Context Appropriate Human Involvement Across the AI System Lifecycle

Michael Boardman and David McNeish
Defence Science and Technology Laboratory UK[1,2], Dstl Porton Down
SP4 0JQ
UNITED KINGDOM

mjboardman@dstl.gov.uk

ABSTRACT

There is considerable national and international emphasis on the regulation and responsible use of Artificial Intelligence (AI) [1] [2] [3] [4] as well as growing public interest in the ethical and societal implications and risks associated with its use. This comes at a time when States and defence industries are accelerating the development of AI across a broad spectrum of military use cases, with some systems reportedly already being fielded operationally. Consequentially, there is considerable focus on how humans remain in control of, and accountable for the actions of AI-based systems, as well as increased awareness of the legal, moral and ethical implications of AI in military applications [5].

Previous work conducted by NATO [6], governmental [7] and non-governmental bodies [10] have explored the concept of human control / involvement in the delivery and responsible use of AI enabled systems within military applications. The nature of and reason for human involvement varies across use cases and system types, but typically includes: system performance, resilience, safety, legality and to meet the ethical commitments of the organisations and States developing and using those systems. Many defence applications of AI will employ a Human Machine Teaming [8] approach, where functions are allocated between human and AI agents as required to meet these objectives.

Based on work conducted by the NATO HFM330 Research Task Group over a four year period together with insights and developments in the wider field [9], this paper revisits the concept of meaningful/context appropriate human control and how it can be operationalised within the systems lifecycle to support effective, legal and responsible AI use.

1.0 INTRODUCTION

There is considerable national and international emphasis on the regulation and responsible use of Artificial Intelligence (AI) [1] [2] [3] [4] as well as growing public interest in the ethical and societal implications and risks associated with its use. This comes at a time when States and defence industries are accelerating the development of AI across a broad spectrum of military use cases, with some systems reportedly already being fielded operationally. Consequentially, there is considerable focus on how humans remain in control of, and accountable for the actions of AI-based systems, as well as increased awareness of the legal, moral and ethical implications of AI in military applications [5].

[2] It should be noted that this paper is an overview of UK Ministry of Defence (MOD) sponsored research and is released for informational purposes only. Its content should not be interpreted as representing the views of the UK MOD, nor should it be assumed that they reflect any current or future UK MOD policy.

Previous work conducted by NATO [6], governmental [7] and non-governmental bodies [10] have explored the concept of human control / involvement in the delivery and responsible use of AI enabled systems within military applications. The nature of and reason for human involvement varies across use cases and system types, but typically includes: system performance, resilience, safety, legality and to meet the ethical commitments of the organisations and States developing and using those systems. Many defence applications of AI will employ a Human Machine Teaming [8] approach, where functions are allocated between human and AI agents as required to meet these objectives.

Based on work conducted by the NATO HFM330 Research Task Group over a four year period together with insights and developments in the wider field [9], this paper revisits the concept of human involvement and control in relation to AI enabled systems and how it can be operationalised across the systems lifecycle to support effective, legal and responsible AI use. In the following sections there are a number of terms used to describe the concepts around the necessary human interaction with AI enabled systems to ensure safe, legal and responsible use. For the purposes of brevity and readability the terms human involvement and control will be used to refer to these concepts in general unless specific terms are being referred to.

2.0 WHAT IS HUMAN INVOLVEMENT / CONTROL IN AI ENABLED SYSTEMS AND WHY DOES IT MATTER?

Human involvement and control of AI enabled systems in the military domain, especially when applied to autonomous or semi-autonomous functions, has been a significant area of research and discussion within the research literature, policy landscape, and media coverage. The United Nations (UN) Group of Governmental Experts (GGE) on Lethal Autonomous Weapon Systems (LAWS) is a good example of this, with control being a central focus of discussion for several years. While this paper considers human involvement and control of AI enabled systems more broadly than the use of AI in weapon systems, LAWS is an area which has attracted significant attention and is the focus of much of the literature relating to control of military AI. Many of the principles and underpinning concepts can be applied to the responsible use of AI more broadly.

Boulanin et al [10] summarise the main reasons commonly given for exercising human control over LAWS as: legal compliance, ethical acceptability, and the safety and efficiency of military operations. Boardman and Butcher [6] identify two main aspects to the desire for human involvement and control over AI systems: meaningful human control (focussed on compliance with ethical and legal obligations) and effective human control (enabling system performance and military effectiveness). These two interrelated motivations underpin their description of human control as *"the ability to make timely, informed choices to influence AI-based systems that enable the best possible operational outcomes"*. This legal, ethical and operational categorisation characterises much of the debate around human control within the military application of AI.

