Human-Centred Autonomy: Designing Robots and Intelligent Machines That Work With People

Daniel Okafor, Sofia Petrov

Article type: Editorial

Volume 1, Issue 1 (2026)

Read Current Issue →

Abstract

Autonomous and intelligent systems increasingly work alongside people in factories, vehicles, healthcare environments, service settings, and decision-support contexts. Their success, however, should not be defined by the extent to which human involvement can be eliminated. Highly autonomous systems may perform individual functions accurately while still producing poor operational outcomes when users misunderstand system capabilities, lose situational awareness, become over-reliant, cannot intervene effectively, or experience excessive cognitive workload. This editorial argues that successful autonomy should be evaluated through the quality of human–machine collaboration rather than maximum machine independence. Human-centred autonomy requires coordinated attention to trust calibration, shared control, situation awareness, cognitive ergonomics, explainability, communication, safe transfer of authority, inclusive design, training, and organizational responsibility. Current evidence also demonstrates that explanations, anthropomorphic interfaces, and greater automation do not automatically improve collaboration. Research should therefore prioritize standardized human-autonomy evaluation, longitudinal trust studies, adaptive and multimodal interfaces, safe control transitions, representative end-user testing, and interdisciplinary validation. Autonomous systems should ultimately augment human capability while preserving meaningful understanding, agency, intervention, and responsibility.

Keywords: human-centred autonomy; human–robot interaction; human-autonomy teaming; trust calibration; shared control; cognitive workload; collaborative robotics

Citation

Okafor, D., & Petrov, S. (2026). Human-Centred Autonomy: Designing Robots and Intelligent Machines That Work With People. Journal of AI, Automation & Intelligent Systems, 1(1). https://eldranacademic.com/jaais/articles/2/

Received	Revised	Accepted	Published
05 January 2026	03 March 2026	25 March 2026	09 April 2026

Copyright

© 2026 Daniel Okafor, Sofia Petrov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) ↗. Use, distribution or reproduction in other forums is permitted, provided the original authors and copyright owners are credited and the original publication in this journal is cited in accordance with accepted academic practice.

Disclaimer

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, the publisher, the editors or the reviewers. Any product evaluated in this article, or claim made by its manufacturer, is not guaranteed or endorsed by the publisher.

1.0 Introduction

1.1 Context and Problem Framing

Autonomous technologies increasingly operate in domains where meaningful interaction with people is intrinsic to system performance. Collaborative robots now share industrial workspaces with operators; automated vehicles redistribute driving tasks between human and machine; assistive and healthcare technologies support physical and cognitive activities; intelligent decision-support systems advise professionals; and service robots interact directly with non-expert users. This expansion challenges a technological trajectory that has frequently equated progress with increasing machine independence. Human-centred artificial intelligence instead emphasizes configurations in which strong computational capability and meaningful human control can coexist (Shneiderman, 2020). Human-autonomy research similarly treats autonomous agents not merely as tools but increasingly as interdependent members of teams pursuing shared goals (Lyons et al., 2021; O'Neill et al., 2022). In industrial environments, contemporary safety requirements likewise emphasize integration, operation, maintenance, foreseeable misuse, and protection of people across the robot-system lifecycle rather than robot capability in isolation (International Organization for Standardization, 2025).

The central tension is that greater autonomy can simultaneously improve routine performance and weaken the human capacities required when automation reaches its limits. Function-allocation research has long demonstrated that automation changes human activity rather than simply replacing portions of it (Parasuraman et al., 2000). When operators are removed from active control, situation awareness and readiness to intervene can deteriorate, producing the well-established out-of-the-loop performance problem (Endsley & Kiris, 1995). Meta-analytic evidence further shows that higher degrees of automation can benefit routine system performance and reduce workload while impairing situation awareness and performance after automation failure (Onnasch et al., 2014). The transition from conventional automation to increasingly adaptive autonomy therefore does not eliminate human-factors problems; it changes their form, introducing new requirements for supervision, coordination, interface design, intervention, and comprehension (Endsley, 2017).

1.2 Why the Issue Matters

Successful autonomy should consequently be judged not by how rarely a human participates but by whether the combined human-machine system performs safely, effectively, and intelligibly across routine and abnormal conditions. Appropriate reliance depends on trust being calibrated to actual system capability rather than simply maximized (Lee & See, 2004), and trust itself varies with dispositional, situational, and learned factors that develop through interaction over time (Hoff & Bashir, 2015). Empirical human-autonomy teaming research also indicates that trust is related to team performance and evolves dynamically during collaboration (McNeese et al., 2021). The editorial position advanced here is therefore that autonomy should be designed around human capabilities, limitations, expectations, and responsibility. Technical intelligence is valuable precisely when it strengthens the performance and resilience of the overall human-machine system rather than marginalizing the humans who must understand, supervise, cooperate with, or recover it.

