Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p27839-cd

Autonomous Systems and the Erosion of Adaptation: Challenges for Human-AI Collaboration

Salvatore Massaiu

Human and Organisational Factors, IFE, Norway. E-mail: salvatore.massaiu@ife.no

The rapid development of Adaptive Artificial Intelligence (AAI) has revived ambitions to reduce or remove human involvement from safety-critical systems. Yet, by embedding only partial or simplified models of human adaptive behavior, current engineering approaches risk undermining the very adaptive capacity they aim to achieve. This paper examines how autonomous systems draw on either top-down formalized adaptation models or bottom-up data-driven learning from experts and shows why both approaches struggle to represent real-world operational decision-making. Human adaptation is grounded on expectations, mental simulation, and experimentation, features that dominant models fail to fully capture. Data-driven systems, meanwhile, cannot distinguish normal from adapted behavior unless designers explicitly identify and encode the cues, evaluative criteria, and deviation strategies underlying adaptive decision-making. As increasing autonomy removes humans from practice, opportunities to observe adaptation diminish. The paper argues that maintaining adaptive capacity requires grounding autonomous systems in naturalistic decision-making models and applying ergonomic principles that support human adaptive contributions in collaborative human-AI systems.

Keywords: Autonomous Systems, Human-AI Collaboration, Adaptation, Decision-making.

1. Introduction

Automation has long aimed to reduce human involvement in industrial processes, typically by encoding repetitive human tasks into reliable mechatronic systems. In safety-critical domains, however, technological limitations have historically required humans to remain in the loop as supervisors, optimizing decision-makers, and safeguards. As long as human operators remained in the loop, the systems' adaptive capacity was maintained (see Figure 1 for a classical example).

With the rapid development of Artificial Intelligence (AI), and in particular Adaptive Artificial Intelligence (AAI), the prospect of reducing or even removing human involvement has re-emerged with new force. The concept of self-adaptive systems, referring to autonomous agents capable of altering their behavior according to variations in the operational context or their own internal state (Maes, 1989; Oreizy et al., 1999), extends beyond the notion of complete automation found in traditional industrial settings.

Fig. 1. A famous instance of adaptive decision-making: due to total loss of power, the Fukushima Daiichi nuclear power plant operators used car batteries as power source for the instruments. Foto: Tepco. https://www.tepco.co.jp/en/nu/fukushima-np/review/review1_2-e.html

These have relied on mechatronic design principles that integrate electrical, mechanical, and control components to automate repetitive tasks in static environments reliably. Automated systems generally require manual adjustment to adapt to evolving environmental conditions and

production objectives, whereas autonomous systems are designed to achieve such adaptability independently.

AAI thus promises systems that can operate robustly in dynamic, uncertain, and interactive environments by modifying their behavior, internal representations, or decision policies at runtime. AAI systems are engineered to adjust their behavior, internal models, or decision-making policies during operation in response to changes in the environment, shifting objectives, or system degradation, increasingly taking over the adaptive decision-making function that was formerly allocated to the human operators.

Two broad approaches currently guide the development of self-adaptive systems. Top-down approaches engineer adaptation explicitly in the system at design stage by defining rules, goals, and predefined adaptation strategies, for instance, through Monitor-Analyze-Plan-Execute over a shared Knowledge base (MAPE-K) loops (Kephart & Chess, 2003) and semantic web language (Lam & Haugen, 2019). Bottom-up approaches derive adaptive behavior form data, including observation of human operators, optimization processes, or machine-learning methods such as reinforcement learning or imitation learning.

The two approaches are not exclusive and are often combined (Huuhtanen et al., 2018). Taken together they are consistent with the way human decision-making displays in operating environments that control dynamic processes, that is, combining data-driven, intuitive cue-response activations (e.g., patter recognitions) and knowledge-based, analytical reasoning. Indeed, both approaches encode human decision-making: in the top-down approach by explicitly formalizing the decision rules and knowledge bases; in the bottom-up approach by implicitly learning them from examples or demonstrations from experts.

