Ethics and Information Technology (2021) 23:455–464
https://doi.org/10.1007/s10676-021-09588-w

ORIGINAL PAPER

Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach

Steven Umbrello[1]

Accepted: 29 March 2021 / Published online: 5 April 2021

Abstract

The international debate on the ethics and legality of autonomous weapon systems (AWS), along with the call for a ban, primarily focus on the nebulous concept of fully autonomous AWS. These are AWS capable of target selection and engagement absent human supervision or control. This paper argues that such a conception of autonomy is divorced from both military planning and decision-making operations; it also ignores the design requirements that govern AWS engineering and the subsequent tracking and tracing of moral responsibility. To show how military operations can be coupled with design ethics, this paper marries two different kinds of meaningful human control (MHC) termed levels of abstraction. Under this two-tiered understanding of MHC, the contentious notion of 'full' autonomy becomes unproblematic.

Keywords Meaningful human control · Autonomous weapons · Systems theory · Design for values · Applied ethics

Introduction

Although technological innovations have always played a key role in military operations, autonomous weapons systems (AWS) have received asymmetric attention in public debate as well as academic discussions—and for good reason (Kania, 2017). As these systems are designed to carry out more and more tasks once in the domain of human operators, questions regarding their autonomy and potential recalcitrance have sparked discussion. Debate highlights a potential *accountability gap* between their use and who, if anyone, can be held accountable. At the international level, discussions about how to exercise control over the development and deployment of these autonomous military systems have been underway for over a decade. Still, there remains very little consensus as to what constitutes a sufficient level of control.

The concept of *meaningful human control* (MHC) has emerged in discourse to encompass this ideal of human control over autonomous systems. Various approaches have been taken to define a sufficiently robust notion of MHC that addresses technical requirements (Arkin, 2008), proper training for use (Article36, 2015; Asaro, 2009), designer-user engagement (Leveringhaus, 2016), operations planning (Ekelhof, 2019), design requirements, and the responsibility of designers (Elands et al., 2019; Mecacci and Santoni de Sio 2019; Santoni de Sio and Van den Hoven 2018). Each of these approaches provide insight into how MHC over these types of systems can be understood and attained. Although they are generally proposed as isolated frameworks for attaining MHC, they share some underlying precepts. Approaches that emphasize the operational planning and military context of use, such as that of Ekelhof (2019), provide a strong contextual landscape for understanding MHC. Other approaches, such as that of Santoni de Sio et al. (2018, 2019), focus on design histories, designer intentions and plans, or the responsibilities of designers and supraindividual agents. They provide cogent arguments for designing these systems with both backward- and forward-looking responsibility. Still, they largely focus on a single level of abstraction at the opportunity cost of the other.

This paper aims to employ the concepts of systems theory (theoretical lens) and systems engineering (applied lens) to understand MHC across these levels of abstraction (LoA). As a way of marrying these often-isolated projects of defining MHC, I propose a two-tiered approach to understanding MHC. First, it does not make sense to divorce discussions of AWS from actual and often trivial military operations; AWS exist within this landscape, not outside it. One must

✉ Steven Umbrello
steven.umbrello@unito.it

[1] Institute for Ethics and Emerging Technologies, University of Turin, Via Sant'Ottavio, 20, 10124 Torino, TO, Italy

therefore situate AWS within their operational context—the operational LoA—in order to understand them. However, this does not mean there are no accountability gaps in terms of technical (fully)AWS. The question of design remains important for determining the responsiveness of a system to the relevant moral reasons of the relevant agents, thereby creating the design LoA. By coupling these two LoA, we can account for the technical and full autonomy of certain types of AWS. Many of the issues associated with (fully)AWS can thus be rendered non-issues.

It should be noted that this paper does not advocate *for* the development of (fully)AWS. Rather, it focuses on the notion of control over certain types of AWS in light of how current military operations actually function and design practices contribute to control. At the very least, this paper highlights a potential gap that theorists and policymakers can address when formulating their own arguments for whether/how AWS are ethically problematic or for the prohibition of certain types of AWS.

The paper is divided into the following sections. The first section briefly outlines both systems theory and systems engineering, which provide the conceptual lenses for understanding the coupling of operational and design LoA. The second section details the operational LoA, while the third section details the design LoA. Section four, which forms the bulk of this paper, discusses how both LoA are fundamental to a holistic understanding of MHC—and that the notion of full autonomy is actually unproblematic. Section five presents some of the limits of this approach, as well as some directions for future research. The final section concludes the paper.

Systems theory and systems engineering

The term 'systems theory' is *prima facie* self-explanatory. Still, its spelt-out definition merits mentioning *why* a paper conceptualizing a theory of MHC and its application (i.e., design) warrants any discussion of more abstract ontology. There are multiple reasons for drawing an ontology. First, the primary reason for adopting systems theory as the ontological framework for this investigation is that it (implicitly) characterizes the two levels of abstraction for understanding MHC discussed in the following sections. The operational level of control is characterized by a plurality of actors and networks that complicates—yet also constitutes—how military operations are structured, planned, and carried out. Likewise, the design level of control is fundamentally built on the notion of tracking and tracing networks of systems and actors in both the design histories and use of those systems. Second, systems theory is the theoretical framework from which systems engineering derives. As discussed below, systems engineering developed in the domain of defence. It is essentially the practical and managerial implementation of a systems thinking[1] ontology.

