Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
©2025 ESREL SRA-E 2025 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P6888-cd

Automation criterion for Small Modular Reactors

Roberto Mascherona

Mott MacDonald, France: E-mail: roberto.mascherona@mottmac.com

Romanas Puisa

Mott MacDonald, United Kingdom. E-mail: romanas.puisa@mottmac.com

Martina Artioli

Mott MacDonald, France. E-mail: martina.artioli@mottmac.com

Alan Nelson

Mott MacDonald, United Kingdom. E-mail: alan.nelson@mottmac.com

As it happened with basic software algorithms in 80s, Artificial Intelligence (AI) powered algorithms are slowly but surely becoming integral part of control systems in safety critical industries, including the risk averse nuclear field. The level of control will surely vary across different domains, and given the conservatism of the nuclear industry, we should not expect in a Nuclear Power Plant (NPP) fully autonomous control systems soon. However, the introduction of higher levels of automation is likely, where AI-based automation is given more responsibility, with the human still being able to take control when necessary. We argue that the essential criterion for high level automation is the assurance that the control transition between AI and human (in either direction) is achieved with acceptable risk. The paper attempts to explain what this risk is and how to make it acceptable in the nuclear settings, and beyond.

Keywords: AI, risk, nuclear, autonomous control, decision-making, safety critical functions, human, Small Modular Reactors

1. Introduction

Although energy generation by nuclear processes is well-established, it is constantly subject to innovation within strict regulatory regime. At its heart, innovation aims to improve the business model of energy production, making it more competitive vis-à-vis conventional carbon intensive energy sources, including higher social responsibility through safer and environmentally conscious designs. The Small Modular Reactor (SMR) [Rowinski et al., 2015] is the relatively new outcome of such innovation that is inherently safer, not least due to effective use of passive safety measures relying on fundamental physical phenomena. However, the passive safety measures, although highly preferable, are insufficient to deal with all accident scenarios. To fill the gaps, active measures, which rely on electrical power and/or human action, are also used. Nowadays, the automation of safety control functions in NPPs is standard, mainly limited to basic functions [Huang et al., 2023] as a decision support system processing the large amount of available data (e.g., looking for anomalies, managing alarms, and others). More advanced forms of automation, where the control of some hazardous processes (e.g., criticality control) is shared between the human and the machine (computer) or is solely controlled by machine is yet to be seen. However, this is the direction of travel, in the long term, despite the inherent conservativism of the nuclear domain (risk and uncertainty aversion, suspicion of the new and trust in the old, the use of high levels of redundancy, etc.) that has kept the human in/on the loop and as the fallback (i.e. "the big red button"). To advance to higher levels of automation, technological, social, legal, and other barriers must be lifted. Thus, the practice shows automation can lead to unintended

consequences, some of them ironically due to the very presence of the human [Bainbridge, 1983]. The human presence introduces a precarious phase when the control is transferred from the machine to the human, or vice versa. This transient phase (control takeover) is considered highly risky [Baum, 2024], and, as such, is one of the main barriers to safe automation. The automation challenge is further exacerbated with the use of Artificial Intelligence (AI), and in particular the Narrow Artificial Intelligence (NAI) [a], for doing 'thinking'. The NAI is a relatively new technology which offers several advantages, albeit at the cost of its limitations that are yet to be fully discovered and understood. Due to these limitations, the trust in adequately controlling radiological release and other hazards solely by the machine is low.

In the nuclear settings, the controlled physical processes (e.g., reactivity, coolant temperature, rate of coolant circulation, positioning of control rods) are highly dynamic, volatile, and unforgiving – precise timing and the accuracy of control are of the essence. A few examples are: Chernobyl Disaster (1986, Ukraine), Three Mile Island Accident (1979, USA) and SL-1 Accident (1961, USA). Hence, the inadequate takeover (prolonged, spontaneous, incorrect, incomplete) is a major hazard associated with automation. An example incident happened at Arkansas Nuclear One (2010, USA) when the control transfer over the control rod from automatic to manual was not duly completed, leading to excessive power (rapidly increased within some 40 seconds) and coolant pressure [NRC Information Notice 2011-02].

