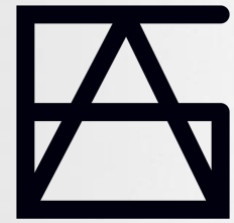


October 2024



ELSA LAB
DEFENCE

No. 6

Research paper series

Team Design Patterns for Meaningful Human
Control in Responsible Military Artificial
Intelligence.

AUTHOR

Van Diggelen, J., Boshuijzen-van Burken, C., & Abbass, H.

Team Design Patterns for Meaningful Human Control in Responsible Military Artificial Intelligence

Jurriaan van Diggelen¹, Christine Boshuijzen-van Burken², Hussein Abbass²

¹ TNO, the Netherlands

² UNSW Canberra, Australia

Abstract. A key requirement for responsible AI in the military is to ensure meaningful human control (MHC) in operational settings. However, MHC is not a binary concept; it could vary from one operational setting to another depending on context and the competencies of available humans and autonomous systems. We propose and discuss five patterns of MHC and highlight some of their key characteristics. We argue that control is only meaningful when humans have sufficient cognitive resources to perform control functions and when the behavior of autonomous systems in their operational context is sufficiently predictable. By anchoring the MHC patterns in human's cognition and situation uncertainty, we propose initial requirements for interaction design, testing and validation, and identify gaps in existing scientific understanding.

1 Introduction

The concept of meaningful human control (MHC) has emerged in the context of discussions on AI enabled and autonomous weapon systems. MHC refers to the idea that as AI-based autonomous technologies advance, humans should remain in control of decisions related to the use of force. Although AI can be responsibly applied for certain military objectives, we advocate that the design of autonomous weapon systems should not proceed unconditionally. One of the conditions is that of human control, which has been recognized in academic, legal, and political debates. The importance of MHC stems from it being a necessary condition for humans to be accountable and responsible for decisions with safety, ethical, and human dignity dimensions, such as those related to the use of force.

Sustaining MHC has been phrased in various ways by NATO [12] and US DoD [15] including as an ethical, and potentially legally binding principle. In 2023, the United Nations Secretary-General and the President of the International Committee of the Red Cross urged States to negotiate a legally binding instrument and state: "We must act now to preserve human control over the use of force. Human control must be retained in life and death decisions." In addition to its significance in military contexts, the concept has been recognized as important in civilian scenarios, such as automated driving systems [4].

Despite these statements and relevant discussions, it remains unclear how to operationalise MHC into concrete, implementable, and verifiable design requirements and specifications. Whereas MHC is frequently studied from the perspective of philosophy [20], artificial intelligence, and law [2], we believe that the fields of Cognitive Engineering and Human Factors are underutilized, especially when they could offer an array of consistent theories, models and methods that are critical to accomplishing MHC [3]. The purpose of this paper is to explore how theories of situation awareness and human decision making can be used to operationalise MHC by developing *team design patterns* [25]. We propose five MHC patterns of increasing complexity as assessed from the perspective of interaction design, testing and validation, and accountability. Instead of offering a thorough and exhaustive design blueprint of the complex issues under discussion, the patterns will form a foundation for proposing a multi-disciplinary research agenda aimed at designing and evaluating MHC.

In Section 2, we provide an overview of the legal and ethical background surrounding MHC, associating it to existing literature on this topic. Section 3 delves into the five patterns of MHC, accompanied by examples drawn from real or fictional military systems. Section 4 conducts a more detailed analysis of these patterns and describes practical applications. Finally, Section 5 offers some conclusions drawn from our exploration.

2 Background

The question of how humans should retain and exercise responsibility in practice has been a central issue in the debate on responsible military AI. The United Nations Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems states: “Although there is agreement on the importance of the human element . . . , [f]urther clarification is needed on the type and degree of human–machine interaction required, including elements of control and judgement” [5]. States use different terminologies; for example the USA’s Department of Defense Directive 3000.09, *Autonomy in Weapon Systems* requires that “[a]utonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.” [15]. The UK DoD states the following about the principle of “*Human Control*”: when using AI-enabled systems for Defence purposes, the need to understand the appropriate form of human involvement required for any given application or context.” [14].