Systems theory is broadly understood as the interdisciplinary study of organized and complex systems (Whitchurch & Constantine, 2009). A system can be conceived as a connected cluster of co-constitutive and co-varying parts, which can be synthetic and/or biological. A system is fundamentally constrained by spatiotemporal vectors, altered by its context (environment), and defined by its architecture and teleology. Its teleology, for instance, is expressed in the operation of the system itself (Adams et al., 2014). Systems are thus often characterized as being more than the sum of their constituent parts when expressing emergent behavior (Dudo et al., 2011; Wan, 2011) or synergy (Haken, 2013). For this reason, an alteration at any given node(s) of the system can result in alterations at other node(s), as well as in resulting behavior (if applicable). One of the aims of systems theory, then, is to map out behavior patterns in these complex systems in order to better predict future behaviors given environmental inputs.

The above is particularly true of systems that adapt and learn (i.e., machine learning) from their environmental contexts (Aliman, 2020; Ivanov, 1993; Wernaart, 2021). Similarly, any single system can both support and constrain other systems to make them more or less robust. Overall, systems theory seeks to understand the kinetics of systems, their pressures and conditions, and the general methods or tools that can be extrapolated for use in understanding other systems at all levels of recursion (Graham et al., 1994) across a variety of fields (i.e., biology, chemistry, ecology, engineering and psychology). Such understanding aims to optimize equifinality (Beven, 2006).

Unlike approaches specific to a single system or domain, general systems theory (GST) intends to develop tools and methods for a general understanding of complex systems (Von Bertalanffy, 1972). GST distinguishes between system types in terms of activity and passivity. Active systems are characterized by structures or components that engage in processes or otherwise exhibit active behavior. Passive systems are those structures that are engaged or processed. Any given system can be both passive and active at any given spatiotemporal vector. An AWS is a passive system when it is powered down or lacks a power source. However, it becomes an active system when it is booted and deployed in the field. Systems can also be composed of passive and active systems. This framing becomes particularly relevant for an ontological understanding of complex AI systems that employ what are often considered opaque algorithmic

[1] Here, the term 'systems thinking' is used in the verbial sense. It refers to conceptualizing things in terms of systems or, more pointedly within the axioms of systems theory.

processes. Such processes result from hybrid machine learning and neural network systems (Boscoe, 2019; Turilli & Floridi, 2009; Wachter et al., 2017). Given the complexity and necessity for directing optimal systems design, the applied domains of GST become particularly relevant—especially the domain of systems engineering.

Systems engineering takes this multidisciplinary approach to understanding systems and applies it the understanding, design, management, and deployment of engineered systems to ensure optimized equifinality over their lifecycles (Adams et al., 2014; Thomé, 1993). More specifically, engineered systems are designed to ensure that their constituent parts work synergistically so emergent behaviors are beneficial. Aside from this, systems engineering draws on many overlapping, human-centric disciplines such as risk analysis, organizational studies and project management (these disciplines parallel the operations planning approach in Ekelhof's, 2019 conception of MHC). It is also informed by technical disciplines such as requirements engineering, cybernetics, software and electrical engineering, and industrial engineering, among others. It thus holistically frames engineering processes as part of the larger system that conditions the project being undertaken.

On a pragmatic level, systems engineering involves anticipating client needs and specific design requirements early on in the development cycle. When these have been accounted for, engineers can then move on to design synthesis and system validation while always maintaining a holistic picture of the lifecycle of system development (i.e., systems *thinking*). To do this successfully, designers must consider all of the stakeholders potentially implicated by the system along with their values as pertaining to the design project. This latter point on stakeholders, which will be discussed in greater detail in the following sections, is directly in line with theories of responsible innovation—particularly value sensitive design (VSD). The conception of MHC detailed by Santoni di Sio et al. arises from and aligns with VSD, which unfolds at the design LoA. By a similar token, systems thinking in general (i.e., systems theory + systems engineering) provides a reasonable tool for framing common ground and the need to combine the two LoA for an equally holistic understanding of MHC. The following section begins by outlining the operational LoA for MHC.

Meaningful human control: the operational level

Ekelhof (2019) approaches MHC by predicating it on military operation practice that both supports and constrains targets in areas of operations. Her approach relates to MHC in that it is a function of the role of designers [as with Santoni di Sio et al. (2018, 2019)] and of technical targeting procedures (as with Leveringhaus (2016)]. But Ekelhof's approach differs in its level of abstraction by focusing on higher level of organization and operational control of the military as a supraindividual agent. This addresses the fact that the 'autonomy' of AWS (and of any human agent in the military, such as soldiers) is necessarily constrained by such operations. The result of these constraints is that 'full' autonomy, which is often construed in discussions on AWS, is not 'full' in the sense that is often implied (e.g., self-determining agents). Instead, it is restricted to various operational decisions and planning a priori to deployment and operations.