We argue that the key criterion (possibly there are others) of getting to high levels of automation—where the control over hazardous processes is shared between the human and the machine or done by the machine alone under human's supervision—is the acceptability of risk of the transient phase of control takeover. The non-trivial question is how to achieve the takeover risk acceptability? This paper aims to answer this question.

Note, the paper falls outside any research on AI per se, including on its suitability in safety critical settings. The paper merely assumes that AI is in principle suitable. In this way, our work contributes towards a better understanding of the risks involved in the use of AI to co-control safety critical processes.

2. Methodology

The purpose is to explain the pathway towards the acceptability of the takeover risk. First, we define the adopted notions of risk, uncertainty, and hazard. Then we define the risk acceptability criteria in the nuclear settings (mainly, but not exclusively, in the UK) and beyond. Following that, we elaborate how the risk acceptability can be achieved.

2.1. *Definitions*

In this paper, the definition of risk conforms to ISO 31000:

Definition 1: Risk is uncertainty in (achieving) hazard control.

This definition is equivalent to the one used by the functional safety standard IEC 63187 [Inge et al., 2023]. The Society of Risk Analysis (SRA) Glossary defines uncertainty as "imperfect or incomplete information/knowledge …" [Aven et al., 2018]. The adopted definition of the hazard is as follows [Levenson, 2004]:

Definition 2: A hazard is a system state or set of conditions that, together with a particular set of worst-case environmental conditions, will lead to a loss.

The reference to worst-case assumptions is native in the nuclear settings. The use of conservative, precautionary principles, and

[a] Narrow AI (NAI) can be defined as the production of systems displaying intelligence regarding specific, highly constrained tasks, like playing chess, facial recognition, autonomous navigation, or locomotion.

methodologies to address uncertainties and risks is a must in this domain. The principle of Defence in Depth (DiD) is a case in point. DiD implies the use of multiple defences against operational anomalies, faults, and hazards, making sure the issues are avoided in the first place, then controlled and effectively mitigated if it comes it [ONR SAP, 2020].

Following from the assertion that the inadequate takeover is a major hazard associated with automation, we can assume the definition of *automation hazard*:

Definition 3: Automation hazard is inadequate transfer of responsibilities between the machine and human (or vice versa) that leads to the loss (temporal or permanent) of control over radiological hazards.

The radiological hazards in question are essentially these (see IAEA Fundamental Safety Principles and Safety Standards):

- Inadequate control of radiation exposure.
- Inadequate confinement of radioactive material.
- Inadequate control of radioactive waste.

The underlying assumption adopted in this paper is that all reasonably practical Automation Levels (ALs) in the nuclear domain will involve the human in (co-controlling) or on (supervising) the loop, i.e. the case of fully autonomous operations is ruled out.

2.2. Risk Acceptability

'So Far As Is Reasonably Practicable' (SFAIRP), also known as 'As Low As Reasonably Practicable' (ALARP), is a statutory requirement for risk acceptance in the UK. The principle makes no reference to the risk definition, its metric/mathematical construct, its magnitude, nor requires it the risk to be quantified. To declare the risk ALARP, one must demonstrate, by providing compelling and cogent argument supported by robust evidence, that all reasonably practicable risk controls are in place and that the risk is demonstrably reduced to the point when any further attempt to reduce it would be grossly disproportional to the benefits gained (i.e., further reduction in risk). The risk reduction to ALARP involves the application of Relevant Good Practice (RGP) to reduce such risks, if such practice is available, and/or the introduction of various safety features (design and/or operational) to address known weaknesses (i.e., potential causal factors) in the given design and implement precautionary measures (safety margins, DiD etc.) against the uncertainty in successful hazard control (recalling the risk definition in Section 2.1).

In addition to reducing the risk to ALARP, it must be acceptably low (or at least tolerable), considering that the ALARP can also be achieved for relatively high risks. The risk tolerability is normally demonstrated through probabilistic (quantitative) safety analysis (PSA) against the numerical targets (limits) [ONR SAP, 2020].

