Validating Kasparov's Law Through Human–AI Collaboration in Clinical Diagnosis

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

[1] *Università degli Studi di Milano-Bicocca, Milan, Italy*

[2] *IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milan, Italy*

Abstract

Artificial intelligence (AI) is increasingly integrated into clinical practice through Clinical Decision Support Systems (CDSSs). While such systems can approach or even surpass human expert performance, their true value lies in how they enhance human–AI teams. This study investigates whether structured interaction protocols can improve diagnostic accuracy in a simulated radiology double-reading task. Sixteen radiologists collaborated with an AI system under eight different coordination strategies. Results demonstrate that protocols such as Accuracy-Oriented, Confidence-Oriented, and Presumptuous produced the highest overall accuracy (up to 97% among strong clinicians and 92% among weak ones), outperforming majority voting and single-metric-optimized approaches. Critically, weaker clinicians with superior protocols outperformed stronger clinicians with inferior ones, validating Kasparov's Law: *Weak Human + Machine + Better Process > Strong Human + Machine + Inferior Process.* These findings highlight process design as central to effective CDSS deployment, advocating for a paradigm shift toward process-centric evaluation and design.

Keywords

Interaction Protocols, Kasparov's Law, Human–AI Collaboration, Coordination, Hybrid Intelligence, Double Reading, Clinical Decision Support Systems

1. Introduction

AI-based CDSSs are increasingly deployed in healthcare, achieving expert-level accuracy in diagnostic tasks [1]. However, algorithmic performance alone does not determine clinical impact. As Friedman's *Fundamental Theorem of Biomedical Informatics* notes, the true unit of evaluation is the human–AI partnership [2]. This principle urges a shift from evaluating standalone AI to assessing how it integrates with human workflows.

Despite progress, many studies treat AI as an autonomous Greek oracle rather than a decision support tool [3]. In contrast, the reorientation towards hybrid intelligence has spurred interest in human–AI interaction protocols—structured mechanisms for exchanging information, managing disagreement, and coordinating decisions. These range from parallel schemes, where human and AI independently produce full decisions later integrated in either human-first or AI-first orderings, to collaborative models that decompose complex tasks, assigning subtasks to humans and AI according to their respective strengths [4, 5].

Kasparov's Law, articulated by Brynjolfsson and McAfee [6], formalizes that process quality can be a decisive determinant of human–AI team effectiveness:

$$\text{Weak Human} + \text{Machine} + \text{Better Process} > \text{Strong Human} + \text{Machine} + \text{Inferior Process}$$

Kasparov's Law offers a compelling conceptual anchor for advancing process-centered evaluation frameworks, emphasizing that synergy emerges not from the mere aggregation of individual competencies, but from the quality of their orchestration within the interactional system.

HAIC 2025 - First International Workshop on Human-AI Collaborative Systems, editors Michele Braccini, Allegra De Filippo, Michela Milano, Alessandro Saffiotti, Mauro Vallati; October 25, 2025, Bologna, Italy.

[*] Corresponding author.

[†] These authors contributed equally.

0009-0007-8210-686X (A. Papale); 0009-0004-9409-3865 (G. Lopiano); 0000-0002-0027-5157 (A. Campagner); 0000-0002-4065-3415 (F. Cabitza)

The study empirically examines this principle in radiological diagnosis, testing whether superior coordination protocols can enable weaker clinicians to outperform stronger peers operating under inferior processes.

2. Background and Related Works

CDSSs have evolved from early rule-based systems like MYCIN [7] to contemporary machine learning models able to detect complex patterns in large-scale data [1, 8, 9, 10]. However, adoption remains limited due to barriers such as automation bias [11, 12], trust deficits, interpretability challenges [13], and risks of professional deskilling [14].