Ekelhof uses a case of conventional air operations to frame human involvement in operations through a dynamic targeting process. By framing the role of human agent decision-making within distributed systems, she outlines ways policymakers and theorists can determine how military planning and operations *actually* function. AWS can then be deployed within the context of use of these practices. Characterizing the human role in military decision-making, she outlines a six-part preoperational briefing package followed by a six-step landscape for mission execution. I briefly summarize them below.

Pre-mission

The briefing

Before the mission is undertaken, the air component receives a briefing with information on mission execution. Such briefings are often highly detailed with information such as "target location, times, and munitions"; however, they are less detailed when we consider dynamic targeting in situ (Ekelhof, 2019, p. 345). Such information is distributed to various domains of operations to specialists, who then vet and use it in more detailed planning. The executers of the mission, in this case fighter pilots, are then brought in for briefing on the mission details. The pilots take the time to study the information provided while also taking care of any last-minute preparations for execution.

The following six components can be included in the briefing package:

1. Target (a military compound) description consisting of all available knowledge;
2. Target coordinates;
3. A collateral damage estimation (CDE) to give operators an estimate (not certainty) of expected collateral damage (NATO 2016). In this example, the risk of collateral damage is low as long as predetermined mitigating techniques are applied;
4. Recommendations for the quantity, type, and mix of lethal and nonlethal weapons needed to achieve the

desired effects (i.e., weaponeering solution) (USAF, 2017). In our example, these are GPS-guided munitions;

5. The joint desired impact, which is used as a standard to identify aim points; and
6. A weather forecast that, in this case, describes a night with overcast condition (clouds cover most or all of the sky) and heavy rainfall (Ekelhof, 2019, p. 345).

Coupled with other information such as the rules of engagement, the operator can then leave to execute the mission.

In situ operations

Step 1: Find

To find the target for operations, intelligence and data are required. Such targets are pre-programmed in the navigation systems of both the fighter jet and the payload. Whereas a dynamic target requires in situ data collection, the task here involves arriving at the preprogrammed "weapon's envelop (i.e., the area within which the weapon is capable of effectively reaching the target)". This process is displayed on the operations heads-up display (Ekelhof, 2019, p. 345).

Step 2: Fix

Once the operator arrives within the weapon's envelope, onboard systems aim to positively identify the target confirmed during operational planning. This ensures payload delivery complies with relevant military and legal norms (e.g., NATO, 2016). In this case, targets were preplanned and confirmed. For positive target identification, the operator usually does not engage in visual confirmation; instead, they refer to onboard systems and the validation that took place during operational planning to ensure lawful engagement of the identified target. Even in this fixed case of pre-planning, the human pilot does not need to attend to anything else during this phase other than arriving within the weapon's envelope (Ekelhof, 2019, pp. 345–346).

Step 3: Track

The operator tracks the target within the weapon's envelope to ensure the continuity of positive identification. This also provides concurrent updates regarding the position and status of the target. In the case of a static target (e.g., a military compound), tracking is relatively straightforward and involves simply entering the weapon's envelope as in the fix phase (Ekelhof, 2019, p. 346).

Step 4: Target

During this phase, the relevant rules of engagement (RoE), laws of armed conflict (LoAC) and other targeting rules are invoked to ensure lawful targeting and deployment. These also address other considerations, such as issues related to collateral damage and risk factors that may result to one's own forces. In this predetermined and validated target case, the legal and military experts who vetted the target permit the pilot to simply input relevant data into the vehicle and weapons payload delivery systems to ensure proper execution. Given the visually impairing weather conditions, any further collateral damage estimates cannot be attained. Planning at pre-mission stages validated that collateral damage estimates were low and were conducted according to the norms that govern them. The human pilot thus does not actively participate or intervene beyond piloting the vehicle into the weapon's envelope (Ekelhof, 2019, p. 346).

Step 5: Engage

Once the operator enters the designated weapon's envelope, the onboard computer suggests to the pilot the most opportune time for releasing the payload to ensure effectiveness. This suggestion is based on its knowledge of the capabilities of the equipped weapons system. Given that the payload system itself is GPS guided, there is no need for any other forms of targeting based on visual identification. Once the pilot authorizes the release of the weapon, the munitions guide themselves to the target (Ekelhof, 2019, p. 346).

Step 6: Assess

At this point, the results from the previous stage are assessed to determine the effects of the strike. Of course, a visual assessment from the pilot can be impaired by a number of factors (weather conditions, in this case). Similarly, visual assessments of collateral damage from the vantage point of a pilot may fail to accurately reflect the efficacy of the strike and its consequences. In the case of aerial engagements such as this, ground support forces may be required for a more accurate assessment of engagement (Ekelhof, 2019, p. 346).

In considering MHC then, it appears that most (if not all) of the performance latent to each step is beyond the pilot's control. It could be argued that this is emblematic of contemporary aerial operations more generally. While the pilot can be seen as in direct operational control of some of the operation, piloting the craft to the weapon's envelope and engaging in weapons release, this type of control is not sufficiently meaningful. This is because the pilot lacks full "cognitive clarity and awareness" of the situation within which they are participating (Article36, 2015). The privation begs the underlying question of whether the pilot actually possesses levels of clarity and awareness sufficient enough to be deemed substantial in a meaningful way.