The above examples are provided by military organizations, with a strong command hierarchy, where human control is always assumed and closely tied to human responsibility, which is visibly or verbally expressed through ranks, directives, or orders, at all levels in the organization. Others argue for MHC over military AI for moral and legal purposes, such as: “*safety and precision, responsibility and accountability, morality and dignity, democratic engagement and consent, and institutional stability*” [7]. While we acknowledge that all five purposes are important in the case of military AI, some are closely related to

the requirements of International Humanitarian Law (IHL), which sets limits on the means and methods of warfare. MHC is relevant for safeguarding traditional ethics of war, in particular the *jus in bello* principle of distinction between combatants and non-combatants and indirectly to the principle of proportionality (balancing military advantage *ex ante* against civilian casualties), which are ethical principles that can be interpreted as moral duties for belligerents, that are formalized in IHL.

The purpose of *safety and precision*, in particular, connects the principle of distinction to proportionality. Although AI-enabled systems may in principle have technical capabilities to ‘discern’ between a combatant and a non-combatant, military operations are intrinsically volatile, chaotic, deceptive, and unstable, even in terms of how a combatant can be distinguished from a non-combatant. Many have argued that at least “states must ensure that any future AWS meet the requirements of distinction and proportionality in its targeting capabilities.” [13].

The purpose of *responsibility and accountability* is articulated in many policies on military AI; for example “human responsibility for AI-enabled systems must be clearly established, ensuring accountability for their outcomes” [14, p.9]. The purpose of *institutional stability* is important as it must be clear when violations of customary law or internationally legally binding instruments occur when AI enabled systems get used in an armed conflict.

While this section aims at providing the background on the topic of MHC, the complexity of defining concepts such as meaningful human control, autonomy, responsibility, and value alignment suggests that debating these definitions falls outside the scope of this work.

3 Meaningful Human Control Patterns

We will focus on the design and evaluation of systems that allow humans to retain the ability to influence, understand, and have agency over the decisions and actions of AI-based and autonomous systems. We propose five patterns of MHC of increasing complexity and, when possible, ground them in established results from Human Factors research.

3.1 Pattern 1: Real-time Meaningful Human Control

Under MHC-P1 (i.e. Meaningful Human Control - Pattern 1), the robot’s behavioral response occurs immediately after the human control directive. An example of this is teleoperation, in which a robot is remotely controlled by a human operator. Other examples may require less human involvement, such as designing interaction according to a certain level of automation [16], or by placing the human in the role of supervisor, allowing them to step in as necessary. Each of these forms qualifies as Pattern 1 MHC; the human operator acts on real-time information and executes real-time control on the robot. This form of human-robot interaction is relatively well understood, with numerous models and theories in

the literature [18]. Applying this to morally sensitive domains such as a military operation³, we adopt the model presented in Figure 1.

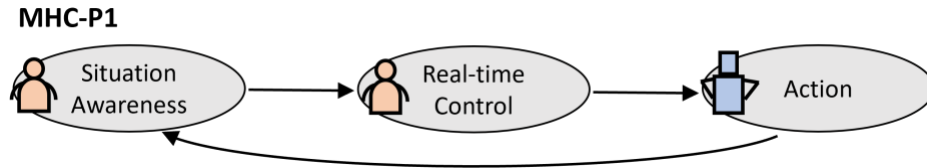


Fig. 1. MHC-P1 real-time meaningful human control

The process starts from a state where humans possess quality **Situation Awareness (SA)**, which is a prerequisite mental state for the operator to conduct situation assessment, course of action generation and selection, then execute necessary control [18]. This means that the human maintains an accurate understanding of the robot's current environment and activity. Based on this, the human can execute **Real-time Control** i.e. making decisions on what would be the right behavior in the current situation, and instructing the robot accordingly. The robot is expected to exactly follow up on these human instructions and to carry out the human-determined **Action**. The robot's icon in the figure can also refer to a group of robots (e.g., swarm), as long as they can be thought of as a single system. The results of the Action (implicitly represented by its outgoing arrow) then feed back to the human operator's SA allowing them to act on it again, forming a control loop.

Summarizing, the mental capacity of *situation awareness* is key to achieving MHC-P1. This entails a thorough human comprehension of the work, the surroundings, the robot's behavior, and ethically and legally significant components⁴.

Example MHC-P1

Recent advances in robotics have resulted in a variety of low-cost medium and small-sized drones, such as the black hornet and quadcopter drones. These drones are primarily employed for reconnaissance and surveillance and may be operated by a single mobile operator. Nonetheless, these tasks may be morally sensitive

³ Examples include: reconnaissance, target selection, identification or confirmation, and application of lethal force.

