Human-Under-Test and Continual Bidirectional Assessment for Co-development of Human-AI Systems

Roberto Casadei[1,*], Giovanni Delnevo[1] and Silvia Mirri[1]

[1]Alma Mater Studiorum—*Università di Bologna, Cesena, Italy*

Abstract

Recent developments in artificial intelligence (AI) and large-language models (LLM) promote collaboration of humans and AI-based agents. However, the use of AI has risks, e.g., related to over-reliance and possibly unintended consequences stemming from structural issues and mistakes from both sides. Given that AI is a tool with intrinsic strengths and weaknesses, there are also responsibilities on the human side regarding how the tool is used. For the human-AI system to be effective, both actors should understand the limitations and risks of both players and adopt strategies to mitigate them. Therefore, in this position paper, we propose a model and process for continual bidirectional assessment and co-development of human-AI systems. Though research has mostly focussed on the evaluation of AI agents, we especially focus on the human. Through an analogy with software testing, we propose a "human-under-test" schema, where the AI agent proactively inspects the human user to identify potential issues (e.g., in knowledge, expectations, or process consistency) that might negatively affect the collaboration.

Keywords

human-AI collaboration, LLM, AI agents, workflows, hybrid intelligence, co-development, human-AI interaction

1. Introduction

Context. The recent developments on artificial intelligence (AI), generative AI (GenAI), and large language models (LLMs) are revolutionising *human-AI collaboration*, creating new opportunities and challenges for the implementation of systems and processes fostering *hybrid/collective intelligence* [1, 2]. Our focus is on human-AI collaboration in "projects", loosely defined as multi-step activities aimed at solving information-intensive tasks and producing deliverables (e.g., software projects).

Problem, state of the art, and gap. Limitations in AI tools, in user knowledge and expectations, and in their *human-AI interaction (HAI)*, implies risks that may undermine performance and give way to unintended consequences. Works on *AI maturity models* have been proposed to comprehensively assess an organisation's ability of leveraging AI, suggesting properties that humans and AI should possess for mature HAI and ecosystems. Several studies have been carried out on *human-AI teams* [3, 4, 5, 6, 7], supporting collaborative problem solving up to human-AI *co-evolution* [8, 9]. Crucial concepts include *feedback loops* [8, 10] for tuning and incremental co-development, *shared mental models* [11], to properly frame expectations and promote effective interaction through human-AI mutual understanding, and *meta-cognitive scaffolding* [12, 13], supporting user reasoning through *thinking assistants* [14]. Though proposals for co-evolution and reciprocal learning exist [10], we observe limited contributions on integrated end-to-end frameworks for human-AI teams with bidirectional assessment of mental model and knowledge gaps. This view is also shared by other researchers [8], mentioning "investigation of bidirectional causality" and "modelling of the feedback loop" as open challenges in human-AI systems.

Contribution. In this position paper, we review contributions on human-AI collaboration, and propose and discuss two ideas to foster structured research on effective human-AI collaboration. The

HAIC'25: 1st International Workshop on Human-AI Collaborative (HAIC) Systems, October 25-30, 2025, Bologna, Italy
*Corresponding author.
roby.casadei@unibo.it (R. Casadei); giovanni.delnevo@unibo.it (G. Delnevo); silvia.mirri@unibo.it (S. Mirri)
0000-0001-9149-949X (R. Casadei); 0000-0001-6640-5746 (G. Delnevo); 0000-0002-5385-4734 (S. Mirri)

general idea is to consider an *integrated, end-to-end process framework for bidirectional scaffolding*, aimed at proactively reducing risks by assessing mental models and knowledge about several entities (knowledge, behaviour, configuration of the human and AI actors, goals, context, process outputs and history). The specific idea, covering one direction of the overall schema, is to consider *humans-under-test*, namely the human actors as "test subjects"—following a software testing analogy. This also suggests an alternative to the common "human-in-the-(AI-)loop" viewpoint, which could be referred to as *"AI-in-the-human-loop"*.

Paper structure. Section 2 covers background and related work. Section 3 presents the contribution. Section 4 provides a discussion with pointers to future research.

2. Background and Related Work

For proper human-AI collaboration, human users and AI tools should interact (and possibly collaborate) in order to understand, assess, and support each other. We review works related to different research directions contributing to this view.