In summary, the takeover risk acceptability (or at least tolerability) can be achieved by:

I. Identifying and addressing the potential causal factors behind the automation hazard (as defined in Section 2.1) for a given plant design. The causal factors can be identified via a hazard analysis (see Section 2.3).
II. Identifying and applying the RGP (incl. industrial standards) to those causal factors and beyond.
III. Introducing precautionary measures (safety margins, DiD etc.) against the uncertainty (associated, inter alia, with the incomplete knowledge of causal factors) in successful hazard control.
IV. If required, modelling the safety control function towards the application of PSA to demonstrate risk tolerability.

The above criteria for risk acceptability have chiefly been informed by the engineering practice in the nuclear domain. However, it is quite universal and would apply more widely.

Given that the identification of causal factors in the lead up to the automation hazard is not trivial, the paper further elaborates this point in

the subsequent sections, listing generic scenarios for inadequate takeover of control.

2.3. *Hazard analysis*

We believe, the Systems-Theoretic Process Analysis (STPA) [Levenson, 2004] is a 'made to measure' method for hazard analysis (HA) of control functions that may give rise to the automation hazard (Section 3.1). We do not explain the STPA per se in this paper, for the reader should refer to the amply available information on STPA (e.g., [STPA Handbook, 2018]). With the help of STPA, we have identified unsafe control actions (UCAs), and their common causal factors, during the control takeover between the human and the NAI-powered machine, as summarised in Section 2.3.3. Note, the HA took cognisance of fundamental limitations of both NAI and humans in the context of control takeover (Section 2.3.1). Other input to the HA was:

- The automation hazard as a system-level hazard that the UCAs are identified against.
- Automation levels (AL) relevant to the nuclear domain, and their requirements along with derived responsibilities for the controllers (Section 2.3.2).
- Safety control diagrams for ALs, showing how the co-control is implemented at various ALs; only the diagrams for AL3 and AL4 are shown in this paper (Section 2.3.3).
- Generic control actions—on the part of the human and the machine—subjected to analysis (Section 2.3.4).

2.3.1. Limitations

Both NAI and humans have known abilities (cognitive capacities) and limitations. The abilities are used to inform the design of various functions, whereas the limitations guide the hazard analysis. This section lists the relevant limitations of NAI known to date. The relevant NAI limitations are [Sabry, 2023]:

- *Lack of Generalisation:* NAI are highly specialised and cannot transfer knowledge from one domain to another.
- *Data Dependence:* The performance of NAI heavily relies on the quality and quantity of the training data.
- *Limited Understanding:* These systems lack human-like understanding, empathy, and common-sense reasoning. They can process and analyse data but do not comprehend context or make nuanced decisions.
- *Inflexibility:* NAI cannot adapt to new or unforeseen situations without explicit reprogramming or retraining. This inflexibility limits its usefulness in dynamic or complex environments.
- *Interpretability:* It can be challenging to understand and explain how NAI systems arrive at their decisions, making it difficult to trust and validate their outputs.
- *Resource Intensive:* Developing and maintaining NAI systems can be resource-intensive, requiring significant computational power and expertise.

To these more obvious limits, we should add others such as the legal and regulatory barriers, the ethical and social concerns and the integration challenges.

The limitations associated with human controllers are numerous and relatively well known, hence not furthered in this section (see for example [Reason, 1995]). Anyway, we can mention the most relevant such as cognitive limitations (as limited attention, information overload, biases and memory constraints), physical limitations (as fatigue, stress and reaction time), performance variability (unlike machines human performance can vary significantly based on several factors as health and motivation) and subjectivity to emotional and psychological factors.