⁴ Cognitive theory of SA does not always delve into defining the boundary of what a situation is and/or the information expected to consider an agent is aware of. In this paper, we emphasize a novel aspect of situation awareness, which is closely related to the concept of moral awareness. Rest [17, p.3] defines moral awareness as the "interpretation of the particular situation in terms of what actions (are) possible, who (including oneself) would be affected by each course of action, and how the interested parties would regard such effects on their welfare." Our perspective is that a situation encompasses not only physical and task-related dynamics but also contains ethical and legal factors.

since they may infringe on civilian privacy, which must be balanced against mission's objectives. Following SA-oriented guidelines for user interface design [18] helps to design interactions that provide the operator with enough SA and the right control options to make these decisions responsibly.

3.2 Pattern 2: Distributed Real-time Meaningful Human Control

In complex military operations, the operator's *workload* to sustain SA and control of the system, escalates. A solution to this challenge is to distribute the control task among a team of operators as depicted in Figure 2.

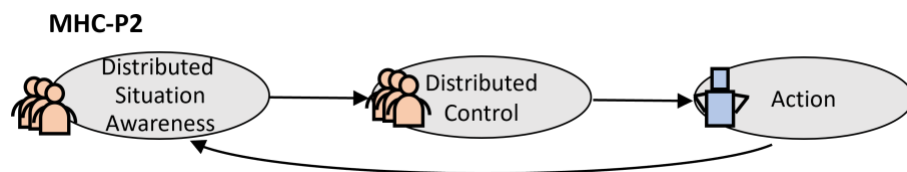


Fig. 2. MHC-P2 Distributed Real-time Meaningful Human Control

The process of maintaining MHC-P2 is largely similar to that of MHC-P1, albeit the reliance on multiple humans and their need to acquire and share SA and jointly execute control. The term **Distributed Situation Awareness** is used in the literature [19] to refer to the collective understanding of a situation that results from collaboration and information exchange across various actors. We use this concept as a necessary condition for executing distributed control. With the right mental state, the group can perform **Distributed Control** to directly influence the robot's **Action**.

In summary, the collective mental capacity of *Distributed Situation Awareness* is key to achieving MHC-P2 by a group of humans. This allows group members to have partial and localized SA, but requires complete SA by the collective.

Example MHC-P2

The Reaper MQ-9 is a remotely piloted aircraft primarily used for intelligence, surveillance, and reconnaissance missions. It is capable of carrying out precision strikes using Hellfire missiles and laser-guided bombs. The Reaper is often deployed in areas of conflict or where terrorist activities are suspected. It can be operated anywhere in the world, by a pilot and a sensor operator located at a ground control station (GCS). The GCS is equipped with advanced technologies that allow operators to control the aircraft, its sensors and weapons systems. Besides the pilot and sensor operator, the human team in control of the aircraft often includes a *mission coordinator* who oversees the overall mission, and an *intelligence analyst* who assists in interpreting data and imagery collected by the drone. While each team member has their own point of view, communications among them are essential for meaningful control.

In many ways, the task is morally charged. An obvious example is carrying out an attack while minimizing collateral harm. Furthermore, flying over specific regions and being visible to people (or not) violates privacy and may induce panic among the inhabitants. Humans can anticipate and interpret such situations when they occur. This is why the Reaper must be kept under MHC. Specific care was given in the design of the Reaper to allow the human operator team to have distributed SA, and preserve MHC-P2. The operators are well-trained; the pilot and sensor operator are in the same location and can communicate remotely with the mission coordinator and intelligence analyst; the high resolution video footage allows operators to inspect if civilians are present nearby a target; the satellite connection is highly secure and reliable; and the control is real-time.

It is important to recognize that using drones in warfare brings up other moral concerns, such as lowering the threshold for war and creating a *cowardly and unfair playing field* [6]. While these are significant issues, they fall outside the scope of the MHC paradigm.

3.3 Pattern 3: Prior Meaningful Human Control

Real-time MHC (as in MHC-P1 or MHC-P2) may not always be feasible for a variety of reasons. To begin with, a technical reason exists when remote connectivity may be insufficient to allow human teleoperation or supervisory control. This could be due to environmental circumstances (e.g., underwater, underground, etc.) or to the enemy purposefully interfering with radio signals. Second, there is a tactical reason, where the operation may require *silent mode* to prevent detection by the enemy. Third, there is a cognitive reason, as the *speed* may not be *manageable* for real-time human intervention, for example, cyber activities at machine speed or provision of air defense for hypersonic missiles. Finally, there is the reason of scale; that is, when the number of robots and AI-based actors exceeds the threshold for teleoperation (as in the case with robot swarms).