2.1 Concepts Associated with Human Involvement and Control

The most widely used term used in relation to the control of military AI is *"meaningful human control"*, which was first popularised by the organisation Article 36 [13]. More recently 'nominal human control' has been a topic of discussion within the GGE on LAWS. It is used to describe the risk of humans being included within the system predominantly to satisfy policy requirements or act as a 'moral crumple zone' [12], but in actuality have limited or no agency and therefore could not be responsible or accountable for the resulting actions of the system. This is an interesting additional perspective to meaningful human control (MHC); where humans have insufficient time, understanding or ability to impact on the behaviour of the system and its effects in an informed manner resulting in an illusion of human control.

A concern cited by Kwik [13], also expressed by a number of States in the GGE, is that MHC is problematic as a policy or law-making tool due to its ambiguity and lack of a unifying theory. Consequently this has led to a wide variety of terminology used to describe the nature of the human role in AI-based systems to ensure

that they are used responsibly (see the non-exhaustive list in Table 1) and the international community are yet to reach consensus around a definition of human control over LAWS, or any other AI-enabled military systems for that matter [14].

Together these insights indicate the diversity of opinion regarding the nature of human control and that human control as a concept is far more complex than it might seem at first glance.

Table 1: Examples of the diverse terminology used to describe human control of LAWS.

(Maintaining) (Ensuring) (Exerting) (Preserving)	(Substantive) (Meaningful) (Appropriate) (Context Appropriate) (Sufficient)	Human	(Participation) (Involvement) (Responsibility) (Supervision) (Validation) (Control) (Judgment) (Decision) (Agency)
(Avoiding) (Preventing)	Nominal	Human	Control

2.1.1 Human Involvement and Control in the United Kingdom's Responsible AI Approach

While this paper does not reflect the views of the MOD, nor current or future UK MOD policy it is useful to provide an example of how human involvement and control is articulated within a broadly recognised, responsible AI approach. The United Kingdom's Ministry of Defence approach to responsible AI adoption highlights a number of concepts related to human involvement and control across the system lifecycle. With regard to responsible use, it states that *"Human responsibility for the use of AI-enabled systems in Defence must be underpinned by a clear and consistent articulation of the means by which human control is exercised, and the nature and limitations of that control. While the level of human control will vary according to the context and capabilities of each AI-enabled system, the ability to exercise human judgement over their outcomes is essential."* [15].

Within the context of LAWS *"We strongly believe that AI within weapon systems can and must be used lawfully and ethically. Sharing the concerns of Governments and AI experts around the world, we therefore oppose the creation and use of systems that would operate without meaningful and context-appropriate human involvement throughout their lifecycle."* and "*We believe the best approach is to focus on building norms of use and positive obligations to demonstrate how degrees of autonomy in weapons systems can be used in accordance with international humanitarian law – with suitable levels of human control, accountability and responsibility" [15].* The UK approach highlights the important role that context, in terms of system capability, operational environment and nature of employment plays in establishing where, when and how human involvement should be applied within the system lifecycle.

2.1.2 NATO Research into Meaningful Human Control of AI Enabled Systems

A NATO Exploratory Team (NATO HFM-ET-178 Meaningful Human Control over AI-Based Systems) was established in an attempt to better understand the nature of human involvement and control in relation to AI use in military systems and to address the ambiguity of terms such as MHC. Notably this exploratory team developed six dimensions of human control of AI enabled systems (see Table 2) [6] emphasising that these dimensions are not discrete – either present or absent – but instead describe the dynamic and multi-dimensional nature of human control over AI.

Table 2: Dimensions of human control of AI [6].

Dimension of Human Control	Description
The human has freedom of choice	The degree to which the human user can choose from all of the possible courses of action available.
The human has ability to impact the behaviour of the system	The extent to which the user is provided with the functionality to change the behaviour of the system. This could be in real time, or in advance through the setting of bounds or constraining allowable actions and behaviours.
The human has time to decide to engage with the system and alter its behaviour	The temporal aspect of user interactions with the system i.e. does the system allow the user sufficient time to process information, make decisions and impact on its behaviour if required.
The human has sufficient situation understanding	The extent to which the human has sufficiently accurate situational understanding to make an informed choice.
The human has sufficient system understanding	The degree to which a human has a sufficient understanding of the system state, in order to understand the provenance, quality and accuracy of the information and the rationale of the decisions and recommendations made.
The human is capable to predict the behaviour of the system and the effects of the environment (physical and information)	The extent to which the user is able to predict how the system will behave in different circumstances.

