Human Error, Reliability, Resilience, and Performance, Vol. 132, 2024, 8–15
https://doi.org/10.54941/ahfe1004875

Challenges in Applying Human Reliability Analysis in Systems Containing Artificial Intelligence

Martin Rasmussen Skogstad[1], Ronald Boring[2], and Jan Hayes[3]

[1]NTNU Social Research, Studio Apertura, Trondheim, Norway
[2]Idaho National Laboratory, Idaho Falls, ID, USA
[3]RMIT University, GPO Box 2476, Melbourne VIC 3001 Australia

ABSTRACT

Can we use the same methods to analyze human reliability, if actions, tasks, and interactions change? This paper discusses three challenges of using traditional static human reliability analysis (HRA) on systems that include AI-elements: 1) how to incorporate and include AI in the quantification of human reliability, 2) how to apply HRA to changing tasks and working conditions, and 3) how to include indirect effects to human reliability.

Keywords: Human reliability analysis, Artificial intelligence

INTRODUCTION

Through the history of artificial intelligence (AI) there have been several cycles of (AI)-summers, with attention, optimism and promises of a new future, followed by AI-winters, as interest fades, funding dries up and promises were not fulfilled. We are currently in an AI-summer, and many are (once again) convinced that this time winter is not coming. Many industries have implemented AI-elements over the last few years through aspects such as computer vision, natural language processing, speech processing, image and text generation, and while industries with major accident potential apart from the automotive industry are not among the early adopters, it seems likely that suitable applications will be found here as well. AI can fundamentally change the role of humans in the system, leading to the question: Can we use the same methods to analyze human reliability, if actions, tasks, and interactions change?

This paper is based on the authors' experiences with risk, safety and human reliability analysis and interviews conducted in the "Consequences of fundamental changes in risk regulation (RISKY)"-project.

This paper discusses three challenges of using traditional static HRA on systems that include AI-elements: 1) how to incorporate and include AI in the quantification of human reliability, 2) how to apply HRA to changing tasks and working conditions, and 3) how to include indirect effects to human reliability.

Human Reliability Analysis

Human reliability analysis (HRA) is defined as, "Any method by which human reliability is estimated" (Swain, 1990, p. 301), generally resulting in a description of one or more human actions and the potential for human error presented in a qualitative and/or quantitative manner. HRA is generally applied to critical human actions in industries with major accident potential (Rasmussen, 2016). As the name implies the focus in an HRA is the human; however, a systems approach (Reason, 2000) is generally applied to consider how context and conditions influence human performance and human error probability (HEP). HRA methods can be used both to retroactively investigate events (Boring et al., 2017), or prospectively as part of a risk assessment (e.g. Quantitative Risk Assessment (Falck, no date), Probabilistic Risk Assessment (U.S.NRC, 2020), Probabilistic Safety Assessment (IAEA, 1992)). This paper discusses traditional static HRA as part of a prospective analysis, assessing human reliability in accident scenarios.

The definition of HRA used in this paper is from Alan Swain, whose work within the HRA field has had tremendous impact, with the method, Technique for Human-Error Rate Prediction (THERP; Swain, 1963, 1964; Swain & Guttmann, 1983), published in 1983 for assessing the reliability of operators in nuclear power plant control rooms being the largest contribution. THERP is (arguably) the first HRA method and has been an inspiration to many of the methods that have been developed since (Boring, 2012). It is, however, not just of historic interest, as numbers from THERP are still cited and used today (Arigi, Park and Kim, 2021; de Morais, Moura and Ramos, 2023). The continued use of THERP highlights that the challenge presented in this paper – validity of methods once the context and conditions change – is not a new one within HRA, as THERP has been used across many industries and with newer technology than was available in a 1983 nuclear power plant control room. While the longevity of THERP demonstrates the relevance of this issue, it should be mentioned that many HRA methods have been developed since THERP and have been made with both newer technology and other industries in mind (e.g., Boring et al., 2016; Boring & Rasmussen, 2016; Parry et al., 2013; Ramos et al., 2020).

