Human-Centred Evaluation Approaches for Autonomous Agents: From Review to Practice

Benjamin Bowers, Catherine Harvey & Robert Houghton

University of Nottingham

SUMMARY

The present research explores existing approaches and methods used to study and evaluate novel system components with increasingly autonomous capabilities, such as AI agents, in safety-critical domains. We report a large-scale review of the human–autonomy teaming (HAT) literature and extend the input-mediator-output model of team effectiveness to guide human-centred assessment, culminating in the IMO-A framework. We then test existing metrics and understand approaches to evaluation used by Human Factors researchers using a novel method which leverages an AI-generated video generation tool to develop underwater maritime scenarios across levels of autonomy. Our final recommendations serve as a roadmap for progressing HAT evaluation from fragmented, study-specific measurement choices toward a standardised, IMO-A guided, autonomy-appropriate and multi-method evidence base that can be translated into practitioner-ready early-phase HSI protocols for the safe integration of autonomous agents in safety-critical systems.

KEYWORDS

Human-autonomy teaming, human-centred evaluation, safety-critical systems, team effectiveness

Introduction

Recent advances in artificial intelligence (AI) have the potential to alter how team-based work is organised and executed in safety-critical domains. Contemporary autonomous systems have attracted growing research interest for their increasing capacity for collaborative, rather than purely assistive, human-machine relationships (Behymer et al, 2015; Chen et al, 2016). This shift has given rise to the concept of human-autonomy teaming (HAT), which describes system configurations in which human and autonomous agents operate in interdependent roles to pursue shared objectives (Lyons et al, 2021; O'Neill et al, 2022).

However, existing test and evaluation approaches retain a technology-centric perspective with little regard for real-world effectiveness beyond technical performance benchmarks (e.g. Wang et al, 2019; see Dehgani et al, 2021). If existing measures become targets for developers, they cease to become good measures (Goodhart, 2015; Thomas and Uminsky, 2022), and operators will continue to "surprise designers with use (or disuse)" of these technologies (Helmer et al, 2024, p. 4). A related concern is that few evaluation tools involve human-AI interactions, prompting the need to "create grounded thresholds for what good real-world performance looks like" (Ruah et al., 2024, p1205).

Ergonomics and Human Factors (EHF), with its long-standing commitment to understanding and optimising interactions between humans and other system elements through a human-centred approach, is uniquely positioned to address this gap. Responding to observations that the EHF community has been "strangely quiet" on AI developments (Grote et al, 2023, p1702) and must

reaffirm its relevance in an era of rapid technological advancement (De Winter and Eisma, 2024; 2025), this paper intends to bring HAT research to the forefront as a critical site for contemporary human-centred evaluation.

Despite the drastic growth of recent HAT interest, the underlying idea of humans collaborating with intelligent or autonomous systems is not new. Early Human Factors work anticipated such arrangements decades ago, such as in the context of intelligent cockpit systems and supervisory control (McNeese, 1986). More recently, HAT has often been framed as an extension of human–automation interaction, drawing heavily on concepts and findings from automation research that examined how humans supervise, rely on, and adapt to increasingly capable systems (Endsley, 2017).

Autonomous agents, however, introduce considerations that extend beyond traditional automation paradigms. Unlike automation constrained to predefined behaviours, autonomous agents warrant novel considerations for teammate-like characteristics such as communication (Guzman and Lewis, 2020), situation awareness (SA) (Demir et al, 2017), and trust (McNeese et al, 2021). Though McNeese et al (2023) highlight how autonomous teammates do not have to replicate humans for successful teaming, suggesting team cognition and trust are generated differently and, hence, must be "measured differently" (p. 2). This shift has prompted researchers to focus on the broader question of "team effectiveness" in HAT (Wynne and Lyons, 2019, p.1; De Visser et al, 2020, p. 8; p. 14).

Earlier HAT work identified the need for consistent measures of effectiveness that go beyond performance-based indicators (Strybel et al, 2018); such indicators are mere system capability measures and should not "be considered a single indication of HAT effectiveness" Richards (2020). Thomas and Uminsky (2022) similarly argued for the combination of complementary quantitative metrics with qualitative accounts for a more complete approach. While many HAT studies combine performance with behavioural and subjective measures (e.g., McNeese et al, 2018), and qualitative interviews (Flathmann et al, 2023), the selection and integration of measures is frequently shaped by local study design rather than a shared theoretical and conceptual basis. As a result, Roberts and colleagues' question "applicability of existing human-centred measures to teams comprising human and increasingly intelligence non-human team members" (Roberts et al, 2022, p.179).

In response to these challenges, the present paper examines how effectiveness is currently evaluated within the HAT literature, addressing two guiding questions: (1) what constructs are being measured, and (2) how are they being measured? (Richards, 2020, p. 237). Drawing on a systematic review of empirical HAT studies in safety-critical domains, this work synthesises evaluation practices across methodological traditions to identify patterns, commonalities, and areas of divergence.