Building on these dimensions a working description of human control was developed by HFM 330 RTG to provide a common foundation for its research [16]:

> *"Humans have the ability to make informed choices in sufficient time to influence AI-based systems in order to enable a desired effect or to prevent an undesired immediate or future effect on the environment."*

This working description, along with the six dimensions, provides a foundation that we build on throughout the rest of this paper as we explore how the concept of meaningful/context appropriate human control can be operationalised across the AI systems lifecycle.

2.2 Types of Control

A typical focus of discussions regarding human involvement and control in the responsible use of AI is how and when human control is exercised in-use. An analogy of a control loop is frequently cited where an operator can be in, out or on the loop. Being 'on-the-loop' normally refers to an operator who is monitoring the performance of a system and able to intervene in real-time when necessary. This is also referred to as supervisory control. Tsamados and Taddeo [17], in their critical review of the literature relating to human control of AI, identify supervisory control as the prominent paradigm associated with control of AI during use. However, there is a risk with supervisory control, and the associated loop analogy, that despite its conceptual clarity it oversimplifies what is in reality, a more complex set of relationships.

It is also the case that real-time supervisory control may not always be feasible, or indeed, desirable, for a variety of reasons. Communications may be denied by adversarial action or environmental conditions. Required response speeds may be too high, for example for point defence systems. The number of systems requiring human supervision may exceed human capacity. System complexity and/or information quantity and complexity may be beyond human capability to process and understand. In such cases prior or indirect human control may be the only feasible option. This concept has been proposed in various forms, such as advance control directives, social contracts, or work agreements associated with advanced mission planning. A key element of prior control is the need to account for potential moral, ethical and legal contingencies in order for MHC to exist. It also places a much greater onus on how and when control measures can be exercised throughout the lifecycle of an AI system and consideration of the complex network of people responsible for decisions relating to control of military AI and where accountability lies.

Control of the effects of the system can also be exercised through other, non-technical means. For example decisions can be made within a responsible chain of command about where, when and how to use a system, imposing temporal, geographic and behavioural bounds based on a wider appreciation of the operational situation (including the entities operating in the environment, how quickly the situation is changing, and risk present) and by setting pre-programmed responses to specific situations and disablement or abort criteria.

3.0 SYSTEMS OF CONTROL ACROSS THE LIFECYCLE

Tsamados and Taddeo [17] highlight the inadequacies of the supervisory control paradigm for dealing with the complex interactions between humans and AI, especially when considering emerging AI approaches like foundation models. So while the in-use phase is important, establishing the conditions for realisation and maintenance of human control extend across the lifecycle of a system must take into account the wide spectrum of scale, intensity and complexity in conflict and operational environments. Development, assessment, evaluation, deployment and revision of AI enabled systems is not a linear, sequential process, rather it is a continuous cycle conducted within a wider political, legal, regulatory, ethical and systems of systems context. As such, a combination of approaches applied at the appropriate points across the whole lifecycle of a system are required to ensure human control is realised in operation. These approaches include, but are not limited to:

- National and international regulation, which drives wider organisational behaviours and processes so should set norms and expectations for human control;
- National and organisational policy;
- Specification of system requirements that enable human control to be realised during the acquisition of AI enabled capabilities;
- Design of system functionality and user interfaces that enable human control with particular focus on human-machine interaction, explainable and trustworthy AI;

- Test, Evaluation, Verification and Validation (TEVV) and certification processes including legal review, ethical review, assessment of potential bias and acceptance testing at system and systems of systems levels;
- Selection of context appropriate training data for AI models;
- Operating procedures and processes, including, command and control structures, clear lines accountability for the effects of AI enabled systems, doctrine and Rules of Engagement (ROE) that support the realisation of human control;
- Training of personnel, across all roles involved in the deployment and use of AI enabled systems, to understand the behaviours, capabilities and limitations of AI enabled systems; allowing them to make informed choices about their use;
- Re-assessment when individual systems are integrated into a system of systems and appropriate monitoring for undesired emergent behaviours or properties;
- Organisational culture that supports accountability and responsibility together with mechanisms for reporting mistakes and loss of human control to allow these to be addressed in a timely manner.

3.1 Alternative Perspectives on Human Involvement and Control Across the System Lifecycle

There are numerous approaches to describing the AI system lifecycle; this section presents three different perspectives on how human involvement and control can be considered within different views of the system lifecycle.

