Human Factors and Systems Interaction, Vol. 230, 2026, 137–144
https://doi.org/10.54941/ahfe1007985

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, 4600 Olten Switzerland
[2]University of Applied Sciences and Arts Northwestern Switzerland (FHNW), School of Business, 4600 Olten Switzerland

ABSTRACT

From a human factors' perspective, the combination of humans and autonomous AI presents a number of challenges. A major problem is the black box nature of AI. Humans are faced with the impossible task of evaluating AI-generated suggestions that they can no longer understand and taking responsibility for them. An effect even appearing when AI provides explanations. Further challenges include difficulties in developing adequate situation awareness, de-skilling, de-motivation, or automation complacency. In our research, we assume that these negative effects on humans are exacerbated by the black box nature of autonomous AI in conjunction with the passive role assigned to humans in terms of supervisory control. To address these two problems while still leveraging the benefits of autonomous AI, we turn to the concept of interpretable primitives. A primitive is an autonomous AI agent with reduced scope, so that its purpose and functioning are easy for humans to understand. To avoid the black box problem, many primitives that are understandable to humans are used instead of a comprehensive but incomprehensible AI. The human's role is to orchestrate the primitives by defining strategies, setting priorities, or directing their deployment. In this way, humans are assigned an active role that includes task characteristics that are considered prerequisites for human engagement and up-skilling. The paper presents operationalized criteria and a method for identifying primitives.

Keywords: Human systems integration, Human-AI collaboration, Automation transparency, Autonomous AI, Primitives

INTRODUCTION

The HORIZON project AI4REALNET (cf. ai4realnet.eu), develops AI-based solutions addressing critical systems (electricity, railway and air traffic management) that are traditionally operated by humans, and where AI systems complement and augment human abilities. In the project, various types of human-AI function allocation are explored: (i) AI assisting humans, (ii) human-AI hybrid intelligence and co-learning, and (iii) full

Received February 3, 2026; Revised April 3, 2026; Accepted April 23, 2026; Available online July 20, 2026

autonomous AI under human supervision. The research presented in this paper focuses on the latter scenario, where a combination of autonomous AI and humans is envisioned.

Autonomous AI in safety-critical control work raises a central challenge from a human factors perspective: as decision logic shifts to automation, human work often moves from active control to monitoring and intervention. This creates a mismatch: responsibility stays with humans, but they may no longer be able to understand, influence, or justify key decisions (Bainbridge, 1983; Endsley, 2023). Against this background, it is common to call for more explanations: if AI decisions are hard to follow, explanations seem like a direct way to make outputs more interpretable and support validation. However, explanations alone do not necessarily resolve the underlying role shift. People can still struggle to mentally reconstruct AI reasoning, and explanations do not automatically change the human decision role (Buçinca et al., 2025). In this work, we therefore take a broader work-design perspective. We assume that many risks in autonomous settings are shaped by how tasks and roles are allocated. From this perspective, explanations can be helpful, but their impact depends on whether the collaboration setup creates an active human role in process control.

This work addresses the methodological challenge outlined above by shifting the focus from "more explanations" to task and role allocation. We propose interpretable primitives (Wahde & Virgolin, 2021) as a way to structure autonomous AI into bounded, human-understandable components. Humans orchestrate these primitives and remain active, rather than only approving AI outputs. The paper contributes a methodological approach for deriving primitives from real work using hierarchical task analysis and theory-based evaluation criteria.

THEORETICAL FOUNDATION

The black-box problem is often described as limited insight into the technical processes that link inputs to outputs, even when inputs and outputs are observable (Fomin, 2022). Beyond explanation-based approaches (Brożek et al., 2024; Dieber & Kirrane, 2020; Fomin, 2022; Lundberg & Lee, 2017), there are proposals that manage black-box risk trough supervision architectures - for example, the "arguing machines" uses two independently trained systems and flags cases of disagreement for human supervision (Fridman et al., 2017). These contributions are very useful. However, they mainly focus on explanation, justification, or other functions of the technical system. In contrast, we add a work-design perspective and focus on how task and role allocation can support meaningful human involvement in autonomous settings. We build on interpretable primitives (Wahde & Virgolin, 2021) as a way to structure AI functionality into bounded building blocks that can be orchestrated by humans.