To provide structure to an otherwise fragmented evidence base, findings are interpreted through the Input–Mediator–Output (IMO) framework, which has been proposed as a valuable lens for understanding team effectiveness in HAT contexts (O'Neill et al., 2023). As the field currently lacks consistent theoretical models that individual primary studies can collectively contribute to testing, this review seeks to consolidate existing practices within a coherent conceptual framework. In doing so, the paper aims to support the development of more systematic, human-centred approaches to evaluating team effectiveness as AI-enabled systems become increasingly embedded in operational work.

Research Design Overview

A two-phase, sequential design is adopted to characterise and advance human-centred evaluation of HAT in safety-critical domains, defined here as contexts where system error can cause harm to

people, property, or the environment (Knight, 2002). In Phase One, a comprehensive systematic review is conduced to establish an evidence base describing how HAT effectiveness is currently conceptualised and measured in research literature. Findings are subsequently interpreted through an Input-Mediator-Output (IMO) lens to support theory-grounded categorisation of measures (McGrath, 1964; Ilgen et al, 2005; see also O'Neill et al, 2020; 2023). Phase Two examines HAT evaluation in practise by using AI-generated vignettes of underwater maritime HAT scenarios to elicit interview responses Human Factors researchers and test the sensitivity of contemporary rating scales.

Method

Phase One – Systematic Review and Framework Development

We conducted a systematic review of recent HAT literature in safety-critical settings (2010–2025). Following PRISMA 2020 guidelines (Page et al, 2021), overarching search objectives were fourfold; (1) identify human-centric constructs frequently considered critical to the success of HATs in safety-critical domains; (2) review and categorise the HAT measurement methods that are used to assess these constructs; (3) recommend adaptions to the IMO framework of team effectiveness to accommodate autonomous agents; (4) set out a research agenda aimed at moving HAT evaluation toward a standardised, generalisable science capable of producing early-stage HSI evaluation protocols for AI-based technology.

A search was conducted across seven databases, selected to capture research spanning Human Factors, Human–Computer Interaction (HCI), and AI/engineering. Specifically, ACM Digital Library and SAGE Journals were included for strong representation of HCI and applied psychology; ScienceDirect for core Human Factors outlets; IEEE Xplore for advanced technology and engineering scholarship; and Web of Science, Scopus, and PubMed to provide broad interdisciplinary indexing and domain coverage. Eligibility criteria required empirical studies involving human participant(s) working with a real or simulated AI-based autonomous system on a defined task. Records were excluded if they lacked human participants, were conducted outside safety-critical domains, were conceptual/non-empirical, or did not involve a clear HAT task.

The search yielded 641 records; after deduplication and screening in accordance with PRISMA 2020 guidance (Figure 1), 90 studies met criteria. A large language model (ChatGPT; OpenAI; GPT-4; June 2025) was used to flag records that were clearly irrelevant against exclusion criteria; all flagged records were then carefully reviewed by the primary reviewer, and no final inclusion/exclusion decisions were made by the model. Searches were executed on 3 June 2025 and limited to publications from 2010–2025 to reflect the shift from predominantly rule-based automation toward modern machine learning-enabled autonomy. The search date sits five years after O'Neill and colleagues state-of-science review (See O'Neill et al, 2020).