3.1.1 Human Machine Touchpoints Across the System Lifecycle

Figure 1 presents a framework submitted to the UN GGE on LAWS [18] for considering “touchpoints” throughout the weapon system lifecycle (including the targeting cycle) where human control may be exerted or enabled. Known colloquially as the ‘sunrise’ diagram it provides a structure for considering the broad range of factors throughout the system lifecycle that directly or indirectly influence the ability to exercise human control. The inner layer of the diagram describes the weapon lifecycle using six stages, the second layer provides example activities, which might fit within each stage, and the third layer highlights wider influences on human control and system development such as regulation and standards. One of the main aims of this diagram is to illustrate the wide range of actors who can contribute to the control of AI-enabled systems, in different ways and at different times, with varying levels of influence on the behaviour and effects of the system. It serves to highlight the complexity of the system of control that is necessary for the responsible development and use of AI in defence, encompassing activities across the whole system lifecycle.

3.1.2 Human Oversight Framework

Verdiesen, de Sio, and Dignum [19] propose a framework *Figure 2* which consists of three different perspectives on control (Governance, Socio-technical, and Engineering) paired with three time periods (Before, During and After deployment). This is perhaps the most comprehensive framework available in the research literature capturing the broadest range of control mechanisms for military AI across the lifecycle and from different perspectives. Although the framework was originally developed for LAWS, it can be generalised to other military AI-enabled systems as the categories are not specific to weapon systems but instead reflect the general types of control measures applicable to all military technologies.

This framework highlights that the behaviour of AI-enabled systems can be influenced and limited through actions taken by the engineering community, the user community, and the wider governance community

(layered approach). Decisions can be made before, during or even after deployment that effect the current or future behaviour of the system (lifecycle approach). To some extent this approach is scalable; it can be used to consider the means of control across the whole lifecycle of a system, or applied to the use of a system within a specific environment, deployment or operation. It is also possible to subdivide the layers e.g. the Governance layer could be decomposed into international law, national law, joint (e.g. NATO) and national policy etc. Figure 3 provides examples of the types of activities through which human oversight can be applied across the different layers and timeframes.

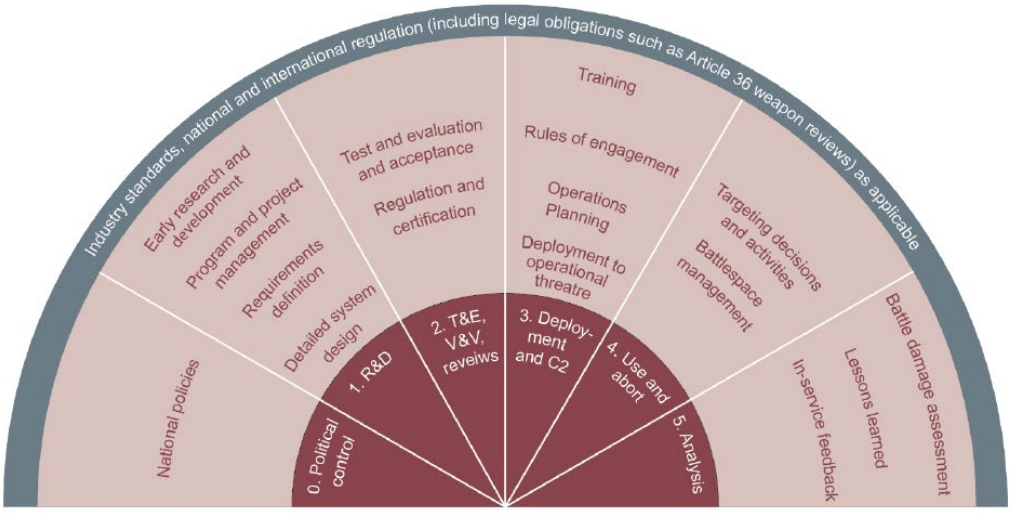

Figure 1: 'Sunrise' diagram illustrating touchpoints across the weapon lifecycle where human control can be exerted or enabled. [18]