Recent research emphasizes the paradigm of hybrid intelligence [15, 16, 17], where AI contributes computational precision, scalability, and pattern recognition, while humans contribute contextual reasoning and ethical judgment [18, 19]. Still, systematic evidence on structured coordination is scarce. Meta-analyses indicate that genuine human–AI synergy, where the joint performance surpasses that of both the human and the AI individually, remains relatively rare, whereas instances of human augmentation, in which the combined performance merely exceeds that of the human alone, are considerably more frequent [20].

Few studies have tested structured interaction protocols as key determinants of performance [21]. This gap motivates our empirical focus.

3. Methods

3.1. Participants and Data

We recruited 16 board-certified radiologists, each evaluating 18 orthopedic X-rays for fracture presence and confidence. Accuracy divided participants into "strong" (above-median accuracy) and "weak" (below-median accuracy) raters, confirmed by Intraclass Correlation Coefficients [22]. AI advice was simulated with fixed accuracy (0.89) and calibrated confidence.

3.2. Interaction Protocols

Once we collected the responses and performance metrics for both the human raters and the simulated AI-based DSS, we implemented eight coordination strategies inspired by double-reading frameworks [23]:

- **Simple-Majority**: Two observers judge; if they disagree, a third intervenes. Final decision by majority vote.
- **Accuracy-Oriented**: Three observers judge. Agreement is final; disagreement resolved by higher average accuracy (pair vs. dissenter).
- **Confidence-Oriented**: Three observers judge. Agreement is final; disagreement resolved by higher average confidence (pair vs. dissenter).
- **Specificity-Oriented**: Second judges only if the first flags the case as abnormal; third intervenes if disagreement. Final decision by higher confidence-weighted specificity (pair vs. dissenter).
- **Sensitivity-Oriented**: Second judges only if the first flags the case as normal; third intervenes if disagreement. Final decision by higher confidence-weighted sensitivity (pair vs. dissenter).
- **Cautious**: Two observers judge. Agreement accepted only if both exceed accuracy and confidence thresholds; else third intervenes. Final decision by higher confidence-weighted accuracy (pair vs. dissenter).
- **Presumptuous**: Two observers judge. If both exceed accuracy threshold, final decision is from the more confident; otherwise from the more accurate.
- **AND_rule**: Second observer judges only if first flags the case as abnormal; final decision is second's judgment.

- **OR_rule**: Second observer judges only if first flags the case as normal; final decision is second's judgment.

Through these simulations, we investigated how alternative cooperation models affect diagnostic accuracy, sensitivity, and specificity, and how the design of interaction protocols shapes clinician performance across different levels of expertise.

3.3. Testing Kasparov's Law

Protocols were classified as superior or inferior based on diagnostic performance (accuracy, sensitivity, and specificity), considering if they exhibited at least two metrics higher or lower than the baseline performance of human clinicians.

We compared outcomes of weak clinicians using superior protocols against strong clinicians using inferior protocols, testing the hypothesis that process quality outweighs individual skill.

4. Results

4.1. Overall Performance

Accuracy-Oriented, Confidence-Oriented, and Presumptuous protocols achieved the highest accuracy (0.91–0.92), significantly outperforming Simple-Majority (0.82). Single-metric strategies achieved extreme sensitivity or specificity but at the cost of balanced performance. AND/OR rules also performed well in their targeted dimensions.

4.2. Strong vs. Weak Clinicians

Strong clinicians outperformed their weaker counterparts, but showed limited gains, with occasional declines under Sensitivity- and Specificity-Oriented strategies. In contrast, weak clinicians exhibited consistent improvements.

4.3. Kasparov's Law

As illustrated in Figure 1, low-performing clinicians using better-designed interaction protocols achieved significantly superior outcomes compared to high-performing clinicians using less effective protocols.

5. Discussion

Findings highlight that coordination design decisively shapes human–AI team outcomes. While strong clinicians show ceiling effect [24], weaker clinicians benefit significantly from structured protocols, narrowing expertise gaps. This suggests that CDSS design should focus not only on predictive accuracy but also on process optimization.