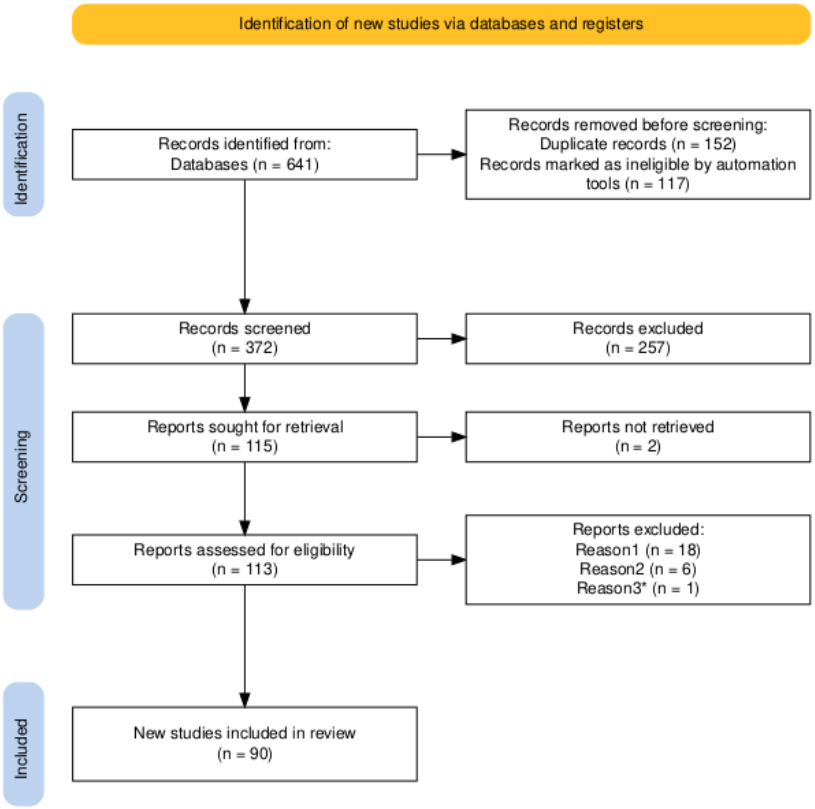

Figure 1. Systematic Review Process Chart. *Note.* This was generated using Haddaway et al's (2022) tool, depicting the article selection process for the systematic review in accordance with PRISMA 2020 guidelines (Page et al, 2021). *One article summarised the results of three individual studies, which were each located and reviewed separately in place of the original summary paper.

Data was extracted through a standard process developed and iteratively refined in accordance with the review's objectives. Extracted data included: title, author(s), year of publication, domain, task, independent variables, dependent variables, measurement method used, and main findings. Extraction was managed using Microsoft Excel by 1 reviewer. Any discrepancies were resolved through discussion or consultation with 2 more reviewers.

Phase Two – Further Exploration of Human Factors Researcher's Approaches to HAT Evaluation Using AI-Generated Vignettes

A vignette-based elicitation design was used to examine how Human Factors analysts evaluate HAT effectiveness in practise. Three AI-generated video recreations of a real-world safety-critical maritime incident (the *Karen* collision; MAIB, 2016) were developed using Google Veo 3 (September 2025). Each vignette comprised a four-minute narrative sequence with brief informational slides to provide operational context. Conditions varied team composition and the AI teammate's level of autonomy (LOA): a human-only baseline reflecting a hierarchical command structure; a low-LOA AI condition in which a voice-based agent provided advisory input to one human only when queried; and a high-LOA AI condition in which the agent issued proactive recommendations, with authority distributed between the agent and commanding officer and information broadcast to multiple crew members.

Seventeen participants (7 male, 10 female; *M* age = 28.2, range 21–54) were recruited from the Human Factors community, spanning Master's/PhD students, postdoctoral researchers, and senior research fellows, primarily from a range of UK institutions. Prior to the session, participants provided consent and completed the Negative Attitudes toward Robots Scale (NARS; Nomura et

al., 2006; reworded for AI). Vignettes were presented in counterbalanced order via Microsoft Teams. After each condition, participants completed a post-vignette evaluation battery comprising items from the HAT Cohesion Scale (Neubauer et al., 2025), the Autonomous Agent Teammate Likeness Scale (Wynne & Lyons, 2019; agency, benevolence, interdependence, synchrony), and a Team Affect Questionnaire (Nass et al., 1996; openness to influence, information quality), followed by a brief verbal probe on what most influenced their ratings. Sessions concluded with a short semi-structured interview examining perceived effectiveness across conditions, perceived ways the AI supported or hindered teamwork, preferred AI roles in human teams, and evaluation-relevant criteria not captured by the surveys.

Verbal responses and interviews were transcribed using Microsoft Teams' automated transcription and analysed using inductive, reflexive thematic analysis with primarily semantic coding and theme development prioritising interpretive meaningfulness over frequency (Burr, 1995). Survey ratings were analysed using one-way repeated-measures ANOVA with Condition (Human, Low LOA, High LOA) as the within-subjects factor for constructs measured across all three conditions. Planned paired comparisons (Human vs Low; Human vs High; Low vs High) used paired *t*-tests with Holm correction within each construct family; effect sizes are reported as Cohen's *dz*. To test for any AI presence effect, Human ratings were compared against an "AI-average" score (mean of Low and High LOA). Agency and benevolence (AI-only) were compared via paired *t*-tests (Low vs High), and Spearman correlations between NARS and condition ratings were computed with Holm correction.

Results

Phase One Findings: How is HAT effectiveness evaluated?

Firstly, we note a large increase in HAT publications across the last five years. Of the 90 total studies review, around one third explicitly mentioned 'effectiveness', with others adopting a targeted focus, such as on organisational outcomes (e.g. Lester et al, 2025), performance (e.g. De Visser & Parasuraman, 2011; Chen & Barnes, 2012), or usability (e.g. Linder & Shulte, 2020). 16 of 90 studies adopted a full breadth of performance, subjective, and behavioural measures (e.g. Grigsby et al, 2017) recommended for HAT research (Strybel et al, 2018). Beyond this, many employed narrower designs with 27 subjective-only (e.g. Jung et al, 2025), 19 subjective + performance (Bogg et al, 2021), and two performance-only evaluations (Lakhmani et al, 2019; Barber et al, 2019).