In some cases, we argue that the system can still be controlled under MHC-P3. The primary distinction between real-time MHC and MHC-P3 is that real-time MHC includes the human during task execution (human-in-the-loop, human-on-the-loop), whereas P3 involves the human *prior* to the task [24] (human-as-delegator, human-before-the-loop). Under P3, the robot does not respond promptly to human control commands, but instead with a significant delay, of hours or days. For example, the human provides the robot with a precise plan of action on how to deal with morally laden situations, after which the robot carries out that plan autonomously. This only works in a *predictable environment*, ensuring that the predefined plan continues to align with the actual situation during execution. This concept has been proposed in various forms, such as play-book interaction [11], or *direct, set and forget control*. If the human directive is precise enough, there is no need for the robot to make moral judgements during execution because they have already been made by humans. Under the right conditions, this would qualify as MHC-P3.

MHC-P3 alleviates the need for some requirements of P1 and P2, such as the need for assured connectivity, or manageable reaction times, but also introduces some other requirements. The model of MHC-P3 is presented in Figure 3.

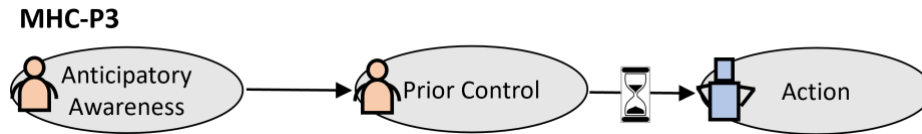


Fig. 3. MHC-P3 Prior Meaningful Human Control

Whereas MHC-P1 and MHC-P2 require humans to have a real-time situational understanding of the current situation (i.e. situation awareness), MHC-P3 requires humans to have imaginative skills about possible future scenarios. We will call this **Anticipatory Awareness (AA)**⁵. In the context of moral decision making, AA allows the operator to anticipate on all morally salient aspects of a potential future situation, including the environment and the robot's behavior. Based on AA, the human hands over instructions for **Prior Control** that prescribe the morally acceptable **Actions** the robot is permitted to undertake. During the robot's action-execution, there is no feedback to the human instructor due to the setup being a set and forget system.

The question of whether this process actually establishes MHC, depends on whether the human's AA at the time of providing the instructions was an accurate and correct representation of the operational environment. Consider the example that a person stationed at a base instructs a drone to *capture aerial images over a village at low altitude following a predetermined route which avoids sacred buildings*. While doing so, they are holding a mental model of the anticipated scenario, such as the village containing sacred buildings along with the proposed routes avoiding these. The moral decision not to fly over sacred buildings has been made by humans prior to the operation. If this mental model of the anticipated situation is an accurate reflection of the actual situations faced during execution, the person indeed exercised MHC over the robot. We would argue that if unforeseen circumstances arise in the scenario, for which no prior instructions were provided (e.g. if a funeral procession would be encountered along the route), this would indicate errors in Anticipatory Awareness that could cause loss of MHC. As a cognitive state requiring human effort, AA is harder to achieve when the environment is less-predictable. In the extreme case, when the environment is unpredictable, achieving AA is near impossible.

⁵ Note that the famous model by Endsley [8] also mentions projection as an element of situational awareness. However, this concerns the ability to project from current events into the near future, e.g. spatial extrapolation of moving objects. However, *anticipatory awareness* is more far reaching, and includes the ability to imagine scenarios further into the future that are not directly connected to the ways things are now.

In summary, MHC-P3 allows the human to execute control *before* the robot's action takes place. The mental capacity of *Anticipatory Awareness* is key to achieving MHC-P3. This means that the human must be capable of anticipating all moral risks in the space of possible scenarios.

Example MHC-P3

Consider, for example, the loitering munition IAI Harpy. This type of weapon is designed as a *fire and forget weapon*. After the human instructs the system to attack a radar system (with specific signature), the Harpy loiters for two hours without any human intervention searching for a radar with that signature, and then attacks it fully autonomously. Although the morally sensitive event (i.e. attacking the radar) takes place up to two hours after human intervention, the human was still in control in the sense that they were providing detailed instructions beforehand and had sufficient and correct AA. The Harpy is designed with a specific maximum loitering time to facilitate AA of the human user, i.e. the human only has to predict the situation in advance of two hours. Moreover, the system is not designed for utilization in highly unpredictable environments, such as urban areas.