Our results align with prior findings on augmentation [20], showing that process design elevates weaker performers and ensures more equitable outcomes. This has implications for medical training, regulation, and the allocation of responsibility in AI-assisted care.

Limitations include reliance on simulations, simplified protocols, and a limited dataset. Nevertheless, results provide strong conceptual and empirical grounding for process-centered evaluation. Future research should validate these findings in real-world clinical settings, extend protocols to diverse domains and multi-label tasks, and integrate richer interaction designs. Broader ethical and regulatory implications also warrant attention.

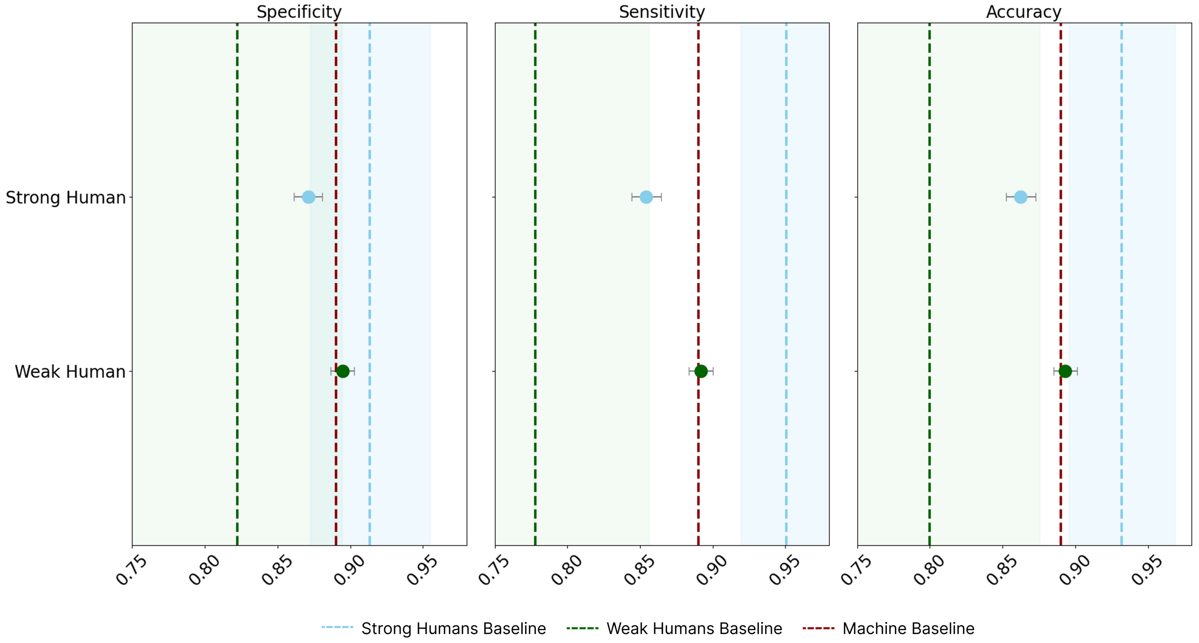

Figure 1: Accuracy, sensitivity, and specificity scores of the group of “weak” people evaluated with strong protocols and the group of “strong” people valuated with weak protocols, along with their 95% confidence intervals represented as error bars. Dashed lines represent the baseline performance of the AI and the average baseline performance of human clinicians when operating independently. The lighter-colored bands surrounding the lines indicate the corresponding confidence intervals (95%).

6. Conclusions

This study demonstrates that structured interaction protocols significantly influence human–AI diagnostic performance. Superior processes enabled weaker clinicians to outperform stronger ones using inferior strategies, providing direct empirical support for Kasparov’s Law. These findings advocate for a paradigm shift in CDSS design: from model-centric approaches to process-centric frameworks, where structured collaboration is treated as a core determinant of clinical effectiveness.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o for checking grammar and spelling, paraphrasing and rewording, improving writing style. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

Disclosure of Interests

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 supported by other PNRR funds (CUP: H41J24000400002) and partially supported by the MUR under the grant “Dipartimenti di Eccellenza 2023-2027“ of the Department of Informatics, Systems, and Communication of the University of Milano-Bicocca, Italy.

