coordinated ensemble outperforms a more proficient but worse coordinated one, such outcomes may arise through qualitatively different mechanisms—either through suppression of expert performance or through genuine enhancement of less experienced clinicians—and should therefore be interpreted in light of the underlying behavioral and performance dynamics of the coordination process. Understanding these mechanisms is essential for translating Kasparov's framework into a meaningful tool for evaluating collective decision-making in clinical practice. Future research should continue to employ Kasparov's Law as a valuable evaluative framework for studying human–AI collaboration, but always alongside an explicit assessment of the baseline performance of the decision-makers involved, accounting for how coordination protocols affect participants relative to their standalone performance.

These findings suggest that thoughtful coordination design should be considered alongside algorithmic optimization, as it can enhance human performance and reduce performance gaps between clinicians of differing expertise, improving consistency of care, reducing inter-clinician variability, and elevating "weak" performers without risking AI-induced deskilling. Such a result reinforces the value of process-centric approaches in the design, development, and deployment of CDSSs. However, the choice of coordination protocol governing how humans and AI integrate their individual decisions, as well as its adaptability to specific clinical situations, is also crucial for overall performance. A one-size-fits-all approach to human–AI coordination is therefore suboptimal. Rather, coordination protocols should be selected strategically according to the diagnostic task, the clinical objective, and the proficiency of the participating clinicians. To support this perspective, we operationalized Kasparov's Law as a context-sensitive, relational, and comparative framework for empirically identifying coordination processes that enable less proficient but better-coordinated ensembles to outperform more proficient but less effectively coordinated

ones. This conclusion is further supported by the contrast with the findings reported by Cabitza and colleagues [41], who tested Kasparov's Law in a different use case and found majority voting to be the most effective coordination strategy for enabling less proficient clinicians to outperform more proficient ones. Future work should further investigate whether certain coordination protocols exhibit stable advantages across tasks, or whether their effectiveness remains closely tied to the specific clinical objective, evaluation metric, and operational context in which they are deployed.

A promising direction suggested by these findings concerns adaptive protocol assignment. Since "strong" clinicians exhibit ceiling effects or even performance decrements under some protocols, a proficiency-aware deployment strategy where the coordination protocol is dynamically selected or adjusted based on the clinician's estimated baseline performance would be preferable to a uniform approach. In practice, this could be operationalized through a pre-deployment calibration phase in which a clinician's accuracy on a validated reference set is estimated, and the coordination protocol is selected or dynamically adjusted accordingly. From a technical perspective, such adaptivity could be implemented via rule-based thresholding (e.g., assigning Accuracy-Oriented to clinicians above a given accuracy quantile and Presumptuous or Confidence-Oriented otherwise) or via reinforcement learning mechanisms [50] that update protocol selection in response to observed performance trajectories. Such an architecture would preserve the performance gains observed among less proficient clinicians while mitigating the performance degradation observed among more proficient clinicians. More broadly, this approach represents a natural extension of the process-centric design philosophy advocated in this work, aligning with the broader vision of adaptive CDSSs and personalized human–AI collaboration.

Yet, some limitations must be acknowledged:

1. The simulated environment, although methodologically controlled, does not capture the variability present in real clinical workflows, thereby limiting the ecological validity of the findings. First, we modeled the AI system as a fixed-performance support with known calibration properties, whereas real CDSSs exhibit variable performance across image subtypes, patient demographics, and acquisition conditions. Coordination protocols should ideally be robust to such performance heterogeneity. Second, although some protocols in our implementation relied on fixed thresholds, in real-world deployments these parameters should remain configurable by clinicians and healthcare organizations, allowing adaptation to contextual and operational constraints such as workload, risk tolerance, clinical urgency, available expertise, or institutional policies. Indeed, a key strength of protocol-based human–AI collaboration lies in the possibility of tailoring coordination mechanisms to the sociotechnical environment in which they are embedded, rather than treating protocols as static or universally optimal solutions. Third, the simulated setting abstracts away the temporal dynamics of clinical practice: in real workflows, the sequential activation of observers and arbiters introduces latency costs that may render certain protocols impractical in high-throughput settings. Future work should therefore evaluate not only the diagnostic effectiveness of different coordination protocols, but also their operational efficiency. Such analyses would help identify coordination strategies that achieve an appropriate balance between diagnostic performance and practical feasibility. Fourth, the grouping of "strong" and "weak" clinicians is based on a single task and a limited set of X-ray images, and may therefore oversimplify the true spectrum of clinical expertise. Future studies should also extend protocol design to incorporate additional factors, such as clinicians' reliance behaviors and cognitive load.
2. The present study considers only binary diagnostic outcomes (fracture present/absent), whereas real clinical workflows often involve multi-label or ordinal classifications.