3.4 Pattern 4: Goal-based Meaningful Human Control

In highly unpredictable environments, it is impossible to provide the robot with a precise set of actions beforehand, as these actions depend on circumstances that will only become clearer during the operation. Furthermore, a precise prescription of which actions to take may prove an excessively time consuming task for the human operator. In these cases, the system might be controlled under MHC-P4.

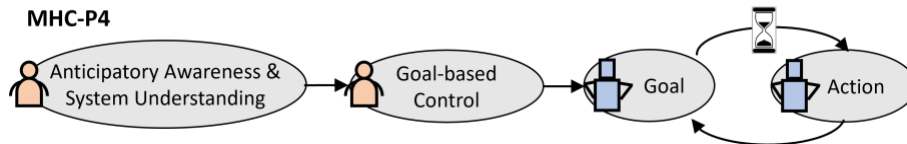


Fig. 4. MHC-P4 Goal-based Meaningful Human Control

In MHC-P4, the Robot is controlled via **Goal-based control** where the human indirectly influences the Robot's **Action** by providing it with a **Goal**. Just like MHC-p3, MHC-p4 is a prior control pattern and therefore requires operator's **Anticipatory Awareness**. Furthermore, the operator must understand how the system translates its goal into actions in various circumstances. We call this property **System Understanding**, which encompasses an operator's knowledge of the algorithms, data structures, decision-making mechanisms, and potential limitations of the AI system. While goal-based controllable systems do

require some level of AI to translate their goals into actions, this AI does not necessarily need to be complex. For instance, consider a navigation system in which users input their destination (i.e., the goal), after which the system autonomously maps out the route (i.e. selects the actions). Using such a system would likewise not lead to radically unexpected behaviors: the user knows that the road infrastructure allows for reaching the destination (i.e. anticipatory awareness of the environment) and that the system obeys basic traffic rules (i.e. system understanding). This example also illustrates that understanding a system from a user's perspective doesn't require being an AI expert with detailed knowledge, for example, of neural network architecture or machine learning parameters; a grasp of the system at a more abstract and behavioral level is sufficient.

The distinction between a goal and an action could spark a deep philosophical debate. Within the context of control, we emphasize that while an action usually specifies precise behavior, a goal provides the agent with significant behavioral freedom, i.e. freedom to deploy a planner to generate and choose actions to reach a goal. Differentiating between the two may not always be straightforward. We suggest considering them as existing along a continuum rather than as strictly separate categories.

In summary, to instantiate MHC-P4, the humans executing prior goal-based control need the ability to anticipate operational conditions. Furthermore, they must possess sufficient understanding of the system to predict the robot's actions in those conditions in order to prevent immoral behavior.

Example MHC-P4

Slaughterbots [23, ?], released in 2017, is an awareness-raising video addressing arms control issues. It depicts a fictional scenario set in the future, where swarms of low-cost microdrones equipped with artificial intelligence and facial recognition technology are utilized to target and assassinate political adversaries based on a pre-uploaded facial images. In several aspects, the system resembles the anti-radar loitering munition (the MHC-P3 example). Like the loitering munition, it is a suicide drone designed to deploy lethal force based on prior control. However, a significant distinction is that the anti-radar munition receives detailed action-level instructions (MHC-P3), while the slaughterbot operates with a more loosely defined goal (MHC-P4), in the form of a facial image. To meet the requirements of Anticipatory Awareness and System Understanding, the operator must understand how the system would behave in all likely scenarios. This includes understanding whether the system would engage a target surrounded by innocent bystanders, how it would classify individuals who resemble each other, and how it would handle situations where a person's face is not clearly visible and thus difficult to identify. Given the complexity of these scenarios and the facial recognition algorithm, it is unlikely that the operator will meet these

standards, rendering the slaughterbot insufficiently controllable and unethical by our standards⁶.

Note that goals can be much more abstract than those discussed in these examples, and one can imagine systems that accept higher level goals like *target-all-hostiles* or *win-the-battle*. Clearly MHC-P4 would not suffice to keep these systems under control due to escalation in ambiguity.