The review identified 81 constructs, operationalised using 169 distinct measurement methods, that can be split into inputs, mediators (explain input-output relationships; see Ilgen et al, 2005), and outcomes of team effectiveness. Four dominant constructs emerged, namely trust (61/90), workload (42/90), individual and/or team performance (41), and situation awareness (15). Performance is treated as an outcome, whereas trust spans both mediator and outcome roles (team viability). Workload and SA are framed as mediating processes in their contribution to team effectiveness.

Beyond this, mediating constructs could be cleanly split into *team*- and *individual-level* constructs. The former group captures collective functioning rather than individual states (e.g. coordination, resilience, cohesion, and team cognition) through a range of subjective and behavioural measures. The latter construct group operationalises the human operator's experiences and perceptions in isolation (emergent states), including attitudes and engagement (e.g. confidence, satisfaction, motivation), usability, safety and risk perceptions, and human process variables such as attention, bias, reliance tendencies. However, mediating relationships are scarcely formally modelled in statistical analyses despite being framed as such in discussions.

Team effectiveness inputs were grouped into five categories. *Autonomous agent characteristics* incorporate variations in the agent's design or interaction mode, such as transparency (Wohleber et al, 2023), LOA (Bogg et al, 2021), and control modes (Chiou et al, 2021). *Task characteristics* were manipulated, often by varying task difficulty or complexity. *Team composition* described who the teammates were and how the teams were structured, such as incorporating a single vs multiple autonomous agents. *Training* interventions, such as trust calibration exercises (Johnson et al, 2021), team-building activities (Walliser et al, 2019), and workflow-specific training (Wenderott et al, 2024), were tested for their effect on effectiveness constructs. Finally, *individual differences* were measured as inputs with a few significant effects reported, though the majority produced non-significant results. Where significant relationships were reported, they linked propensity to trust technology to higher trust in the AI (Küper et al., 2025), attentional control to improved multitasking performance (Chen & Barnes, 2012), and locus of control to greater comfort delegating control to autonomy (Chiou et al., 2021). Additional predictors included decision-making style (rational style positively related to trust; van Arum et al., 2025) and perfect automation schema as a predictor of trust in robots (Matthews et al., 2019).

Phase Two Findings: What criteria do analysts prioritise in HAT evaluation?

Across within-subjects comparisons (Human-only, Low-LOA, High-LOA; $n = 17$), Synchrony was the only construct showing a reliable condition effect, $F(2,32) = 3.39$, $p = .046$; Holm-corrected follow-ups indicated Human > Low-LOA ($t = 3.20$, $p\text{Holm} = .017$, $dz = 0.78$). Other constructs showed no robust condition differences after correction (Table 2). To identify what shaped analysts' overall judgements, a subject fixed-effects model predicting a standardised Effectiveness composite found Information Quality ($\beta = .446$, $p = .026$) and Openness to Influence ($\beta = .206$, $p = .048$) to be the strongest unique predictors, corroborated by relative-importance estimates (IQ semi-partial $R^2 = .032$; OI semi-partial $R^2 = .010$). Mixed-effects extensions were not estimable due to model singularity in the current dataset, so these predictor findings are interpreted as indicative.

Table 2. Phase two quantitative findings

Metric	Human M (SD)	Low-LOA M (SD)	High-LOA M (SD)	F (2, 32)	p	Key pairwise
Cohesion	4.80 (1.00)	4.43 (1.11)	4.19 (1.12)	1.49	0.240	
Interdependence	3.74 (0.77)	3.16 (0.64)	3.43 (0.61)	3.04	0.062	
Synchrony	3.76 (0.83)	3.25 (0.73)	3.55 (0.75)	3.39	0.046	Human > Low-LOA (pHolm=.017, dz≈0.78)
Openness to Influence	3.98 (0.86)	4.20 (1.20)	4.39 (0.54)	1.37	0.269	
Information Quality	4.01 (0.85)	4.00 (0.76)	4.10 (0.64)	0.22	0.802	
Agency		2.94 (1.16)	3.18 (0.92)		0.459	t=-0.76, p=.459, dz=-0.18
Benevolence		3.22 (0.81)	3.22 (0.66)			t=0.00, p=1.000, dz=0.00

Qualitative accounts obtained through post-vignette semi-structured interviews were analysed using thematic analysis. Across transcripts for participants, 345 codes were applied to excerpts, grouped into 91 initial theme, and eventually eight final themes (Table 3).

