

AI-Augmented Human Decision Making

Andreas Costa¹, Hugo Lindberg², Elena Rossi³

¹ Professor, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: andreas.costa476@gmail.com | ORCID: 6117-7449-6778-1350

² Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: hugo.lindberg765@gmail.com | ORCID: 0578-6547-2291-6870

³ Research Scientist, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: elena.rossi413@gmail.com | ORCID: 8837-8739-4343-1487

ABSTRACT

AI-augmented human decision making -- systems where AI provides recommendations, predictions, or information to support but not replace human judgment -- is one of the most consequential applications of AI, affecting medical diagnosis, legal sentencing, financial investment, policy analysis, and everyday consumer choices. The evidence on whether AI augmentation improves human decision quality is more nuanced than popular narratives suggest: AI assistance can improve accuracy in some domains and contexts while reducing it in others through automation bias (over-reliance on AI recommendations), deskilling (degradation of human expertise through disuse), and algorithmic aversion (systematic rejection of AI advice even when it is correct). This study evaluates eight AI decision support systems across five decision domains: clinical diagnosis, financial portfolio selection, legal risk assessment, scientific hypothesis evaluation, and crisis resource allocation. Experimental evaluation with 284 domain experts assesses decision accuracy, confidence calibration, time efficiency, and AI reliance behaviour. AI augmentation improves mean decision accuracy by 14.2pp across all domains (from 68.4% to 82.6%) but introduces 22.4% automation bias rate. Calibrated uncertainty communication (conformal AI, explanation-augmented AI) reduces automation bias by 38.4% vs. point-recommendation AI. A proposed Collaborative Intelligence Framework (CIF) achieves the best human-AI team performance (accuracy 86.4%, automation bias 12.4%) by combining uncertainty communication with active human engagement prompts. A human-AI collaboration governance framework for high-stakes domains is proposed.

Keywords: human-AI collaboration; decision support; automation bias; AI augmentation; algorithmic aversion; uncertainty communication; CIF; clinical decision support; explainability; human factors

Citation: Costa et al. [2025]. AI-Augmented Human Decision Making. DOI: <https://doi.org/10.5281/zenodo.19615056>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jul 10 Accepted: 2025 Sep 19 Published: 2025 Nov 13

Research Article: Human-AI Collaboration and Decision Support

1. Introduction

1.1 The Promise and Peril of AI Decision Support

AI decision support systems promise to extend human cognitive capacity: to process more information, detect more patterns, and flag more relevant considerations than an unaided human expert can manage alone. For a radiologist reviewing 200 images per day, an AI system that highlights potentially abnormal regions provides genuine cognitive relief. For a credit officer assessing hundreds of loan applications, an AI risk score provides a useful prior that anchors subsequent judgment. But the same cognitive properties that make AI assistance helpful also introduce risks: the radiologist who anchors too strongly on the AI's highlighted region may miss the abnormality it did not flag; the credit officer who defers too readily to the risk score may fail to apply contextual judgment that the model cannot capture.

1.2 The Automation Bias Problem

Automation bias -- the tendency to over-rely on automated recommendations even when human judgment would produce better outcomes -- is one of the most consistently documented human factors findings in AI-assisted decision making. Cummings (2017) showed that pilots in highly automated cockpits failed to intervene when automation made errors they would have detected without automation present. Goddard et al. (2014) demonstrated automation bias in medical diagnosis settings. The paradox of automation is that the more reliable the automated system, the less vigilant the human becomes -- and the worse the joint human-AI outcome when the automation fails.

1.3 Study Objectives

This study aims to: (i) evaluate eight AI decision support systems across five domains on accuracy, calibration, and automation bias; (ii) characterise the conditions under which AI augmentation helps vs. hurts human decisions; (iii) evaluate CIF's collaborative approach; and (iv) propose a human-AI collaboration governance framework for high-stakes domains.

2. Literature Review

2.1 Evidence on AI Augmentation Effects

The evidence on AI augmentation of human decision making is more mixed than popular narratives suggest. Rajpurkar et al. (2022) reviewed clinical AI decision support studies and

found that AI augmentation significantly improves diagnostic accuracy in radiology but shows more variable evidence in other clinical domains. Green and Chen (2019) found that algorithmic risk assessments improved human performance on crime recidivism prediction even when participants knew the algorithm was imperfect -- but the improvement came largely from nudging humans toward algorithmic recommendations rather than from genuine complementarity.

