Keeping the organization in the loop – a concept to support oversight as part of AI-governance

Thomas Herrmann
Information Management and Technology Management, IAW
University of Bochum
Bochum, Germany
thomas.herrmann@rub.de

ABSTRACT

Oversight is a crucial aspect of AI-governance that can only be enacted by a socio-technical integration of technical measures and organizational practices. This is underlined by the concept of "keeping-the-organization in the loop" that complements "keeping the human in the loop". This concept contains seven basic principles of reciprocal exchange that must be implemented and continuously maintained through organizational practices to make human oversight of AI a reality: Managerial activities have to support, prepare and encourage the workforce to exercise oversight. Contextual factors and continuous changes need to be regarded to coordinate continuous evolution of the workforce as well as of AI-technology.

CCS CONCEPTS

Human-centered computing, Human computer interaction (HCI), Interaction paradigms

KEYWORDS

Human-centered AI, AI governance, organizational practices, human-in-the-loop, socio-technical design

1 Background

We understand AI-Governance as a framework that helps to mitigate risks, unfair bias or misuse [1] and in this way supports trust calibration [2], i.e. avoiding over- and under-trust by understanding the potential and limits of AI. Without an understanding of the limits of AI, people's motivation to contribute to AI governance will fade. A key element here is to keep the human in the loop within the workflows in which AI-supported decision making takes place. Humans need to be encouraged and empowered to exercise control and oversight [3]. A prerequisite are interaction modes that allow users to intervene into AI driven procedures [4], veto on AI decisions [5], refine results [6], etc. However, these kinds of interaction modes need to be complemented by appropriate organizational practices. Keeping the human in the loop [7] will not work if people are not allowed to truly exercise oversight and are not supported, trained and encouraged to do so. Job design and organization of work practices must give workers sufficient time, offer appropriate entry points for interventions and provide encouraging feedback, so that they are willing and capable to avoid the problems addressed by AI-governance.

Thus, we have developed the concept of "keeping-the organization in the loop" [8], [9]. Based on empirical research on predictive maintenance we have demonstrated that the management has to provide certain activities that lead to a series of organizational practices and mutual interactions that help to keep the human in the loop. Fig. 1 demonstrates the intertwinement of both loops [10].

The left side of Fig. 1 presents central activities of human oversight when being in the loop during AI-usage:

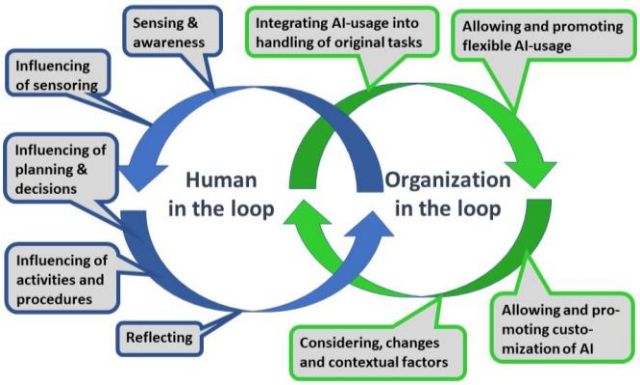

Figure 1: Keeping the Organization in the Loop [10]

1. Maintaining an appropriate degree of sensing and awareness and understanding about the activities of AI and results provided by AI.
2. Based on this awareness and understanding, various ways of influencing the activities and outcomes of AI are possible, such as
   - adaptation of how AI perceives its environment via sensors and how it interprets this input,
   - having an influence on AI's planning of further activities or of procedures of decision making,
   - influencing AI-based procedures and their outcome such as controlling an autonomous vehicle or running a workflow for purchasing goods.
3. Ex-post reflection on the activities and outcomes of AI and their appropriateness.

The right side of Fig. 1 represents activities that must take place by the management if humans should be in the loop on the various

levels of the left side. These activities need to instantiate a series of organizational practices. We suggest that there is an original task to be carried out that is supported by AI and that this support has to be properly aligned with the requirements of the original task. This alignment includes the coordination of the collaboration between humans and AI also covering possibilities for oversight. ("Planning and managing AI-based task handling," Fig. 1). Furthermore, the flexible dealing with AI, such as rejecting or modifying results and procedures must be explicitly allowed and promoted ("Allowing & promoting critical and flexible AI-usage," Fig. 1). This also applies to the customization of AI. Furthermore, the organization must be aware of "…changes and contextual factors" (see Fig. 1) that might influence the way of using AI and task sharing with AI.

