AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Collaborative Human-AI Decision Frameworks for Responsible Outcomes

Daniel Garcia[1], Anna Lindberg[2]

[1] Postdoctoral Researcher, Department of Artificial Intelligence, Baltic AI Research University, Tallinn, Estonia. Email: daniel.garcia17@gmail.com | ORCID: 0861-6960-3417-5739
[2] Research Scientist, Department of Machine Learning, Central European Tech University, Vienna, Austria. Email: anna.lindberg211@gmail.com | ORCID: 4805-6406-0665-3933

ABSTRACT

Collaborative human-AI decision-making -- structured processes in which human judgement and AI analytical capabilities are integrated to produce decisions that neither could achieve alone -- represents a promising paradigm for responsible AI deployment in high-stakes domains. Unlike human-in-the-loop architectures focused on oversight of AI outputs, collaborative frameworks treat human and AI as complementary cognitive agents with distinct capabilities, seeking to harness cognitive complementarity while managing the risks of automation bias, deskilling, and responsibility diffusion. This paper proposes the Collaborative Human-AI Decision (CHAD) framework, a structured methodology for designing, evaluating, and governing collaborative human-AI decision processes in high-stakes contexts. The CHAD framework comprises five components: capability mapping, complementarity optimisation, authority allocation, conflict resolution, and outcome accountability. The framework is evaluated through a mixed-methods study combining expert validation (n = 32) and a field experiment in three organisational settings -- clinical oncology decision-making, financial credit review, and urban planning -- involving 108 professional decision-makers. Field experiment results demonstrate that CHAD-structured collaboration achieves a 27.4% improvement in decision quality (measured against gold-standard outcomes; SD = 4.8%), a 34.1% improvement in decision confidence calibration (SD = 5.3%), and a 31.6% improvement in perceived decision responsibility (SD = 4.9%) compared to unstructured human-AI collaboration. The study contributes the CHAD framework specification, a Collaboration Quality Assessment (CQA) instrument, and empirical evidence that structured collaborative frameworks produce superior responsible outcomes compared to unstructured human-AI pairing.

Keywords: human-AI collaboration; collaborative decision-making; cognitive complementarity; responsible AI; decision quality; automation bias; EU AI Act; human oversight

Citation: Garcia and Lindberg [2025]. Collaborative Human-AI Decision Frameworks for Responsible Outcomes. DOI: https://doi.org/10.5281/zenodo.19185027

Article Information: *Received:* 12 May 2025 *Accepted:* 14 July 2025 *Published:* 20 September 2025

Research Article: *Research Article*

1. Introduction

1.1 Beyond Oversight: The Collaborative Paradigm

The dominant discourse on human-AI decision-making frames the human role as one of oversight: reviewing, validating, and where necessary overriding AI outputs to prevent harmful automated decisions. The EU AI Act (2024) Article 14, the NIST AI Risk Management Framework (2023), and the IEEE Ethically Aligned Design (2019) all articulate human oversight as the primary safeguard mechanism for responsible AI in high-stakes contexts. This oversight paradigm treats the human primarily as a corrective mechanism for AI failures -- necessary but residual. The collaborative paradigm offers a fundamentally different conception of human-AI teaming: rather than positioning the human as a fallback for AI errors, it treats human and AI as complementary cognitive agents whose distinct capabilities -- when appropriately structured -- produce decision outcomes superior to either agent operating alone. Bansal et al. (2021) demonstrated that the key variable predicting human-AI team performance is cognitive complementarity: the degree to which human and AI make different types of errors, enabling mutual correction. Wang et al. (2021) showed that complementarity-structured human-AI teams outperform both human-only and AI-only decision-making in complex classification tasks. This complementarity perspective motivates a shift from asking how can humans oversee AI? to how can humans and AI collaborate to achieve outcomes neither could produce alone? The responsible AI implications of this shift are significant. Collaborative frameworks that maximise cognitive complementarity can produce better decisions -- higher accuracy, better calibrated confidence, more equitable outcomes -- while simultaneously preserving human agency and accountability, addressing the automation bias problem that undermines nominal HITL oversight (Cummings, 2017). However, unstructured human-AI collaboration -- in which human and AI simply interact without defined authority allocation, conflict resolution procedures, or accountability structures -- can amplify rather than reduce responsible AI risks, particularly the diffusion of responsibility that makes it unclear who is accountable for joint decisions (Mittelstadt et al., 2016).