Interpretable Primitives as a Solution Approach

Whade and Virgolin (2021) describe primitives as components whose purpose and mode of operation are "easily human-understandable". In

their definition, primitives can be compared to program constructs (e.g., sorting, comparing, fetching), which can be combined to carry out complex operations. In this way, complex tasks are completed by many autonomous primitives, each of which is understandable, so that the overall behaviour remains readable and inspectable for the human. This shifts the focus of design from providing explanations of a complex model (explainable AI) to designing the system in such a way that it remains understandable by construction, which enables transparency in automation and, consequently, human control.

This perspective implies a different role allocation between humans and AI. Rather than assigning humans a passive supervisory role (e.g., mainly monitoring and accepting/rejecting AI outputs), primitives enable a collaboration pattern where humans remain responsible for strategic control (e.g., defining goals, setting priorities, choosing between strategies), while the AI executes operational subtasks autonomously. In other words, primitives are not mainly a technical way for breaking down otherwise complex agents into smaller parts, but a structural way to keep the human role active: humans orchestrate, primitives execute.

What Makes a "Good" Primitive: Criteria Derived From the Definition

A challenge in applying the primitive concept to real-world work is that the definition is conceptual. It describes what primitives should be like, but not how to identify them from work domains in a traceable way. For this reason, we translate the conceptual definition into evaluation criteria that can be applied to candidate primitives. These criteria are design-oriented decision aids for determining whether a possible primitive can be seen as a suitable primitive for which interpretability and meaningful human control are preserved.

Based on Wahde and Virgolin's (2021) definition, we derived three core criteria that capture what a primitive must provide:

- Clarity of purpose: A primitive should be immediately understandable in terms of its intended use and the goals that can be achieved with it.
- Process transparency: A primitive's process transparency (Waefler et al., 2003) allows that its operation can be described and checked at the process level, so that humans can follow how the output was produced without needing to understand the internal AI mechanism.
- Granularity: Granularity describes how detailed a primitive is, whether it is a meaningful action that can be combined with actions of other technical or human agents. If the primitive is too comprehensive, there is a risk of bringing back the black box problem. If it is too detailed, it becomes unmanageable and no longer reflects how humans think and make decisions.

To make these criteria useful for systematic assessment, we operationalize them into observable indicators on a 3-level rating scale (low/medium/high). See Tables 1–3 for a detailed set of indicator and rating scale per criteria.

METHODOLOGICAL APPROACH

This paper proposes a stepwise method to derive interpretable primitives from real work. The method follows three steps, which are illustrated using the task of railway traffic controllers during disruption management, an example from the AI4REALNET project: (1) describing the work by eliciting its tasks and subtasks with their goals and subgoals, (2) representing this work as a hierarchical task analysis (HTA; Stanton et al., 2017), (3) Deriving primitives from the HTA.

Table 1: Clarity of purpose criteria derived from the primitive definition.

Criteria: Clarity of Purpose *Immediate understanding of what the primitive is for (goal it addresses)*				
Indicator	**Example Evaluation Question**	**Rating Scale**		
		Low	**Medium**	**High**
Logical link to the goal	Does the unit support the higher-level goal logically and directly (clear contribution), or is the link only loose/ indirect?	The contribution to the higher-level goal is implausible, contradictory, or very indirect; the unit could just as well serve a different goal.	The contribution is generally plausible, but not unambiguous; there is overlap with other goals or the benefit remains unclear.	The contribution is logical and direct; it is clear why this unit supports exactly this goal (clear contribution).
Clearly defined object of reference	Is it clear what the unit refers to (object/elements/ decision object)?	It is unclear what the unit refers to (object/ elements/ decision object); too broad or ambiguous.	The object is recognizable, but not cleanly bounded yet (multiple possible interpretations).	The object is unambiguous and clearly bounded; it is clear which elements/objects are affected.
Purpose is directly understandable from the label	Is the purpose of the primitive recognizable from its label when the usage context is known?	Even with known context (input), the purpose remains unclear or ambiguous from the label.	With known context, the purpose is broadly understandable, but there is still room for interpretation.	With known context, the purpose is unambiguous from the label; the contribution to the goal is immediately clear.