2 Organizational practices and reciprocal influences supporting human oversight.

To elaborate more on the managerial tasks indicated in Fig. 1, we refer to a more detailed diagram of the organizational embeddedness of AI, where Herrmann and Pfeiffer [8, p. 1537] present several interacting organizational practices. An adaptation of this diagram represents seven aspects of reciprocal influence being relevant for oversight of AI. "Reciprocal influence" means that organizational practices involve and develop a back and forth between management decisions and the way these decisions are interpreted and adopted through work processes. We conceive organizational practices as interactions and negotiations within an organization that are based on structures and processes that are subject to their own logic [11]. From our point of view, oversight includes two basic aspects:

1. Understanding what is going on, e.g. by means of explainable AI, explorative experimenting, asking human experts etc.
2. Influencing what is going on by intervention, vetoing, triggering teams to adapt AI-solutions etc.

Apparently, to enable oversight by and for an AI-governance framework, a socio-technical intertwinement of technical interaction modes and organizational practices is necessary. Supporting human oversight when using AI is a good example to demonstrate the relevance of including organizational aspects by socio-technical design that can be understood as focusing on the interplay of humans, organization and technical artefacts [12].

The seven aspects of reciprocal influences in Fig. 2 can be described as follows:

1. In the center of Fig. 2 is the managerial task of offering and coordinating the possibilities for oversight into AI based processes, and for the evolution of AI. This coordination is important for organizational units where AI is part of a socio-technical system and where interventions into regular processes or decision-making can influence other workers or organizational units. This may include risks if others do not understand whether an intervention is still active or is already terminated. Thus, managerial coordination establish rules and conventions [13] that guide the workforce by clarifying what kinds of exercising oversight can be initiated by whom and under which conditions. Furthermore, the coordination must specify how the interplay between oversight of and evolution of AI (reciprocal influence #6 in Figure 2), is implemented and which role the technical project team needs to take over for reconfiguring AI.