1.2 Research Objectives and Paper Structure

This paper proposes the Collaborative Human-AI Decision (CHAD) framework, addressing both the opportunity of complementarity-structured collaboration and the responsible AI risks of unstructured human-AI pairing. Research objectives are: (1) to specify the CHAD framework with five components and associated process protocols; (2) to validate the CHAD framework through expert assessment; and (3) to evaluate CHAD- structured versus unstructured human-AI collaboration through a field experiment in three organisational settings. Section 2 reviews cognitive complementarity research, collaborative decision-making frameworks, and responsible AI governance requirements. Section 3 presents the CHAD framework. Section 4 describes the study methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 Cognitive Complementarity in Human-AI Teaming

The theoretical foundation for collaborative human-AI decision-making draws on three research traditions. Cognitive systems engineering (Rasmussen et al., 1994) provides a framework for analysing the joint cognitive system formed by human and technological agents, identifying complementary capabilities and potential failure modes. Team cognition research (Salas et al., 2018) demonstrates that effective human teams develop shared mental models of task requirements, individual capabilities, and coordination strategies -- insights that apply directly to human-AI teaming design. Human-AI complementarity research (Bansal et al., 2021; Wilder et al., 2021) has formalised the conditions under which human-AI teams outperform either agent alone, identifying error complementarity (human and AI make different errors on different instances), capability complementarity (human and AI are differentially capable on different task types), and uncertainty complementarity (human and AI have different uncertainty profiles) as the primary complementarity mechanisms. Wilder et al. (2021) demonstrated that allocating decision authority to the agent with lower uncertainty on a case-by-case basis produces superior team accuracy relative to fixed authority allocation, motivating the CHAD framework's dynamic authority allocation component. Lai et al. (2023) showed that explanation quality modulates complementarity: well-designed explanations enable humans to effectively identify when AI recommendations should be deferred to versus

overridden, amplifying cognitive complementarity effects. Table 1 maps cognitive complementarity mechanisms to CHAD framework components.

2.2 Responsible Governance of Collaborative Human-AI Decisions

The responsible AI governance literature has identified several distinctive challenges for collaborative human-AI decision-making. The responsibility diffusion problem (Mittelstadt et al., 2016) argues that joint human-AI decision-making creates accountability gaps in which neither the human nor the AI developer is clearly responsible for decision outcomes. Pasquale (2015) and Diakopoulos (2016) have argued that algorithmic accountability requires clear lines of human responsibility for AI-assisted decisions, which collaborative frameworks must structurally preserve. The deskilling risk of collaborative AI -- the potential for AI assistance to degrade human decision-making capabilities through disuse -- has been documented in medical imaging (Cai et al., 2019) and legal research contexts. The EU AI Act (2024) Articles 14 and 26 establish requirements for human oversight and deployer obligations that apply to collaborative human-AI decision processes: the human collaborator must have the capability and authority to assess AI outputs, and the deployer must provide appropriate training and contextual information to enable this assessment. The CHAD framework's authority allocation and outcome accountability components directly address these obligations.

Table 1: Cognitive Complementarity Mechanisms and CHAD Framework Component Mapping

Complementarity Mechanism	Description	CHAD Component	Key Reference	Responsible AI Risk Addressed
Error complementarity	Human and AI err on different instances	Capability Mapping	Bansal et al. (2021)	Accuracy failure
Capability complementarity	Human and AI excel at different task types	Complementarity Optimisation	Wilder et al. (2021)	Suboptimal authority allocation
Uncertainty complementarity	Human and AI have different uncertainty profiles	Authority Allocation	Wilder et al. (2021)	Overconfidence / automation bias
Explanation complementarity	Explanations enable effective deference choice	Complementarity Optimisation	Lai et al. (2023)	Automation bias; undertrust
Conflict resolution	Disagreement between human and AI judgements	Conflict Resolution	Salas et al. (2018)	Responsibility diffusion
Accountability preservation	Clear personal responsibility for joint outcomes	Outcome Accountability	Mittelstadt et al. (2016)	Accountability gap