2.3.2. Automation Levels

According to NUREG, automation refers to the use of technology to perform tasks that were previously carried out by human operators. Building on NUREG definition and other works [NUREG-0700, Alberti et al., 2023], where there is a description of the possible Als involving

incremental responsibility and relevance to automation, we adopt a definition of AL focused on functional process control:

- Level 1- Manual Control: the human controller is fully responsible for monitoring, decision-making and control actions.
- Level 2 – Basic Automation (Decision Support): the human controller is still the primary decision-maker but supported by tools such as alarms, displays or trend analysis - processed information by the machine.
- Level 3 - Intermediate Automation (Shared Control): the human controller shares control with the machine but retains authority to override and/or intervene upon defined fallback triggers (e.g., exceedance of safety margins, automation faults, departure from design envelope). This AL can allow the implementation of control transfer in both directions, albeit the human is always the fallback.
- Level 4 - High Automation (Supervisory Control): The machine is in control under constant human supervision (locally or remotely), who can intervene in emergency or upon request from the machine.
- Level 5 – Full Automation (Autonomous): The machine is in full autonomous mode of control—the sole decision-maker. The human is only involved in strategic, high-level decisions (e.g., maintenance).

2.3.3. Safety Control Diagrams

These are functional models of how the co-control at various ALs can be implemented. derived from the generic control model published in Figure G-2 (179p) of the STPA Handbook [STPA Handbook, 2018]. We assume that the automation is realised as a closed loop control system, analogous to the one assumed in functional safety standards (e.g., the IEC 61508 family).

As explained in Section 2.3, the co-control only really happens at AL3 and AL4, and hence the issue of control takeover is only relevant there. Below, the AL3 and AL4 diagrams are outlined and briefly explained.

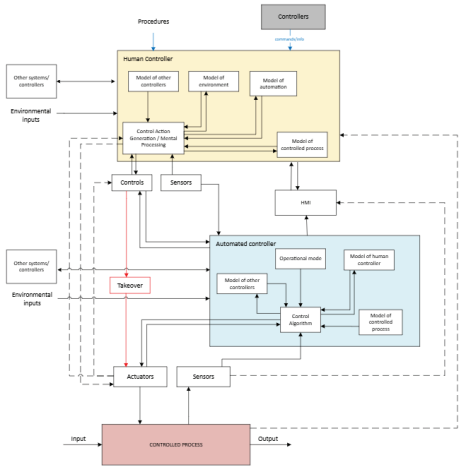

Figure 1: Level 3 Control Diagram

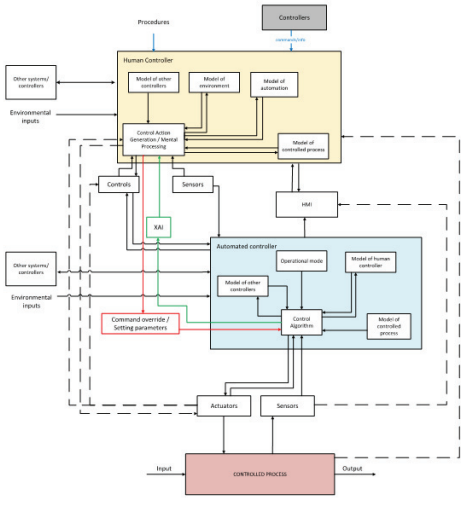

Figure 2: Level 4 Control Diagram

In Level 3 (Figure 1) the machine performs routine tasks and informs the human of significant events. AI augments automation by enabling adaptive decision-making (e.g. real-time optimization) and identifying precursors to abnormal conditions. Human and machine in this Level have a comparable level of responsibility, nevertheless the human must monitor automation

behaviour and intervene during abnormal situations and oversee AI-driven recommendations maintaining situational awareness and avoiding over-reliance on automation. Note that in this Level human and automation have similar levels of responsibilities (the placement of the tiles shows them one under the other for size issues).

In Level 4 (Figure 2) the human supervises automation, intervening only in emergencies or when it is requested. Machine controls most routine operations independently, relying on predefined logic and algorithms. AI acts as a decision-making agent, performing predictive control, real-time anomaly detection and complex optimization. Moreover, it must provide summaries of its actions and status updates to the human and rise alerts when human intervention is necessary. On the other hand, automation still relies on human input for highly ambiguous/critical situations or regulatory compliance. The human agent monitors performance metrics and system health, approve and can override AI decisions and conduct periodic reviews of automated actions to ensure compliance with safety protocols. The main challenge for the human is to ensure readiness to intervene during emergencies and understanding complex AI-driven decisions, especially under time pressure.