The extension of displacement protocols to such settings is non-trivial, as agreement and disagreement are no longer binary states, and aggregation rules must be redesigned accordingly.

3. Broader ethical and regulatory implications also warrant attention. If process design can improve the performance of less experienced diagnosticians, this may affect how responsibility is assigned in AI-assisted care and how competence is assessed or augmented through technology.
4. The relatively limited dataset and the fact that all 16 radiologists were recruited from a single institution constrain the generalizability of the baseline performance distribution and the resulting proficiency stratification. However, our aim was not to identify universally superior coordination processes. Rather, our goal was to investigate whether Kasparov's Law could be evaluated in a specific clinical use case and under specific pairs of coordination protocols. Consequently, the relevance of the framework is not diminished by the absence of broad generalizability, as its purpose is to assess coordination quality on a case-by-case basis, taking into account the specific task, performance objectives, and population of decision-makers under consideration.

In conclusion, this study underscores that the effectiveness of CDSSs depends not only on predictive accuracy or algorithmic optimization, but also on the design of human–AI coordination itself, supporting a shift toward a process-centric perspective as a central concern in medical AI. By placing structured coordination at the core of system design and evaluation, we move closer to realizing the potential of hybrid and collective decision-making in healthcare.

Author Contributions: Conceptualization: A.P.; methodology: A.P., G.L. and A.C.; original draft preparation: A.P., G.L. and F.C.; review and editing: A.P. and F.C.; visualizations: G.L.; coordination and supervision: A.P. and F.C. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge funding support provided by the Italian projects AMAR (CUP: D53C22002380006) and InXAID (CUP: H53D23008090001), both funded by the European Union (Next Generation EU). This work was also partially supported by the MUR under the grant "Dipartimenti di Eccellenza 2023–2027" (REGAINS) of the Department of Informatics, Systems, and Communication of the University of Milano-Bicocca, Italy, and Fastweb SpA.

Institutional Review Board Statement: According to the regulations of our Department, the responsibility for assessing the ethical risk level of a study lies with the Principal Investigator. In cases where the study is assessed as minimal risk, formal approval from the University Ethics Committee is not required under our departmental guidelines.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data supporting the findings of this study are publicly available in a GitHub repository at https://github.com/G1oLo/Collaboration-Strategies (accessed on 11 June 2026).