2.0 Current Landscape

2.1 Shared Autonomy

Human-centred autonomy draws on a mature but increasingly interconnected body of research spanning human factors, human–robot interaction, automation, ergonomics, cognitive science, artificial intelligence, and control engineering. Trust remains a dominant theme. Early quantitative synthesis established that robot performance and attributes strongly influence human trust (Hancock et al., 2011), while more recent review work confirms that both over-trust and under-trust can undermine collaboration and that trust measurement remains methodologically heterogeneous (Campagna & Rehm, 2025). Importantly, trust attitudes do not translate mechanically into behaviour. Human responses depend on the form of robot failure, experience, context, and interaction characteristics, making static trust scores an incomplete proxy for effective collaboration (Roesler et al., 2024).

2.2 Human-Robot Interaction

Human performance is equally shaped by workload, vigilance, situation awareness, and the distribution of responsibility. Automation complacency and automation bias arise partly from attentional dynamics, particularly where people supervise highly reliable systems while simultaneously managing other tasks (Parasuraman & Manzey, 2010). Similar concerns now appear in AI-assisted decision making: interventions that force users to engage analytically with machine recommendations can reduce over-reliance, although they may impose usability costs (Buçinca et al., 2021). In automated driving, takeover performance deteriorates when drivers must return rapidly from non-driving activities, illustrating the difficulty of treating human intervention as an instantaneous safety fallback (Weaver & DeLucia, 2022).

2.3 Trust Calibration

Collaborative robotics has broadened the discussion beyond cognitive trust to include physical safety, ergonomics, mental stress, workload, and workspace organization. Reviews of shared-space human–robot collaboration identify trust and cognitive workload among the most frequently studied human factors but also reveal substantial variability in measurement practice (Hopko et al., 2022). Ergonomic research increasingly seeks to adapt robot behaviour to users' physical and cognitive states (Lorenzini et al., 2023), while evidence from industrial workspace studies demonstrates that safe collaboration requires integrated consideration of technical, organizational, ergonomic, and social factors (Simões et al., 2022). Mental stress and perceived safety further influence how people behave around collaborative machines, yet real-world evidence remains more limited than laboratory experimentation (Lu et al., 2022).

2.4 Collaborative Decision-Making

Communication and explainability form another important frontier. Multimodal human–robot interaction combines visual, verbal, haptic, gestural, and physiological channels to support more natural coordination (Wang et al., 2024). Human-centred explainable-AI research likewise shows that explanations must be evaluated through user understanding, usability, trust, and collaboration rather than algorithmic properties alone (Rong et al., 2024). Systematic evaluations reinforce that explanation quality is inherently contextual and should be judged by whether it improves meaningful human-AI interaction and performance (Kim et al., 2024).

3.0 Critical Challenges

3.1 Inappropriate Reliance

The first challenge is inappropriate trust. Effective collaboration requires neither maximal trust nor permanent skepticism. Over-trust can produce complacency, monitoring failures, and acceptance of incorrect recommendations, while under-trust can cause users to ignore systems that would improve performance. The longstanding distinction among appropriate use, misuse, and disuse of automation remains directly relevant to contemporary autonomous systems (Parasuraman & Riley, 1997). Trust calibration is particularly difficult because users observe only fragments of system competence and may generalize successful performance in one context to situations where the automation is less capable.

Explainability does not automatically solve this problem. Explanations can clarify machine reasoning, but they can also increase confidence in incorrect outputs if users interpret plausibility as evidence of correctness. Experimental work shows that explanations may increase over-reliance when they reinforce user intuition even when the AI recommendation is wrong (Chen et al., 2023). Research on appropriate reliance similarly emphasizes that users must accept beneficial machine advice while rejecting erroneous recommendations, a requirement that is more demanding than merely increasing perceived transparency (Schemmer et al., 2023). Human-centred explanation must therefore support discrimination between trustworthy and untrustworthy machine behaviour rather than persuasion.