Artificial Intelligence

AI can be defined as: "Technology that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations or decisions influencing the environments they interact with" (EASA, 2023, p. 6). Approaches adopted include machine learning (algorithms improving as they are exposed to data), deep learning (a subset of machine learning using multilayer neural networks and often large amounts of data), or logic- and knowledge-based approaches (using a combination of logic programming and a knowledge database) (EASA, 2023).

In the context of this paper an AI-system is one that assists a human either through decision support or through performing tasks on its own, e.g., as a decision support system for control room operators to mitigate task overload as presented by Mietkiwicz et al. (2024).

CHALLENGES

Challenges in Including AI in the Analysis

HRA methods generally produce an HEP through a combination of standard values (e.g., task types in SPAR-H (Gertman et al., 2005), generic task types in HEART (Williams, 2015), or control modes in CREAM (Hollnagel, 1998)) modified through a set of context factors (e.g., Performance Shaping Factors [PSFs] in SPAR-H, Error Producing Conditions in HEART, or Common Performance Conditions in CREAM).

The challenge in including AI-systems in the analysis has at least three components:

1) The technical part of including AI in the method. HRA methods have clear boxes of task types and context factors and including aspects that are not specifically included in a box is difficult.
2) We are unlikely to understand the full extent of the AI-system and how it will impact human reliability.
3) To access the full capabilities of an AI system it must be allowed to learn and change, leading to any analysis of system behavior and human-system interaction being quickly outdated.

Current AI systems, especially those based on deep learning, involve millions or billions of parameters. This complexity makes it challenging (if not impossible) to understand how decisions are made or to reliably predict behavior. This can potentially lead to unexpected new error types and unknown impacts in terms of human reliability.

Challenges Due to AI Altering the Human Role

The challenges of performing HRA on systems that include AI go beyond the simple challenge of how to include them in an analysis. The inclusion of AI can fundamentally change work in a way that requires changes to how HRA – and risk analysis in general – is performed.

AI can be introduced and included in many forms. In this paper we are discussing AI-systems that assist humans in or alleviate them from tasks or decisions. This would change how work is performed, how the systems are interacted with and which potential errors traps humans are facing. The combination of human and AI in decision-making is often described metaphorically in terms of a human placed somewhere in regard to a decision-making loop; Human In The Loop (HITL), Human in the Loop For Exceptions (HITLFE), Human On The Loop (HOTL) and Human Out Of The Loop (HOOTL) (Ross and Taylor, 2021).

In industries with major accident potential, it is unlikely that the human will leave the decision loop (HOTL and HOOTL) in the near future. In HOTL the human is removed from the decision loop, but humans provide feedback to the AI system as a way of training it over time. HOTL requires a system that can operate fairly well on its own and can fail without large consequences. Here, humans act as trainers or teachers for the AI guiding it toward gradually better outcomes. As this is effectively a form of trial and error learning it is unlikely to be found in industries with a major accident

potential, HOTL systems are unlikely to be used where an HRA would be applied (although analysis of the reliability of the trainer could in itself be interesting). In HOOTL, the human is removed from the decision loop. Once this stage is reached, there will be no more human actions that require reliability assessments. There are, however, still many obstacles to overcome before this becomes a common occurrence.

In the short to medium term, solutions like HITL and HITLFE are more realistic and are already seen as tasks are being automated in control rooms. From a human reliability perspective, there is a risk that humans may become overly reliant on AI systems as they improve and gradually take a large role in decision making. This could potentially lead to a lack of critical oversight, lack of confidence, skill degradation, boredom and motivational loss among the users. This can result in failure to detect errors made by the AI-systems or failure to intervene when necessary. Potentially, this could lead to a situation where humans are intended to be in the loop, but as reliance grows and system understanding decreases, we are gradually entering an unintended HOOTL-situation. This could add a few digits to the known saying of 99% boredom and 1% sheer terror, making it 99.9999% boredom and 0.0001% not realizing you should have taken control. From an HRA perspective, it is challenging to quantify human performance in these cases, as it would be a gradual process drifting into an unsafe space as system trust grows, reaction time increases and likelihood of taking control decreases. In some situations workload could also increase due to automation. An example of this is made by Sullenberger (known for his remarkable landing on the Hudson River) when he explains that a last-minute runway change now requires multiple systems to be reset instead of the pilot just working things out for himself, which for a skilled pilot would be faster and involve a lower workload (Sullenberger and Zaslow, 2009).