Discussions at the pilot level could provide some future insight both for operations employing AWS as well as modern aerial crafts. But these would converge on the operator, which is the wrong vector. Alternatively, such discussions should emphasize how the military as a supraindividual agent (i.e., an organization) can have MHC over targeting operations. Because of this, the ongoing international debate on AWS focuses overly much on the deployment stage of AWS and their relations to individual operators. In doing so, the debate attempts to locate the vector for MHC between those two agents (AWS-human). But it ignore the broader covariance of the distribution of labour between agents within a military complex that determines decision-making practices. The steps outlined above, particularly the pre-mission briefing stage with its collateral damage and proportionality assessments, are largely sidelined in these discussions.

This approach shows the need for a distributed notion of MHC to accurately account for numerous decision and measures undertaken by different agents in the broader decision-making mechanism before deployment. Different agents have different levels of control over any given vector in the process. Any sufficient conception of MHC must therefore reflect this. Of course, this does not negate the role that human operators play. Rather, it positions the role within the larger distributed network of decision-making. Here, 'full autonomy' is not full in the sense that is commonly intuited. It is constrained by the larger apparatus within which it forms a part.[2]

Meaningful human control: design level[3]

The second level of abstraction is drawn from the account for MHC by Santoni di Sio et al. (see Mecacci and Santoni de Sio 2019; Santoni di Sio and van den Hoven 2018). Their view strays from existing approaches to describing MHC to provide a philosophical account of MHC. For them, MHC is the co-variance between the behavior of a system and the intentions or reasons behind an agent's decisions and actions. Systems can be designed in ways that permit agents to forfeit some of their direct operational control while still possessing global control over the system itself. This means that more, rather than less, levels of autonomy may (in certain cases) permit more salient control of a system. As mentioned in the preceding section, more direct operational control has little meaning in the desired sense for autonomous systems. In their approach, clearer lines of accountability can be drawn when humans remain 'in-the-loop' over a system. As tracking the relevant reasons behind an agent's decisions is a necessary condition for MHC, the retention of humans 'in-the-loop' allows MHC.

Their approach to MHC is functionally comprehensive in its breadth, which looks beyond individual systems to the whole sociotechnical infrastructure wherein systems are embedded. Although the specific design and deployment of systems implicate important factors in understanding MHC, they cannot be understood in isolation from the infrastructures, organizations, and other agents who are inextricably connected to their design, deployment and use (Umbrello, 2020). The approach focuses on the design level because it describes MHC as something that can be designed *for* by engineers. In other words, MHC are technical design requirements—not only for the system itself, but also for relevant sociotechnical infrastructures. In order to design for MHC, two necessary conditions must be met: tracking and tracing. Satisfaction of these two conditions permits a more comprehensive conception of MHC that reaches beyond that of solely end users. Here, a level of meaningful control is extended to agents such as designers and policymakers along with organizations and states. With this control comes clearer lines for attributing responsibility.

Tracking and tracing conditions

The tracking condition deals with how responsive a system is to the actions consequent of human reasons.[4] It is more comprehensively defined as:

> *First necessary condition of meaningful human control*. In order to be under meaningful human control, a decision-making system should demonstrably and verifiably be *responsive* to the *human* moral reasons relevant in the circumstances - no matter how many system levels, models, software, or devices of whatever nature separate a human being from the ultimate effects in the world, some of which may be lethal. That is, decision-making systems should *track* (relevant) human moral reasons. (Santoni de Sio & van den Hoven, 2018, p. 7)

[2] This echoes (and Ekelhof repeats it as well) the Defence Science Board's statement that "there are no fully autonomous systems just as there are no fully autonomous soldiers, sailors, airmen or Marines" (USSB, 2012, p. 23).

[3] Much of description in this section is adapted from a paper I published previously, which offers a similar recounting of Santoni di Sio et al.'s version of MHC (Umbrello 2020).

[4] Here, the term 'reasons' is understood as any element that can both prompt and demonstrate human behavior, such as objectives, programs, and strategies.

The tracing condition is different given that it asks whether it is possible to delimit the human agent(s) along the design and deployment history of the system. This means designers, manufacturers, users, and others who are capable of: (1) understanding the system's potential; and (2) recognizing their moral responsibility for the deployment and use of a system (i.e., the liability of moral consequence). Santoni de Sio and van den Hoven (2018) define tracing more thoroughly as:

> *Second necessary condition of meaningful human control*: in order for a system to be under meaningful human control, its actions/states should be traceable to a proper moral understanding on the part of one or more relevant human persons who design or interact with the system, meaning that there is at least one human agent in the design history or use context involved in designing, programming, operating and deploying the autonomous system who (a) understands or is in the position to understand the capabilities of the system and the possible effects in the world of the its use; (b) understands or is in the position to understand that others may have legitimate moral reactions toward them because of how the system affects the world and the role they occupy. (p. 9)

MHC is attained by agents who can satisfy both of these conditions; only then can they be said to have MHC over a system. This means AWS can *prima facie* fall under the MHC of (an) agent(s) when they are designed to support the values of accessibility and explicability (explainability and transparency), which manifest in system behavior, as much as possible. If a system is able to explain its internal decision-making (explicability) and if such systems are themselves transparent (also a factor of explicability), then they can be brought under MHC more easily at least in theory. This is because agent understanding of the use and deployment of a system can be more easily attributed to the design architecture of the system.