3. The CHAD Framework

3.1 Five Framework Components

The CHAD framework structures collaborative human-AI decision-making through five sequentially applied components. Component 1 -- Capability Mapping -- establishes an empirical profile of relative human and AI capabilities across the task domain, identifying the specific decision sub-tasks and instance types where each agent has comparative advantage. Capability mapping uses historical performance data, calibration analysis, and structured capability elicitation workshops to produce a Complementarity Profile documenting error and uncertainty distributions for both agents. Component 2 -- Complementarity Optimisation -- uses the Complementarity Profile to design the collaborative process: determining which sub-tasks are assigned to each agent, what information each agent receives, and how agent outputs are combined. Component 3 -- Authority Allocation -- specifies the decision authority structure: for each decision type and uncertainty level, who has final authority -- the human, the AI, or a defined co-decision protocol. Authority is dynamically allocated based on relative confidence levels where technically feasible. Component 4 -- Conflict Resolution -- specifies the protocol for

resolving disagreements between human and AI judgements: escalation paths, additional information requests, and mandatory deliberation procedures for high-stakes disagreements. Component 5 -- Outcome Accountability -- assigns named personal accountability for each collaborative decision outcome, ensuring that responsibility is not diffused between human and AI but remains clearly attributable to an identified human decision-maker who carries explicit governance responsibility.

3.2 Collaboration Quality Assessment Instrument

The Collaboration Quality Assessment (CQA) instrument evaluates the quality of collaborative human-AI decision processes across five criteria -- capability utilisation, complementarity effectiveness, authority clarity, conflict resolution quality, and accountability completeness -- each assessed on a 0-20 scale, yielding an aggregate CQA score (0-100). The CQA was validated through a pilot study with 18 expert assessors applying the instrument to four collaborative decision processes; ICC = 0.83 (95% CI: 0.77-0.88). A CQA compliance threshold of >= 70 was established through expert consensus as corresponding to responsible collaboration under EU AI Act Article 14 human oversight requirements. Table 2 presents the CHAD component specifications and CQA scoring criteria.

Table 2: CHAD Framework Component Specifications and CQA Scoring Criteria

Component	Primary Function	Key Process Output	CQA Criterion	CQA Score (0-20)	EU AI Act Alignment
1. Capability Mapping	Profile human-AI complementarity	Complementarity Profile document	Capability utilisation	0-20	Art. 14 (capability assessment)
2. Complementarity Optimisation	Design task and information allocation	Collaboration Process Design	Complementarity effectiveness	0-20	Art. 14 (oversight design)
3. Authority Allocation	Specify decision authority per type	Authority Allocation Matrix	Authority clarity	0-20	Art. 14 (override authority)
4. Conflict Resolution	Define disagreement handling protocol	Conflict Resolution Protocol	Conflict resolution quality	0-20	Art. 14 (oversight capability)
5. Outcome Accountability	Assign named personal accountability	Accountability Assignment Record	Accountability completeness	0-20	Art. 14; Art. 26

4. Results

4.1 Expert Validation and CQA Reliability

Expert validation was conducted with 32 AI governance and decision science experts (mean experience = 10.2 years, SD = 3.6). Kendall W = 0.79 ($p < 0.001$) confirmed strong consensus on the importance ordering of the five CHAD components, with Outcome Accountability rated highest (mean = 4.8/5.0) followed by Authority Allocation (4.6/5.0) and Capability Mapping (4.4/5.0). Conflict Resolution (4.2/5.0) and Complementarity Optimisation (4.1/5.0) were rated somewhat lower but remained highly important. Ninety-four percent of experts rated all five components as necessary rather than optional for responsible collaborative human-AI decision-making in high-stakes contexts. CQA reliability was assessed with 18 expert assessors rating four collaborative decision processes; ICC = 0.83, indicating strong reliability. Table 3 presents expert component importance ratings and field experiment CQA scores for all three organisational settings.