References

[1] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, K. I. Kroeker, An overview of clinical decision support systems: benefits, risks, and strategies for success, NPJ digital medicine 3 (2020) 17.

[2] C. P. Friedman, A “fundamental theorem” of biomedical informatics, Journal of the american medical Informatics association 16 (2009) 169–170.

[3] R. A. Miller, F. E. Masarie, The demise of the “greek oracle” model for medical diagnostic systems, Methods of information in medicine 29 (1990) 1–2.

[4] I. Seeber, E. Bittner, R. O. Briggs, T. De Vreede, G.-J. De Vreede, A. Elkins, R. Maier, A. B. Merz, S. Oeste-Reiß, N. Randrup, et al., Machines as teammates: A research agenda on ai in team collaboration, Information & management 57 (2020) 103174.

[5] F. Cabitza, A. Campagner, L. Ronzio, M. Cameli, G. E. Mandoli, M. C. Pastore, L. M. Sconfienza, D. Folgado, M. Barandas, H. Gamboa, Rams, hounds and white boxes: investigating human–ai collaboration protocols in medical diagnosis, Artificial Intelligence in Medicine 138 (2023) 102506.

[6] E. Brynjolfsson, The second machine age: Work, progress, and prosperity in a time of brilliant technologies, volume 236, WW Norton Company, 2014.

[7] E. Shortliffe, Computer-based medical consultations: MYCIN, Elsevier, 1976.

[8] J. Van der Laak, G. Litjens, F. Ciompi, Deep learning in histopathology: the path to the clinic, Nature medicine 27 (2021) 775–784.

[9] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, jama 316 (2016) 2402–2410.

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

[11] R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, Human factors 39 (1997) 230–253.

[12] R. Khera, M. A. Simon, J. S. Ross, Automation bias and assistive ai: risk of harm from ai-driven clinical decision support, Jama 330 (2023) 2255–2257.

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

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

[15] E. Kamar, Directions in hybrid intelligence: Complementing ai systems with human intelligence., in: IJCAI, 2016, pp. 4070–4073.

[16] D. Dellermann, P. Ebel, M. Söllner, J. M. Leimeister, Hybrid intelligence, Business & Information Systems Engineering 61 (2019) 637–643.

[17] Z. Akata, D. Balliet, M. De Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, et al., A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence, Computer 53 (2020) 18–28.

[18] P. Hemmer, M. Schemmer, M. Vössing, N. Kühl, Human-ai complementarity in hybrid intelligence systems: A structured literature review., PACIS 78 (2021) 118.

[19] K. Donahue, A. Chouldechova, K. Kenthapadi, Human-algorithm collaboration: Achieving complementarity and avoiding unfairness, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1639–1656.

[20] M. Vaccaro, A. Almaatouq, T. Malone, When combinations of humans and ai are useful: A systematic review and meta-analysis, Nature Human Behaviour (2024) 1–11.

[21] F. Cabitza, A. Campagner, L. M. Sconfienza, Studying human-ai collaboration protocols: the case of the kasparov’s law in radiological double reading, Health information science and systems 9 (2021) 1–20.

[22] T. K. Koo, M. Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, Journal of Chiropractic Medicine 15 (2016) 155–163.

[23] H. Geijer, M. Geijer, Added value of double reading in diagnostic radiology, a systematic review, Insights into imaging 9 (2018) 287–301.

[24] D. A. Cotter, J. M. Hermsen, S. Ovadia, R. Vanneman, The glass ceiling effect, Social forces 80 (2001) 655–681.