Step 1: Data Collection as a Foundation for the HTA

Data collection methods like interviews or observations are used to collect two types of information needed for an HTA: what people do (tasks) and why they do it (goals). First the overall goal for a selected scenario is identified. Then the tasks carried out to reach this goal are described. During data collection, tasks were structured into main task and subtasks. In addition, participants are asked to describe the subgoals they pursue across phases, so that tasks can be linked to intentions. The data material is then coded for

goals and corresponding tasks. This coding serves as the basis for constructing the HTA in the following step.

Table 2: Process transparency criteria derived from the primitive definition.

Criteria: Process Transparency *Traceability of how the primitive works (steps/rules leading to the result)*				
Indicator	**Example Evaluation Question**	**Rating Scale**		
		Low	**Medium**	**High**
Clear action type	Is it clear what type of action the unit performs (e.g., check, classify, compare, prioritize, generate)?	It is unclear which action type is meant; the unit feels like a container (mixing several different actions) or remains abstract.	The action type is generally recognizable, but not unambiguous, or it contains multiple functions.	The action type is unambiguous and clearly bounded (one clear function type).
Required inputs can be specified	Can you clearly state which data/ information is needed?	It is not clear which information is needed; essential inputs remain implicit.	Some key inputs can be named, but important ones are missing or only described vaguely.	The required inputs are clearly and sufficiently specifiable; it is plausible that this allows the function to be specified in a traceable way.

Step 2: Constructing the HTA

An HTA is a structured method to describe work by decomposing an overall goal into sub-goals and the tasks and subtasks that contribute to them (Stanton et al., 2017). We chose this modelling of work because its hierarchical structure supports a systematic decomposition of complex work into subtasks. This matches the idea of interpretable primitives: a primitives can be defined as an autonomous unit with a limited scope, and HTA helps to identify such units at the subtask level. In addition, the goal-structure makes the intention behind each subtask explicit, which supports defining a primitive‘s purpose. In this study, the coded data (goals and corresponding tasks) is used to reconstruct the overall goal, refine sub-goals, as well as tasks and subtasks in a consistent hierarchy.

Step 3: Deriving Primitives From HTA

Primitive candidates are identified by a bottom-up scanning procedure within the HTA. The screening starts with detailed units at the lower subtask levels and rates each unit against the three evaluation criteria (see Table 1-3 for the full indicators and anchors). Whenever a sub-task does not meet the criteria, the procedure is repeated at the next higher level of the HTA. This continues

until sub-tasks are identified, which achieve high rating across all three criteria and thus can be taken as suitable primitive candidates. In railway dispatching for example a task such as "select the best rerouting strategy" is a poor candidate because it bundles multiple actions and trade-offs, whereas a task such as "identify trains affected by the disruption at the location X during time window Y" is a stronger candidate because its purpose is clear in context and it can be specified with named inputs and understandable rules (e.g., include trains whose planned route intersects segment X within time window Y).

Table 3: Granularity criteria derived from the primitive definition.