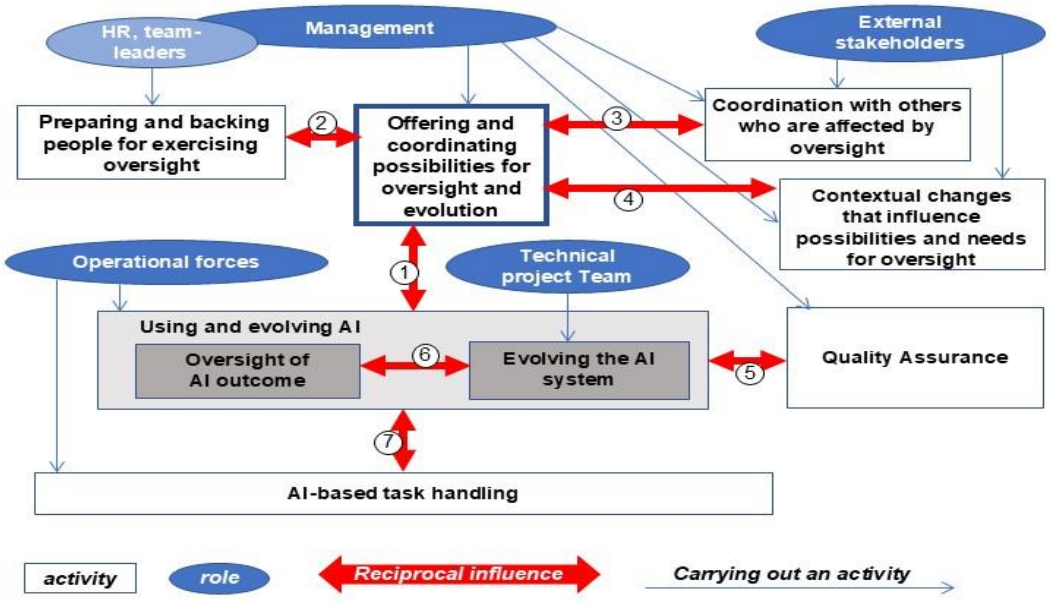

Fig. 2: Management of reciprocal influences to maintain AI-oversight

2. The handling of the original tasks that are supported by employing AI must be organized and prepared in a way that allows for exercising oversight. While preparing the stuff, the rules for coordinating the handling of AI might be negotiated, adapted, and eventually adopted. The technical infrastructure – including artificial intelligence – is not only the subject of oversight, but can itself support the coordination of oversight, e.g. through documentation, by establishing workflows specifying who is involved in an action, etc.
   Making oversight happen requires not only that people are allowed to influence decision making of AI. They also need to be prepared, encouraged and safeguarded to do so by an intensive communicational exchange between AI users and HR. This is an example that the coordinative measures of a governance framework must be coupled with HR development and supported by team leaders so that flexible handling of AI results takes place, such as rejecting or adjusting them by exercising oversight. It must also be part of organizational practice that both, AI's and employees' capabilities, are continuously and reciprocally developed with support of HR for the human side.
3. An organizational unit interacts with other stakeholders who must be informed about the effects of oversight that affect them. It must be specified who in the environment will be aware of measures of oversight, whose interests might be affected, and how the reactions of others need to be taken into account when interventions as part of oversight take place. For example, if the energy management of an AI-based smart home [14] is temporally adapted by its owner, then the installation and maintenance service must be able to be aware of this modification. AI by itself can help to support this awareness.
4. The continuous changes in the context of an AI application must be considered by the management when coordinating the possibilities and needs for oversight. Changes may be triggered by new technologies or by the market. Changes in contracts or laws might require standard procedures that prevent certain types of interventions that violate the regulations. However, the exercising of oversight could also require intervention in order to immediately meet the requirements of new regulations or ethical discourses on AI. Apparently, not only shortcomings of a certain AI application but also the continuous changes in the context of AI usage are further triggers that make adaptation and continuous evolution of AI and its usage necessary.
5. From an organizational point of view, carrying out oversight by adapting AI-outcome and its interplay with the adaptation of AI itself needs supervision and quality assurance on a meta-level. This includes testing of AI and its reconfiguration, evaluating and mutual reflecting whether actions of oversight were reasonable, and assessing the workflows for oversight, for example whether people being allowed to veto against an AI-outcome are sufficiently skilled to do so. In particular, it is a management task to consider whether the same kind of intervening into AI-driven processes is repeated too often and should be avoided by reconfiguring the whole AI-based decision support. Not only the ways of how AI generates its outcome, but also activities of oversight must be explainable to allow for quality assurance. As a consequence of quality assurance, re-configuration or customizing of AI might be initiated and coordinated with the technical project team.
6. Possibilities for oversight of AI-based outcome on the one side and a continuous evolution of AI on the other side are intertwined. Exercising oversight has a twofold possible effect: an immediate adaptation of AI-outcome and a more general adaptation or customization of how an AI-system works. This customization is carried out by the technical project team. AI itself can contribute to the customization by self-adaptation or by making proposals of how it could be customized. Self-adaptation again must be a subject of oversight. Explainability is needed to understand whether the need for correcting AI-outcome should be followed by measures of re-configuring AI-technology. The reconfiguration itself changes the possibilities and conditions of oversight. These changes must be fed back to the workforce and must be regarded by the managerial decision making, organizational practices and the governance framework into which oversight is embedded.
7. Using and influencing AI must be closely related to the task handling that AI should support. The task-handling procedures might have to be adapted and adjusted to support oversight by coordinative measures. For example, employees can be encouraged to first consider for themselves what decision they would make before looking at the AI's proposed decision [15]. By such a measure, evaluating AI results and possibly rejecting them becomes more likely, as well as the calibration of trust in AI. Explainability of AI-reasoning and feedback about the success of task handling must be intertwined for supporting oversight. Considering the actual task helps to assess the effects of AI on the efficiency and quality of task handling.

The reciprocal influences of Figure 2 illustrate that possibility for oversight is a criterion that must be implemented within a holistic, socio-technical approach that integrates organizational practices, individual activities and capabilities, technical artifacts and their evolution, as well as the development of human competencies.

Conclusive remarks

To allow for human oversight in the context of AI governance requires a socio-technical integration of human-centered technical measures [16] with organizational practices. The instantiation of these organizational practices starts with management activities of coordinating the way of task completion with AI and of preparing the staff to exercise oversight. Using AI is closely intertwined with a continuous evolution of AI-technology and the way of its use. This need for evolution is a basic concept of socio-technical design that suggests that socio-technical systems are never complete [17]. Continuous evolution is particularly relevant for AI since it continuously can adapt to changing context such as newly available data, legal regulations or ethical discourses. Furthermore, in the case of AI we have the remarkable constellation that AI is as well the subject of oversight but can also help to conduct oversight by means of documentation, explanation, triggering reflection or proposing measures of customization by itself.