Indirect Effects on Human Reliability

The third challenge of using traditional static HRA on systems that include AI-elements is including indirect effects. Ensuring that humans are adequately trained to work with all systems in the workplace is essential. AI-systems are no exception, but as they represent a new type of system and a changing system, it is likely that there could be additional difficulties in ensuring that users have adequate training. There can be a gap between the capabilities of the AI system and the capabilities expected by the users, leading to misuse or misinterpretation of AI outputs (e.g. Bunz, 2019; Long & Magerko, 2020). As AI systems learn and adapt, humans need to continuously update their understanding and skills to interact effectively with these systems. This constant adaptation can be challenging to analyze and quantify in terms of reliability.

The integration of AI can change the work culture and dynamics within an organization. It could alter the number of employees needed, status changes within the organization and status changes for the entire professions involved. Kongsvik et al., (2020) found several safety challenges as new technology eroded the role of traditional seamanship, while Sullenberger talks about how

the perceived status of airline pilots have dropped from one step below an astronaut to being one step above a bus driver (Sullenberger and Zaslow, 2009). These changes can have indirect effects on human reliability, which might be overlooked in traditional HRA models.

DISCUSSION

This paper presents three main challenges with applying conventional HRA to systems that have AI-elements: 1) how to incorporate and include AI in the quantification of human reliability, 2) how to apply HRA to changing tasks and working conditions, and 3) how to include indirect effects to human reliability.

The first challenge could in part be met through including AI in the existing task types and context factors, as most methods include software and HMI in one or more of these. Many HRA methods are generic in terms of what technology is used (e.g., Petro-HRA: Blackett et al., 2022; Bye et al., 2017; SPAR-H: Gertman et al., 2005) and the input to the analysis is generally on whether the human-technology interaction works well and whether it has a positive or negative impact in terms of human performance. In SPAR-H the PSF Ergonomics/HMI states that: "Aspects of human-machine interaction (HMI) are included in this category. The adequacy or inadequacy of computer software is also included in this PSF" (Gertman et al., 2005, p. 24). In Petro-HRA the HMI PSF "refers to the quality of equipment, controls, hardware, software, monitor layout, and the physical workstation layout where the operator/crew receives information and carries out tasks" (Blackett et al., 2022, p. 62). A second possibility would be to include it in a factor of (task) complexity (Rasmussen, Standal and Laumann, 2015; Rasmussen and Boring, 2016). A third possibility would be to include it in through context factors that include whether or not the operator has sufficient training to deal with the system and situation (e.g., the Experience/Training PSF in SPAR-H or PetroHRA; Laumann and Rasmussen, 2016a). If we at some point start to consider the AI-system more like a person than software (even though that seems like science fiction at the moment) it could be included in context factors that cover interaction and team work (e.g., Teamwork in PetroHRA; Laumann & Rasmussen, 2016b). While perhaps the easiest challenges to deal with, including AI-systems through existing task types and context factors, would still have limitations as the values in the methods would not have been set with this system in mind. The ideal solution, if we are still going to use static HRA, would be to create new task types and context factors intended for use on AI-systems.

New working conditions where humans are moved to a supervisory role fit well within many current HRA uses. However, situations where humans very rarely do anything and still are expected to take control will create dangerous situations where it would be easy to be overly optimistic in analyzing human reliability.

The indirect effects of AI-systems in terms of dynamic training requirements, dynamic system knowledge needs, changed status hierarchies and

changes to culture will be new challenges for human reliability analysts to capture through the qualitative data collection part of the analysis.

In a retrospective analysis, including AI is much more straightforward. In a retrospective analysis we already know the outcome and we already know how all the systems behaved.

Traditional static HRA methods have survived many technological innovations and while they could still be applied in AI-systems, we should be critical about whether that would be the best solution, or if the introduction of AI could be a good opportunity to look towards new ways of including human reliability like dynamic or computation-based HRA (Rasmussen et al., 2017; Ulrich et al., 2017, 2020; Li and Mosleh, 2019).