With these two necessary conditions, this approach to MHC ultimately entails a definition of control that is more nuanced and stringent than operational control. The latter demands full, direct control. But control over design is more stringent than direct operational control because it precludes the attribution of human control to systems just because they have an agent 'in-the-loop' (e.g., a soldier co-commanding a field operation with an AWS). Even if a commander has a kill switch or can visibly see the current status and activities of an AWS, this does not necessarily mean they are equipped to understand why the system does what it does. In such cases, MHC by the end user is unattainable because the tracing condition cannot be fulfilled due to the opacity of the system. Other agents, such as designers, programmers, the military institution, or even the state. may very well understand what is going on in the 'black box' of the AWS. The system could successfully track the reasons of those agents. Those agents may indeed be capable of understanding the behavior exhibited by the system. Based on the tracking of their primary and more proximal reasons, they could then be held responsible for the behavior exhibited by the system the system along with the way it acts. In the end, these agents have MHC over the AWS. It is here that we can begin to see how the design level can help navigate the distributed nature of military operations planning that was discussed regarding the operational level of MHC.

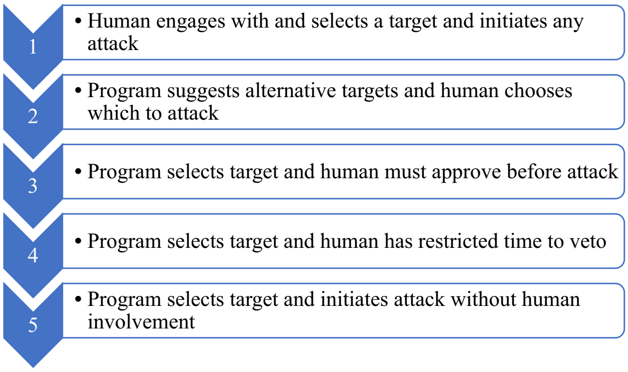
1
• Human engages with and selects a target and initiates any attack
2
• Program suggests alternative targets and human chooses which to attack
3
• Program selects target and human must approve before attack
4
• Program selects target and human has restricted time to veto
5
• Program selects target and initiates attack without human involvement

Fig. 1 Level of Autonomy. Source: (Sharkey, 2014)

The Santoni de Sio et alia account for MHC, which is far more nuanced than presented here, delves into various types of reasons (e.g. proximal and distal ones) (Mecacci & de Sio, 2019). This paper aims to take a more meta-normative approach of combining these theories into a unified notion of MHC. The following section begins this project by discussing how the two LoA are complementary. Both are underscored by a systems thinking perspective and both can be optimized via a systems engineering approach to operational and design innovation.

MHC as design thinking and design engineering as MHC

Technical full autonomy and AWS

As mentioned in the introduction, the central premise of a possible ban on AWS is grounded in a certain level of autonomy that results in an accountability gap in the event of recalcitrance. Sharkey (2014) aptly describes five levels of technical autonomy that can describe AWS targeting (Fig. 1). The least problematic stage is level 1 (although Ekelhof's analysis arguably brings into question which human). Levels 4 and 5 are arguably the most problematic. Both are seen as dangerous due to how an AWS selects a target (i.e., systemic opacity, computer vision, etc.) and its technical ability to do so as a function of various targeting norms and the rules of engagement. Level 4 questions the cognitive clarity of a human operator who has veto power when determining the validity of the target(s) chosen by the system. Regardless, Level 5 is typically the subject of debate as it is considered the descriptor of full autonomy in terms of AWS.

Here, we can already begin to tease out some of the issue with problematizing autonomy per se. There are convincing arguments contra AWS other than the supposed accountability gap proposed by the above ordinance, such as the dehumanization of war and its deleterious effects on human dignity. But it appears that *actual* military operations planning and deployment intuitively constrains the autonomy of any given agent, soldier, or AWS to being a function of a larger a priori plan. This plan bears little, if any, intrinsic operational value outside their functional capacity to carry out such plans. Of course, this does not extricate the AWS deployed within such constraints from the possibilities of limitless actions or wanton recalcitrance. As a predicate for technical design requirements, technical design must reflect both the proximal and distal intentions and goals of relevant agents within the deployment envelope. These would be the commanders who employ such weapons in their area of operations as well as potential human operators who may be engaging with them symbiotically on the ground (e.g., aerial AWS such as fully autonomous drones). Regardless, the system capacity to respond to the relevant moral reasons of relevant agents must be considered a foundational variable in the weaponeering decision-making process for any given context of deployment in the pre-mission stages.