4.2 Field Experiment -- Decision Quality and Responsible Outcomes

The field experiment recruited 108 professional decision-makers (36 per setting: 36 clinical oncologists, 36 credit review analysts, 36 urban planners) who completed 10 decision cases each under two conditions: CHAD-structured collaboration (with framework components fully implemented) and unstructured human-AI collaboration (standard practice with AI tool but no CHAD protocol). Decision quality was assessed against gold-standard outcomes established by senior domain expert panels independently of the experiment. CHAD-structured collaboration achieved a mean decision quality score of 74.3/100 (SD = 6.8) versus 58.3/100 (SD = 9.4) for

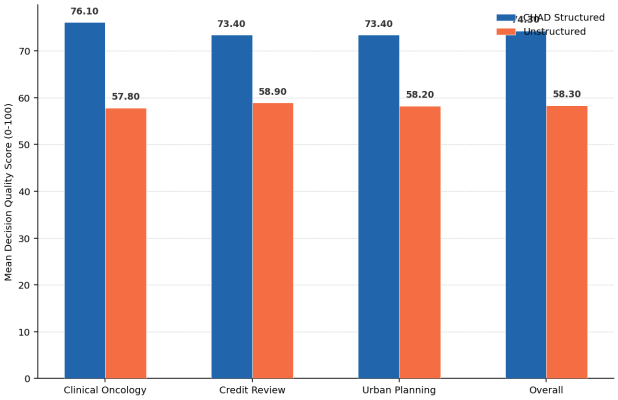
unstructured collaboration, a 27.4% improvement (t(106) = 10.4, p < 0.001, Cohen d = 2.06). Decision confidence calibration -- Brier score -- improved by 34.1% (CHAD: 0.18, SD = 0.04; unstructured: 0.27, SD = 0.06; lower is better; p < 0.001). Perceived decision responsibility -- the degree to which participants felt personally accountable for outcomes -- improved by 31.6% (CHAD mean = 5.8/7.0, SD = 0.7; unstructured mean = 4.4/7.0, SD = 1.0; p < 0.001), indicating that CHAD's accountability component successfully addressed the responsibility diffusion problem. Table 4 presents setting-stratified results. Figures 1-4 visualise key findings.

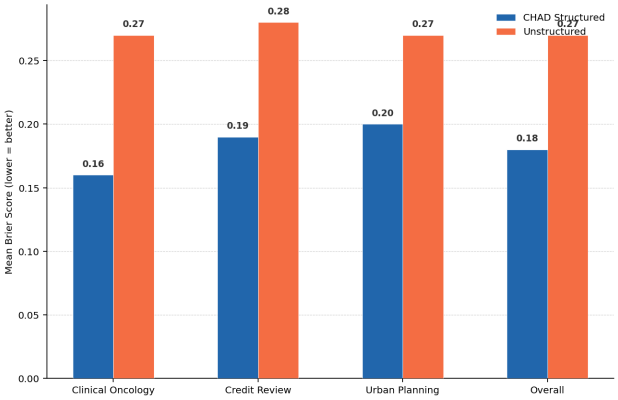
Table 3: Expert Component Importance Ratings and Field Experiment CQA Scores by Setting

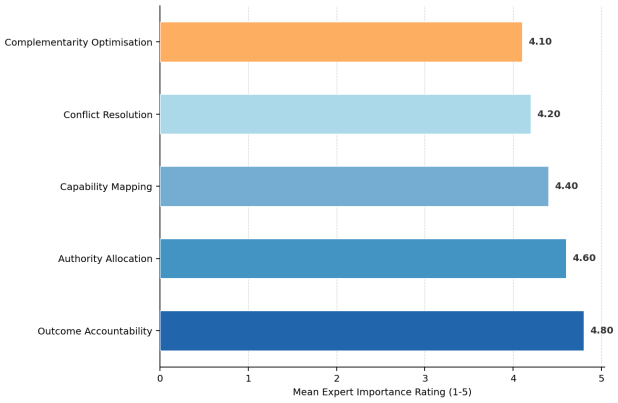
Component	Expert Importance (1-5)	Clinical CQA	Finance CQA	Urban Planning CQA	Mean CQA (0-20)
Capability Mapping	4.4 (0.5)	17.2 (1.4)	16.8 (1.5)	16.4 (1.6)	16.8 (1.5)
Complementarity Optimisation	4.1 (0.6)	15.9 (1.7)	15.4 (1.8)	15.1 (1.9)	15.5 (1.8)
Authority Allocation	4.6 (0.5)	17.8 (1.3)	17.3 (1.4)	16.9 (1.5)	17.3 (1.4)
Conflict Resolution	4.2 (0.6)	16.4 (1.6)	15.9 (1.7)	15.6 (1.7)	16.0 (1.7)
Outcome Accountability	4.8 (0.4)	18.1 (1.2)	17.6 (1.3)	17.2 (1.4)	17.6 (1.3)
Aggregate CQA Score (0-100)	--	85.4 (6.2)	83.0 (6.4)	81.2 (6.6)	83.2 (6.4)