3.2 Interpretability

Control transfer constitutes a third challenge. Highly autonomous systems often assume that a human supervisor can resume authority when machine capability becomes insufficient. Yet the operator may have been cognitively disengaged for an extended period, may not possess sufficient situational awareness, or may have only seconds to understand why intervention is required. This problem is particularly visible in conditionally automated driving but also applies to industrial supervision, teleoperation, aviation, and autonomous infrastructure. Effective handover requires advance warning, intelligible status information, sufficient transition time, and interfaces that maintain the human's mental model during automated operation rather than reconstructing it only when failure occurs.

3.3 Safety Around People

The social presentation of autonomous systems introduces further complexity. Anthropomorphic design can influence expectations, emotional responses, and perceived competence, but its effects are inconsistent. Meta-analytic evidence indicates that anthropomorphic robot features can improve some human-related outcomes without consistently improving task performance or perceived safety (Roesler et al., 2021). Recent evidence from automated vehicles similarly shows that simply adding human-like characteristics does not reliably increase trust and that effects vary across interface forms and user groups (Arabian et al., 2026). Design preferences also differ across cultural contexts, as demonstrated in comparative research involving German and Arab participants (Korn et al., 2021). A universally "intuitive" autonomous interface should therefore not be assumed.

3.4 Evaluating Human-Machine Teams

Finally, the field lacks sufficiently standardized measures of effective human-machine collaboration. Accuracy, completion time, subjective trust, workload, safety, situation awareness, intervention quality, resilience, and team viability capture different dimensions of performance. A system can increase task efficiency while degrading user understanding or long-term skill. Laboratory studies may also recruit technically literate participants who do not

represent older adults, operators with varying educational backgrounds, people with disabilities, or culturally diverse end users. Human-centred autonomy therefore requires broader definitions of success and more representative validation populations.

4.0 Discussion and Priorities

4.1 Integrative Discussion

Technical performance and human factors are not separable layers that can be optimized independently. Machine reliability shapes trust; trust shapes reliance; reliance influences monitoring; monitoring affects the probability that a human detects machine failure; and interface design determines whether detection can be converted into effective intervention. A nominally accurate autonomous system can consequently produce unsafe outcomes if its accuracy encourages excessive reliance or if its interface hides the conditions under which that accuracy deteriorates.

This interdependence changes the meaning of autonomy. The design problem is not simply whether a function should be performed by a human or by a machine. It is how information, authority, workload, and responsibility should be dynamically allocated across the combined system. Human-autonomy teaming research emphasizes interdependence, communication, team cognition, and coordination rather than unilateral machine substitution (Lyons et al., 2021; O'Neill et al., 2022). Greater autonomy may reduce routine workload yet simultaneously create passive supervision, skill degradation, delayed intervention, or ambiguity about responsibility. Conversely, excessive human control may prevent intelligent systems from providing the speed, precision, or consistency that justified automation in the first place.

Human-centred autonomy should therefore be treated as a system-design philosophy encompassing machine behaviour, interfaces, organizational procedures, user training, task allocation, and operational governance. Trust calibration cannot be achieved exclusively through interface messaging when organizational incentives encourage unquestioned automation use. Safe handovers cannot be solved only by alarm design if staffing and workload make timely intervention impossible. Explainability cannot improve decisions when users lack the domain knowledge or time required to interpret explanations. The objective is not human control at every moment but meaningful human agency: people should understand the role of automation, remain able to identify consequential limitations, possess feasible mechanisms for intervention, and operate within organizations that assign responsibility clearly.

Complementary scholarship informing this synthesis includes International Organization for Standardization. (2025). These works broaden the evidentiary base considered across the field.

4.2 Research and Practice Priorities

A first priority is standardized evaluation of human-autonomy collaboration. Studies should measure combined-system performance using coordinated indicators of task effectiveness, safety, situation awareness, cognitive workload, trust calibration, intervention quality, communication, and recovery from failure. Domain-specific metrics will remain necessary, but common measurement principles would make results more comparable and expose situations in which higher machine performance degrades team performance.

A second priority is longitudinal research on trust and reliance. Trust changes through repeated success, failure, repair, learning, and organizational experience; short laboratory sessions cannot capture these dynamics adequately. Longitudinal studies should examine how people recalibrate after automation errors, how skills evolve

under persistent autonomy, and whether apparently effective interfaces remain effective after novelty disappears. Evidence of human-centred design should therefore include behavioural reliance rather than subjective trust alone.