Table 3. Summary of qualitative themes

Theme	Observation	Quote

Authority and Challenge Boundaries	When the AI communicated confidently, especially with one-to-all broadcasts, participants observed decision authority shifting toward the AI and human challenge behaviour decreasing.	"They… don't have many discussions after the suggestion of AI."
Trust and Reliance Calibration	Participants observed a "collapse of verification culture" as autonomy increased, with teams accepting AI outputs with minimal checking.	"Very quick to trust it… over reliance on it."
Communication Ecology and Shared Understanding	Human-only teams naturally negotiated and cross-checked, whereas AI-supported teams, especially high LOA, showed shorter, more transactional exchanges with less discussion.	"Didn't communicate between team members … good teamwork is more of a discussion."
Human Agency, Skill Use and Role Clarity	Analysts reported that higher autonomy nudged humans toward passive monitoring (following the AI) rather than thinking through options.	"(HighLOA) actual member of the team, (LowLOA) more a tool to provide a function."
AI Role Framing and Cognitive Fit	Participants preferred the AI framed as an assistant/decision aid that provides information on request, rather than acting as a coordinator or quasi-leader.	"…an intelligent support aid… speak when they're spoken to."
Transparency, Explainability, and Access	Limited visibility into the AI's reasoning led teams to appear aligned without shared understanding; analysts wanted rationale, evidence, and confidence accessible to all.	"(LowLOA) wasn't good … AI didn't provide evidence or explainability at all."
Integration Maturity and Sociotechnical Readiness	Analysts focused on workflow fit, proceduralisation, training, and auditability, emphasising that deployment hinges on integration maturity and clear responsibility for risk.	"Actual technicalities… level of integration of AI."
Safety and Ethics Orientation	Strong preference for human having final sign off on safety-critical actions. Concerns about legal accountability and need to document who approved what, and why.	"The safety critical things… in [HighLOA]… were done by the human leader [here]."

Discussion

As autonomous agents take on roles once reserved for people, making teaming more than a purely human endeavour, test and evaluation approaches must evolve accordingly. Established schools of thought, namely human teaming and human-automation interaction have scaffolded HAT research, but simultaneously risk casting a shadow that obscures the fields distinct challenges and possibilities (McNeese et al, 2023). This risk is particularly salient in safety-critical domains that demand not only technical excellence but social fluency from contemporary machines with autonomous capabilities. The contributions from the present investigation intended to progress a fragmented field toward generalisable science capable of supporting the effective integration of autonomous agents into safety-critical systems.

Phase One, an updated comprehensive state-of-science review, five years on from the previous of its kind (O'Neill et al, 2020), surfaced three main gaps: (1) over-reliance on single-modality metrics (subjective-only common); (2) inconsistent definitions/operationalisations of "effectiveness"; (3) limited attention to autonomy-specific perceptions that shape trust calibration and appropriate use with a preference for outdated metrics such as usability. Four Key Effectiveness Variables (trust, workload, situation awareness, performance) dominate yet are inconsistently operationalised and rarely integrated with behavioural or physiological indicators. Moreover, we apply and propose an

extension to the IMO model of team effectiveness (Ilgen et al, 2005) to capture the unique dynamics of HATs. The framework recognises that some members are non-human agents by splitting mediators (processes and emergent states) into human-only (individual-level) and team-level constructs with KEVs at the intersection (accounting for SA & Team SA; workload & workload distribution; collective team trust).

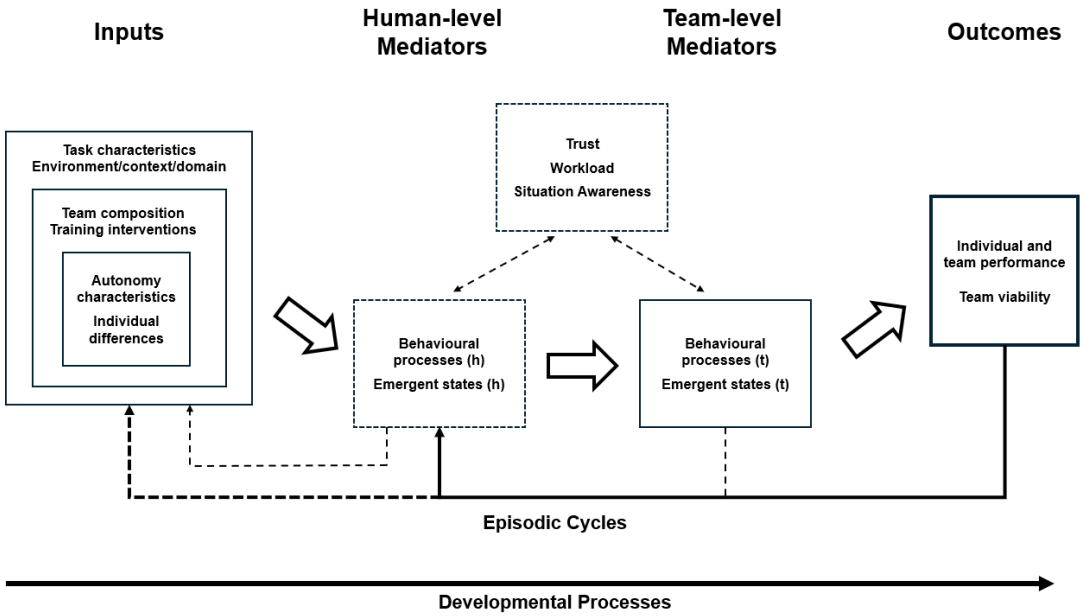

Figure 3. The Input-Mediator-Output for Autonomy (IMO-A) Framework. *Note.* This is an extension of the I-M-O framework (Ilgen et al, 2005) and is intended as a structured lens through which the human-autonomy team effectiveness can be viewed.