Yet, both approaches face challenges when dealing with adaptive decision-making. Human adaptation aligns in real time operational objectives with real-world constrains based on highly variable, subtle, even chaotic, cues and rich, multi-level representations of the work domain. Formalization of such highly complex decision rules into AI algorithms is a formidable challenge (Nguyen, 2025). In the top-down approach the system learns adaptative rules from the training dataset, but unlike humans that have the capacity of learning and generalizing to novel situations, adaptive behavior is confined to the training set.

A further yet frequently overlooked consideration is that adaptive systems remain inherently linked to the human participants within their environment. Machine learning components, in isolation, cannot form fully adaptive socio-technical systems without engagement with human stakeholders. The capacity for adaptation in real-world contexts remains for the foreseeable future dependent on effective coordination and a shared situational understanding between technological elements and human decision-makers. This underscores the need to frame adaptation as a joint human-AI property and to apply ergonomic principles that enable this collaboration in practice.

Hence, the incorporation of human adaptation into autonomous systems reveals a critical challenge: current engineering approaches risk eroding the very adaptive capacities they seek to achieve, because they create systems that incorporate only a small subset of meaningful adaptive behavior, either because they rely on simplified cognitive models, or because they rely on training data that without explicitly knowing the structure of expert adaptation that they contain will never learn. And to complicate further: as human adaptive behavior is based on mental models that are tightly coupled to the specifics of the work-domain, the more the autonomy increases and humans are removed from practice the harder it is to know what real adaptation is in the specific contexts.

This paper argues that avoiding the erosion of adaptation in autonomous systems requires a shift in how adaptation is conceptualized, studied, and engineered. Specifically, three claims are made:

(i) Adaptive systems must be grounded in naturalistic models of operational decision-making, in addition to than standard problem-solving models.

(ii) Human-AI collaboration needs specific ergonomic principles for joint adaptive systems, where humans and technology share control and support each other's adaptive capacities.

Correspondingly, the paper has two sections. First, it examines the nature of human adaptive behavior and explains why sequential models such as MAPE-K fail to capture its continuous, expectation and action-driven character. Second, it explores how adaptive capacity can be maintained at the system level through ergonomics principles for collaborative human-AI interaction.

2. Adaptive decision-making

Human adaptive behavior displays in all socio-technical systems that operate under uncertainty, time-pressure, and evolving situational constraints. While traditional engineering often assumes stable conditions and optimal ways of performing the tasks, and codifies these in instructions and procedures, modern research across cognitive systems engineering, ergonomics, and resilience engineering recognizes that the operators often adapt prescribed and proceduralized tasks to situational contingencies that could not be anticipated by the designer (Jeffroy & Charron, 1997; Roth et al., 1994; Suchman, 1987; Vicente, 1999, Zsambok & Klein, 1997).

Traditional engineering aligns with classical decision-making (problem solving) and standards information processing, often via the computer metaphor of the mind. In these, decisions can be decomposed in discrete elements (e.g., problem formulation, evidence gathering, choice) and sequential processing stages with fixed order (e.g., starting form incoming data, interpreting them, choosing a plan, and executing actions).

A good illustration of this is the MAPE-K (Monitor, Analyze, Plan, and Execute-Knowledge) loop (Kephart & Chess, 2003), the most common architecture for self-adaptive systems. MAPE-K is a cyclical framework of four sequential main stages (Monitor, Analyze, Plan, and Execute) that rely on a shared knowledge base (K) that mimics the traditional sequential information processing model of human cognition. It keeps the strictly sequential flow of information and the inherently reactive start point of the sequence in the incoming data (with limited ability to anticipate changes, explore future scenarios and act accordingly). In MAPE-K adaptation is a discrete event in the analysis (A) and planning (P) stages. Another classical assumption in MAPE-K is that decision is about choice among options (possible configurations) that are examined for their feasibility relative to the goals of the system.

Traditional approaches do not satisfactorily describe the operational decision-making seen in naturalistic, real-work settings which differ from classical approaches in several respects (Table 1). Firstly, operational decision-making has more the character of a continuous control of an activity in which it is difficult to separate the decision-making moments of problem formulation, evidence gathering, and choice (Rasmussen et al., 1994).