CONCLUSION

If the future still includes risk analysis and we still have humans included in the system, we will need ways of including the human aspect. We cannot expect current HRA methods to stay relevant in every possible future. We need to develop methods so that we can stay relevant.

ACKNOWLEDGMENT

The authors would like to acknowledge The Research Council of Norway for funding this research through the Consequences of Fundamental Changes in Risk Regulation (RISKY; project number 315302).

REFERENCES

Arigi, A. M., Park, G. and Kim, J. (2021) 'An approach to analyze diagnosis errors in advanced main control room operations using the cause-based decision tree method', Energies, 14(13), p. 3832.

Blackett, C. et al. (2022) The Petro-HRA Guideline Revision 1 Vol. 1. IFE/E-2022/001. Halden, Norway: Institute for Energy Technology.

Boring, R. L. (2012) 'Fifty years of THERP and human reliability analysis', in. Probabilistic Safety Assessment and Management (PSAM11), Helsinki, Finland: Idaho National Lab. (INL), Idaho Falls, ID (United States).

Boring, R. L. et al. (2016) 'Human Unimodel for Nuclear Technology to Enhance Reliability (HUNTER): A Framework for Computation-Based Human Reliability Analysis', in 13th International Conference on Probabilistic Safety Assessment and Management (PSAM 13).

Boring, R. L. et al. (2017) 'Retrospective Application of Human Reliability Analysis for Oil and Gas Incidents: A Case Study Using the Petro-HRA Method', Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61(1), pp. 1653–1657. Available at: https://doi.org/10.1177/1541931213601900.

Boring, R. L. and Rasmussen, M. (2016) 'GOMS-HRA: A method for treating subtasks in dynamic human reliability analysis', in L. Walls, M. Revie, and T. Bedford (eds) Risk, Reliability and Safety: Innovating Theory and Practice, pp. 956–963.

Bunz, M. (2019) 'The calculation of meaning: on the misunderstanding of new artificial intelligence as culture', Culture, Theory and Critique, 60(3–4), pp. 264–278.

Bye, A. et al. (2017) The Petro-HRA Guideline. IFE/HR/E-2017/001. Halden, Norway: Institute for Energy Technology.

EASA (2023) Artificial Intelligence Roadmap 2.0 – Human-centric approach to aviation. European Union Aviation Safety Agency.

Falck, A. (no date) Quantitative Risk Assessment. Available at: https://www.dnv.com/services/quantitative-risk-assessment-1397.

Gertman, D. et al. (2005) The SPAR-H Human Reliability Analysis Method. NUREG/CR-6883 INL/EXT-05-00509. Idaho Falls, ID: Idaho National Laboratory.

Hollnagel, E. (1998) Cognitive reliability and error analysis method (CREAM). Elsevier.

IAEA (1992) Probabilistic safety assessment: a report by the International Nuclear Safety Advisory Group. Internat. Atomic Energy Agency.

Kongsvik, T. et al. (2020) 'Re-boxing seamanship: From individual to systemic capabilities', Safety science, 130, p. 104871.

Laumann, K. and Rasmussen, M. (2016a) 'Experience and training as performance-shaping factors in human reliability analysis', in European Safety and Reliability Conference, ESREL 2016. European Safety and Reliability Conference, ESREL 2016, Glasgow, Scotland.

Laumann, K. and Rasmussen, M. (2016b) 'Suggested improvements to the definitions of Standardized Plant Analysis of Risk-Human Reliability Analysis (SPAR-H) performance shaping factors, their levels and multipliers and the nominal tasks', Reliability Engineering & System Safety, 145, pp. 287–300. Available at: https://doi.org/10.1016/j.ress.2015.07.022.

Li, Y. and Mosleh, A. (2019) 'Dynamic simulation of knowledge based reasoning of nuclear power plant operator in accident conditions: Modeling and simulation foundations', Safety Science, 119, pp. 315–329.