3.5 Pattern 5: Human-Machine Teaming

MHC-P4 places a tall order on the operator's mental resources; they must be able to comprehend all relevant potential scenarios and have an accurate idea of how the system would react under those conditions. This is infeasible for more complex scenarios. To allow human-robot interaction in complex morally sensitive domains, a much more dynamic control pattern is needed which is independent of real-time control (such as in P1, and P2), and independent of set-and-forget control (such as in P3 and P4). Instead, the control pattern should support multi-level coordination when needed, and should allow communication-free periods when possible. Furthermore, the control should focus on the joint interdependent activities of the human and the robot. In the literature, this is known as human-machine teaming [9], a paradigm of human machine interaction inspired by the way humans collaborate in teams.

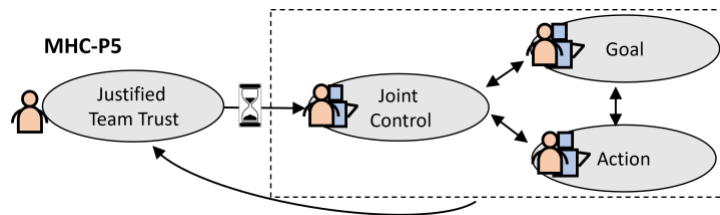


Fig. 5. MHC-P5: Human-machine teaming

Whereas teamwork comes intuitively for humans, programming it into a computer has proven to be an elusive endeavour. The most sophisticated AI systems are only beginning to show elementary forms of team behavior [24]. Therefore, MHC based on human machine teaming is not yet feasible. From a control perspective, a critical moment occurs when the human enters the team relation. The human requires **Justified Team Trust** that the robot and the human will do the right things, coordinate when necessary, and pursue the right goals. This

⁶ One might argue that, even if operators cannot predict the system's behavior in certain circumstances, they trust the manufacturer and certification agency to have constructed the system in a morally correct manner (i.e. trust in design [1]). However, slaughterbots don't fit MHC because neither the operator made the moral decision, nor did the manufacturer or certification agency due to their distance from the moral incident and unawareness of the specifics of the context during use.

trust does not form instantly but grows gradually over time. Thus, there is a feedback loop from the teamwork back to the human instantiating the work, which reinforces the trust-building process. Because both the human and the machine are involved in the team-based task execution, the human cannot fully disengage as with the set-and-forget style interaction, but must remain in an *at the ready* state [25]. This aspect might be perceived as a drawback of this pattern by those advocating for the substitution of humans with machines. Nonetheless, it is essential if we aim to deploy AI under MHC in more complex scenarios beyond the relatively simple set-and-forget scenarios. Despite intermittent involvement, the human experiences significantly less workload compared to real-time control, as they can manage multiple systems simultaneously and offload tasks to the machines when possible.

In summary, MHC-P5 can be regarded as a dot on the horizon where humans and machines collaborate similar to human teams. Achieving this level of cooperation relies on humans having a justified trust in the robots' abilities and their behavior as team members.

Example MHC-P5

During a military house search, two soldiers and two dog-shaped robots work as a team to clear a building. The soldiers hold position at the rear, while the robots scout ahead in tight and dangerous spaces. One robot detects a hidden compartment with its advanced sensors, alerting the soldiers who then secure the area and investigate further. Meanwhile, the other robot employs thermal imaging to identify a hidden insurgent in a closet, guiding the soldiers to arrest the individual.

4 Using patterns of control

So far, we have introduced five patterns of MHC and explored the different cognitive demands they place on humans in control. Our discussion is not exhaustive, as there are additional requirements such as self-control and alignment of intentions [22]. We concentrated on the patterns and properties we deemed most significant in the military MHC context, particularly those relevant to designing effective human-machine teams. A summary is presented in Fig. 6.

The grey bars in the figure show the primary design choices for MHC for the dimensions of **Control directedness** (what is being controlled?), **Control timing** (when is control executed?), and **Controlling actors** (who is executing control?). Given that each control dimension contains three options, this leads to 27 possible patterns. In fact, each of these patterns are viable options, and many of these have not been discussed in this paper. For example, **Machine Action** control at **Real-time** by **Humans& Machines** is applied in the lane assist function of semi-autonomous vehicles, where both the driver and the vehicle contribute to steering the wheel. **Machine goal** control at **Real-time** by **One human** is applied in supervisory control stations where humans direct the motion of a ship or airplane by setting simple waypoints and thresholds. In this