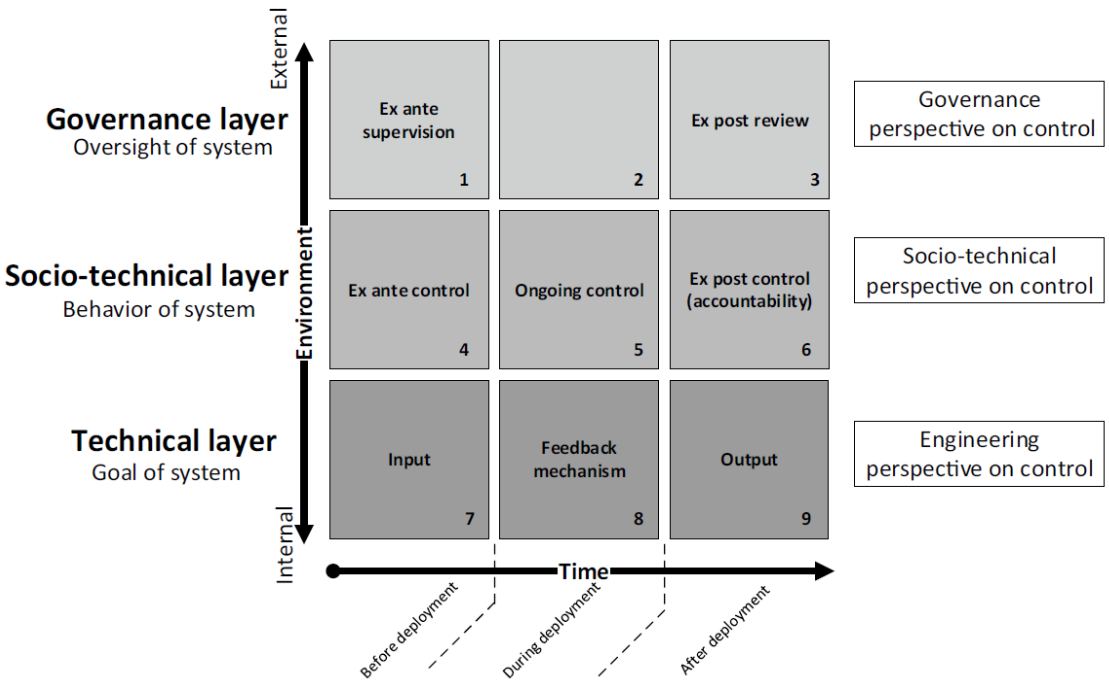

Figure 2: Comprehensive human oversight framework proposed by Verdiesen, de Sio, and Dignum [19].

Governance Layer	• International Humanitarian Law • Civil Law • National AI policies • NATO ethical principles • Article 36 legal review • Standards • Military doctrine	• Rules of Engagement • Command and Control structures • Legal and policy advice to Commanders	• Evaluation of effectiveness of policy • Sharing of good practice • Sharing of information with civil society to build confidence
Socio-technical Layer	• Training and education of personnel involved in the use of systems • Human-Centred Design approaches and human factors assessments • Function allocation • Organisational culture in design and development	• Organisational culture in-use • Standard Operating Procedures • Monitoring of compliance • Apportionment and recording of who is responsible/accountable for AI effects • Provision of expert advice on AI use	• Post-hoc review of system behaviour to inform explainability and future training • Critical incident analysis • Training review • Reporting system failures and misuse of AI system
Technical Layer	• Test Evaluation Verification and Validation (TEVV) • System Requirements • Failure modes and effects analysis • Training data quality review	• TEVV of Systems of Systems • Rapid TEVV when systems are refined, retrained, or adapted • Monitoring for emergent/ unpredicted behaviours • Mission specific behaviour bounding	• Use of data collected to retrain future iterations of AI systems • Optimisation of models for future operations
	Before Deployment	**During Deployment**	**After Deployment**

Figure 3: Examples of how human oversight can be applied across layers and timeframes.

3.1.3 NATO HFM 330 RTG Lifecycle Framework for Human Control

Based around the national approaches of the contributing nations to the NATO HFM 330 RTG and its own development activities, the following lifecycle framework was adopted within its work (see Figure 4) [16]. It should be noted that this lifecycle is not static and is highly likely to be iterative in nature. For example, systems may need to be revaluated and reassessed in response to changes in policy, or where the context of use or operational environment changes radically. In these situations Design and TEVV many need to be conducted rapidly in the field in response to operational requirements and in-use feedback etc.

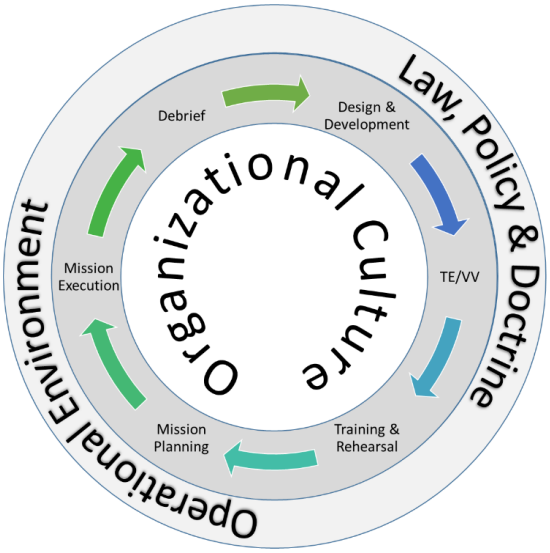

Figure 4: NATO HFM 330 RTG System Lifecycle for MHC.