2.1. AI Maturity Models

In general, *maturity models* are models aimed at evaluating and supporting the optimisation of processes by humans or organisations [15]. *AI maturity models* specifically aim to evaluate an organisation's readiness in taking advantage of AI, by using or developing it. Sometimes, the proposed maturity models have a different focus, e.g., on the *human organisation* adopting AI, or on the *AI system itself*, considered as an entity with different levels of maturity.

Surveys on AI maturity models. A number of surveys provide insights on the AI maturity models themselves: their focus, goals, and development methods. In 2023, Akbarighatar et al. [16] reviewed AI maturity models, with a focus on *responsible* development and use of AI, under the lenses of a sociotechnical perspective. The extracted capabilities for responsible AI include: (i) continuous evaluation of the effects of AI decisions, (ii) employee's awareness of ethical issues in AI, (iii) security and privacy, (iv) fairness evaluation, (v) transparency/understandability of AI models, (vi) accountability. A 2021 systematic literature review on AI maturity models is also provided by Sadiq et al. [17]. They adopt a taxonomy based on: research objectives (development, validation, or application of AI maturity models), scope (domain generality, analysis scope), method (what analytical and empirical methods), design approach (top-down vs. bottom-up), architecture (stage-based vs. continuous), purpose of use (descriptive vs. prescriptive vs. comparative), typology (maturity grids, structured models, Likert-like questionnaires, or hybrid models), and maturity model components (levels and elements). They extract seven critical dimensions: (i) data, (ii) analytics, (iii) technology and tools, (iv) intelligent automation, (v) governance, (vi) people, and (vii) organisation.

Examples of recent AI maturity model proposals. Hartikainen et al. [18] propose a (preliminary) maturity model to guide the *development* of human-centred AI systems (HCAI-MM), based on six building blocks: (i) working with AI uncertainty, (ii) collaboration and human control, (iii) accountability, (iv) fairness, (v) transparency, and (vi) explainability. We share similar motivations, though we emphasise the uncertainty related to human actors. Fukas et al. [19] provide an AI maturity model tailored to the auditing domain (A-AIMM). The A-AIMM consists of five levels for eight dimensions (technologies, data, people and competences, organisation and processes, strategy and management, budget, products and services, ethics and regulation). At the most advanced level, the organisation (i) leads the development of AI technologies, (ii) audit is data-driven, (iii) people have leading AI competences, (iv) processes are AI-enabled and -driven, (v) the AI strategy is decided, (vi) AI has dedicated budget, (vii) AI supports products and services, and (viii) AI is trustworthy and explainable. Hansen et al. [20] also propose an AI

capability maturity model to understand and guide adoption of AI in organisations. Their framework uses two technological (data, infrastructure), three organisational (strategy, people, culture), and two external dimensions (ethics & regulations, and pressures & motivation). At the sixth, top-most level, AI is a core integrated part of the organisation's business model and culture.

Relationship with our work. We share the vision on the critical aspects of AI use and development. We focus on the people, culture, and human-AI collaboration aspects.

2.2. Human-AI Interaction and Collaboration: Mental models, Teaming, Co-evolution

Another thread of topics focus on the *interaction* between humans and AI. Research could be distinguished mainly in terms of the scope of analysis: *co-evolution* is more long-term and broad in scope, whereas human-AI *teaming* tends to focus more on short-term decision-making or specific aspects (e.g., mental models, trust). Specifically, *mirroring* mental models and *scaffolding* user understanding and reasoning are two key goals and means for human-AI collaboration. In mirroring, the AI reflects back cognitive and affective states to users. In (metacognitive) scaffolding, the AI supports users in self-regulation and -reasoning during interaction with AI.

Human-AI teams. Enssley [21] discusses aspects affecting human-AI team performance, including decision making, coordination, interaction methods, team training, trust, transparency, explainability, and the role of bias. The topic is vast and there are not broad surveys yet, beside a scoping review on human-centred human-AI by Berretta et al. [4] and a conceptual outlook from Lou et al. [3]. Contributions on this topic are indeed quite diverse. For instance, Lancaster et al. [6] investigate human-AI team *training*, identifying that users tend to value *cross-training* (understanding more about other roles) and *adaptive roles* of AI. To help humans understand AI systems, Cabrera et al. [5] propose to use *AI behaviour (pattern) descriptions* for sub-groups of problem instances.