Secondly, the information processing flow has no fixed order and the decision sequence does not necessarily start from detection of external data: real-work tasks decision sequences can start at any point of the information processing sequence. For instance, a schedule assigns a task, so that task execution is the start of the decision sequence, which in turn directs the observations, followed by the identification of the system state and the selection of a written procedure. Particularly, observation (data-gathering) is not a passive, reactive start of the sequence but an active process of search and selection.

A distinction is also made, between intuitive, automatic decisions, often called heuristics or recognitions, and analytical decision. Heuristic decisions are intuitive cue-response activations, immediate perceptions of familiar features in the environment (e.g., pattern recognition) that activate actions that have been previously successful. Analytical decisions rely on knowledge-based (semantic) reasoning involving mental models (explicit mental representations of the end-means relations and the constraints of a work domain).

The distinction is particularly important for understanding the logic of adaptation. Adaptation requires reasoning and domain knowledge to develop new rules and plans of actions. But the activation typically occurs when heuristic decisions fail. Cue-response associations work well in familiar situations, but might not be effective in unfamiliar contexts, in which case a mismatch between the state of affairs in the environment and the expectations of the actor will be experienced. If the mismatch is not again solved by finding discriminating information, knowledge based (semantic) analysis involving metal models is needed and will take place.

Another important difference is that action is not the final, separate stage of the decision process (its outcome) but an integral part of the process. Decisions that require analysis (problem solving) imply experimentation. This can occur internally, by mental simulation, evaluating different plans for their effects against the chosen goal. But it can also consist in physical experiments (e.g., trial and error, diagnostic test) in which the system or environment are acted upon for producing real effects. In both cases, the resulting data will be compared with the expected data (the data expected by the simulated plan or by the experimental test).

This means that decision-making involves expectations about what and where to look for data, probing the environment (acting to see what happens, and generating new data), and testing hypotheses (about what is happening, included root causes) (Massaiu, 2023; Rasmussen, 1986, Schön, 1983).

It is reassuring to see that several of these insights are incorporated in the AWARE (Assess, Weigh, Act, Reflect, Enrich) architecture (Sanwouo et al, 2025), an attempt to surpass the reactivity and lack of learning of the MAPE-K approach. AWARE does not only react to data but incorporates anticipation. Also, the impact of actions on goal achievement is anticipated and evaluated, alternative strategies are experimented and their outcomes assessed, future scenarios simulated, and preventive action taken. AWARE claims to immediately assess the impacts of actions and update its models iteratively (so that patterns can be recognized when handling similar tasks in the future or decisions adapted to new requirements).

It is unlikely that the developers of AWARE started with insights from research into operational, naturalistic decision-making. Yet their efforts to overcome the limitations of the MAPE-K system led them to identify recurring tropes in naturalistic decision-making research. That human cognition provides a practical foundation for designing adaptive system is a well-established concept. This paper suggests that even more foundational insights may come from the study of real-world decision-making in operational domains.

Table 1. Classical vs. Naturalistic Decision-Making

Dimension	Classical DM	Naturalistic DM
Decision structure	Fixes sequence of discrete stages (problem formulation, evidence. gathering, choice).	Continuous control activity without predefined sequence.
Information processing flow	Reactive; begins with detection of external inputs.	May begin at any point. Observation is active and guided by expectations.
Cognitive modes	Analytical reasoning (problem solving).	From immediate recognitions to analytical (triggered by expectation / data mismatches).
Role of action	Action as outcome of decision.	Action integral to decision. Includes experimentation and hypothesis testing.
Adaptation logic	Selection among options.	Adaptation arises from mismatch detection. If no discriminating cues found, mental simulation and experimentation.
Environment	Assumes stable, predictable, optimizable tasks and conditions.	Assumes dynamic, uncertain, variable conditions that cannot be fully anticipated at design stage.
View of the operator	Executor of procedures.	Executor of procedures and active problem solver who interprets cues and revises plans dynamically.