The following paragraphs describe the main activities and influencing factors within the NATO HFM 330 RTG lifecycle framework.

Organisational Culture, Operational Environment, Law, Policy and Doctrine – This consists of the decisions made by politicians, policy makers and military leadership and will influence all other stages of the system life-cycle. These are routinely revised and therefore should be reviewed regularly to understand their potential implications on Human Involvement and Control. Within a NATO context this consists of: National AI Policies, from which National Defence AI Policies will be derived, NATO Policy and for some system types International Humanitarian Law (IHL).

Design and Development – This is the stage that has the most apparent and potentially greatest impact on the nature of human involvement with the system. It is during this stage that the system requirements are defined based on an analysis of user needs, context of use and the application of relevant law, regulations and policies. This stage draws on Science and Technology research, Operational Analysis and human factors methods, such as task analysis, to identify performance requirements and allocation of function between human and machine, which form the foundation of the nature of human involvement with the system. The application of a human-centred approach to system design together with human factors design guidance is critical to ensure that human involvement and control are appropriately considered throughout the system development process. An important consideration in the risk management process is the potential for the loss of context appropriate human involvement. Therefore regular reviews of function allocation, system design and the results of usability testing should be used to inform risk assessment and mitigation measures related to human involvement.

Testing and Evaluation, Verification and Validation – These activities typically focus on compliance with system requirements as well as relevant regulation, certification and legal requirements. However, the nature of human involvement and wider control in the design of the system and whether this is appropriate for the intended use cases should also be formally assessed through appropriate analysis, test and evaluation methods. Assessments of human control should consider the full range of use cases, intended contexts of use and operational environments. Assessments should consider human involvement and control in both individual systems and the system as part of the wider system of systems, where emergent system behaviours may impact on the nature of human involvement and control of the system.

Training, Rehearsal and Mission Planning – This stage involves taking into consideration the specific threat, environmental, legal and operational context and using this to inform the preparation of the force elements to be deployed as well as the constraints under which they must operate. Crucially this includes the training and testing of operators and commanders in safe and effective use of the system is a critical component in ensuring appropriate human involvement in AI-based systems. This is likely to include an understanding of the capabilities and limitations of the system and any modes of operation. Training must take into consideration the specific challenges associated with the theatre of operations including the requirements of IHL.

Mission Execution – This stage focuses on decisions made at the tactical level by commanders and operators as well as the pre-defined Standard Operating Procedures (SOPs) and Tactics, Techniques and Procedures (TTPs) governing the way in which a system is to be used within the specific operational context. A range of methods can be used to exert control over system use and its effects on the environment – these may be (near) real-time approaches such as monitoring and intervention in autonomous functions or in advance through the use of bounding or limiting AI behaviours e.g. applying behaviour, geographic and temporal constraints and varying the degree of system autonomy or functions under rea-time human control.

Debrief – This stage includes activities to consider whether a system is being used and operating as intended. This should take a socio-technical approach considering not just the performance of the technology, but how that technology is used in operation and how in-use processes and controls support the achievement of human control and prevention of undesired effects on the environment. It is unlikely that single system based reviews will be sufficient, rather a systems of systems approach to review will be required. This in-service feedback and lessons learned is particularly important in the period following the introduction into service of a new system and is critical in identifying any undesirable behaviour, design issues, incompatibilities between existing processes and practices and capabilities of the system. However, it is essential that this feedback process continues through the life of the systems, in particularly following the retraining of a model, or where the system is being used in a new operating environment or operating context. Mechanisms to support reporting of concerns over AI performance, loss of human control, ethical issues or misuse of AI-based systems are also required.

3.2 Operationalising MHC in Systems Design

Having established that a whole lifecycle approach to human involvement and control of AI is important, now we focus on some of the critical activities occurring prior to system use. The specification, development and test, evaluation, verification and validation of the system occurring *before* use are all critical enablers of context appropriate human involvement *during* the in-use phase of the system lifecycle. These activities define the allocation of function between human and machine, system functionality and behaviours. This is also where the Human Machine Interface is developed, providing a means through which users can interact with the AI agent(s) and wider technological components, enabling them to understand the system state, behaviour and build a mental model of how the system behaves in order to calibrate their trust in the system.

Boulanin et al [20] and Umbrello [20] emphasise the importance of considering how human control can be implemented during the early phases of the lifecycle before use, including research and development, design and acquisition. Umbrello [20] suggests that control over the design of AI-enabled systems is more stringent than control during use because it avoids the assumption that just having an operator “on-the-loop” who can intervene if necessary is sufficient, when in reality they may not understand why the system behaves as it does.