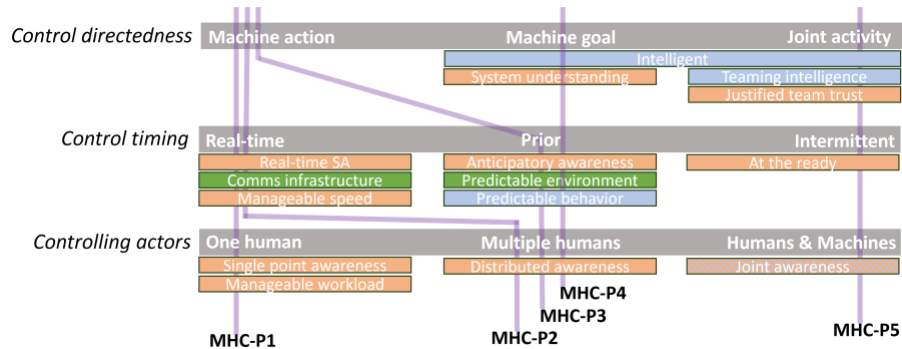


Fig. 6. Overview of five patterns of MHC in terms of their primary design choices (the grey bars) and secondary requirements (the colored blocks). Red blocks indicate human requirements; blue blocks indicate machine requirements; green blocks refer to requirements on task or environment; blue/red striped blocks are requirements on the human-machine team.

sense, the five patterns outlined in this paper may seem somewhat arbitrary. However, we’ve chosen them because they represent the most significant classes of military systems that are under debate in the MHC discussions.

Also, note that among the different primary design options, there is often a spectrum of values. For instance, between action-based and goal-based control (as elaborated in Section 3.4), or between one or many controlling actors, or prior control in terms of minutes versus months or years. For reasons of simplicity, these nuances have been omitted from the figure.

The patterns of MHC are illustrated by the purple lines in Fig. 6. These lines also indicate the secondary human, machine, and environment requirements that follow from the primary design choices. This provides a first visual clue to choose the right pattern of control: a control pattern should be chosen such that a viable set of secondary requirements is established. As discussed, MHC-P1 should not be chosen in absence of reliable communications. MHC-P3 circumvents this requirement, but it raises additional requirements, such as the need for anticipatory awareness.

4.1 Meeting the secondary requirements

To better understand the choice of the appropriate MHC design patterns, we delve deeper into the secondary requirements and their subsequent implications below.

The MHC patterns with higher numbers require more sophisticated technologies. Action-based control doesn’t require AI since the machine isn’t tasked with converting goals into actions. Goal-based control, on the other hand, requires explainable AI [21] to support human system understanding, introducing an extra layer of complexity. For controlling joint activity in a Human-machine

team, teaming intelligence [10] is required, enabling the system to grasp mental models, employ suitable communication strategies, manage task interdependencies, etc. While there has been some advancement with modern Large Language Models [26], this form of AI remains limited and is currently still insufficient to enable human-level teaming.

Regarding attributing responsibility and accountability, MHC patterns with higher numbers pose greater challenges. Firstly, the higher and more diverse the number of controlling actors, the harder it becomes to designate someone as responsible. Secondly, less direct forms of control (s.a. goal-based or joint-activity based) open opportunity for the controller to evade responsibility.

As a general design principle, there must be a compelling reason for using higher patterns of MHC because doing so necessitates a far more sophisticated system architecture. As was previously discussed, four task factors that encourage employing higher patterns of MHC include a fast operational tempo, the need to operate in silent mode, degraded communication links and scale of system elements. On the other hand, it is more challenging to meet the design requirements of higher MHC patterns when a task has a pattern with high unpredictability and risk. Lower pattern MHC may be more suited to tackle such jobs. When properly designed, MHC-P5 would be able to obtain the best of both worlds.

5 Conclusion

In this paper, we introduce five meaningful human control (MHC) patterns that, depending on context, can fulfill a crucial requirement of responsible AI in the military: ensuring meaningful human control. It's crucial to recognize that no single pattern should be seen as inherently superior to the others. Each pattern has its own applications depending on the conditions. The increasing numbering indicates a rise in complexity in terms of design, testing, and assignment of responsibility. Depending on various factors, one level may be better suited to a situation than another.

Starting with the human requirements for the MHC patterns, one can drive practical requirements for use conditions, interaction design, testing and validation, and ensuring accountability. The first two patterns of MHC are based on the relatively well understood concept of Situation Awareness. We propose that MHC-P3 and MHC-P4 should be based on human anticipatory awareness and system understanding, which are far less well-known. More research is needed to understand their relation to MHC, and how to design and assess user interfaces that facilitate anticipatory awareness (e.g. Course of Action Simulation Analysis) and system understanding (e.g. using explainable AI). This requires interdisciplinary approaches and experimentation, which includes human factors and ethics, as well as AI experts and military operators. The fifth pattern of MHC adds an additional layer of complexity where the human and the machine form a team. Understanding how to develop machines as teammates to sustain MHC could be seen as a focal point on the research horizon.