3. Adaptation in collaborative Human-AI

Autonomous systems are set to expand the boundaries of technological agency compared to automated systems, but they will still be embedded in a broader environment in which interaction with humans occur. Much of the ergonomic issues know in human-automation research are still valid for human-AI interaction. In both cases, the human must understand what the system is doing and intervene when needed. The classical themes of automation research of

transparency, trust, out-of-the-loop, and loss of expertise remain in Human-AI interaction. However, AI systems that exhibit self-adaptive behavior are intrinsically less predictable, as their functioning and internal decision rules are not explicitly known, even to their designers, and may shift with learning. This transforms the interaction from a human monitoring/supervising a technology into a collaborative relation between humans and autonomous or semi-autonomous agents. Humans and artefacts become now elements of a joint system in which each influences the other's behavior. New ergonomics requirements arise: visibility of intentions, predictability of behavior, negotiation of authority, awareness of each other's constraints and assumptions, to name some.

The question with adaptability in human-AI collaboration is twofold: (1) how to maintain the human adaptive capacity within the system, and (2) how to ensure the overall, system-level adaptive capacity. In both cases human adaptive capabilities must be ensured and supported.

Designing systems that support human adaptive capabilities is inherently challenging because adaptation operates according to principles very different from those that guide traditional engineering. Most design and engineering methods are about system optimization, which for real-work tasks translates into anticipating possible disturbances, prescribing optimal ways to perform the tasks, and controlling against deviations. This approach optimizes reliability but suppresses adaptation rather than supporting it.

Nonetheless, engineering approaches for supporting people's adaptive capabilities, and specifically the ability to prevent, detect and recover deviations from normal functioning, exist. Cognitive Systems Engineering, Naturalistic-Decision-Making, Activity Theory and other modern approaches to the study of human work with technologies, emerged primarily because traditional design and analysis approaches could not account for the adaptive, situated nature of human work.

Rather than assuming that systems can be fully specified at design time, these modern approaches recognize that complex socio-technical systems operate in environments that produce conditions that cannot be exhaustively anticipated. In such settings, the operators make autonomous decisions in real time, recognizing anomalies and creating ad-hoc responses. CSE's express this point effectively by stating that the operators complete the design, are "the designer's representative on site" in peculiar, although infrequent, situations (Rasmussen and Goodstein, 1987). Hence the design should focus on creating conditions that enable adaptation, e.g., by making constraints visible, by supporting hypothesis testing and exploration, by exposing uncertainties, and by creating representations that match the operators decision strategies and cognitive control modes.

Creating conditions (technologies, organizations and training programs) that support the ability to adapt is a principle still valid for human-AI systems. Yet it needs to expand to cover the collaborative dimension between human and AI, because task allocation is no longer static but reconfigures itself as demands, uncertainties, and operator states change. Key design principles here are:

- Adjustable autonomy. Operator-selectable autonomy ranges visualizes current mode, actionable explanations, and clear indicators of "who is driving" to help the human understand what the system is doing and how to step in when adaptation becomes necessary.
- Shared control. Human and AI contribute simultaneously, the system mediates with arbitration rules.
- Functional overlap and redundancy. Enable the system to continue operating under degraded conditions by shifting responsibilities dynamically.
- Fail-safe at boundaries. When the system approaches the boundaries it adopts conservative behaviors and provide the operator information on time to intervene.

Table 2 below provides an illustrative example in the field of autonomous driving of how to maintain adaptive capacity by applying NDM-based insights and human-AI collaboration design principles.

Table 2. Illustrative autonomous-driving case

In autonomous driving the vehicle controls steering and speed while the driver continuously supervises and takes over when requested (SAE Level 2 and 3).
Adaptive transition of control
In SAE levels 2/3 automated driving adaptive capacity depends less on emergency take-over performance than on how drivers anticipate and manage the limits of automation. Experienced drivers actively monitor whether the vehicle remains within its design envelope and disengage automation before known problem areas, such as roundabouts, minor roads, temporary road works, or socially negotiated traffic situations.
Human-AI Collaboration
Shared control includes graded shifts, like small steering corrections or speed modulation that probe traffic behavior and test driver expectations. Adaptation emerges as mixed-initiative interaction, preserving human adaptive capacity rather than treating all interventions as a failure of automation.
Links to Theory
-Adaptive decision-making: Anticipatory thinking keeps operation within the operational design envelope and avoids last-second takeovers.
-NDM grounding: Expectation/reality mismatches trigger adaptation; probing actions serve as hypothesis tests.
-Ergonomics for Human-AI collaboration: Shared control and adjustable autonomy enable graded transitions, preserving the human's role in adaptation.