Coupling levels of abstraction for MHC

Systems theory (and in practice, systems thinking) provides salient ground for thinking about these various LoA. The procedural process of operational planning and target identification form the higher, or meta-, level of MHC as clearer lines of causality can be conceptualized. This culminates in weapons release and efficacy assessments. Similarly, the design level of MHC is functionally dependent on a systems understanding of both *tracing* design histories as well as *tracking* the responsiveness of autonomous systems to the relevant moral reason(s) of the relevant agent(s) in the design and use chains for such systems. Theoretically speaking, both LoA are predicated on systems or networks of interconnected nodes (Fig. 2). Similarly, both LoA feed into one another despite their different scopes. Within the operational level, the bounds in which weaponeering decisions are made prior to deployment are contingent on the functionality of the system itself in order for it to be chosen as the most salient means for carrying out the intended mission. But how such technical responsiveness to the on-the-ground needs for successful mission completion is not contingent on those types of pre-mission assessments. System-level recalcitrance can jeopardize the overall level of MHC even when the system is bound by the operational level of control.

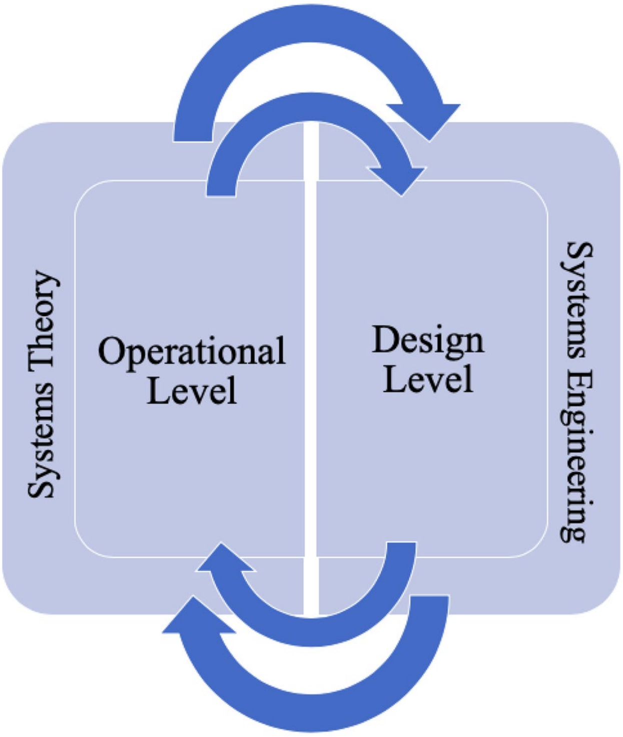

Fig. 2 The superordinance of Systems Theory and Engineering over the two levels of abstraction of MHC

Weaponeering decisions must thus be reflected at the design level in order for those decisions to be sufficiently salient prior to deployment. In this sense, the operational level feeds down into the design level by supplying the norms, objectives, and intentions necessary for deployment to be lawful. These are also necessary for the operational level to be holistic in terms of the sufficiency of control. Likewise, the various agents that are essential to pre-mission planning operations form part of the population of relevant moral agents (or collectively as a supraindividual agent); these agents permit the design level to *actually* design AWS so they are sufficiently responsive to the reasons and intentions of the actor(s) who makes the weaponeering of AWS permissible—and thus under a priori MHC on both LoA. Of course, this would mean a closer military-industrial partnership that uses these agents as stakeholders for whom systems can be designed (coupled with the relevant RoE and LoAC).

One scenario that is often discussed in the literature contra AWS is that of an AWS killing civilians. Within this scenario, we can begin to trace reasons for dismissing such a *prima facie* objection. In order for an AWS to kill a civilian on the ground, the civilian must fall within the weapons envelope delimited prior to deployment. The killing is not *mala in se* to the extent that collateral damage assessments are agreed upon pre-deployment under existing norms for proportionality. To some degree, the killing of civilians is not necessarily equivocal to recalcitrance as it can be traced back to the briefing information. If we imagine that an AWS kills civilians disproportionately even within the weapons envelope and even against explicitly acceptable damages determined in pre-planning, this can be construed as technical recalcitrance. This is because it can be traced back to the relevant agents within the design and use histories of the AWS to determine whether the system was designed in such a way as to be maximally responsive to the relevant intentions of those agents.

If such is shown to not be the case, then the AWS was under MHC. It is thus not a viable option for weaponeering decisions and its deployment was unlawful overall (this is a good vector for thinking about ban criteria). If relevant agents such as designers and users (commanders, AWS designers/programmers, proportionality specialists, etc.) are capable of understanding the capabilities and consequences of the system, then they may be said to be in possession of MHC. They have MHC both in their weaponeering decisions on the operational level, as well as in design decisions at the design level. Divorcing one level from the other leaves open vectors from which accountability gaps can arise.

Systems engineering can be understood as the design and application of both of these levels of MHC. It seems (and perhaps is) more appropriate to speak of systems engineering in terms of the design level, given its explicit focus on building autonomous systems responsibly in a holistic and anticipatory way—particularly in alignment with the values of stakeholders. But the operational level is necessary for that reason. As mentioned above, all of the agents within the complex network distributed across the process of military target acquisition and deployment are relevant moral ones. They are all part of the larger system within which the AWS exists. Likewise, the AWS themselves are embedded in the larger sociotechnical network of operations involving those human agents. In order for these human agents to make salient and hopefully lawful decisions in terms of weaponeering, whether early on or throughout the development cycle, their needs must be analyzed and elicited. This step is critical. At the same time, the entirety of the lifecycle of the system as relates explicitly to those weaponeering decisions must be kept in mind. The system is not a discrete technological artifact divorced from its use-context. To put it more simply, AWS should not be built and marketed as a discrete and novel weapons platform. Designers and experts who plan operations must be part of the design teams to weaponeer the design decisions themselves.