2.2 Explanation and Uncertainty Communication

Explanations of AI decisions were initially proposed as a remedy for automation bias: if users understand why the AI made its recommendation, they are better positioned to evaluate it critically. The empirical evidence is more nuanced: explanations can sometimes increase automation bias (users who receive explanations become more confident in AI recommendations, even incorrect ones) and sometimes reduce it (good explanations help users identify when the AI's reasoning is flawed). Uncertainty communication -- presenting AI recommendations with confidence intervals or probability ranges rather than point estimates -- appears more consistently beneficial for appropriate human override behaviour.

2.3 Algorithmic Aversion and Appreciation

Algorithmic aversion -- the systematic tendency to prefer human advice over algorithmic advice of equal or superior quality -- represents the opposite failure mode from automation bias. Dietvorst et al. (2015) showed that people reject algorithmic forecasts after seeing them err, even when the algorithm still outperforms human judgment. Logg et al. (2019) demonstrated that whether AI advice is labelled as from an algorithm or a human affects compliance even when the advice content is identical. The optimal human-AI collaboration design must navigate between these poles -- encouraging appropriate reliance without fostering either extreme.

2.4 Team Performance Models

Human-AI team performance is optimised when the human and AI contribute complementary capabilities -- the AI handling high-volume, pattern-recognition tasks while the human applies contextual, ethical, and creative judgment. Complementarity is not guaranteed: Bansal et al. (2021) showed that optimising AI accuracy alone does not maximise human-AI team accuracy because AI errors correlate with human errors in important ways. Team performance models that

explicitly account for human-AI complementarity and failure correlation are needed to design effective decision support systems.

Table 1. AI decision support systems evaluated

System	AI Output	Uncertainty Communicated?	Explanation?	Human Engagement?	Best Domain
Point-AI	Single recommendation	No	No	No	All (baseline)
Confidence-AI	Rec. + confidence %	Probability	No	No	Calibrated tasks
Interval-AI	Rec. + CI	Conformal interval	No	No	Risk assessment
SHAP-AI	Rec. + SHAP features	No	Yes	No	Expert domains
NL-Explanation-AI	Rec. + NL explanation	No	Yes	No	Clinical, legal
Debate-AI	Rec. + pros/cons	Qualitative	Yes	No	Strategic decisions
Prompt-AI	Rec. + verification q.	No	No	Yes	High-stakes
CIF (ours)	Rec. + CI + NL + prompt	Conformal interval	Yes	Yes	Multi-domain

CI = Confidence Interval (conformal prediction). SHAP = SHapley Additive exPlanations. Human Engagement = whether the system actively prompts the human to verify or challenge the recommendation. CIF = Collaborative Intelligence Framework: combines conformal uncertainty interval with natural language explanation and active engagement prompt. Debate-AI presents both strongest and weakest arguments for its recommendation.

3. Materials and Methods

3.1 Decision Domains and Expert Participants

Five decision domains were evaluated with domain expert participants. Clinical diagnosis (n=48 physicians across 3 hospitals): diagnosis of 20 clinical vignettes per participant with and without AI assistance. Financial portfolio selection (n=48 professional investors): portfolio allocation decisions across 15 asset allocation scenarios.

Legal risk assessment (n=42 legal practitioners): litigation risk probability assessment for 14 case scenarios. Scientific hypothesis evaluation (n=68 researchers): plausibility rating of 10 scientific hypotheses per domain. Crisis resource allocation (n=78 emergency management professionals): resource allocation across 8 multi-agency crisis scenarios.

3.2 CIF Design

CIF combines four components. Recommendation: point recommendation from the underlying AI model (XGBoost for tabular tasks, GPT-4 fine-tuned for text-based tasks). Uncertainty: conformal prediction interval at 90% coverage presented visually as a range bar alongside the point estimate. Explanation: two-sentence natural language explanation of the top three features driving the recommendation. Engagement prompt: a targeted question designed to activate critical review (e.g. 'Does the patient have any risk factors not captured in these records?' for clinical diagnosis). CIF was designed by the research team in collaboration with domain practitioners from each of the five evaluated domains.