Acknowledgments: During the preparation of this manuscript/study, the authors used GPT-4o for the purposes of checking grammar and spelling, paraphrasing and rewording, improving writing style. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sutton, R.T.; Pincock, D.; Baumgart, D.C.; Sadowski, D.C.; Fedorak, R.N.; Kroeker, K.I. An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success. *npj Digit. Med.* **2020**, *3*, 17. [CrossRef] [PubMed]
2. Shortliffe, E. *Computer-Based Medical Consultations: MYCIN*; Elsevier: Amsterdam, The Netherlands, 1976.
3. Miller, R.A.; Pople, H.E., Jr.; Myers, J.D. Internist-I, an Experimental Computer-based Diagnostic Consultant for General Internal Medicine. In *Computer-Assisted Medical Decision Making*; Springer: Berlin/Heidelberg, Germany, 1985; pp. 139–158.
4. Clark, A.; Chalmers, D. The Extended Mind. *Analysis* **1998**, *58*, 7–19. [CrossRef]
5. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef] [PubMed]
6. Topol, E.J. High-performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]
7. Ash, J.S.; Berg, M.; Coiera, E. Some Unintended Consequences of Information Technology in Healthcare: The Nature of Patient Care Information System-related Errors. *J. Am. Med. Inform. Assoc.* **2004**, *11*, 104–112. [PubMed]
8. Cabitza, F.; Rasoini, R.; Gensini, G.F. Unintended Consequences of Machine Learning in Medicine. *JAMA* **2017**, *318*, 517–518. [CrossRef] [PubMed]
9. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of Methods for Explaining Black Box Models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [CrossRef]
10. Schemmer, M.; Kuehl, N.; Benz, C.; Bartos, A.; Satzger, G. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*; Association for Computing Machinery: New York, NY, USA; pp. 410–422.
11. Campagner, A.; Fregosi, C.; Natali, C.; Cabitza, F. Under what influence: Measuring AI influence to fit user profiles in decision-making. *Int. J. Hum.-Comput. Stud.* **2026**, *193*, 103763. [CrossRef]
12. 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. Impact of a Deep Learning Assistant on the Histopathologic Classification of Liver Cancer. *npj Digit. Med.* **2020**, *3*, 23. [CrossRef] [PubMed]
13. Parasuraman, R.; Riley, V. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum. Factors* **1997**, *39*, 230–253. [CrossRef]
14. Skitka, L.J.; Mosier, K.; Burdick, M.D. Accountability and Automation Bias. *Int. J. Hum.-Comput. Stud.* **2000**, *52*, 701–717. [CrossRef]
15. Khera, R.; Simon, M.A.; Ross, J.S. Automation Bias and Assistive AI: Risk of Harm from AI-driven Clinical Decision Support. *JAMA* **2023**, *330*, 2255–2257. [CrossRef] [PubMed]
16. Natali, C.; Marconi, L.; Dias Duran, L.D.; Cabitza, F. AI-induced deskilling in medicine: A mixed-method review and research agenda for healthcare and beyond. *Artif. Intell. Rev.* **2025**, *58*, 356. [CrossRef]
17. Berretta, S.; Tausch, A.; Ontrup, G.; Gilles, B.; Peifer, C.; Kluge, A. Defining human-AI teaming the human-centered way: A scoping review and network analysis. *Front. Artif. Intell.* **2023**, *6*, 1250725. [CrossRef] [PubMed]
18. National Academies of Sciences; Medicine, E. *Human-AI Teaming: State of the Art and Research Needs*; The National Academies Press: Washington, DC, USA, 2021.
19. Cuevas, H.M.; Fiore, S.M.; Caldwell, B.S.; Strater, L. Augmenting team cognition in human-automation teams performing in complex operational environments. *Aviat. Space Environ. Med.* **2007**, *78*, B63–B70. [PubMed]
20. Friedman, C.P. A "Fundamental Theorem" of Biomedical Informatics. *J. Am. Med. Inform. Assoc.* **2009**, *16*, 169–170. [CrossRef] [PubMed]
21. Kamar, E. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA; pp. 4070–4073.
22. Dellermann, D.; Ebel, P.; Söllner, M.; Leimeister, J.M. Hybrid Intelligence. *Bus. Inf. Syst. Eng.* **2019**, *61*, 637–643. [CrossRef]
23. Akata, Z.; Balliet, D.; De Rijke, M.; Dignum, F.; Dignum, V.; Eiben, G.; Fokkens, A.; Grossi, D.; Hindriks, K.; Hoos, H.; et al. A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect with Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer* **2020**, *53*, 18–28. [CrossRef]
24. Hemmer, P.; Schemmer, M.; Vössing, M.; Kühl, N. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. In Proceedings of the 25th Pacific Asia Conference on Information System, Dubai, United Arab Emirates, 12–14 July 2021; Volume 78, p. 118.
25. Donahue, K.; Chouldechova, A.; Kenthapadi, K. Human-algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA; pp. 1639–1656.
26. Wilson, H.J.; Daugherty, P.R. Collaborative Intelligence: Humans and AI are joining Forces. *Harv. Bus. Rev.* **2018**, *96*, 114–123.
27. Vaccaro, M.; Almaatouq, A.; Malone, T. When Combinations of Humans and AI are Useful: A Systematic Review and Meta-analysis. *Nat. Hum. Behav.* **2024**, *8*, 2293–2303. [CrossRef] [PubMed]