Mental models, mirroring, and meta-cognitive scaffolding. Andrews et al. [11] review the role of *shared mental models* in human-AI teams, based on the intuition that when teammates' mental models align, then the team will perform better due to improved prediction accuracy backed by reciprocal understanding. Bansal et al. [7] focus on AI-advised human decision making in high-stakes domains. They show that humans with accurate mental models of AI systems (e.g., their *error boundary*) improve team performance and, more interestingly, that updating AI to increase accuracy, at the expense of *compatibility* (coherence with mental model based on previous experience), may degrade team performance. Te'eni et al.[10] proposed the notion of *reciprocal human-machine learning (RHML)* in which both humans and machines iteratively update internal representations through cycles of feedback. Their *Fusion* system exemplifies this principle by supporting experts' message classification tasks through cognitive mirroring and mutual adaptation.

In a broader critique, Lewis[22] argues that most current AI systems lack true reflective capacity, a core cognitive faculty in humans, thus failing to manage ambiguity, context, and emergent meaning. They propose a *reflective AI architecture* grounded in complex systems and cognitive science to address these shortcomings. Tankelevitch et al.[12] and Lim [13] extend these ideas into practical design strategies for *scaffolding* user metacognition when interacting with GenAI. These include explainability features, adaptive prompting, and bias-aware nudges, as seen in the DeBiasMe platform. Levin [23] introduces the concept of *Meta-AI skills*, a new class of metacognitive competencies essential for co-reasoning with GenAI. These skills include reflective prompt engineering, multimodal synthesis, and tacit knowledge articulation, which together enable learners to engage AI not just as a tool, but as a cognitive partner. An interesting concept relevant in this context is that of *thinking assistants* [14], namely AI agents that *help users think* by fostering self-reflection in the user.

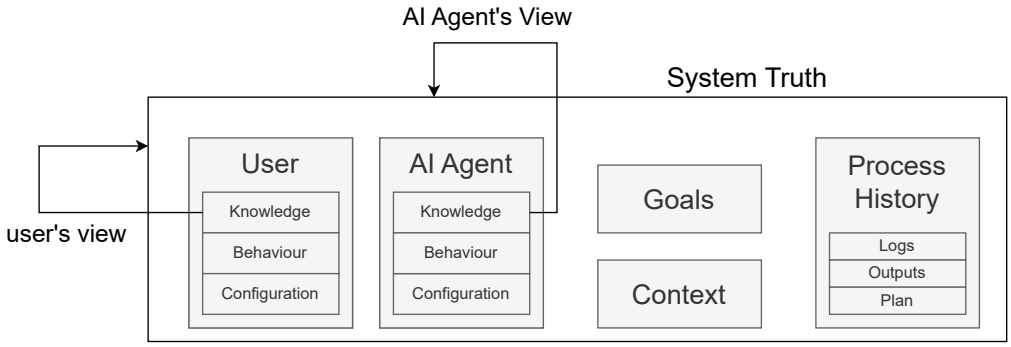

Figure 1: System model.

Human-AI coevolution. Defined by Pedreschi et al. [8] as *"a process in which humans and AI algorithms continuously influence each other"*, it affects human-AI ecosystems and the broader society. It is based on a *feedback loop* where, iteratively, the human user feeds data into AI *recommenders* or *assistants* which re-train/tune and then provide a suggestion influencing the user for the next iteration. Among the open challenges, the authors mention the "investigation of bidirectional causality" and the "modelling of the feedback loop", which align with the focus of this paper. Some studied focus on particular directions of the team relationships: e.g., Schut et al. [9] investigate human learning from AI, even as a means for advancing human knowledge.

Relationship with our work. This paper focusses on human-AI teaming, especially on collaboration aimed at identifying issues with mental models and knowledge gaps through bidirectional assessment. Though extensive research has been carried out on mental models [4, 11, 7], to the best of our knowledge, limited contributions exist on bidirectional knowledge gap assessment, which is related but not the same as reciprocal learning as in [10]. Bidirectional knowledge gap assessment is a diagnostic method for identifying missing information, whereas reciprocal learning is a collaborative teaching strategy that fosters mutual knowledge transfer. We explore how metacognitive scaffolding is appropriated and transformed by users over time, and how mirroring strategies can serve not only as feedback mechanisms but as foundations for shared cognitive ground.

3. A Model for End-to-End Integrated Bidirectional Assessment and Scaffolding for Human-AI Teams

This section introduces a model (Section 3.1) and process framework (Section 3.2) for bidirectional assessment and scaffolding in human-AI systems.

