The Human Component of Safety in Defence

Katie J Parnell[1], Georgina Mason[1], Oliver Malpass[1], Isabel Holtby[1], Nicola Turner[2] & Andrew Leggatt[2]

[1]Defence Science Technology Laboratory, Porton Down, Wiltshire, UK, [2]Trimetis, Bristol, UK

SUMMARY

Autonomous and Artificial Intelligence (AI) systems are being integrated into Defence at speed, with likely significant safety implications to the human component. This paper provides an overview of the human component of safety with respect to current Defence guidance and policy. Key areas of current UK Ministry of Defence (MOD) Human Factors (HF) guidance have been identified that need to keep pace with the rapid developments in Autonomous and AI systems. These areas will need to be addressed to tackle future safety challenges that impact human performance within safety critical systems. The practical implications for safety critical areas and ways to mitigate them through updates to government-based guidance will be presented. In addition, an exciting, novel and innovative Defence specific Human Reliability Assessment (HRA) tool is under development to capture and assess human performance in the context of risk in current Defence systems, to support the assessment of a broad range of military equipment and tasks.

KEYWORDS

Safety, Human-AI interaction, Defence, Human Reliability Assessment, Artificial Intelligence

Introduction

In contrast to civilian sectors, the trade-off between risk and benefit in Defence can be less clear and within an operational setting, high risks may need to be taken without clearly identified short term benefits. Safety is a key concern for the Ministry of Defence (MOD): activities it conducts either during training exercises or when deployed are often high risk and can cause harm (MOD, 2018). Within the MOD a significant safety concern is the performance and reliability of military equipment when in operation and the service personnel who operate it. Humans obviously play a crucial role in ensuring safety within complex and high-risk environments throughout the MOD. The Introduction to the MOD System Safety Management document states that "*the user must be involved in safety throughout the lifecycle, from setting appropriate safety requirements through to managing residual risk and feeding back information on changes of capability requirement, desired changes of use or problems in service*" (MOD, 2018, p.16).

Artificial Intelligence (AI) initiatives are rapidly advancing and Defence must embrace them safely and responsibly to gain military advantage). The Joint Services Publication - JSP 936 on Dependable AI in Defence states that "*our understanding of related risks, safeguards and assurance standards continues to evolve" (MOD, 2024, page i)*. Also, within their 'Ambitious Safe and Responsible AI Policy' the MOD state that "*Realising the benefits of AI – and countering threats and challenges associated with the use of AI by others – is one of the most critical strategic challenges of our time*." (MOD, 2022, p. 1).

This paper presents ongoing work that is being conducted to better understand, assess and design for the human component of safety within Defence with respect to current systems as well as future requirements; against the backdrop of increasing advancement and adoption of AI and autonomous

agents. There are two components to this work: Part 1 of this work presents critical knowledge gaps in Defence guidance surrounding the human component of safety in relation to the increased adoption and integration of AI and autonomous agents; Part 2 presents the development of a Human Reliability Assessment (HRA) tool specifically designed for application in Defence.

Artificial Intelligence and Safety in Defence

MOD policy is grounded in the importance of meaningful human control and context-appropriate human involvement in AI (e.g. MOD, 2022). Part 1 of this work aims to identify future Human Factors (HF) safety challenges that may arise with the increased advancement and integration of AI into Defence.

A review of academic literature and Defence guidance was conducted to identify areas of the guidance that surround the human component of future safe and resilient systems that need to be updated to keep pace with the integration of AI and autonomous agents. The replacement of humans with AI and autonomous systems changes the potential risks, the types of accidents and the requirement for assurance (Endsley, 2023). The integration and interoperation of multiple automated / autonomous systems will lead to highly complex challenges and will change the role of the humans involved. The implications that this has for safety need to be fully understood in order to provide effective safety management plans for future systems.

Six safety critical areas were identified:

1. Safety in human–autonomy teams (HAT): Defence lacks frameworks for managing the added complexity of human–AI collaboration, including trust calibration, mental models, oversight and responsibility. Without these, operators may under or over rely on autonomous systems, thereby creating safety risks.
2. The future human role in system safety: As autonomy increases, human roles shift toward monitoring, supervision and recovery—these are tasks that humans perform poorly without targeted design support. Issues such as mental underload, workload spikes, complacency and skill fade introduce new hazards.
3. Modelling the human component in system safety: Current modelling tools cannot represent the adaptive, evolving or collaborative behaviours of AI-enabled systems. Defence needs validated methods to predict dynamic human–AI interactions and anticipate emerging risks.
4. The human component of safety assurance: Existing assurance processes do not account for model bias, unpredictable AI behaviour or the sociotechnical context of human–AI interaction. Defence requires assurance approaches that integrate human, organisational and technical factors across the full lifecycle.
5. Systems of systems (SoS) and safety: Autonomous agents operating across interconnected platforms introduce emergent risks that current tools cannot assess. Defence lacks a consistent SoS approach to understand cross system dependencies and human roles within them.
6. The human component of system resilience: Humans remain essential for detecting anomalies, responding to disruption, anticipating future events and learning from experience. Defence needs guidance on designing AI-enabled systems that support these resilience behaviours and mitigate human variability.

