BOUNDED TRUST IN ARTIFICIAL INTELLIGENCE IN ORGANIZATIONS: OUTPUT VERIFICATION AND RETAINED HUMAN ACCOUNTABILITY

Feruza Abdivalieva
Saint Petersburg State University,
Graduate School of Management Saint Petersburg, Russia
https://doi.org/10.5281/zenodo.22326022

Abstract. This paper examines how managers define the boundaries of appropriate reliance on artificial intelligence in organizational practice. The empirical material comprises nine semi-structured interviews with leaders and managers directly involved in AI implementation or use in organizations with different levels of digital maturity. The data were examined through theory-informed thematic analysis and cross-case comparison. The findings show that participants understood trust in AI neither as unconditional acceptance nor as rejection of the technology, but as bounded and differentiated reliance. Such reliance depends on output verification, task sensitivity, user competence, data-security restrictions, and the retention of accountability by a human decision-maker. The findings further connect trust in AI with leader legitimacy: critical and transparent use may strengthen confidence in managerial decisions, whereas uncritical copying of AI outputs and attempts to transfer responsibility to the system may weaken managerial authority. The paper therefore proposes shifting the managerial focus from building generalized trust in AI to organizing justified, controlled, and accountable AI use.

Keywords: artificial intelligence; bounded trust; AI overreliance; human accountability; leadership; legitimacy; AI governance.

Introduction. The diffusion of generative artificial intelligence has made interaction with AI part of everyday organizational work. AI is now used to search for and structure information, prepare documents and presentations, conduct analysis, support forecasting and prototyping, process documents, and assist decision-making. Yet the growing accessibility of these systems does not resolve two fundamental questions: for which tasks is reliance on AI appropriate, and who remains accountable for the final outcome?

Research frequently treats trust as a precondition for technology acceptance. This perspective is important because algorithm aversion may prevent users from benefiting from an otherwise useful system after observing a single error (Dietvorst et al., 2015). However, the opposite extreme - overreliance on AI - also creates costs. Users may follow an algorithmic recommendation even when it conflicts with contextual information or professional judgment (Klingbeil et al., 2024). The managerial challenge is therefore not to maximize trust, but to establish a justified degree of reliance on AI.

This interpretation is consistent with a socio-technical perspective. The reliability of organizational AI use depends not only on model properties but also on rules of use, employee competence, verification procedures, and the distribution of authority and accountability. Lahusen et al. (2024) show that trust, trustworthiness, and AI governance emerge through interactions among technologies, people, and institutions. Herrmann and Pfeiffer (2023) extend the human-in-the-loop principle by arguing that the organization must also remain in the loop through defined roles, feedback procedures, and control mechanisms.

The issue is particularly significant for leaders. Algorithmic systems can redistribute visibility, knowledge, control, and authority within organizations (Jarrahi et al., 2021). AI use therefore affects not only the quality of a particular output but also the legitimacy of the managerial decision built around it. Employees may evaluate whether a leader understands the limitations of the technology, verifies AI-generated materials, and is prepared to accept responsibility. This study asks how managers establish the boundaries of appropriate reliance on AI while retaining human accountability for AI-supported decisions.

Methodology

The study followed a qualitative, exploratory, and interpretive design. The empirical material consisted of nine semi-structured interviews, each lasting approximately one hour, with leaders and managers directly involved in AI implementation, AI use, or related organizational change. Participants represented organizations in Russia, Uzbekistan, Kazakhstan, and the United Arab Emirates across banking and fintech, engineering, energy, manufacturing, construction, marketing, IT recruitment, and digital-product development. Their organizations ranged from individual experimentation and pilot projects to the regular integration of AI into key processes.

Participants were selected purposively, with additional elements of snowball sampling. The interviews were conducted in Russian and addressed organizational context, AI-use practices, trust, transparency, information security, the allocation of decision rights and responsibility, and the influence of AI on perceptions of leaders. The materials were anonymized using codes R1-R9. The analysis combined theoretically informed categories - trust, accountability, legitimacy, and socio-technical boundaries - with themes emerging from the data. Individual interviews were coded first, after which related codes were grouped and compared across cases. The study aims for analytical rather than statistical generalization. All interview excerpts presented below were translated from Russian into English by the author.

Findings

The first consistent finding concerns mandatory verification. Participants did not view AI use and managerial control as opposites. Time was saved because the technology accelerated initial processing, while final evaluation remained a human responsibility. R5 summarized this position as follows: "AI requires verification. It requires regular monitoring of its outputs and verification by the leader." AI implementation therefore does not remove managerial control; it redirects control toward assessing whether an output fits the context, quality standards, and organizational objectives.

The second finding is that reliance varies by task type and the cost of error. Participants were more willing to use AI for drafting, meeting notes, initial exploration, translation, data structuring, and routine processing. Strategic decisions, critical infrastructure, formal justification for organizational documents, and sensitive personnel processes were treated differently: these areas required more intensive verification and a final human decision. R8 described an employee who attempted to justify an official document with a ChatGPT exchange: "GPT [...] simply generated the answer you wanted. That is not a valid basis." The example illustrates that a plausible-looking response does not become legitimate evidence merely because it is presented confidently.