A criticism of MHC is that it is an ambiguous concept and establishing whether is exists or not, or is sufficient for a given system in a given situation is somewhat complex. In an effort to operationalise human control within systems design and testing the NATO HFM330 RTG [16] drafted examples of Systems Requirements, Table 3, developed around the six dimensions of meaningful human control described in

Table 2 which could be used as a means to drive system design and acceptance testing[3]. System requirements play major roles in systems engineering, as they:

- Form the basis of system architecture and design activities.
- Form the basis of system integration and verification activities.
- Act as a reference for validation and stakeholder acceptance.
- Provide a means of communication between the various technical staff that interact throughout the project.

Table 3 presents examples of the types of requirement that might support the development of a system that is capable, technologically, of supporting human control in-use. Projects would need to undertake their own analysis to determine the specific detail of the requirements and populate measurable criteria indicated by square brackets [X] together with relevant acceptance tests. For many of these requirements the acceptance tests are likely to include some form of user testing in a live / simulated environment with relevant user and system performance metrics.

Table 3: Example Systems Requirements based on the Dimensions of Meaningful Human Control.

Dimension of Human Control	Example of Potential Systems Requirements
To ensure that appropriately trained humans are sufficiently involved in achieving desired **and** preventing undesired immediate or future effects on the environment, systems containing AI functions **SHALL**:	
The human has freedom of choice	• Provide [sufficient freedom of choice] to allow an appropriately trained human(s) to impact the behaviour of the system.
The human has ability to impact the behaviour of the system	• Provide [sufficient functionality] to allow an appropriately trained human(s) to impact the behaviour of the system. • Provide the ability for an appropriately trained human(s) to constrain the systems effects on the environment. For example by including one or more of the following constraints: temporal, geographic, permitted behaviour/action.
The human has time to decide to engage with the system and alter its behaviour	• Provide [sufficient time] for an appropriately trained human(s) to understand the situation, understand system state, predict system behaviour, make informed decisions regarding necessary interactions with the system and enact them.

[3] A requirement is a statement that identifies a product or process operational, functional, or design characteristic or constraint, which is unambiguous, testable or measurable, and necessary for product or process acceptability.

Dimension of Human Control	Example of Potential Systems Requirements
The human has sufficient situation understanding	• Provide the information required to allow an appropriately trained human(s) to develop [sufficiently accurate] understanding of the current situation to make informed decisions regarding necessary interactions with the system. • Provide the information required to allow an appropriately trained human(s) to develop [sufficiently accurate] understanding of the likely future situation during which the system will act outside of human supervision to make informed decisions regarding necessary interactions with the system.
The human has sufficient system understanding	• Provide the information required to allow an appropriately trained human(s) to develop [sufficiently accurate] understanding of system state to make informed decisions regarding necessary interactions with the systems.
The human is capable to predict the behaviour of the system and the effects of the environment (physical and information)	• Provide the information required to allow an appropriately trained human(s) to predict, with [sufficient accuracy], system behaviour in the context of the situation and environmental conditions. • Suitable training environments SHALL be provided allow personnel involved in the fielding and operation of AI Enabled Systems to understand their behavioural characteristics, capabilities and limitations within the operational environments and conditions within which they would reasonably be expected to operate.

4.0 CONCLUSIONS

The realisation of appropriate human involvement and control in AI enabled systems will be dependent on activities across the entire system lifecycle. Given the broad range of AI technologies and potential applications across defence there will be no single approach to human involvement that is appropriate for all applications. Each application is subject to specific contextual factors including the purpose of use, physical and digital environment, nature of possible threats, time pressures, risks associated with system behaviour, regulatory environment, and so on. These contextual factors together with the technological capabilities being employed will shape the combination of controls that are necessary to meet military, safety, legal and ethical objectives.

Significant further work is required to develop specific methodologies and create good practice approaches that encompass both technical and non-technical aspects of human involvement control and are appropriate to the lifecycle phase.

In addition to this proposed area of future work there are several other specific challenges that are not discussed in this paper, but that require further exploration. Two of the potentially most important include the following:

Systems of Systems Complexity: Considering the nature of human involvement and control within single user - single systems can be complex in its own right, but this complexity increases significantly when wider systems of systems aspects are considered. Interactions between multiple AI-enabled systems, humans, and Human Machine Teams are likely to lead to complex and unanticipated emergent behaviours and systems properties. This complexity increases the potential for human control to be lost with associated detriments in system performance and with risks of accountability, legal and moral issues arising. There is also a risk that human control at an individual system level is lost at a system of systems level.