Criteria: Granularity *Level of action/abstraction: bounded scope; suitable as a composable unit*				
Indicator	**Example Evaluation Question**	**Rating Scale**		
		Low	**Medium**	**High**
Completeness of the action unit	Is it a self-contained unit with clear boundaries (start/end, input/output), not "in the middle of a process"?	The unit feels fragmented or "in the middle of a process"; start/end and/or input/output cannot be clearly delineated.	The unit can generally be defined, but boundaries or output remain implicit in some cases.	The unit is clearly self-contained and well bounded; it is plausible where it starts/ends and what result it produces.
Combinability with other primitives	Can the unit combine well with other units (clear interfaces, no hidden side effects)?	The unit is tightly coupled or has unclear interfaces; interplay with other units is difficult or leads to hidden side effects.	Combination is possible, but dependencies/interfaces are not fully clear or require additional assumptions.	The unit is easily combinable; interfaces are clear, and there are no indications of hidden side effects.
Cognitive manageability	Is the unit (and its combinations) realistically manageable without overloading the human?	The unit (or the required combinations) is hard to manage; it would likely overload people (too complex or too many building blocks needed).	Generally manageable, but with noticeable effort or risk of overload when combinations are frequent.	Easy to manage; the unit is manageable, and its combinations remain understandable and controllable.

CONCLUSION

Human supervisory control over autonomous AI is a task that normally exceeds human capabilities. This is due both to the black box nature of such AI and to the passive nature of the humans' supervisory control task, which leads to automation complacency as well as de-skilling. However, to exploit

autonomous AI potential and still avoiding negative impact on human performance, we suggest to de-compose complex tasks and to assign simple tasks to autonomous AI. In this way, each AI agent is a primitive that covers a limited scope, so that its purpose and functionality are understandable to humans. This is to avoid the black box problem. On the other hand, humans have the task of defining strategies, setting priorities, and controlling the deployment of primitives. This makes the human role active, which is a precondition for human engagement (e.g. Parker & Knight, 2024).

Compared to explanation-oriented approaches to the black-box problem and the monitoring problem, our approach shifts the focus away from explaining or monitoring black-box behavior toward breaking down technical autonomy into discrete units that humans can intuitively understand and control in their everyday work. While prior work primarily mitigates risk trough explanations, justification modules, or disagreement-based escalations, we contribute a work-design-oriented method for deriving and assessing building blocks from the task structure, with the explicit objective of supporting humans in playing an active role in the process control.

In this paper, we present evaluation criteria and a method to identify suitable primitives for complex tasks, which were developed as part of the HORIZON project AI4REALNET. This approach is intended to guide the subsequent development of concrete primitives for critical system control, which will be evaluated in later project stages in terms of system performance and human acceptance. As a methodological contribution, this work does not yet include an implementation or an empirical evaluation.

ACKNOWLEDGMENT

AI4REALNET has received funding from European Union's Horizon Europe Research and Innovation programme under the Grant Agreement No 101119527 and from the Swiss State Secretariat for Education, Research and Innovation (SERI). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union and SERI. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

Bainbridge, L. (1983). Ironies of automation. Analysis, design and evaluation of man–machine systems, 129–135.

Brożek, B., Furman, M., Jakubiec, M., & Kucharzyk, B. (2024). The black box problem revisited. Real and imaginary challenges for automated legal decision making. Artificial Intelligence and Law, 32(2), 427–440. https://doi.org/10.1007/s10506-023-09356-9

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

Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2012.00093

Endsley, M. R. (2023). Ironies of artificial intelligence. Ergonomics, 66(11), 1656–1668. https://doi.org/10.1080/00140139.2023.2243404

Fomin, V. V. (2022). The Black Box Problem. https://doi.org/10.13140/RG.2.2.31841.99687

Fridman, L., Ding, L., Jenik, B., & Reimer, B. (2017). Arguing Machines: Human Supervision of Black Box AI Systems That Make Life-Critical Decisions (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1710.04459

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1705.07874

Parker, S. K., & Knight, C. (2024). The SMART model of work design: A higher order structure to help see the wood from the trees. Human Resource Management, 63(2), 265–291. https://doi.org/10.1002/hrm.22200

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

Waefler, T., Grote, G., Windischer, A., & Ryser, C. (2003). KOMPASS: A method for complementary system design. In Handbook of cognitive task design (S. 477–502). Lawrence Erlbaum Associates Publishers. https://doi.org/10.1201/9781410607775.ch20

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