Long, D. and Magerko, B. (2020) 'What is AI literacy? Competencies and design considerations', in Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–16.

Mietkiewicz, J. et al. (2024) 'Enhancing Control Room Operator Decision Making', Processes, 12(2), p. 328. Available at: https://doi.org/10.3390/pr12020328.

de Morais, C. P. M., Moura, R. N. and Ramos, M. (2023) 'Which human reliability analysis methods are most used in industrial practice?–A preliminary systematic review', in. ESREL 2023, Southampton, England.

Parry, G. et al. (2013) IDHEAS–a new approach for human reliability analysis. Idaho National Lab. (INL), Idaho Falls, ID (United States).

Ramos, M. A. et al. (2020) 'A human reliability analysis methodology for oil refineries and petrochemical plants operation: Phoenix-PRO qualitative framework', Reliability Engineering & System Safety, 193, p. 106672.

Rasmussen, M. (2016) The development of performance shaping factors for the PetroHRA method: A human reliability method for the petroleum industry. Department of Psychology, Faculty of Social Sciences and Technology Management, Norwegian University of Science and Technology.

Rasmussen, M. et al. (2017) 'The Virtual Human Reliability Analyst', in Advances in Human Error, Reliability, Resilience, and Performance: Proceedings of the AHFE 2017 International Conference on Human Error, Reliability, Resilience, and Performance, July 17–21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA. Springer, pp. 250–260.

Rasmussen, M. and Boring, R. L. (2016) 'The implementation of complexity in computation-based human reliability analysis', in Risk, Reliability and Safety: Innovating Theory and Practice. Proceedings of the 26th European Safety and Reliability Conference, ESREL. Risk, Reliability and Safety: Innovating Theory and Practice. Proceedings of the 26th European Safety and Reliability Conference, ESREL, Glasgow, Scotland.

Rasmussen, M., Standal, M. I. and Laumann, K. (2015) 'Task complexity as a performance shaping factor: A review and recommendations in Standardized Plant Analysis Risk-Human Reliability Analysis (SPAR-H) adaption', Safety Science, 76, pp. 228–238. Available at: https://doi.org/10.1016/j.ssci.2015.03.005.

Reason, J. (2000) 'Human error: models and management', BMJ, 320(7237), pp. 768–770. Available at: https://doi.org/10.1136/bmj.320.7237.768.

Ross, M. and Taylor, J. (2021) 'Managing AI decision-making tools', Harvard Business Review [Preprint].

Sullenberger, C. and Zaslow, J. (2009) Sully: The untold story behind the miracle on the Hudson. William Morrow.

Swain, A. D. (1963) A method for performing a human-factors reliability analysis. SCR-685. Sandia Corporation.

Swain, A. D. (1964) THERP. SC-R-64-1338. Sandia Corporation.

Swain, A. D. (1990) 'Human reliability analysis: Need, status, trends and limitations', Reliability Engineering & System Safety, 29(3), pp. 301–313. Available at: https://doi.org/10.1016/0951-8320(90)90013-D.

Swain, A. D. and Guttmann, H. E. (1983) Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report. NUREG/CR-1278, SAND-80-0200, 5752058, p. NUREG/CR-1278, SAND-80-0200, 5752058. Available at: https://doi.org/10.2172/5752058.

Ulrich, T. A. et al. (2017) 'Operator Timing of Task Level Primitives for Use in Computation-Based Human Reliability Analysis', in Advances in Safety Management and Human Performance. AHFE 2017, Los Angeles, CA: Springer.

Ulrich, T. A. et al. (2020) 'Dynamic Modeling of Field Operators in Human Reliability Analysis: An EMRALD and GOMS-HRA Dynamic Model of FLEX Operator Actions', in Advances in Safety Management and Human Performance. AHFE 2020, Springer, pp. 346–352.

U. S. NRC (2020) Probabilistic Risk Assessment (PRA). Available at: https://www.nrc.gov/about-nrc/regulatory/risk-informed/pra.html.

Williams, J. (2015) 'HEART—a proposed method for achieving high reliability in process operation by means of human factors engineering technology', in Safety and Reliability. Taylor & Francis, pp. 5–25.