3.1. Model

With no loss of generality, we consider a system involving a *team* consisting of a single human user and a single AI agent. The *ground-truth system state* $\mathcal{S} = (U, Ag, G, Ctx, P)$ (see Figure 1) is given by:

- *user state* $U = (K_U, B_U, C_U)$, including its knowledge K_U, behaviour B_U, and configuration C_U (which can be interpreted as a set of explicit predictors or factors affecting the behaviour);
- *AI agent state* $Ag = (K_{Ag}, B_{Ag}, C_{Ag})$, including its knowledge K_{Ag}, behaviour B_{Ag}, and configuration C_{Ag};
- *goals* G, modelling what the system is trying to achieve;
- *context* Ctx, modelling all relevant information that can be exploited to achieve the goals;
- *process* $P = (L, O, W)$, modelling the entire process history in terms of a *log object* L, as well as the produced *output* O (cf. project deliverables), and the *plan* (or *workflow*) W of future activities.

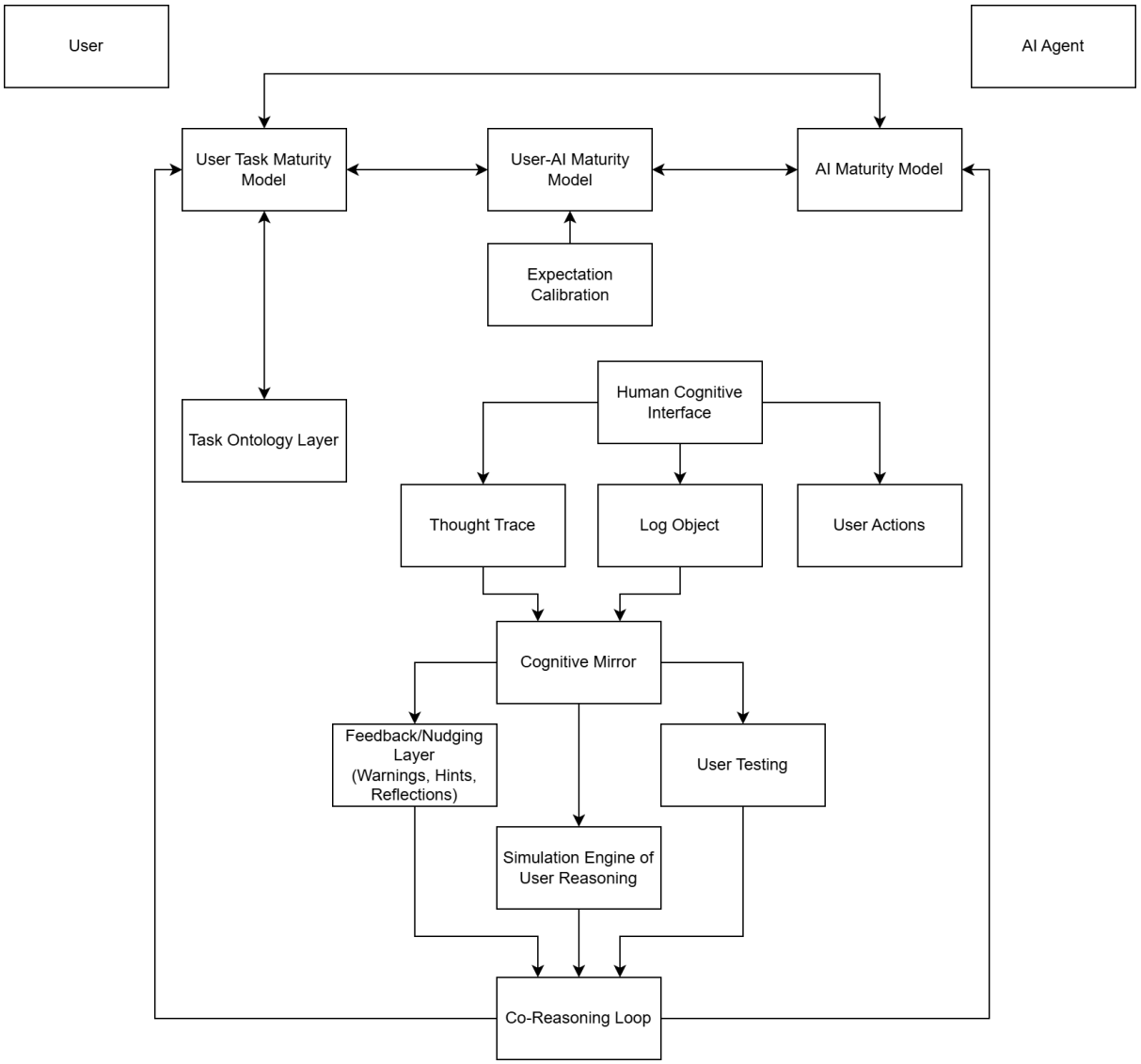

Figure 2: Human-AI interaction process and co-reasoning loop.