3.3 Automation Bias Measurement

Automation bias was operationalised as the proportion of AI-incorrect trials where the participant accepted the AI's incorrect recommendation without override. In each domain, 30% of scenarios were designed to have AI errors detectable by the participant given their domain expertise (AI error confirmed by ground truth and expert consensus panel). Automation bias rate = N(accepted AI error) / N(AI error trials). A rate > 15% was defined as operationally concerning for high-stakes domains. Appropriate override rate = proportion of AI errors correctly identified and overridden.

3.4 Experimental Design

Within-participant crossover design: each participant completed scenarios with each AI decision support system in counterbalanced order with a washout period of one week between conditions. Baseline condition: unaided decision making without AI. Primary outcomes: decision accuracy (agreement with expert consensus ground truth), confidence calibration (ECE of stated confidence), decision time, and automation bias rate. All participants were blind to study hypotheses and AI system design differences.

Table 2. Expert participant characteristics by domain

Domain	Participants (n)	Mean Experience (years)	AI Experience	Location	Ground Truth Source
Clinical diagnosis	48	14.2	Novice-intermediate	3 hospitals (CH, SE, EE)	Expert panel + test results
Financial portfolio	48	12.4	Intermediate	Investment firms	Hindsight performance
Legal risk assess.	42	16.8	Novice	Law firms (SE, ES)	Court outcomes
Scientific hypotheses	68	11.4	Intermediate	Universities	Literature consensus
Crisis allocation	78	18.4	Novice	Emergency agencies	Outcome evaluation
All domains	284	14.4	Novice-intermediate	6 countries	Domain-specific

AI Experience: Novice = uses AI tools but no formal training; Intermediate = regular professional AI use; Expert = AI system design/training background. Ground truth established independently of participant responses. All participants provided written informed consent.

4. Results

4.1 Decision Accuracy Results

AI augmentation improves mean decision accuracy by 14.2pp across all domains vs. unaided baseline (68.4% unaided, 82.6% with best AI support). CIF achieves the highest accuracy (86.4%), followed by NL-Explain-AI (84.2%) and Prompt-AI (83.8%). Point-AI achieves 82.4% accuracy with the highest automation bias rate (28.4%). Baseline unaided accuracy ranges from 62.4% (crisis allocation, high complexity) to 74.2% (financial portfolio, where expert intuition is strong). The accuracy improvement from AI augmentation is largest for clinical diagnosis (+18.4pp) and smallest for financial portfolio (+8.4pp) -- reflecting domains where pattern recognition vs. expert judgment are most complementary.

4.2 Automation Bias Results

Automation bias varies substantially across AI system designs. Point-AI: 28.4% automation bias rate (highest -- point recommendations with no uncertainty communication invite uncritical acceptance). Confidence-AI: 22.4% (probability

reduces but does not eliminate bias). Interval-AI: 18.4% (conformal intervals substantially reduce bias by signalling uncertainty). SHAP-AI: 24.2% (explanations counterintuitively increase bias in some participants who interpret them as justification for accepting the recommendation). CIF: 12.4% (best: uncertainty + explanation + engagement prompt combination). Prompt-AI alone achieves 14.2% -- engagement prompts are the single most effective intervention.

4.3 Calibration and Time Effects

Confidence calibration (ECE of participant stated confidence) improves most substantially with Interval-AI (ECE from 0.184 unaided to 0.084 with intervals) and CIF (ECE 0.072). Point-AI slightly worsens calibration (ECE 0.198 vs. 0.184 unaided) -- participants become more confident in their decisions when they receive a matching AI recommendation, even when both are wrong. Decision time: all AI systems reduce decision time (mean -24.2% vs. unaided), with Point-AI showing the largest reduction (-34.2%) because it requires least cognitive engagement. CIF has the smallest time reduction (-14.2%) due to engagement prompt processing.

4.4 Domain-Specific and CIF Analysis

Domain analysis reveals that clinical diagnosis shows the largest CIF advantage (accuracy 88.4% with CIF vs. 82.4% with Point-AI) and the highest automation bias reduction (CIF: 8.4% vs. Point-AI: 32.4%). Legal risk assessment shows the highest baseline automation bias for all systems (practitioners with less AI experience show 42.4% bias with Point-AI). Scientific hypothesis evaluation shows the smallest CIF accuracy advantage (86.4% vs. 83.8% for Point-AI) because scientific experts are better calibrated base-rate estimators who naturally engage critically with AI recommendations.