4. Conclusion

Existing approaches to engineering adaptive autonomous systems incorporate only a limited subset of the adaptive capabilities required for resilient functioning in complex socio-technical settings. Both rule-based and data-driven methods struggle to represent the continuous, situated, and anticipation-driven nature of human adaptative decision-making as they are not capitalizing on models and findings from operational, naturalistic decision-making research. Increasing levels of autonomy also diminish opportunities to observe human adaptation, thereby limiting the empirical basis for modelling adaptive behavior.

To counteract these tendencies, the design of autonomous systems must be informed by naturalistic models of operational decision-making and supported by methodological approaches that can uncover the cues, heuristics, and evaluative criteria employed by expert practitioners. Furthermore, as adaptability becomes a joint property of human-AI systems, ergonomic principles such as adjustable autonomy, shared control, and functional overlap become essential for maintaining system-level resilience. Preserving adaptive capacity therefore requires designing technologies, organizations, and training programs that support, rather than displace, the human contribution to adaptation.

References

Huuhtanen, A., Mäkitalo, N., & Mikkonen, T. (2018). Architecting Self-adaptive Software Systems. In C. Pautasso, F. Sánchez-Figueroa, K. Systä, & J. M. Murillo Rodríguez (Eds.), Current Trends in Web Engineering (Vol. 11153, pp. 59–70). Springer International Publishing.

Jeffroy, F., & Charron, S. (1997). From safety assessment to research in the domain of human factors: The case of operation with computerised procedures. Human Factors and Power Plants, 1997. Global Perspectives of Human Factors in Power Generation., Proceedings of the 1997 IEEE Sixth Conference On, 13–15.

Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. Computer, 36(1), 41–50.

Lam, A. N., & Haugen, Ø. (2019). Applying semantics into service-oriented iot framework. 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), 1, 206–213.

Massaiu, S. (2023). Operator Decision Strategies in Nuclear Control Centres: A Domain-Specific Information Flow Map. Proceeding of the 33rd European Safety and Reliability Conference, 2388–2395.

Nguyen, H. T. (2025). Continuous Learning and Adaptation in Autonomic Systems for Industrial Applications: A Human Centric Approach. University of Oslo, Norway.

Maes, P. (1989). How to do the Right Thing. Connection Science, 1(3), 291–323.

Oreizy, P., Gorlick, M. M., Taylor, R. N., Heimhigner, D., Johnson, G., Medvidovic, N., Quilici, A., Rosenblum, D. S., & Wolf, A. L. (1999). An architecture-based approach to self-adaptive software. IEEE Intelligent Systems and Their Applications, 14(3), 54–62.

Rasmussen, J. (1986). Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering. North Holland.

Rasmussen, J., & Goodstein, L. P. (1987). Decision Support in Supervisory Control of High-Risk Industrial Systems. Automatica, 23(5), 663–671.

Rasmussen, J., Pejtersen-Mark, A., & Goodstein, L. P. (1994). Cognitive systems engineering. Wiley.

Roth, E. M., Mumaw, R. J., & Lewis, P. M. (1994). An empirical investigation of operator performance in cognitively demanding simulated emergencies

(NUREG/CR-6208). Nuclear Regulatory Commission.

Sanwouo, B., Quinton, C., & Temple, P. (2025). Breaking the Loop: AWARE is the New MAPE-K. Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering. Association for Computing Machinery, New York, NY, USA, 626–630.

Schön, D. A. (1983). The Reflective Practitioner: How Professionals Think in Action. Ashgate Publishing Limited.

Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge University Press.

Vicente, K. J. (1999). Cognitive work analysis: Toward safe, productive, and healthy computer-based work. Routledge.

Zsambok, C. E., & Klein, G. A. (1997). Naturalistic decision making. Routledge.