Such state $\mathcal{S}$ is of course *dynamic*, meaning that all its components can evolve in time: from $\mathcal{S}^0$ to $\mathcal{S}^T$ (current time).

The system includes two actors: the *user agent* $\mathcal{U}$ and the *AI agent* $\mathcal{A}$. Crucially, each actor has its own view of the system and its components, $\mathcal{U}(\mathcal{S}) = K_U$ and $\mathcal{A}(\mathcal{S}) = K_{Ag}$, which may not be aligned with the ground truth and with each other's view. Notice that, while certain elements such as the project outputs can be directly inspected by the actors, other elements are directly inaccessible and can only be *estimated* in terms of one's knowledge and context information.

3.2. Process framework

The proposed framework, summarised in Figure 2, envisions an end-to-end, bidirectional process in which humans and AI systems co-develop understanding, strategies, and alignment. This process is grounded in a dynamic, developmental view of human-AI interaction and is operationalised through a set of interacting architectural components that adapt over time to user behaviour and task complexity.

Grounding and Calibration. At the outset, the system initiates a grounding and calibration phase. The Expectation Calibration Layer aligns the user's mental model with the actual capabilities, limitations,

and intended epistemic role of the AI system. This step ensures a realistic understanding of how the AI will support cognitive work. The Task Ontology Layer decomposes the user's activity into hierarchical structures comprising phases (e.g., planning, execution), subtasks, and the cognitive goals driving them. The User Task Maturity Model estimates the user's proficiency and fluency across subtasks, while the User-AI Maturity Model assesses the user's awareness, adaptability, and strategic competence in collaborating with AI. These models inform both the human-side scaffolding strategy and the initialisation of the AI Maturity Model, which characterizes the AI's current ability to adaptively assist across task contexts.

Tracking Cognition. As the interaction unfolds, the user engages through the Human Cognitive Interface to express decisions, strategies, hesitations, and reasoning outputs. These expressions materialize as Human Actions (observable physical or cognitive activities) and are captured as structured data in the Log Object. Simultaneously, the user's inner reasoning processes are modelled via the Thought Trace, which captures the fine-grained structure of cognitive operations, including inferred intentions, logical steps, and reflective loops.

This cognitive and behavioural data is monitored and interpreted by the Cognitive Mirror. It is the AI function that acts as a reflective twin of the user's reasoning. It monitors, maps, and surfaces latent cognitive patterns, mirroring back the inferred logic and prompting metacognitive reflection. These insights are shared with three other core entities. When deviations, misconceptions, or epistemic breakdowns are detected, the Feedback/Nudging Layer is activated. This layer does not merely provide corrective responses; it generates metacognitive scaffolds in the form of warnings, reframing questions, strategic hints, or reflective prompts. The Simulation Engine of User Reasoning, that generates hypothetical reasoning trajectories based on the current task state and historical behaviour. It detects discrepancies, biases, or blind spots and supports adaptive diagnostics. This engine functions as both a diagnostic tool and an anticipatory agent, supporting proactive, personalised scaffolding. The User Testing module, that is responsible for actively probing the user's capabilities and understanding via targeted assessments and strategic interruptions. These interventions help dynamically update the user models and inform future support. The intensity and type of feedback are calibrated based on the joint analysis from User Task Maturity Model and User-AI Maturity Model, ensuring that support is neither redundant nor cognitively intrusive.

Feedback Loop. As interaction progresses, both the human and the AI system evolve. The Co-Reasoning Loop embodies the long-term synergy between the two agents, allowing for the mutual development of cognitive models, strategy refinement, and trust calibration. The human's models (User Task Maturity Model and User-AI Maturity Model) are dynamically updated based on performance and adaptation, while the AI Maturity Model evolves to fine-tune its scaffolding behaviours and metacognitive prompts, reinforcing the shift from static assistance to collaborative intelligence.