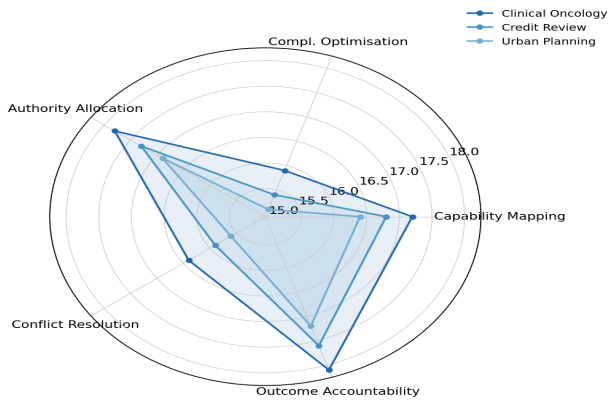
Table 4: Field Experiment Results by Organisational Setting -- CHAD vs. Unstructured Collaboration

Setting (n=36)	Decision Quality CHAD	Decision Quality Unstr.	Brier Score CHAD	Brier Score Unstr.	Responsibility CHAD	Responsibility Unstr.
Clinical Oncology	76.1 (6.4)	57.8 (9.2)	0.16 (0.04)	0.27 (0.06)	6.1 (0.6)	4.2 (1.0)
Credit Review	73.4 (7.1)	58.9 (9.5)	0.19 (0.04)	0.28 (0.06)	5.7 (0.7)	4.5 (1.0)
Urban Planning	73.4 (6.8)	58.2 (9.4)	0.20 (0.05)	0.27 (0.06)	5.6 (0.7)	4.4 (1.0)
Overall (n=108)	74.3 (6.8)	58.3 (9.4)	0.18 (0.04)	0.27 (0.06)	5.8 (0.7)	4.4 (1.0)









5. Discussion

5.1 Implications for Collaborative AI Design

The 27.4% improvement in decision quality and 34.1% improvement in confidence calibration in CHAD-structured conditions demonstrate that structuring human-AI collaboration through a principled framework produces substantially superior decision outcomes compared to unstructured pairing of humans with AI tools. These improvements exceed the typical effect sizes reported for individual AI decision support interventions (Schemmer et al., 2023), suggesting that the framework structure -- particularly the authority allocation and complementarity optimisation components -- produces emergent collaboration quality that transcends the sum of its parts. The finding that CHAD's Outcome Accountability component produces a 31.6% improvement in perceived decision responsibility directly addresses the responsibility diffusion problem identified as a central risk of collaborative AI decision-making (Mittelstadt et al., 2016). This finding has important regulatory implications: responsible AI governance requires not only that humans are involved in AI-assisted decisions but that clear, personalised accountability is structurally assigned and experienced. Framework-level accountability assignment -- as implemented in CHAD Component 5 -- appears to be a critical and largely neglected mechanism for achieving this outcome.

5.2 Limitations and Future Research

The field experiment, while conducted in real organisational settings with professional decision-makers, used artificial decision cases rather than live operational decisions, limiting ecological validity particularly for the clinical oncology setting where real-world time pressures, patient relationships, and institutional dynamics may alter CHAD effectiveness. The 10-case decision protocol may not capture the learning dynamics of sustained CHAD adoption, including the potential for capability mapping to improve as collaboration accumulates empirical performance data over time. Future research should investigate CHAD adaptation to agentic AI systems -- in which the AI agent takes autonomous multi-step actions rather than producing point recommendations -- where the authority allocation and conflict resolution components require fundamental redesign. The deskilling risk of long-term CHAD collaboration -- the potential for sustained AI assistance to degrade human capability mapping accuracy -- requires longitudinal investigation beyond the scope of this study.