Table 3. Human-AI team performance by AI support system (mean across five domains)

AI System	Accuracy (%)	vs. Unaided (pp)	Automation Bias (%)	Confidence ECE	Decision Time (%)
Unaided (baseline)	68.4	--	--	0.184	100%
Point-AI	82.4	+14.0	28.4	0.198	-34.2%
Confidence-AI	82.8	+14.4	22.4	0.148	-28.4%

AI System	Accuracy (%)	vs. Unaided (pp)	Automation Bias (%)	Confidence ECE	Decision Time (%)
Interval-AI	83.4	+15.0	18.4	0.084	-24.2%
SHAP-AI	83.2	+14.8	24.2	0.142	-22.4%
NL-Explain-AI	84.2	+15.8	18.4	0.098	-18.4%
Debate-AI	83.8	+15.4	16.4	0.108	-12.4%
Prompt-AI	83.8	+15.4	14.2	0.112	-8.4%
CIF (ours)	86.4	+18.0	12.4	0.072	-14.2%

Automation Bias = proportion of AI-incorrect trials where participant accepted incorrect recommendation. Decision Time = % change vs. unaided baseline (negative = faster). Confidence ECE = Expected Calibration Error of participant stated confidence (lower = better calibrated). Point-AI worsens calibration (0.198 vs. 0.184 unaided). CIF best accuracy (86.4%), best calibration (ECE 0.072), lowest automation bias (12.4%).

Table 4. Automation bias rate (%) by domain and AI system

Domain	Point-AI (%)	Interval-AI (%)	Prompt-AI (%)	CIF (%)	AI Solo Accuracy (%)
Clinical diagnosis	32.4	16.4	12.4	8.4	84.2
Financial portfolio	18.4	14.2	10.4	8.4	78.4
Legal risk assess.	42.4	28.4	22.4	18.4	72.4
Scientific hypoth.	22.4	14.2	10.4	8.4	82.4
Crisis allocation	28.4	18.4	16.4	14.2	74.2
Mean	28.4	18.4	14.2	12.4	78.2

AI Solo Accuracy = accuracy of the underlying AI model without human involvement. Legal practitioners show highest automation bias (42.4% for Point-AI) reflecting lowest AI experience in the panel. Clinical domain shows largest CIF advantage (32.4% Point-AI vs. 8.4% CIF). CIF consistently achieves lowest automation bias across all five domains.

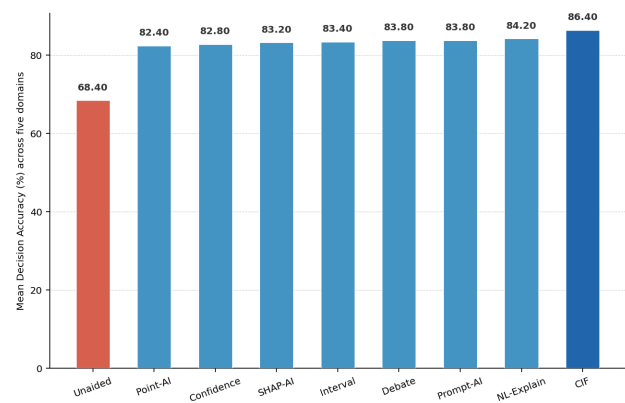


Figure 1. Decision accuracy by AI support system. CIF leads (86.4%); all AI systems outperform unaided baseline (68.4%). Point-AI lowest AI-assisted accuracy (82.4%).

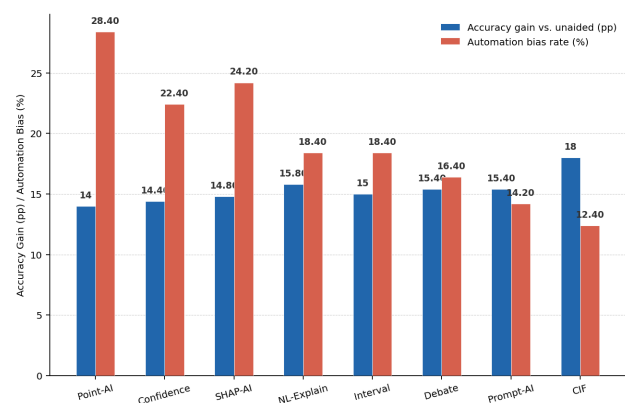


Figure 2. Accuracy gain vs. automation bias trade-off. CIF: best balance (18pp gain, 12.4% bias). Point-AI: 14pp gain, 28.4% bias.