Adaptive interfaces constitute a third priority. Autonomous systems should adjust information presentation, explanation depth, alerts, and interaction modality according to task demands and observable user state without covertly manipulating behaviour. Multimodal systems are particularly promising because communication can be distributed across speech, vision, gesture, haptics, and other channels rather than overloaded into a single display (Wang et al., 2024). Adaptation, however, should remain predictable enough that users can form stable mental models of system behaviour.

Fourth, control transfer must be engineered as a process rather than an emergency event. Systems should support progressive re-engagement, communicate why intervention is required, indicate available time, and maintain enough operator awareness during autonomous operation to make transfer feasible. Automated-driving research demonstrates that engagement in competing activities can impair takeover performance (Weaver & DeLucia, 2022), making human readiness a design variable rather than an assumption.

Fifth, explainability should be evaluated through human decision quality. Human-centred XAI studies should test whether explanations improve recognition of system limitations, appropriate reliance, error detection, and collaborative performance across realistic workloads (Rong et al., 2024). Explanations that merely increase perceived trust should not be considered successful. Representative users must also replace expert-only convenience samples wherever systems will be used by diverse populations. Age, accessibility requirements, cultural expectations, technological literacy, professional expertise, and language can materially shape interaction.

Finally, interdisciplinary testing should become routine. Robotics engineers, AI researchers, control specialists, ergonomists, psychologists, cognitive scientists, safety engineers, and domain experts should evaluate the same human-machine system rather than treating technical and human validation as sequential activities. Industrial safety standards already reinforce a system-level approach to robot integration (International Organization for Standardization, 2025). Comparable thinking should extend to intelligent systems in transportation, healthcare technologies, infrastructure, service robotics, and other environments where autonomy redistributes rather than eliminates human responsibility.

4.3 Implications for the Journal's Scope

For the Journal of AI, Automation & Intelligent Systems (JAAIS), human-centred autonomy provides an inherently interdisciplinary research agenda. Contributions concerned with autonomous capability should increasingly demonstrate how algorithms, robotic systems, intelligent controllers, and decision-support technologies interact with the people who supervise, operate, collaborate with, or are affected by them. Technical novelty remains important, but operational credibility requires evidence that the human-machine configuration performs effectively as a whole.

JAAIS should therefore encourage research integrating artificial intelligence and robotics with human factors, ergonomics, cognitive science, control engineering, safety, and responsible system design. Relevant contributions include experimental human-autonomy teaming, shared and adaptive control, human-aware robot planning, multimodal interfaces, calibrated trust, explainable decision support, safe handover mechanisms, cognitive-workload adaptation, collaborative industrial robotics, and longitudinal field evaluation. Particularly valuable studies will test representative end users under realistic conditions, report failure as well as success, and distinguish improvements in autonomous capability from improvements in collaborative system performance. Such scholarship can help move autonomy research from a model of machine independence toward one of complementary human-machine competence.

4.4 Conclusion

The future of autonomous and intelligent systems should not be defined by how completely people can be removed from decision loops. In many consequential applications, people remain supervisors, collaborators, beneficiaries, safety backstops, decision makers, or accountable professionals. Increasing autonomy without designing these roles deliberately can reduce situational awareness, distort trust, create unsafe transfers of authority, increase cognitive difficulty, and obscure responsibility even when machine performance is technically impressive.

Human-centred autonomy offers a more defensible standard. The relevant question is whether the allocation of intelligence and control enables humans and machines to accomplish tasks more safely, effectively, intelligibly, and resiliently than either could alone. Achieving that standard requires calibrated reliance, transparent communication, ergonomic interaction, inclusive design, feasible intervention, appropriate training, and evaluation of the complete socio-technical system. Autonomous systems should be judged not by the distance they create between machines and human operators, but by the quality of the collaboration they make possible. The most mature intelligent machines will therefore be those designed not merely to act without people, but to work dependably with them.

5.0 Declarations

5.1 Author Contributions

Prof. Daniel Okafor and Dr. Sofia Petrov contributed to the conceptual development, literature synthesis, critical interpretation, drafting, and revision of the editorial.

5.2 Funding

No external funding was received for the preparation of this editorial.

5.3 Competing Interests

The authors declare that they have no conflicts of interest or competing interests relevant to this editorial.

5.4 Data Availability

No original datasets were generated or analysed for this editorial.

5.5 AI Use Disclosure

Generative AI was used to support literature organization and language drafting during preparation of this editorial. The authors remain responsible for independently verifying the accuracy of the arguments, citations, references, and final submitted text in accordance with the publisher's policy.