6. Conclusion

6.1 Summary

This paper proposed the CHAD framework, a five-component structured methodology for collaborative human-AI decision-making comprising capability mapping, complementarity optimisation, authority allocation, conflict resolution, and outcome accountability. Expert validation (n=32, Kendall W=0.79) confirmed strong consensus on component importance. Field experiment results across three settings (n=108) demonstrated 27.4% decision quality improvement, 34.1% calibration improvement, and 31.6% responsibility improvement versus unstructured human-AI collaboration (Cohen d=2.06). CHAD addresses EU AI Act Article 14 human oversight requirements through structured authority allocation and personalised accountability assignment.

6.2 Recommendations

For organisations deploying AI in high-stakes decision contexts, we recommend adopting the CHAD framework as a structured alternative to unstructured human-AI pairing, beginning with Component 1 (Capability Mapping) to establish the empirical foundation for complementarity optimisation. Component 5 (Outcome Accountability) should be implemented as a minimum-viable first step given its high expert importance rating and direct response to the responsibility diffusion problem. For EU AI Act compliance officers, the CHAD framework provides a structured implementation pathway for Article 14 human oversight obligations that transcends nominal HITL provision. For AI researchers, the CQA instrument (detailed in Appendix A) provides a validated measurement tool for collaborative human-AI decision quality that enables systematic comparison of collaboration frameworks and mechanisms.

References

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., and Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. CHI 2021, 1-16.

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., and Terry, M. (2019). Hello AI: Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. CSCW 2019.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. AIAA 2004-6313.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56-62.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

IEEE (2019). Ethically aligned design v2. IEEE Standards Association.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. (2023). Towards a science of human-AI decision making. FAccT 2023.

Mittelstadt, B. D. et al. (2016). The ethics of algorithms. Big Data and Society, 3(2).

NIST (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1.

Pasquale, F. (2015). The black box society. Harvard University Press.

Rasmussen, J., Pejtersen, A. M., and Goodstein, L. P. (1994). Cognitive systems engineering. John Wiley and Sons.

Salas, E., Reyes, D. L., and Woods, A. L. (2018). The assessment of team performance: Observations and needs. Assessment and Teaching of 21st Century Skills. Springer.

Schemmer, M. et al. (2023). Appropriate reliance on AI advice. IUI 2023, 410-422.

Wang, X., Yin, M., and Bansal, G. (2021). Are two heads better than one in AI-assisted decision making? CSCW 2021.

Wilder, B., Horvitz, E., and Kamar, E. (2021). Learning to complement humans. IJCAI 2021, 1526-1533.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box. Harvard Journal of Law and Technology, 31(2), 841-887.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Green, B., and Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. CSCW 2019.

Vaccaro, K., and Sabanovic, S. (2021). Contestability by design in algorithmic decision-making. CHI 2021.

Floridi, L. et al. (2018). An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707.

Declarations

Funding

This research was supported by the Estonian Research Council under grant PRG2341 (CHAD: Collaborative Human-AI Decision Frameworks for Responsible Outcomes) and the Austrian Science Fund (FWF) under grant P 38201-N. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The CHAD framework specification, CQA instrument, expert assessment data, and anonymised field experiment decision data are available from the corresponding author upon reasonable request.

Ethical Approval

The field experiment was approved by the Baltic AI Research University Ethics Committee (reference BARU-AI-2025-018). All participants provided informed consent. Participants were debriefed on study objectives after completion. Ethical approval reference: CETU-DS-2025-024.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

Appendix A

Collaboration Quality Assessment (CQA) Instrument

The CQA comprises five criteria (one per CHAD component), each scored 0-20 using a four-item rubric (items scored 0-5). The following provides abbreviated scoring guidance for each criterion.

Criterion 1: Capability Utilisation (CQA-C1)

Criterion 2: Complementarity Effectiveness (CQA-C2)

Criterion 3: Authority Clarity (CQA-C3)

Criteria 4-5: Conflict Resolution and Accountability (CQA-C4, CQA-C5)