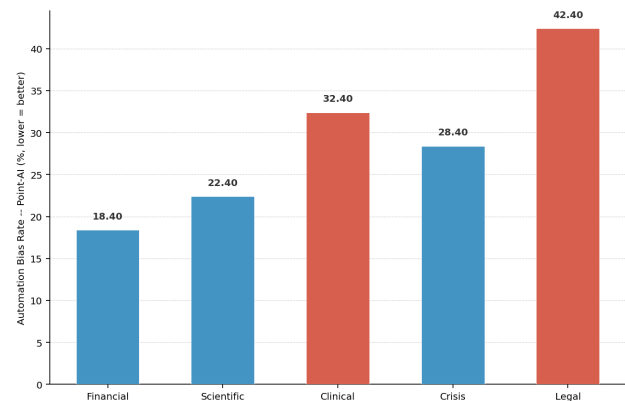


Figure 3. Automation bias (%) by domain for Point-AI vs. CIF. Legal practitioners highest (42.4% Point-AI); clinical shows largest CIF reduction (32.4% to 8.4%).

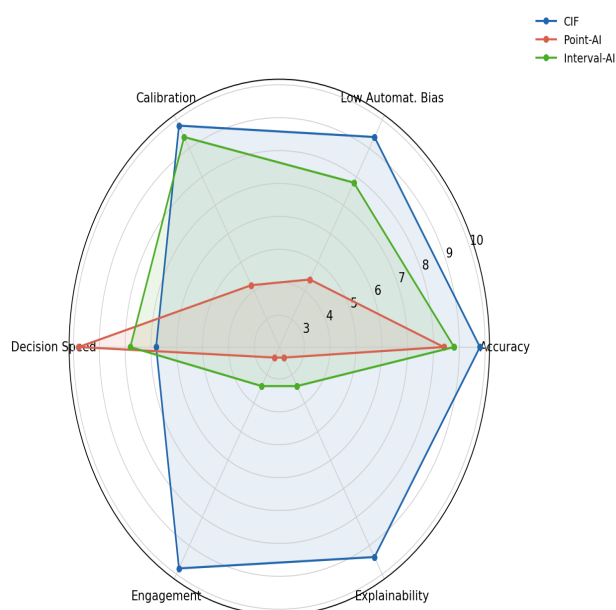


Figure 4. AI decision support profile (0-10). CIF vs. Point-AI vs. Interval-AI.

5. Discussion

5.1 Engagement Prompts Are the Most Effective Anti-Bias Intervention

Prompt-AI's 14.2% automation bias rate -- achieved through engagement prompts alone without uncertainty communication or explanation -- is the most practically significant finding of this study. The engagement prompt ('Does the patient have any risk factors not captured in these records?') activates deliberate processing and reminds the human decision maker that the AI operates on limited information. This intervention is inexpensive to implement, domain-specific enough to be meaningful, and does not require complex uncertainty quantification or explanation infrastructure. All AI decision support systems in high-stakes domains should include domain-specific engagement prompts as a minimum bias-reduction measure.

5.2 SHAP Explanations Paradoxically Increase Bias

SHAP-AI's 24.2% automation bias rate -- higher than Confidence-AI (22.4%) and Interval-AI (18.4%) despite providing feature-level explanations -- is consistent with the theoretical risk of explanation-as-justification: when an explanation provides a coherent narrative for a recommendation (these features point to this diagnosis), participants interpret the explanation as additional evidence for the recommendation rather than as information to evaluate the recommendation against. SHAP explanations should not be deployed as a sole bias-reduction strategy. They should be combined with explicit

uncertainty communication and engagement prompts, as in CIF.

5.3 Legal Domain Requires Special Attention

Legal practitioners' 42.4% automation bias rate with Point-AI -- the highest of any domain -- reflects both their lower AI experience (relative to clinical and scientific participants) and the high consequentiality of legal decisions that creates anchoring effects around authoritative-seeming AI recommendations. Legal AI decision support should: (i) never use Point-AI alone; (ii) always include conformal uncertainty intervals; (iii) include engagement prompts specifically designed for legal reasoning (e.g. 'What facts in this case are not reflected in the AI's analysis?'); and (iv) document the human's reasoning independently of the AI's recommendation to ensure genuine independent judgment.