6.0 References

Arabian, A., Mosaferchi, S., Mortezapour, A., Colley, M., Navarro, J., Goodridge, C. M., Oviedo-Trespalacios, O., Vitiello, G., & Naddeo, A. (2026). How does the anthropomorphism of automated vehicles shape human trust? A systematic review and meta-analysis. Transportation Research Part F: Traffic Psychology and Behaviour, 121, 103720. https://doi.org/10.1016/j.trf.2026.103720

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–21. https://doi.org/10.1145/3449287

Campagna, G., & Rehm, M. (2025). A systematic review of trust assessments in human–robot interaction. ACM Transactions on Human-Robot Interaction, 14(2), Article 30, 1–35. https://doi.org/10.1145/3706123

Chen, V., Liao, Q. V., Vaughan, J. W., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), Article 370, 1–32. https://doi.org/10.1145/3610219

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. Human Factors, 59(1), 5–27. https://doi.org/10.1177/0018720816681350

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human Factors, 37(2), 381–394. https://doi.org/10.1518/001872095779064555

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human Factors, 53(5), 517–527. https://doi.org/10.1177/0018720811417254

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

Hopko, S., Wang, J., & Mehta, R. (2022). Human factors considerations and metrics in shared space human-robot collaboration: A systematic review. Frontiers in Robotics and AI, 9, 799522. https://doi.org/10.3389/frobt.2022.799522

International Organization for Standardization. (2025). Robotics—Safety requirements—Part 2: Industrial robot applications and robot cells (ISO 10218-2:2025). International Organization for Standardization.

Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. Frontiers in Artificial Intelligence, 7, 1456486. https://doi.org/10.3389/frai.2024.1456486

Korn, O., Akalin, N., & Gouveia, R. (2021). Understanding cultural preferences for social robots: A study in German and Arab communities. ACM Transactions on Human-Robot Interaction, 10(2), Article 12, 1–19. https://doi.org/10.1145/3439717

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lorenzini, M., Lagomarsino, M., Fortini, L., Gholami, S., & Ajoudani, A. (2023). Ergonomic human-robot collaboration in industry: A review. Frontiers in Robotics and AI, 9, 813907. https://doi.org/10.3389/frobt.2022.813907

Lu, L., Xie, Z., Wang, H., Li, L., & Xu, X. (2022). Mental stress and safety awareness during human-robot collaboration—Review. Applied Ergonomics, 105, 103832. https://doi.org/10.1016/j.apergo.2022.103832

Lyons, J. B., Sycara, K., Lewis, M., & Capiola, A. (2021). Human–autonomy teaming: Definitions, debates, and directions. Frontiers in Psychology, 12, 589585. https://doi.org/10.3389/fpsyg.2021.589585

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

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. Human Factors, 64(5), 904–938. https://doi.org/10.1177/0018720820960865

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. Human Factors, 56(3), 476–488. https://doi.org/10.1177/0018720813501549

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

Roesler, E., Manzey, D., & Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. Science Robotics, 6(58), eabj5425. https://doi.org/10.1126/scirobotics.abj5425

Roesler, E., Vollmann, M., Manzey, D., & Onnasch, L. (2024). The dynamics of human–robot trust attitude and behavior—Exploring the effects of anthropomorphism and type of failure. Computers in Human Behavior, 150, 108008. https://doi.org/10.1016/j.chb.2023.108008

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards human-centered explainable AI: A survey of user studies for model explanations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(4), 2104–2122. https://doi.org/10.1109/TPAMI.2023.3331846

Schemmer, M., Kühl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. Proceedings of the 28th International Conference on Intelligent User Interfaces, 410–422. https://doi.org/10.1145/3581641.3584066

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Simões, A. C., Pinto, A., Santos, J., Pinheiro, S., & Romero, D. (2022). Designing human-robot collaboration (HRC) workspaces in industrial settings: A systematic literature review. Journal of Manufacturing Systems, 62, 28–43. https://doi.org/10.1016/j.jmsy.2021.11.007

Wang, T., Zheng, P., Li, S., & Wang, L. (2024). Multimodal human–robot interaction for human-centric smart manufacturing: A survey. Advanced Intelligent Systems, 6(3), 2300359. https://doi.org/10.1002/aisy.202300359

Weaver, B. W., & DeLucia, P. R. (2022). A systematic review and meta-analysis of takeover performance during conditionally automated driving. Human Factors, 64(7), 1227–1260. https://doi.org/10.1177/0018720820976476