Human Reliability Assessment in Defence

Part 2 of this work involves the development of a Human Reliability Assessment (HRA) tool appropriate to the Defence context. This tool aims to revolutionise HRA in the Defence sector ultimately contributing to improvements in safety and a more cohesive approach.

Indeed, given the increasing integration of automation, human autonomy teaming, and the use of AI into Defence, the importance of the consideration of human performance in the context of risk for systems remains critical. HRA tools are applied to capture the human performance factors that contribute to risk and unsafe system performance in complex, safety critical domains. A number of safety critical domains have developed specific HRA tools that capture the risks innate to their specific context, such as nuclear power generation, aviation and rail. No Defence specific tool currently exists to reflect its unique operating context. Currently, when HRA is applied in the Defence sector, it involves methods that were developed from other sectors, and each HRA practitioner tends to modify the method and "shoehorn" it in as best as possible (Leggatt et al, 2025). However, these ad hoc approaches, although pragmatic, can lead to inconsistent application of the HRA analysis and the context in which existing HRA methods were developed are often different to the unique contexts faced in Defence.

The Human Error Assessment and Reduction Technique (HEART) (Williams, 1985), for example, is one of the most widely applied HRA techniques in system safety analysis. HEART was developed to assess the human reliability of nuclear power plant operators when implementing standard or emergency operating procedures in a fully airconditioned, well-lit control room when shift patterns and rest breaks are strictly enforced by legislation. Many Defence tasks are conducted in conditions which are mandated by the mission; this may involve a challenging thermal, rest and threat environment which challenges human performance. The newly developed Defence specific HRA technique aims to address these requirement gaps and help safety practitioners apply the technique in a more consistent manner.

The development approach adopted by the research team has been to establish what particular requirements are needed for a Defence specific HRA method and then to systematically develop these aspects to satisfy these requirements. During the requirements elicitation phase, it was identified that adaption / enhancement of an existing method was preferred as this would minimise training needs, have the potential to build on previous validation works, and is likely to be most acceptable to the safety and regulatory community.

Consequently, the team set about modifying the HEART method with suitable Generic Task Types (GTTs) and additional Performance Shaping Factors (PSFs) that are particularly encountered in Defence. This is a significant research and development activity that is currently ongoing.

It is therefore the aim of this work to produce a first of its kind Defence HRA tool that will improve overall systems design by tailoring the assessment of human performance to the Defence context, increasing the uptake of HRA, and reducing costs by providing a tool which is easy to use for Defence risk assessment.

The HRA component of the work will present this cutting-edge tool and identify the opportunities it may bring to improve the MOD's ability to manage risk better in its large, high-cost programmes by supporting the optimisation of system performance through the design of systems which effectively manage and mitigate human performance risk. The tool will also strengthen the consideration of HF in risk assessment and reinforce HF arguments using risk-based data. The tool would be suitable for use from the early design stage. The HRA component of the work will outline the tool's development and its application to current safety challenges in Defence.

Conclusion

Both parts of this work provide an important insight into the current safety challenges facing the human component of safety in Defence, against the backdrop of increasing AI and autonomous agent integration. The key safety areas identified in this work provide avenues for future work that are already underway to enable guidance and safety policy to meet and overcome future challenges.

Alongside this, the development of the HRA Defence tool provides the opportunity to improve MOD's ability to manage risk in their large, high-cost programmes by supporting the optimisation of system performance through the design of systems which effectively manage and mitigate human performance risk.

References

Endsley, M. R. (2023). Ironies of artificial intelligence. *Ergonomics*, *66*(11), 1656-1668

Leggatt, A., Attfield, S., Kirwan, B., Turner, N., Forrest, E., & Hairsine, W. (2025) Defence-sector Human Reliability Assessment Tool Requirements. ACC2160752. Trimetis Ltd.

MOD (2018) An Introduction to System Safety Management https://www.asems.mod.uk/sites/default/files/documents/White%20and%20Green%20Book/SSM%20Whitebook%20PART%201%202020%20update.pdf?_t=1589874441 [Accessed 19/08/2025]

MOD (2024) JSP 936 V1.1 Dependable Artificial Intelligence (AI) in Defence. Part 1: Directive. https://assets.publishing.service.gov.uk/media/5b02f398e5274a0d7fa9a7c0/20180517-concepts_uk_human_machine_teaming_jcn_1_18.pdf [Accessed 20/08/2025]

MOD (2022) Ambitious, safe, responsible: our approach to the delivery of AI-enables capability in Defence. https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [Accessed 20/08/2025].

Williams, J. (1985). HEART-A Proposed Method for Achieving High Reliability in Process Operation By Means of Human Factors Engineering Technology. *In Proceedings of a Symposium on the Achievement of Reliability in Operating Plant, Safety and Reliability Society* (Vol. 16), 5/1–5/21.

Acknowledgement