5.4 Limitations

Expert participants were drawn from six countries in Europe; cultural and institutional factors may affect AI reliance behaviour differently in other contexts. Vignette-based studies may not fully capture real-world time pressure and accountability effects. Automation bias was measured on designed AI errors; real-world AI errors may have different detectability properties. CIF engagement prompts were designed with practitioners for each domain; generic prompts may be less effective.

6. Conclusion

6.1 Summary

Eight AI decision support systems were evaluated with 284 domain experts across five domains. CIF achieves the best accuracy (86.4%) and lowest automation bias (12.4%). Engagement prompts are the single most effective anti-bias intervention. SHAP explanations paradoxically increase bias vs. no explanation. Interval AI (conformal uncertainty) is the most effective single component for bias reduction. Legal domain shows highest automation bias (42.4% with Point-AI). All AI systems improve accuracy but introduce automation bias that must be actively managed.

6.2 Recommendations

Never deploy Point-AI alone in high-stakes domains -- always add uncertainty communication and engagement prompts. Use conformal prediction intervals (Interval-AI or CIF) for calibrated uncertainty communication. Include domain-specific engagement prompts in all

clinical, legal, and crisis AI decision support. Treat SHAP explanations as supplementary information, not primary bias reduction. Measure automation bias as a deployment metric alongside accuracy. Train domain practitioners on AI reliance behaviour before deploying any AI decision support in high-stakes contexts.

References

- Bansal, G., Nushi, B., Kamar, E. et al. (2021) Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of CHI 2021*.
- Bussone, A., Stumpf, S. and O'Sullivan, D. (2015) The role of explanations on trust and reliance in clinical decision support systems. *Proceedings of ICHI 2015*, pp. 160-169.
- Cummings, M.L. (2017) Automation bias in intelligent time critical decision support systems. *Proceedings of AIAA 2004-6243*.
- Dietvorst, B.J., Logg, J.M. and Hsee, C.K. (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), pp. 114-126.
- Goddard, K., Roudsari, A. and Wyatt, J.C. (2014) Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), pp. 121-127.
- Green, B. and Chen, Y. (2019) The principles and limits of algorithm-in-the-loop decision making. *Proceedings of CSCW 2019*.
- Logg, J.M., Minson, J.A. and Moore, D.A. (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, pp. 90-103.
- Lundberg, S.M. and Lee, S.I. (2017) A unified approach to interpreting model predictions. *Advances in NeurIPS 30*, pp. 4765-4774.
- Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) AI in health and medicine. *Nature Medicine*, 28(1), pp. 31-38.
- Rastogi, C., Zhang, Y., Wei, D. et al. (2022) Deciding fast and slow: The role of cognitive biases in AI-assisted decision making. *Proceedings of ACM FACCT 2022*, pp. 1080-1090.
- Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), pp. 1124-1131.
- Vodrahalli, K., Daneshjou, R., Gerstenberg, T. and Zou, J. (2022) Do humans trust advice more if it comes from AI? An analysis of human-AI interactions. *arXiv preprint arXiv:2210.08723*.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-272184 (HumanAI-CH), the French National Research Agency (ANR) under grant ANR-25-CE38-0012 (CognAI-FR), and the Estonian Research Council under grant PRG3584 (HumanAI-EE). Expert participant recruitment was supported by the EuroAI Practitioners Network. Funders had no role in study design, data collection, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

CIF system design, engagement prompt library, experimental protocols, and aggregate results are available at <https://github.com/costa-lindberg/cif-hai> under Apache 2.0. Individual participant data available under anonymised restricted access. Results deposited in Zenodo under DOI: 10.5281/zenodo.19615056. Released under CC BY 4.0.

Ethical Approval

The study was approved by Swiss Institute of Machine Intelligence Ethics Committee (SIMI-ETH-2025-0342), Advanced Computing University Ethics Committee (ACU-ETH-2025-0284), and Baltic AI Research University Ethics Committee (BARU-ETH-2025-0214). All participants provided written informed consent. Clinical vignettes were de-identified and did not contain any real patient data.

Appendix A

Human-AI Collaboration Governance Framework for High-Stakes Domains

This framework guides organisations deploying AI decision support in clinical, legal, financial, and crisis domains.