Phase Two demonstrated evaluation nuances in practise using three AI-generated video reconstructs of a real-world underwater maritime incident (MAIB, 2016). Despite events being kept consistent between conditions, the presence of an autonomous agent of varying LOA (input variable) caused variations in analysts quantitative and qualitative accounts of HAT effectiveness.

Teams containing a low LOA AI agent were rated significantly lower in perceived synchrony, a cornerstone of autonomy agent teammate likeness (Wynne & Lyons, 2018; 2019). Hence, we find support for the use of metrics developed with the unique complexities autonomous agents bring to teaming in mind over outdated legacy metrics. The survey findings, while associational and based on a small repeated-measures vignette dataset, further indicate analysts judgements of HAT effectiveness hinge on whether the team's information flow supports good decisions and whether the team is receptive to influence. Thus, evaluation protocols that focus narrowly on task performance (or even on trust alone) risk missing what evaluators treat as decisive indicators of team functioning.

Moreover, qualitative accounts from EHF researchers shed light onto how evaluation approaches in adapt in practise to the presence of autonomous agents, and the significance of LOA as an input factor. We conclude that unique considerations must be made for (1) authority and challenge boundaries, (2) trust calibration and reliance, (3) communication ecology and understanding, (4) human agency and role clarity, (5) AI role framing and cognitive fit, (6) transparency and explainability, (7) integration maturity and sociotechnical readiness, (8) safety and ethics orientation.

Recommendations and Future Directions

Our recommendations are threefold: (1) HAT research should explicitly adopt 'effectiveness' as an overarching goal to avoid narrow approaches (e.g. performance only) through a three-pronged approach spanning subjective, behavioural, and performance metrics. (2) Apply the aforementioned approach through an IMO-A structure to appreciate the nuances between emergent states and processes across individual and team levels, appreciating that teams no longer containing human-only members need to be evaluated as such. (3) Avoid outdated metrics designed for technology without autonomous capabilities, swapping in contemporary metrics designed with the unique complexities autonomous agents bring to teaming in mind.

Our qualitative investigation produced an early thematic model of how EHF researchers actually reason about HAT effectiveness during human-centred AI evaluation. As a next step, this model can serve as a roadmap for cumulative evaluation science and a meaningful contribution to early-phase human-system integration by (i) formalising the themes into a set of candidate evaluation dimensions with clear construct definitions and boundary conditions and (ii) operationalising each dimension with validated, autonomy-appropriate measures mapped to IMO-A levels (individual, team, and cross-level KEVs). Progress along this pathway would move HAT evaluation from heterogeneous, study-specific measurement choices toward practitioner-ready, early-phase HSI protocols that support the safe and effective integration of autonomous agents into operational systems.

References

Barber, D., Reinerman-Jones, L., & Hidalgo, M. (2019). Optimizing Military Human-Robot Teaming: An Evaluation of Task Load and Modality Switch Cost. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1751–1755. https://doi.org/10.1177/1071181319631352

Behymer, K. J., Mersch, E. M., Ruff, H. A., Calhoun, G. L., & Spriggs, S. E. (2015). Unmanned vehicle plan comparison visualizations for effective human-autonomy teaming. *Procedia Manufacturing*, *3*, 1022–1029.

Bogg, A., Birrell, S., Bromfield, M., & Parkes, A. (2021). Can we talk? How a talking agent can improve human autonomy team performance. *Theoretical Issues In Ergonomics Science*, *22*(4), 488–509. (WOS:000593412700001). https://doi.org/10.1080/1463922X.2020.1827080

Burr, V. (1995). *An Introduction to Social Constructionism*. Routledge.

Chen, J. Y. C., & Barnes, M. J. (2012). Supervisory control of multiple robots: Effects of imperfect automation and individual differences. *Human Factors*, *54*(2), 157–174. Scopus. https://doi.org/10.1177/0018720811435843

Chen, J. Y. C., Barnes, M. J., Selkowitz, A. R., & Stowers, K. (2016). *Effects of Agent Transparency on human-autonomy teaming effectiveness*. 1838–1843. Scopus. https://doi.org/10.1109/SMC.2016.7844505

Chiou, M., McCabe, F., Grigoriou, M., & Stolkin, R. (2021). Trust, shared understanding and locus of control in mixed-initiative robotic systems. *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 684–691. https://ieeexplore.ieee.org/abstract/document/9515476/?casa_token=Ba18h6NFvnkAAAAA:qgPghq3Np02SXz4z5zXG7SgWIKydQgAa94ELkbBAPvfc72VKmLYeAmkKk2vf9absLGz_L_s

De Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

De Visser, E., & Parasuraman, R. (2011). Adaptive Aiding of Human-Robot Teaming: Effects of Imperfect Automation on Performance, Trust, and Workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231. Scopus. https://doi.org/10.1177/1555343411410160

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., & Vinyals, O. (2021). *The Benchmark Lottery* (arXiv:2107.07002). arXiv. https://doi.org/10.48550/arXiv.2107.07002

Flathmann, C., Schelble, B. G., Rosopa, P. J., McNeese, N. J., Mallick, R., & Madathil, K. C. (2023). Examining the impact of varying levels of AI teammate influence on human-AI teams. *International Journal of Human-Computer Studies*, *177*, 103061. https://doi.org/10.1016/j.ijhcs.2023.103061

Goodhart, C. (2015). Goodhart's law. In *The encyclopedia of central banking* (Vol. 227). Edward Elgar Publishing Cheltenham, UK. http://lelibellio.com/wp-content/uploads/2013/02/Pages-29-%C3%A0-33-Goodhart-Ch.-2013-dossier-Goodharts-Law-Libellio-vol.-9-n%C2%B0-4.pdf

Grigsby, S., Crossman, J., Purman, B., Frederiksen, R., & Schmorrow, D. (2017). *Dynamic task sharing within human-UxS teams: Computational situation awareness*. *10285 11th International Conference, AC 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part II*, 443–460. Scopus. https://doi.org/10.1007/978-3-319-58625-0_32

Helmer, D., Boardman, M., Conroy, S. K., Hepworth, A. J., & Harjani, M. (2024). *Human-centred test and evaluation of military AI* (arXiv:2412.01978). arXiv. https://doi.org/10.48550/arXiv.2412.01978

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in Organizations: From Input-Process-Output Models to IMOI Models. *Annual Review of Psychology*, *56*(1), 517–543. https://doi.org/10.1146/annurev.psych.56.091103.070250

Johnson, C., Demir, M., McNeese, N., Gorman, J., Wolff, A., & Cooke, N. (2023). The Impact of Training on Human-Autonomy Team Communications and Trust Calibration. *HUMAN FACTORS*, *65*(7), 1554–1570. (WOS:000705088100001). https://doi.org/10.1177/00187208211047323

Jung, J., Kang, S., Choi, J., El-Kareh, R., Lee, H., & Kim, H. (2025). Evaluating the impact of explainable AI on clinicians' decision-making: A study on ICU length of stay prediction. *International Journal of Medical Informatics*, *201*, 105943. https://doi.org/10.1016/j.ijmedinf.2025.105943

Knight, J. C. (2002). Safety critical systems: Challenges and directions. *Proceedings of the 24th International Conference on Software Engineering - ICSE '02*, 547. https://doi.org/10.1145/581339.581406

Küper, A., Lodde, G. C., Livingstone, E., Schadendorf, D., & Krämer, N. (2025). Psychological Factors Influencing Appropriate Reliance on AI-enabled Clinical Decision Support Systems: Experimental Web-Based Study Among Dermatologists. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/58660

Lakhmani, S. G., Wright, J. L., Schwartz, M. R., & Barber, D. (2019). Exploring the Effect of Communication Patterns and Transparency on Performance in a Human-Robot Team. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 160–164. https://doi.org/10.1177/1071181319631054

Lester, C., Rowell, B., Zheng, Y., Co, Z., Marshall, V., Kim, J. Y., Chen, Q., Kontar, R., & Yang, X. J. (2025). Effect of Uncertainty-Aware AI Models on Pharmacists' Reaction Time and Decision-Making in a Web-Based Mock Medication Verification Task: Randomized Controlled Trial. *JMIR Medical Informatics*, *13*. https://doi.org/10.2196/64902

Lyons, J., Sycara, K., Lewis, M., & Capiola, A. (2021). Human-Autonomy Teaming: Definitions, Debates, and Directions. *FRONTIERS IN PSYCHOLOGY*, *12*. (WOS:000660190600001). https://doi.org/10.3389/fpsyg.2021.589585

MAIB. (2016). *Report on the investigation of the collision between the stern trawler Karen (B317) and a dived Royal Navy submarine in the Irish Sea on 15 April 2015*. Marine Accident Investigation Branch. https://assets.publishing.service.gov.uk/media/57fe2e2ee5274a496200000a/MAIBInvReport20_2016.pdf

Matthews, G., Lin, J., Panganiban, A., & Long, M. (2020). Individual Differences in Trust in Autonomous Robots: Implications for Transparency. *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS*, *50*(3), 234–244. (WOS:000538155900006). https://doi.org/10.1109/THMS.2019.2947592

McGrath, J. E. (1964). Social psychology: A brief introduction. https://cir.nii.ac.jp/crid/1130282272136366208

McNeese, N. J., Demir, M., Chiou, E. K., & Cooke, N. J. (2021). Trust and Team Performance in Human–Autonomy Teaming. *International Journal of Electronic Commerce*, *25*(1), 51–72. https://doi.org/10.1080/10864415.2021.1846854

McNeese, N. J., Flathmann, C., O'Neill, T. A., & Salas, E. (2023). Stepping out of the shadow of human-human teaming: Crafting a unique identity for human-autonomy teams. *Computers in Human Behavior*, *148*, 107874. https://doi.org/10.1016/j.chb.2023.107874

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678.