28. Skitka, L.J.; Mosier, K.L.; Burdick, M. Does Automation Bias Decision-making? *Int. J. Hum.-Comput. Stud.* **1999**, *51*, 991–1006. [CrossRef]
29. Buçinca, Z.; Malaya, M.B.; Gajos, K.Z. To Trust or to Think: Cognitive Forcing Functions can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* **2021**, *5*, 1–21. [CrossRef]
30. Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M.S.; Krishna, R. Explanations can Reduce Overreliance on AI Systems during Decision-making. *Proc. ACM Hum.-Comput. Interact.* **2023**, *7*, 1–38. [CrossRef]
31. Liu, P.; Zhang, J.; Chen, S.; Chen, S. Human-AI teaming in healthcare: 1+ 1> 2? *npj Artif. Intell.* **2025**, *1*, 47. [CrossRef]
32. Cabitza, F.; Campagner, A.; Fregosi, C.; Cameli, M.; Gallazzi, E.; Sconfienza, L.M.; Tontini, G.E. Five Degrees of Separation: Investigating the Unexpected Potential of Displaced Human-AI Collaboration Protocols for Apter AI Support. *Proc. ACM Hum.-Comput. Interact.* **2025**, *9*, 1–28. [CrossRef]
33. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias—A hidden issue for clinical decision support system use. In *International Perspectives in Health Informatics*; IOS Press: Amsterdam, The Netherlands, 2011; pp. 17–22.
34. Arrow, K.J.; Sen, A.; Suzumura, K. *Handbook of Social Choice and Welfare*; Elsevier: Amsterdam, The Netherlands, 2010; Volume 2.
35. Kameda, T.; Toyokawa, W.; Tindale, R.S. Information aggregation and collective intelligence beyond the wisdom of crowds. *Nat. Rev. Psychol.* **2022**, *1*, 345–357. [CrossRef]
36. Dahlgren Lindström, A.; Methnani, L.; Krause, L.; Ericson, P.; de Rituerto de Troya, Í.M.; Coelho Mollo, D.; Dobbe, R. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. *Ethics Inf. Technol.* **2025**, *27*, 28. [CrossRef] [PubMed]
37. Glickman, M.; Sharot, T. How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nat. Hum. Behav.* **2025**, *9*, 345–359. [PubMed]
38. Amershi, S.; Cakmak, M.; Knox, W.B.; Kulesza, T. Power to the people: The role of humans in interactive machine learning. *AI Mag.* **2014**, *35*, 105–120. [CrossRef]
39. Vicente, L.; Matute, H.; Fregosi, C.; Cabitza, F. Machine learning systems as mentors in human learning: A user study on machine bias transmission in medical training. *Int. J. Hum.-Comput. Stud.* **2025**, *198*, 103474. [CrossRef]
40. Kasparov, G. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*; Hachette: London, UK, 2017.
41. Cabitza, F.; Campagner, A.; Sconfienza, L.M. Studying Human-AI Collaboration Protocols: The Case of the Kasparov's Law in Radiological Double Reading. *Health Inf. Sci. Syst.* **2021**, *9*, 8. [CrossRef] [PubMed]
42. Brynjolfsson, E. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*; WW Norton Company: New York, NY, USA, 2014; Volume 236.
43. Lai, V.; Chen, C.; Liao, Q.V.; Smith-Renner, A.; Tan, C. Towards a science of human-ai decision making: A survey of empirical studies. *arXiv* **2021**, arXiv:2112.11471.
44. Geijer, H.; Geijer, M. Added Value of Double Reading in Diagnostic Radiology, a Systematic Review. *Insights Into Imaging* **2018**, *9*, 287–301. [CrossRef] [PubMed]
45. Baier, P.; DeLallo, D.; Sviokla, J.J. Your Organization isn't Designed to Work with GenAI. *Harv. Bus. Rev.* **2024**, *26*.
46. Schenker, N.; Gentleman, J.F. On judging the significance of differences by examining the overlap between confidence intervals. *Am. Stat.* **2001**, *55*, 182–186. [CrossRef]
47. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
48. Cicchetti, D.V. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychol. Assess.* **1994**, *6*, 284. [CrossRef]
49. Cotter, D.A.; Hermsen, J.M.; Ovadia, S.; Vanneman, R. The Glass Ceiling Effect. *Soc. Forces* **2001**, *80*, 655–681. [CrossRef]
50. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]