Many of the technical issues presented as *mala in se* against the development of AWS, such as increased autonomy (particularly level 5 as in Fig. 1) or the targeting of civilians, are only problematic if decoupled from responsible design, actual military planning, and actual operations practices. When these are taken into account, the augmentation of autonomy is necessarily constrained by many—if not all º of these processes. In certain cases, autonomy can increase rather than decrease the ability to have MHC. If these systems are designed so as to be maximally sensitive to the relevant moral reasons of the relevant moral agent(s), then they likewise augment MHC rather than lessen it. Mecacci and Santoni de Sio (2019) demonstrate this seemingly paradoxical paradigm nicely with autonomous vehicles. The marriage of both LoA, then, is teleological towards systemic synergism. In avoiding component friction, it subsequently avoids the unreliability of the design and deployment of AWS.

For systems engineering practices to successfully optimize equifinality across various levels of nesting, complexity has to be modeled as a function—one of not only the technical architecture of a system, but also the logical human organization of data (i.e., planning, target data, proportionality assessments, geography, etc.). Given the volume and quality of the data, variables, and components across technical and human spheres, systems may become increasingly complex over time. Much of this can be addressed through the design and development of smarter control algorithms and environmental systems analyses. Tools such as system architecture modeling, verification and learning simulations, and statistical or reliability analyses along with formal decision-making psychology can all be levied to understand the covariance between technical design and human operations.

Divorcing the operational level from design leaves design impotent and potentially recalcitrant. Divorcing the design level from operations leaves operations with an opaque and nebulous lethal tool that may result in poor (if not unlawful) weaponeering decisions. We can think about systems and, more specifically, these various levels of abstraction as co-constituting one another. This permits their inherent complexity to be modeled more easily. As a consequence, we can design *for* complexity rather than leaving design decisions ad hoc afterthoughts. Doing this allows for the tracing of clearer lines of emergent behaviors and boundaries—provided that systems thinking is employed at all levels of nesting.

Limitations and areas for future research

There is at least one notable limitation on this multi-tiered approach: dynamic engagements of AWS in situ rather than purely pre-programmed engagements pose greater challenges. This is particularly true of ground-based AWS in comparison to aerial ones. Ground-based AWS can find themselves (and perhaps most often will) in dynamic and changing engagement scenarios even within the weapon's envelope. Their ability to adhere to the pre-determined mission and targets while also adapting to changing scenarios gives rise to both technical and ethical issues. Given the decisions and identification that emerges from dynamic war theatres and their proximity (and thus finer grained situational input) to targets, such type of ground-based systems appear to take on more moral agency. In these cases, the operational level may be insufficient for grounding MHC in such cases. Still, the design level can provide possible ways to ensure sufficient control.

Systems can be designed to be maximally responsive to the largest set of moral reasons and intentions of the relevant agents (in this case, perhaps the commanding officer on the ground alongside the AWS and/or the commander supervising the mission/engagement). The recalcitrance of such systems can then be tracked and traced back to these individuals, along with the designers who engineered the autonomous systems. Elands et al. (2019) provide one model route for the design level in conceptualizing exactly how more in situ operations can take place and remain under MHC. Likewise, there are developments on advanced intelligent systems (e.g., in the domain of cyber physical systems, Artificial General Intelligence, and AI safety) that explore how systems can be (technically) self-aware, have hybrid AI, and be oriented towards a well-specified objective, utility function, or goal formulated by humans (Aliman, 2020). An important consequence of this research is that the objective, utility, or value function is not part of the design of the system [orthogonality between goal and intelligence] (Aliman, 2020).

Either way, this limitation shows a further nuance that resists arguments for a blanket ban on (fully)AWS: the difference between aerial (fully)AWS and ground-based (fully)AWS. The operational level seems to tokenize the agent who is in direct operational control of the engagement, stripping them of most (if not all) relevant levels of autonomy necessary for moral responsibility. For this reason, the substitution of such human agents in aerial engagements appears benign. For a ban on (fully)AWS to be effective, then, it seems that targeting autonomy per se is not the right strategy. Instead, a more effective route would involve targeting various specific types of AWS and differentiating them (i.e., ground, aerial, naval AWS). Of course, this risks over-specification and leaves open the possibility of circumventing very specific designations and criteria for banned systems. However, it should not discount the above criticism. Rather, it should rather wrestle with it head on in order to ensure more robust policy-making.

Conclusions

This paper uses systems thinking and systems engineering as conceptual tools to frame the commonalities between two different levels of abstraction in understanding the meaningful human control of autonomous weapons systems. It argues that with AWS in particular, both LoA are necessary to achieve MHC. If this coupling is successful, the result would be increased levels of autonomy. Increased autonomy is often seen as problematic, lying at the core of the rationale for a ban on those types of AWS. But this perception is flawed and perhaps entirely dismissible. Autonomy per se, whether of humans or the AWS, is necessarily constrained by military operations planning and the co-construction of these systems with their relevant moral stakeholders. As long as the strict conditions are met across LoA, then increasing the autonomy of AWS to what is traditionally called 'full' autonomy is not problematic. Such an increase in autonomy can conceivably increase MHC as well.