This bidirectional process not only supports problem-solving but enables a deeper form of metacognitive scaffolding, wherein the human gradually becomes more reflective, strategic, and autonomous, while the AI becomes more context-aware, sensitive to individual variation, and educationally aligned.

4. Discussion and Research Perspectives

"Humans-under-test" and the software testing analogy. Software testing is the overall *process* of *preparing* and *executing* various kinds of *tests* to *verify* (w.r.t. requirements) and *validate* (w.r.t. intended usage) a software system, to promote *quality* and *reduce risks* [24]. A *test case* gives the procedure, conditions, expected results for the assessment of a *subject-under-test (SUT)* by a *tester*.

The idea can be extended to *human-machine systems* and specifically to *humans*[1]. In the latter case,

[1]This is just an analogy meant to convey the similarity of certain goals and procedures between the different domains, in order to provide a starting point for further reflections, and should not be interpreted as a form of "dehumanisation" which would be ethically deplorable. Other ethical issues exist: for instance, terms like "defects" and "test failure" are problematic,

Software Testing Concept	Human Testing Analogue
Basic concepts	
Test	Human test: a test executed by the AI aimed at verifying or validating some aspect of a human's knowledge and behaviour
System/Subject-under-Test (SUT) *	Human-under-Test (HUT)
Failure *	The HUT response "significantly" differs from expectation (more generally, it seems that pass/failure is a limited dichotomy for human test outcomes)
Fault/defect/bug *	The cause for failure (e.g., lack of knowledge, lack of consistency w.r.t. plan, or a mistake)
Coverage	The amount of "relevant" HUT behaviour and knowledge that has been assessed by a collection of tests
Types of tests	
Unit test	Assessment of a minimal unit of the HUT's knowledge/behaviour
Integration test	Assessment of how the HUT interacts with other technical tools or the AI
Regression test	A test aimed at monitoring that the human maintains the learned competences over time
Black-box test	The AI only observes input-output behaviour
White-box test	The AI aims to assess the human's mental model (e.g., beliefs and intentions)
Other concepts and methods	
Test scheduling	The process of planning when human tests are executed
Test-driven development	Gap-directed human-AI co-evolution

Table 1
Software testing analogy for "user assessment" by AI. The asterisk (*) marks terms with strong ethical concerns.

we may talk of *"human-under-test (HUT)"* as the human subject that is "tested" by an AI agent. The goal is to help the human improve its understanding and behaviour (quality), and to reduce risks in human-AI collaboration. In this context, validation could refer to the assessment of *compliance* w.r.t. the AI agent's mental model, whereas verification could refer to the assessment of the human user's mental model and behaviour w.r.t. *shared* requirements and factual knowledge. The AI agent should *plan* what tests to be executed, when, and how, to cover the most significant risks (also checking for *regressions*) according to available resources and the *shared* testing strategy. See Table 1 for a summary of elements of the analogy.

This interpretation raises interesting questions that should be addressed by further research:

- What kinds of "tests" are most suitable for testing different aspects of the HUT? (cf. white- vs. black-box testing, unit vs. integration tests)
- How to identify and generate the *relevant tests* depending on the context?
- How to define the *expectations* for the responses of the HUT? How to quantify the compliance or deviance? How to integrate confidence levels and testing outcomes? (cf. construction of *test oracles*)
- How to estimate the *coverage* of relevant knowledge and behaviour?
- How to deal with the *non-determinism* that characterises (most of) human activity?
- How could tests be planned to avoid bothering the human user?
- How to design a *privacy-* and *ethics-aware* human testing[2] procedure?
- Who evaluates the evaluators? [25] How to measure if the human testing activity carried out by the AI is useful or effective?

and methods for locating failures might be considered "aggressive" or "manipulative". It should be remarked that the goal is to identify risks and promote quality (cf. improvement and learning).

[2]Notice that, in literature, "human testing" refers to "testing carried out by humans" and not our acceptation (where the human is the subject of tests).

From "human-in-the-(AI-)loop" to (the complementary) "AI-in-the-human-loop". As AI becomes increasingly embedded in decision-making, learning, and creative workflows, its role is undergoing a fundamental transformation. Traditionally, AI systems have operated as autonomous agents embedded within human-centred workflows, a paradigm commonly referred to as human-in-the-loop (HITL) [26]. In this configuration, AI outputs are subject to human validation or override, and humans remain the locus of reasoning, with AI serving as a subordinate tool. This unidirectional model of interaction emphasises human oversight and correction of AI-generated content, largely focusing on output quality rather than on mutual understanding or reflective improvement.