Neubauer, C., Forster, D. E., Lakhmani, S., Fitzhugh, S. M., Berg, S., Rovira, E., & Krausman, A. (2025). Development of a team cohesion scale for use in human-autonomy team research. In *Interdependent Human-Machine Teams* (pp. 67–99). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780443292460000109

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006). Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *7*(3), 437–454. https://doi.org/10.1075/is.7.3.14nom

O'Neill, T. A., Flathmann, C., McNeese, N. J., & Salas, E. (2023). Human-autonomy Teaming: Need for a guiding team-based framework? *Computers in Human Behavior*, *146*, 107762.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020

statement: An updated guideline for reporting systematic reviews. *Bmj*, *372*. https://www.bmj.com/content/372/bmj.n71.short

Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Comanescu, R., Akbulut, C., Stepleton, T., Mateos-Garcia, J., Bergman, S., & Kay, J. (2024). Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 1200–1217. https://ojs.aaai.org/index.php/AIES/article/view/31717

Richards, D. (2020). Measure for Measure: How Do We Assess Human Autonomy Teaming? In C. Stephanidis, D. Harris, W.-C. Li, D. D. Schmorrow, C. M. Fidopiastis, P. Zaphiris, A. Ioannou, X. Fang, R. A. Sottilare, & J. Schwarz (Eds), *HCI International 2020 – Late Breaking Papers: Cognition, Learning and Games* (Vol. 12425, pp. 227–239). Springer International Publishing. https://doi.org/10.1007/978-3-030-60128-7_18

Roberts, A. P. J., Webster, L. V., Salmon, P. M., Flin, R., Salas, E., Cooke, N. J., Read, G. J. M., & Stanton, N. A. (2022). State of science: Models and methods for understanding and enhancing teams and teamwork in complex sociotechnical systems. *Ergonomics*, *65*(2), 161–187. https://doi.org/10.1080/00140139.2021.2000043

Strybel, T. Z., Keeler, J., Mattoon, N., Alvarez, A., Barakezyan, V., Barraza, E., Park, J., Vu, K.-P. L., & Battiste, V. (2018). Measuring the Effectiveness of Human Autonomy Teaming. In C. Baldwin (Ed.), *Advances in Neuroergonomics and Cognitive Engineering* (pp. 23–33). Springer International Publishing. https://doi.org/10.1007/978-3-319-60642-2_3

Thomas, R. L., & Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. *Patterns*, *3*(5). https://doi.org/10.1016/j.patter.2022.100476

van Arum, S., Genç, H. U., Reidsma, D., & Karahanoğlu, A. (2025). Selective Trust: Understanding Human-AI Partnerships in Personal Health Decision-Making Process. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*. https://doi.org/10.1145/3706598.3713462

Walliser, J., de Visser, E., Wiese, E., & Shaw, T. (2019). Team Structure and Team Building Improve Human-Machine Teaming With Autonomous Agents. *JOURNAL OF COGNITIVE ENGINEERING AND DECISION MAKING*, *13*(4), 258–278. (WOS:000481283200001). https://doi.org/10.1177/1555343419867563

Wenderott, K., Krups, J., Luetkens, J. A., & Weigl, M. (2024). Radiologists' perspectives on the workflow integration of an artificial intelligence-based computer-aided detection system: A qualitative study. *Applied Ergonomics*, *117*, 104243. https://doi.org/10.1016/j.apergo.2024.104243

Wohleber, R. W., Stowers, K., Barnes, M., & Chen, J. Y. C. (2023). Agent transparency in mixed-initiative multi-UxV control: How should intelligent agent collaborators speak their minds? *Computers in Human Behavior*, *148*, 107866. https://doi.org/10.1016/j.chb.2023.107866

Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, *19*(3), 353–374. https://doi.org/10.1080/1463922X.2016.1260181

Wynne, K. T., & Lyons, J. B. (2019). Autonomous Agent Teammate-Likeness: Scale Development and Validation. In J. Y. C. Chen & G. Fragomeni (Eds), *Virtual, Augmented and Mixed Reality. Applications and Case Studies* (Vol. 11575, pp. 199–213). Springer International Publishing. https://doi.org/10.1007/978-3-030-21565-1_13