Acknowledgements I would like to thank Maurizio Balistreri who provided useful comments on an earlier draft. The views in this paper are those of the author alone and do not represent those of the Institute for Ethics and Emerging Technologies.

Funding Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement. The author received no funding for this research.

Declarations

Conflict of interest The author declares no conflict of interest.

References

Adams, K. M., Hester, P. T., Bradley, J. M., Meyers, T. J., & Keating, C. B. (2014). Systems theory as the foundation for understanding systems. *Systems Engineering, 17*(1), 112–123.

Aliman, N. M. (2020). Hybrid cognitive-affective strategies for AI safety. Utrecht University. https://doi.org/10.33540/203.

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part I:

Motivation and philosophy. In *Proceedings of the 3rd International Conference on Human Robot Interaction - HRI '08* (p. 121). New York, New York, USA: ACM Press. https://doi.org/10.1145/1349822.1349839

Article 36. (2015). Killing by machine: Key issues for understanding meaningful human control. Retrieved January 28, 2020, from http://www.article36.org/weapons/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/

Asaro, P. (2009). Modeling the moral user. *IEEE Technology and Society Magazine, 28*(1), 20–24. https://doi.org/10.1109/MTS.2009.931863.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology, 320*(1–2), 18–36.

Boscoe, B. (2019). Creating transparency in algorithmic processes. *Delphi - Interdisciplinary Review of Emerging Technologies*. https://doi.org/10.21552/delphi/2019/1/5.

Dudo, A., Dunwoody, S., & Scheufele, D. A. (2011). The emergence of nano news: Tracking thematic trends and changes in U.S. newspaper coverage of nanotechnology. *Journalism & Mass Communication Quarterly, 88*(1), 55–75. https://doi.org/10.1177/107769901108800104.

Elands, P. J. M., Huizing, A. G., Kester, L. J. H. M., Peeters, M. M. M., & Oggero, S. (2019). Governing ethical and effective behaviour of intelligent systems. *Military Spectator*, June 2019. Retrieved from https://www.militairespectator.nl/thema/operaties-ethiek/artikel/governing-ethical-and-effective-behaviour-intelligent-systems

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343–348. https://doi.org/10.1111/1758-5899.12665.

Graham, R., Knuth, D., & Patashnik, O. (1994). 1. Recurrent problems. In R. Graham (Ed.), *Concrete mathematics: A foundation for computer science*. (2nd ed., p. 670). Addison-Wesley Professional.

Haken, H. (2013). *Synergetics: Introduction and advanced topics*. . Springer.

Ivanov, K. (1993). Hypersystems: A base for specification of computer-supported self-learning social systems. *Comprehensive systems design: A new educational technology*. (pp. 381–407). Springer.

Kania, E. B. (2017). Battlefield singularity. *Artificial Intelligence, Military Revolution, and China's Future Military Power, CNAS.*

Leveringhaus, A. (2016). Drones, automated targeting, and moral responsibility. In E. Di Nucci & F. Santoni de Sio (Eds.), *Drones and responsibility: Legal, philosophical, and socio-technical perspectives on the use of remotely controlled weapons*. (pp. 169–181). Routledge.

Mecacci, G., & de Sio, F. S. (2019). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-019-09519-w.

NATO. (2016). *NATO Standard AJP-3.9 Allied Joint Doctrine for Joint Targeting*. Retrieved April 15, 2020, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/628215/20160505-nato_targeting_ajp_3_9.pdf

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 15.

Thomé, B. (1993). *Systems engineering: Principles and practice of computer-based systems engineering*. . Wiley.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology, 11*(2), 105–112. https://doi.org/10.1007/s10676-009-9187-9.

Umbrello, S. (2020). Meaningful Human control over smart home systems: A value sensitive design approach. *Humana. Mente: Journal of Philosophical Studies, 13*, 40–65.

USAF. (2017). *Annex 3-60 Targeting*. Retrieved from https://www.doctrine.af.mil/Doctrine-Annexes/Annex-3-60-Targeting/

USSB. (2012). *Defense Science Board Task Force Report: The Role of Autonomy in DoD Systems*. Washington, DC

Von Bertalanffy, L. (1972). The history and status of general systems theory. *Academy of Management Journal, 15*(4), 407–426.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics, 2*(6), eaan6080.

Wan, P. Y. (2011). Emergence à la systems theory: Epistemological Totalausschluss or ontological novelty? *Philosophy of the Social Sciences, 41*(2), 178–210.

Wernaart, B. (2021). Developing a roadmap for the moral programming of smart technology. *Technology in Society, 64*, 101466. https://doi.org/10.1016/j.techsoc.2020.101466.

Whitchurch, G. G., & Constantine, L. L. (2009). Systems theory. *Sourcebook of family theories and methods*. (pp. 325–355). Springer.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.