2.3.4. Generic control actions

The following high-level, generic (design agnostic) control actions (derived from the AL descriptions) can be assumed for the human and the machine for AL3 and/or AL4:

- Human to watch for fallback triggers;
- Machine to watch / self-diagnose for fallback triggers;
- Machine requests takeover by human;
- Human requests from machine to give back control;
- Human takes over control from machine (i.e., overrides the machine);
- Human passes control to machine;
- Machine takes over control from human (ie, overrides human).

2.3.5. Hazardous scenarios (results)

The generic control actions may lead to the following unsafe behaviours (only a fragment is shown) on the part of the human and machine at either AL3 or AL4:

- Human not watching for fallback triggers when machine is in control or continues watching for already appeared fallback triggers rather than taking action of immediate takeover.
- Machine does not watch / self-diagnose for fallback triggers or miss self-diagnoses or miss interprets fallback triggers.
- Machine watches for wrong fallback triggers or stops watching / self-diagnose for fallback triggers too earlier, before the takeover has started or finished.
- Machine does not request takeover by human when fallback triggers are showing or requests takeover by human too late when fallback triggers are already showing, or the control process is already in its critical phase that is supposed to be controlled by human.
- Human does not request from machine to give back control when fallback triggers are showing or requests from machine to give back control when no fallback triggers are showing and the control process is in its critical phase that is supposed to be controlled by machine.
- Human requests from machine to give back control with undue delay, after the fallback triggers started showing or does not complete the takeover.
- Human does not take over control from machine (on machine request) when critical fallback triggers are showing (eg indicating machine faults) or takes over control when no fallback triggers are showing, when not being ready to do so (i.e. not fully cognisant of the current state of controlled process or environment).

- Human does not pass control to machine during critical phases of control when only machine can control safely or passes control to machine when/despite fallback triggers are showing.
- Machine does not take over control from human during critical phases of control when only machine can control safely or does not take over control from human when the human is inactive / showing incapacity.
- Machine takes over control from human too late (with undue delay) when critical phases of control have already started or when the human has been inactive / showing incapacity for safe control.

The summary of causal factors behind these unsafe scenarios is as follows.

NAI related causal factors:

- Feedback: feedback regarding critical process information can be wrong, wrongly delivered, or not delivered, due signal transmission/sensor failures, etc.
- Controller hardware: powering issues (no power, brownouts), random hardware failures, overheating (due to poor control of ambient temperature), or damage by excessive radiation.
- Controller software: wrong assumptions / wrong software requirements, other software glitches, incomplete coverage in NAI training data. etc., controller can be overwhelmed with the intensity of input data or tasks or make wrong priorities (e.g., focusing on less safety critical tasks), can misinterpret (buy in) the noise in feedback / sensor data.

Human related causal factors:

- Trust and confidence: the human can suffer from both the lack and excess of trust in automation.
- Training and skills: lack of proper training or experience can prevent the human from correctly interpreting fallback triggers and other critical information.
- External and Physical Factors: human is subject to external and internal stimuli (physical, cognitive etc.) that can cause hook / steal the attention from the controlled process and/or degrade performance to the point of inability to respond efficiently and effectively.

Finally, there can be causal factors related to the human machine interface (HMI) when the feedback information on fallback triggers, controlled process or the automation is not properly presented (visually or audibly) to the human, due to dashboard design or other issues (e.g., dashboard can be obstructed or switched off).

3. Results discussion

The hazard analysis has identified multiple scenarios in which control takeover failures can compromise the safety of nuclear processes. These failures stem from unsafe control actions, which may result from human errors, AI limitations, or inadequate interface design. Key risks include Human Factors, AI limitations and Human Machine Interface issues. These findings emphasize the need for a rigorous design phase that proactively addresses automation hazards. Specifically, future automation systems should incorporate countermeasures as:

Resilient Control Handover Mechanisms: Ensuring smooth transitions between human and machine control through redundancy, fail-safes, and structured takeover protocols.

