Avoiding Black Box Problems by Assigning an Active Role to Humans in the Control of Autonomous AI: A Methodological Approach

Nerissa Dettling[1], Samira Hamouche[1], Julia Usher[2], Manuel Renold[2] and Toni Waefler[1]

[1] University of Applied Sciences and Arts Northwestern Switzerland (FHNW), School of Applied Psychology
[2] University of Applied Sciences and Arts Northwestern Switzerland (FHNW), School of Business

Abstract. Autonomous AI may shift humans into a passive supervisory role while responsibility remains with them. This paper proposes interpretable primitives as a work-design approach to support meaningful human control. Based on empirical work analysis and Hierarchical Task Analysis (HTA), a method is presented for deriving and evaluating primitive candidates using three criteria: clarity of purpose, process transparency, and granularity. The approach is illustrated using railway disruption management and represents a methodological contribution that will be evaluated in future work.

Keywords: Human-AI Collaboration; Autonomous AI; Interpretable Primitives; Human Control; Hierarchical Task Analysis; Railway Disruption Management.

1 Introduction

This contribution refers to URAI 2026 Track 2 “Journal/Conference-First-Contributions» and is based on a paper published in the AHFE 2026 Conference Proceedings (Human Factors and Systems Interaction):“Avoiding Black Box Problems by Assigning an Active Role to Humans in the Control of Autonomous AI: A Methodological Approach” (Dettling et al., 2026).

Autonomous AI can support humans in complex and safety-critical work. At the same time, it can shift the human role from active control to passive supervision. Humans may then mainly monitor AI outputs and intervene when needed, while still being responsible for the consequences of decisions. This is problematic because humans must still be able to understand, influence, and justify key decisions (Bainbridge, 1983; Endsley, 2023).

Explainable AI is often proposed as a response to black-box problems. Explanations can support understanding, but they do not automatically make the human role more active (Buçinca et al., 2025). A person can receive an explanation and still remain in a passive accept-or-reject role. The original paper therefore takes a broader work-design perspective and focuses on task and role allocation.Table 1. Lorem ipsum dolor sit amet

2 Theoretical Foundation

The paper proposes interpretable primitives as a possible solution approach. Interpretable primitives are small, bounded AI components whose purpose and mode of operation are understandable to humans (Wahde & Virgolin, 2021). Instead of one large AI system that provides complete solutions, several primitives can be combined to support complex work. This changes the role distribution: humans define goals, set priorities, and orchestrate primitives, while AI executes bounded operational subtasks.

A key challenge is to decide what makes a task unit suitable as a primitive. The paper therefore derives three evaluation criteria from the primitive definition. Clarity of purpose means that the primitive has a clear function and supports a recognizable higher-level goal. Process transparency means that the operation can be described at a functional level, including relevant input, operation, and output. Granularity means that the primitive has the right level of detail. It should not be too broad, because then it may become a black box again. It should also not be too small, because then it becomes fragmented and difficult to manage.

3 Methodological Approach

The paper presents a stepwise methodological approach for deriving interpretable primitives from real work. The approach is illustrated with railway disruption management from the HORIZON project AI4REALNET. First, empirical work analysis is used to understand what people do and why they do it. This means that both tasks and goals are collected. Second, the work is represented as a Hierarchical Task Analysis (HTA). HTA structures work into goals, sub-goals, tasks, and subtasks and is therefore useful for linking possible primitives to meaningful task units (Stanton et al., 2017). Third, possible primitives are derived from the HTA through a bottom-up scanning procedure. Each candidate is then assessed against the three evaluation criteria: clarity of purpose, process transparency, and granularity. For example, “identify affected trains” is a strong primitive candidate, whereas “select the best rerouting strategy” is not, because it combines multiple decisions and trade-offs. In this way, the method identifies task units that are meaningful, bounded, understandable, and suitable for human orchestration

4 Conclusion

The original paper contributes a criteria-based method for identifying interpretable primitives for autonomous AI. It argues that black-box problems should not only be addressed by adding explanations, but also by designing a better task and role allocation between humans and AI. Interpretable primitives are intended to support an active human role because humans orchestrate bounded AI functions instead of only approving complete AI outputs. As a methodological contribution, the paper does not yet include an implementation or empirical evaluation. Future work should test whether primitive-based interaction actually avoids the problems of a purely supervisory role.

References

1. Bainbridge, L. (1983). Ironies of automation. Analysis, design and evaluation of man–machine systems, 129–135.
2. Buçinca, Z., Swaroop, S., Paluch, A. E., Doshi-Velez, F., & Gajos, K. Z. (2025). Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills (arXiv:2410.04253). arXiv. https://doi.org/10.48550/arXiv.2410.04253
3. Dettling, N., Hamouche, S., Usher, J., Renold, M., & Waefler, T. (2026). Avoiding black box problems by assigning an active role to humans in the control of autonomous AI: A methodological approach. 17th International Conference on Applied Human Factors and Ergonomics (AHFE 2026). https://doi.org/10.54941/ahfe1007985

4. Endsley, M. R. (2023). Ironies of artificial intelligence. Ergonomics, 66(11), 1656–1668. https://doi.org/10.1080/00140139.2023.2243404
5. Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., & Jenkins, D. P. (2017). Human Factors Methods: A Practical Guide for Engineering and Design (1. Aufl.). CRC Press. https://doi.org/10.1201/9781315587394
6. Wahde, M., & Virgolin, M. (2021). The five Is: Key principles for interpretable and safe conversational AI. 2021 The 4th International Conference on Computational Intelligence and Intelligent Systems, 50–54. https://doi.org/10.1145/3507623.3507632