Rapid Human Control Reassessment In-Use: Context plays an important role when determining the nature of human control that is appropriate for a given system in a particular operational environment. Boardman and Butcher [6] highlight that human control of AI systems needs to be dynamic in response to the changing nature of the operational environment, command and control structure, and the adaptive characteristics of some AI systems. To do this rapidly in theatre during an operation will be challenging, but necessary if control is to be maintained throughout.

ACKNOWLEDGEMENTS

The authors would like to thank the members of NATO HFM 330 RTG the insights of whom this paper draws upon. This includes: Mark Draper (AFRL, USA), Jurriaan van Diggelen (TNO, The Netherlands), Marlijn Heijnen (TNO, The Netherlands), Chris Miller (SIFT, USA), Robert J. Shively (NASA, USA), Emma Parry (Cranfield School of Management, UK), Frank Flemisch (Fraunhofer, Germany), Marcel Baltzer (Fraunhofer, Germany), Rogier Woltjer (Defence Research Agency, Sweden), Michael Boardman (DSTL, UK), Kate Devitt (TASDCRC, Australia), Marie-Pierre Pacaux-Lemoine (LAMIH - UMR CNRS 8201 Université Polytechnique Hauts de France), Sissy Friedrich (Universität der Bundeswehr München, Institute of Flight Systems (IFS), Germany), Diana Donath (Universität der Bundeswehr München, Institute of Flight Systems (IFS), Germany), Axel Schulte (Universität der Bundeswehr München, Institute of Flight Systems (IFS), Germany)

5.0 REFERENCES

[1] Bletchley Declaration (2023) https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

[2] European Union AI Act (2024) https://artificialintelligenceact.eu/

[3] G7 Ministerial Declaration (2024) https://www.gov.uk/government/publications/g7-ministerial-declaration-deployment-of-ai-and-innovation/g7-ministerial-declaration

[4] Governing AI for Humanity (2023) https://www.un.org/en/ai-advisory-body

[5] Resolution adopted by the General Assembly 78/241 (2023) Lethal autonomous weapons systems https://documents.un.org/doc/undoc/gen/n23/431/11/pdf/n2343111.pdf?token=WIvKYKHPikKCwxNOYO&fe=true

[6] Boardman and Butcher (2017) An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenges of Achieving It. NATO HFM-ET-178 https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-IST-178/MP-IST-178-07.pdf

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

[8] Human-Machine Teaming (JCN 1/18) (2018) https://www.gov.uk/government/publications/human-machine-teaming-jcn-118

[9] An Artificial Intelligence Strategy for NATO (2021) https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html

[10] V. Boulanin, N. Davison, N. Goussac and M. Peldán Carlsson, (2020) "Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control," Stockholm International Peace Research Institute and the International Committee of the Red Cross, Stockholm.

[11] Article 36 (2015), "Killing by Machine: Key Issues for Understanding Meaningful Human Control,". [Online]. Available: https://article36.org/wp-content/uploads/2020/12/KILLING_BY_MACHINE_6.4.15.pdf [Accessed 2 August 2023]

[12] M. Elish, (2019) Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology and Society Journal Vol. 5.

[13] J. Kwik, (2022) "A Practicable Operationalisation of Meaningful Human Control," *Laws,* vol. 11, no. 43, 2022.

[14] M. Ekelhof and G. Persi Paoli, (2020) "The Human Element In Decisions About The Use Of Force," United Nations Institute for Disarmament Research, Geneva.

[15] AMBITIOUS, SAFE, RESPONSIBLE Our approach to the delivery of AI enabled capability in Defence. UK MOD https://assets.publishing.service.gov.uk/media/62a9b1d1e90e07039e31b8cb/20220614-Ambitious_Safe_and_Responsible.pdf June 2022.

[16] NATO HFM 330 RTG final report in Draft 2024.

[17] A. Tsamados and M. Taddeo, (2023) "Human Control of Artificial Intelligent Systems: A Critical Review of Key Challenges and Approaches," To be published. [Online]. Available: https://ssrn.com/abstract=4504855 . [Accessed 2 August 2023].

[18] United Kingdom, "United Kingdom Expert paper: The human role in autonomous warfare," United Nations Office for Disarmament Affairs, Geneva, 2020.

[19] I. Verdiesen, A. Aler Tubella and V. Dignum, (2021) "Integrating Comprehensive Human Oversight in Drone Deployment: A Conceptual Framework Applied to the Case of Military Surveillance Drones," *Information,* vol. 12, no. 9, p. 385, 2021Verdiesen, de Sio, and Dignum (2021).

[20] Umbrello, (2021) "Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach," Ethics and Information Technology, vol. 23, pp. 455-464, 2021 Boulanin et al (2020).