In contrast, emerging paradigms, particularly those involving GenAI and LLMs, redefine the nature of human-AI collaboration. In the framework (Section 3), we introduce a notion of *AI-in-the-human-loop*, where AIs is positioned not as a subordinate, but as a reflective partner in cognition. Rather than being evaluated solely by the human, the AI continuously evaluates, adapts to, and scaffolds the user's reasoning. It does so by embedding itself within the human's cognitive loop: observing, simulating, and mirroring human mental processes to foster metacognitive awareness, strategic refinement, and epistemic development. The result is a shift from task completion to co-reasoning, where the AI acts as a cognitive twin that supports reflective learning and knowledge articulation through adaptive feedback and dialogic interaction. Key differences are highlighted in Table 2.

This shift motivates a broader rethinking of how AI systems can be designed to not only support decision-making, but also promote human growth, self-regulation, and sense-making. Traditional approaches to AI design prioritize performance metrics such as accuracy and efficiency; however, such metrics are insufficient when AI is integrated into tasks that involve ambiguity, judgment, or iterative understanding. Instead, we argue that AI systems—particularly those built on LLMs—should be capable of mirroring users' reasoning, surfacing cognitive biases, and scaffolding metacognitive processes. Doing so requires both theoretical reconfiguration and practical innovation. We draw from cognitive science, educational theory, and reflective interaction design to propose new foundations for human-AI collaboration. Our goal is to lay the groundwork for AI systems that can not only perform tasks with users, but also help users better understand themselves through interaction with AI.

Table 2
Comparison between Human-in-the-Loop and AI-in-the-Human-Loop approaches

Human-in-the-Loop	AI-in-the-Human-Loop
Human checks and corrects	AI for co-interpretation and reflection
AI executes a task	AI for cognitive influence
Human as fallback	Human as epistemic driver
AI judged by human	AI judges and mirrors the human

Final remarks. Our conclusion is that an *end-to-end integrated* framework for proactive bidirectional assessment and scaffolding is needed to continuously identify risks, monitor processes, and foster mutual learning and reasoning (co-evolution). Further research is needed to realise this vision. The "human-under-test" analogy and the "AI-in-the-human-loop" concept might represent guiding metaphors for suggesting specific research questions and directions.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT-4) in order to: grammar and spell check, paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. M. M. Peeters, J. van Diggelen, K. van den Bosch, A. Bronkhorst, M. A. Neerincx, J. M. Schraagen, S. Raaijmakers, Hybrid collective intelligence in a human–AI society, AI & SOCIETY 36 (2020) 217–238. doi:10.1007/s00146-020-01005-y.

[2] R. Casadei, Artificial collective intelligence engineering: A survey of concepts and perspectives, Artif. Life 29 (2023) 433–467. doi:10.1162/ARTL_A_00408.

[3] B. Lou, T. Lu, T. S. Raghu, Y. Zhang, Unraveling human-AI teaming: A review and outlook, CoRR abs/2504.05755 (2025). doi:10.48550/ARXIV.2504.05755. arXiv:2504.05755.

[4] S. Berretta, A. Tausch, G. Ontrup, B. Gilles, C. Peifer, A. Kluge, Defining human-ai teaming the human-centered way: a scoping review and network analysis, Frontiers in Artificial Intelligence 6 (2023). doi:10.3389/frai.2023.1250725.

[5] A. A. Cabrera, A. Perer, J. I. Hong, Improving human-ai collaboration with descriptions of AI behavior, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–21. doi:10.1145/3579612.

[6] C. M. Lancaster, W. Duan, R. Mallick, N. J. McNeese, Human-centered team training for human-ai teams: From training with AI tools to training for AI teammates, Proceedings of the ACM on Human-Computer Interaction 9 (2025) 1–38. doi:10.1145/3710998.

[7] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, E. Horvitz, Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff, in: 33rd AAAI Conference on Artificial Intelligence, AAAI, AAAI Press, 2019, pp. 2429–2437. doi:10.1609/AAAI.V33I01.33012429.

[8] D. Pedreschi, L. Pappalardo, E. Ferragina, R. Baeza-Yates, A.-L. Barabási, F. Dignum, V. Dignum, T. Eliassi-Rad, F. Giannotti, J. Kertész, A. Knott, Y. Ioannidis, P. Lukowicz, A. Passarella, A. S. Pentland, J. Shawe-Taylor, A. Vespignani, Human-ai coevolution, Artificial Intelligence 339 (2025) 104244. doi:10.1016/j.artint.2024.104244.