By defining the MHC requirement with greater precision, this paper has sought to enhance the ongoing international dialogues on AI regulation, both within military contexts and beyond. It's only when MHC is defined in terms of specific and measurable criteria that it can truly serve as a meaningful concept serving a higher moral purpose.

References

1. Hussein A Abbass. Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, 11(2):159–171, 2019.
2. Daniele Amoroso and Guglielmo Tamburrini. Autonomous weapons systems and meaningful human control: ethical and legal issues. *Current Robotics Reports*, 1:187–194, 2020.
3. Michael Boardman and Fiona Butcher. An exploration of maintaining human control in ai enabled systems and the challenges of achieving it. In *Workshop on Big Data Challenge-Situation Awareness and Decision Support*. Brussels: NATO STO, Dstl Porton Down, 2019.
4. Simeon C Calvert, Daniël D Heikoop, Giulio Mecacci, and Bart Van Arem. A human centric framework for the analysis of automated driving systems based on meaningful human control. *Theoretical issues in ergonomics science*, 21(4):478–506, 2020.
5. UNGGE CCW. Group of governmental experts on emerging technologies in the area of lethal autonomous weapons system. *Report of the 2019 Session of the GGE on Emerging Technologies in the Area of LAWS*. UN Office, Geneva, 2019.
6. Mark Coeckelbergh. Drones, information technology, and distance: mapping the moral epistemology of remote fighting. *Ethics and information technology*, 15:87–98, 2013.
7. Jovana Davidovic. On the purpose of meaningful human control of ai. *Frontiers in big data*, 5:1017677, 2023.
8. Mica R Endsley, Daniel J Garland, et al. Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 1(1):3–21, 2000.
9. Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69, 2014.
10. Matthew Johnson and Alonso Vera. No ai is an island: the case for teaming intelligence. *AI magazine*, 40(1):16–28, 2019.
11. Christopher A Miller. Delegation architectures: playbooks and policy for keeping operators in charge. *WS3*, page 28, 2005.
12. NATO. Summary of the nato artificial intelligence strategy, 2021.
13. Kevin Neslage. Does meaningful human control have potential for the regulation of autonomous weapon systems. *Nat'l Sec. & Armed Conflict L. Rev.*, 6:151, 2015.
14. UK Ministry of Defence. Ambitious, safe, responsible: Our approach to the delivery of ai enabled capability in defence. *MoD Policy Report*, 2022.
15. US Department of Defense. Dod directive 3000.09: Autonomy in weapon systems, 2023.

16. Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297, 2000.
17. James R Rest. *Moral development: Advances in research and theory*. Praeger, 1986.
18. Jennifer M Riley, Laura D Strater, Sheryl L Chappell, Erik S Connors, and Mica R Endsley. Situation awareness in human-robot interaction: Challenges and user interface requirements. *Human-Robot Interactions in Future Military Operations*, pages 171–192, 2010.
19. Paul M Salmon, Neville A Stanton, and Daniel P Jenkins. *Distributed situation awareness: Theory, measurement and application to teamwork*. CRC Press, 2017.
20. Filippo Santoni de Sio and Jeroen Van den Hoven. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:15, 2018.
21. Timo Speith, Barnaby Crook, Sara Mann, Astrid Schomaicker, and Markus Langer. Conceptualizing understanding in explainable artificial intelligence (xai): an abilities-based approach. *Ethics and Information Technology*, 26(2):40, 2024.
22. Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Roßnsberg, Philip Meinel, and Markus Langer. On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2495–2507, 2024.
23. Stewart Sugg. Slaughterbots. <https://youtu.be/O-2tpwW0kmU?si=nwMr8F9OPxHCf7OY>, 2017. Accessed: May 2024.
24. J. van Diggelen, K. van den Bosch, M. Neerincx, and M. Steen. Designing for meaningful human control in military human-machine teams. In *Research handbook on Meaningful Human Control of Artificial Intelligence Systems*. Edward Elgar Publishing, 2024.
25. Jurriaan van Diggelen and Matthew Johnson. Team design patterns. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 118–126, 2019.
26. Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, et al. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*, 2023.