A crucial issue for further research is how managerial activities and succeeding organizational practices can enable people and

motivate them to exercise oversight. It cannot be taken for granted that people are willing to exercise oversight on their job when using AI. It is a necessity that their context of social relationships, values and organizational practices nudges and encourages them to contribute to oversight. We need to understand which kinds of motivation, feedback and recognition are helpful to make oversight a value that is pursued by the workforce when using AI, and which kinds of interaction modes when using AI positively correlate with these motivational aspects. Furthermore, oversight might disturb the quality of AI involvement since not all kinds of human-AI task sharing provide better results than leaving the lead to AI [18].

REFERENCES

[1] O. O. Garibay et al., "Six Human-Centered Artificial Intelligence Grand Challenges," International Journal of Human–Computer Interaction, vol. 39, no. 3, pp. 391–437, Feb. 2023, doi: 10.1080/10447318.2022.2153320.

[2] K. Okamura and S. Yamada, "Adaptive trust calibration for human-AI collaboration," PLoS ONE, vol. 15, no. 2, p. e0229132, Feb. 2020, doi: 10.1371/journal.pone.0229132.

[3] European Commission, C. and T. Directorate General for Communications Networks, and High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI. 2019. Accessed: May 23, 2021. [Online]. Available: https://data.europa.eu/doi/10.2759/346720

[4] A. Schmidt and T. Herrmann, "Intervention user interfaces: a new interaction paradigm for automated systems," interactions, vol. 24, no. 5, pp. 40–45, 2017.

[5] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, "Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices," Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW1, pp. 1–23, 2021.

[6] C. J. Cai et al., "Human-centered tools for coping with imperfect algorithms during medical decision-making," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–14.

[7] R. Crootof, M. E. Kaminski, and W. N. Price II, "Humans in the Loop," 76 Vanderbilt Law Review, no. 76/2/429, 2023, Accessed: Jun. 03, 2022. [Online]. Available: https://scholarship.law.vanderbilt.edu/vlr/vol76/iss2/2

[8] T. Herrmann and S. Pfeiffer, "Keeping the organization in the loop: a socio-technical extension of human-centered artificial intelligence," AI&Soc, vol. 38, pp. 1523–1542, 2023, doi: 10.1007/s00146-022-01391-5.

[9] T. Herrmann and S. Pfeiffer, "Keeping the Organization in the Loop as a General Concept for Human-Centered AI: The Example of Medical Imaging," in Proceedings of the 56th Hawaii International Conference on System Sciences (HICSS), 2023, pp. 5272–5281.

[10] J. Beringer and T. Herrmann, "The contextual framing of the interplay between human and AI – a socio-technical perspective," in Intelligent Systems in the Workplace: Design, Applications, and User Experience, C. Coursaris, P.-M. Léger, and J. Beringer, Eds., Springer, Cham, 2024.

[11] N. Phillips and T. B. Lawrence, "The turn to work in organization and management theory: Some implications for strategic organization," Strategic Organization, vol. 10, no. 3, pp. 223–230, Aug. 2012, doi: 10.1177/1476127012453109.

[12] C. Kirsch, P. Troxler, and E. Ulich, "Integration of people, technology and organization: the european approach," in Symbiosis of Human and Artifact, vol. 20, Y. Anzai, K. Ogawa, and H. Mori, Eds., in Advances in Human Factors/Ergonomics, vol. 20. , Elsevier, 1995, pp. 957–961. doi: 10.1016/S0921-2647(06)80337-0.

[13] G. Mark, "Conventions for Coordinating Electronic Distributed Work: A Longitudinal Study of Groupware Use," Distributed Work, pp. 259–282, 2002.

[14] P. Rodriguez-Garcia, Y. Li, D. Lopez-Lopez, and A. A. Juan, "Strategic decision making in smart home ecosystems: A review on the use of artificial intelligence and Internet of things," Internet of Things, vol. 22, p. 100772, Jul. 2023, doi: 10.1016/j.iot.2023.100772.

[15] R. Fogliato et al., "Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging," May 19, 2022, arXiv: arXiv:2205.09696. Accessed: Jun. 03, 2022. [Online]. Available: http://arxiv.org/abs/2205.09696

[16] B. Shneiderman, "Human-Centered AI," Issues in Science and Technology, vol. 37, no. 2, pp. 56–61, 2021.

[17] A. Cherns, "Principles of Sociotechnical Design Revisted," Human Relations, vol. 40, no. 3, pp. 153–162, 1987.

[18] M. Vaccaro, A. Almaatouq, and T. Malone, "When combinations of humans and AI are useful: A systematic review and meta-analysis," Nat Hum Behav, vol. 8, no. 12, pp. 2293–2303, Oct. 2024, doi: 10.1038/s41562-024-02024-1.