[9] L. Schut, N. Tomašev, T. McGrath, D. Hassabis, U. Paquet, B. Kim, Bridging the human–ai knowledge gap through concept discovery and transfer in alphazero, Proceedings of the National Academy of Sciences 122 (2025). doi:10.1073/pnas.2406675122.

[10] D. Te'eni, I. Yahav, A. Zagalsky, D. Schwartz, G. Silverman, D. Cohen, Y. Mann, D. Lewinsky, Reciprocal human-machine learning: A theory and an instantiation for the case of message classification, Management Science (2023). doi:10.1287/mnsc.2022.03518.

[11] R. W. Andrews, J. M. Lilly, D. Srivastava, K. M. Feigh, The role of shared mental models in human-ai teams: a theoretical review, Theoretical Issues in Ergonomics Science 24 (2022) 129–175. doi:10.1080/1463922x.2022.2061080.

[12] L. Tankelevitch, V. Kewenig, A. Simkute, A. E. Scott, A. Sarkar, A. Sellen, S. Rintel, The metacognitive demands and opportunities of generative ai, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24, ACM, 2024, p. 1–24. doi:10.1145/3613904.3642902.

[13] C. Lim, DeBiasMe: De-biasing human-AI interactions with metacognitive AIED (AI in Education) interventions (2025). doi:10.48550/ARXIV.2504.16770.

[14] S. Park, C. Kulkarni, Thinking assistants: LLM-based conversational assistants that help users think by asking rather than answering, CoRR abs/2312.06024 (2023). doi:10.48550/ARXIV.2312.06024. arXiv:2312.06024.

[15] W. S. Humphrey, Managing the software process, The SEI series in software engineering, Addison-Wesley, 1989.

[16] P. Akbarighatar, I. Pappas, P. Vassilakopoulou, A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review, International Journal of Information Management Data Insights 3 (2023) 100193. doi:10.1016/j.jjimei.2023.100193.

[17] R. B. Sadiq, N. Safie, A. H. Abd Rahman, S. Goudarzi, Artificial intelligence maturity model: a systematic literature review, PeerJ Computer Science 7 (2021) e661. doi:10.7717/peerj-cs.661.

[18] M. Hartikainen, K. Väänänen, T. Olsson, Towards a human-centred artificial intelligence maturity model, in: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing

Systems, CHI EA 2023, Hamburg, Germany, April 23-28, 2023, ACM, 2023, pp. 285:1–285:7. doi:10.1145/3544549.3585752.

[19] P. Fukas, J. Rebstadt, F. Remark, O. Thomas, Developing an artificial intelligence maturity model for auditing, in: 29th European Conference on Information Systems - Human Values Crisis in a Digitizing World, ECIS 2021, 2021.

[20] H. F. Hansen, E. Lillesund, P. Mikalef, N. Altwaijry, Understanding artificial intelligence diffusion through an AI capability maturity model, Information Systems Frontiers 26 (2024) 2147–2163. doi:10.1007/s10796-024-10528-4.

[21] M. R. Endsley, Supporting human-AI teams: Transparency, explainability, and situation awareness, Computers in Human Behavior 140 (2023) 107574. doi:10.1016/j.chb.2022.107574.

[22] P. R. Lewis, S. Sarkadi, Reflective artificial intelligence, Minds and Machines 34 (2024). doi:10.1007/s11023-024-09664-2.

[23] I. Levin, M. Marom, A. Kojukhov, Rethinking AI in education: Highlighting the metacognitive challenge, BRAIN. Broad Research in Artificial Intelligence and Neuroscience 16 (2025) 250. doi:10.70594/brain/16.s1/21.

[24] P. Ammann, J. Offutt, Introduction to Software Testing, 2 ed., Cambridge University Press, 2016.

[25] P. Liguori, C. Improta, R. Natella, B. Cukic, D. Cotroneo, Who evaluates the evaluators? on automatic metrics for assessing ai-based offensive code generators, Expert Systems with Applications 225 (2023) 120073. doi:10.1016/j.eswa.2023.120073.

[26] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, L. He, A survey of human-in-the-loop for machine learning, Future Generation Computer Systems 135 (2022) 364–381. doi:10.1016/j.future.2022.05.014.