Enhanced Operator Training and Decision Support: Designing automation to support human oversight, with real-time adaptive assistance and intuitive HMI feedback.

AI Trust Calibration: Preventing over-reliance on AI by ensuring system transparency, interpretability, and alignment with established nuclear safety principles.

Despite these insights, limitations exist. The analysis relies on current knowledge of AI capabilities, which are evolving, and does not account for unforeseen emergent behaviours.

Additionally, the uncertainty in identifying all potential causal factors remains a challenge, reinforcing the need for conservative design principles such as Defense in Depth (DiD).

4. Conclusions and possible future developments

This study highlights control takeover as a critical automation hazard in nuclear settings, where failures in transition between human and AI control can lead to loss of oversight over radiological risks.

A primary takeaway is that automation in nuclear control systems should not seek to replace human judgment but rather augment it with well-structured supervisory mechanisms. To achieve risk acceptability, design efforts should prioritize the organization of control transfer protocols, develop AI systems with verifiable safety margins and transparent decision-making and ensuring human operators remain actively engaged, avoiding over-reliance on automation.

A logical next step would be to apply this methodology to a real-world case study. One particularly relevant application is the automated initiation of reactor startup sequences, a process requiring both automation and human oversight due to its complexity and safety implications.

References

Rowinski, Marcin Karol, Timothy John White, and Jiyun Zhao (2015). "Small and Medium sized Reactors (SMR): A review of technology." *Renewable and Sustainable Energy Reviews* 44: 643-656.

Bainbridge, Lisanne (1983). "Ironies of automation." Analysis, design and evaluation of man–machine systems. *Pergamon*. 129-135.

Qingyu Huang, Shinian Peng, Jian Deng, Hui Zeng, Zhuo Zhang, Yu Liu, Peng Yuan (2023) "A review of the application of artificial intelligence to nuclear reactors: Where we are and what's next", *Helyion,* e13883

Baum, Seth D. (2024). "Assessing the risk of takeover catastrophe from large language models." *Risk Analysis*.

NRC Information Notice 2011-02: Operator Performance Issues Involving Reactivity Management at Nuclear Power Plants.

NRC INFORMATION NOTICE 2011-02: Operator Performance Issues Involving Reactivity Management at Nuclear Power Plants.

Inge, James, et al. (2023) "IEC 63187: engineering safety into complex defense systems." *Safety in an Agile Environment: the 2023 Annual International Systems Safety Summit and Training*.

Leveson, Nancy (2004). "A new accident model for engineering safer systems." *Safety science* 42.4: 237-270.

Leveson, Nancy G. (2016) *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

Office for Nuclear Regulation (ONR) (2020). Safety Assessment Principles for Nuclear Facilities 2014 Edition, Revision 1.

Sabry, Fouad. (2023) *Narrow artificial intelligence: fundamentals and applications*. Vol. 167. One Billion Knowledgeable.

Kelley, Troy D., and Lyle N. Long. (2010) "Deep Blue cannot play checkers: The need for generalized intelligence for mobile robots." *Journal of Robotics* 2010.1: 523757.

Bewersdorff, Arne, et al. (2023) "Myths, mis-and preconceptions of artificial intelligence: A review of the literature." *Computers and Education: Artificial Intelligence* 4: 100143.

Sieg, Stephan. (2023) *AI Art and Its Limitations: Narrow Intelligence and Visual Indulgence*. MS thesis.

Leveson, Nancy. (2004) "A new accident model for engineering safer systems." *Safety science* 42.4: 237-270.

Levenson N. G. and Thomas J. P. (2018), STPA Handbook, MIT Partnership for Systems Approaches to Safety and Security (PSASS).

Reason, James. (1995) "Understanding adverse events: human factors." *BMJ Quality & Safety* 4.2: 80-89.

Alberti, Anthony L., et al. "Automation levels for nuclear reactor operations: A revised perspective." *Progress in Nuclear Energy* 157 (2023): 104559.