The third finding is that appropriate reliance depends on user competence. Participants associated the quality of AI-supported work with the ability to define the task accurately, ask follow-up questions, understand model limitations, and compare the output with professional knowledge. AI literacy in this sense involves more than interface proficiency or prompt-writing skills. It includes the capacity to recognize hallucinations, omissions, false confidence, and contextual mismatch. The same tool can therefore produce different levels of value and risk depending on the user's expertise.

The fourth boundary concerns security and confidentiality. Participants mentioned restrictions on uploading sensitive data, the need to follow information-security requirements, and the risk of disclosing access keys or internal information. Trust in output quality is therefore not equivalent to permission to use a system with any type of data. Organizationally acceptable use requires classifying tasks by sensitivity and defining which data may be transferred to external services.

The fifth and central finding is that responsibility cannot be transferred to AI. R2 stressed: "Whatever AI says, you make the decision, and you will be accountable for that decision." In the interviews, delegating individual operations to AI was not interpreted as delegating accountability. Even when analysis, recommendation preparation, or information sorting was automated, leaders and employees retained the obligation to verify the result and make the final decision. This boundary is especially important when an AI output may become a convenient excuse: referring to the system does not release an organizational actor from professional responsibility.

Finally, the data show a relationship between trust in AI and trust in the leader. Employees evaluate not only the tool but also the way it is introduced: whether the leader explains its purpose and limitations, demonstrates critical use, and aligns promises with actual consequences. When AI visibly reduces routine work and the leader remains transparent and accountable, the technology may support managerial credibility. Conversely, credibility may decline when a leader copies AI outputs uncritically, exaggerates system capabilities, or promises that there will be no adverse consequences while simultaneously using AI to reduce headcount. Leader legitimacy is therefore associated not with the mere presence of AI but with the quality of managerial judgment surrounding its use.

Discussion

The findings define bounded trust as differentiated and controlled reliance on AI that varies with the task, user competence, output quality, data sensitivity, and cost of error. This position differs both from algorithm aversion and from unconditional technological dependence. It is close to the idea of vigilant trust, in which the value of a system is acknowledged while its use remains subject to purposeful caution and verification (Lahusen et al., 2024).

The empirical contribution of the study is to shift attention from the individual attitude of whether one "trusts or does not trust AI" to the organizational conditions of justified use. Verification requirements, data restrictions, escalation rules, mandatory human decisions, and the appointment of a responsible person are not external additions to a technology. They are components of its practical organizational reliability. This interpretation is consistent with the NIST AI Risk Management Framework, which treats validity, reliability, security, transparency,

and accountability as interconnected dimensions of AI governance (National Institute of Standards and Technology, 2023).

For practice, the findings suggest five minimum managerial rules. Organizations should classify tasks according to the cost of error, identify outputs that require mandatory verification, define prohibited data categories, develop critical-evaluation skills alongside tool-use skills, and assign the final decision and accountability to a specific person. Such an approach can support productive human-AI collaboration without reducing human oversight to a formality (Kolbjørnsrud, 2024).

Conclusion. This study shows that organizational trust in AI should not be understood as maximum confidence in the technology. In the interviews, productive interaction with AI was based on bounded and conditional reliance: the system could accelerate and extend work, but its outputs remained subject to verification, contextual interpretation, and human decision-making. The central implication is that organizations should replace the goal of "increasing trust in AI" with the goal of organizing responsible AI use. Leaders must create room for technological experimentation while establishing boundaries that preserve verification, security, and human accountability. Bounded trust is not an obstacle to implementation; it is a mechanism for connecting technological usefulness with organizational responsibility.

The study is limited by its small and heterogeneous sample, reliance on self-reported data, and predominant focus on leaders' perspectives. Future studies could compare the evaluations of leaders and employees and examine how the boundaries of reliance change as organizations move from generative assistants to more autonomous AI agents.

Adabiyotlar, References, Литературы:

1. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114-126. https://doi.org/10.1037/xge0000033
2. Herrmann, T., & Pfeiffer, S. (2023). Keeping the organization in the loop: A socio-technical extension of human-centered artificial intelligence. AI & Society, 38, 1523-1542. https://doi.org/10.1007/s00146-022-01391-5
3. Jarrahi, M. H., Newlands, G., Lee, M. K., Wolf, C. T., Kinder, E., & Sutherland, W. (2021). Algorithmic management in a work context. Big Data & Society, 8(2), 1-14. https://doi.org/10.1177/20539517211020332
4. Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior, 160, 108352. https://doi.org/10.1016/j.chb.2024.108352
5. Kolbjørnsrud, V. (2024). Designing the intelligent organization: Six principles for human-AI collaboration. California Management Review, 66(2), 44-64. https://doi.org/10.1177/00081256231211020
6. Lahusen, C., Maggetti, M., & Slavkovik, M. (2024). Trust, trustworthiness and AI governance. Scientific Reports, 14, 20752. https://doi.org/10.1038/s41598-024-71761-0
7. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1