Bounded Autonomy

17

Jan Maarten Schraagen

INTRODUCTION

The aim of this concluding chapter is not so much to summarize what has been stated so well by the authors of the various chapters, but rather to reflect on some common themes that run throughout this book, as well as to highlight some additional issues and research challenges, particularly in the field of Human Factors and Ergonomics (HFE). First, I will address the use of AI in military systems and how this relates to the heated debate on 'killer robots'. Second, I will discuss the concept of 'autonomy' and its use in 'autonomous weapon systems (AWS)'. I will argue that, just as there is no such thing as 'rationality' in humans, only 'bounded rationality' (Simon, 1955; 1957), there is no such thing as autonomy in systems, only 'bounded autonomy'. This then will lead to a discussion of the third concept, that of 'meaningful human control (MHC)', which, as will be shown, is closely related to the concept of 'bounded autonomy'. I will argue that the existence of bounded autonomy makes MHC over military systems possible. In the final section, I will discuss what it could mean to develop AI in a responsible fashion in military systems, which, after all, is the title and main topic of this book.

DEMYSTIFYING DYSTOPIAN VIEWS ON THE USE OF AI IN MILITARY SYSTEMS

In some dystopian visions, AI is seen as a technology that is beyond our control and that enables particular weapon systems to select and engage targets by themselves. This vision has been reinforced by the Future of Life Institute that has funded a movie

DOI: 10.1201/9781003410379-22

showing how swarms of flying robots ('drones') kill large numbers of innocent people (Russell, 2022; Slaughterbots, 2017). Not only is the flying of the robots enabled by AI, but the assumption is that the AI employed in these drones also enables the face-recognition that is required for the targeting process. What is particularly worrisome in this depiction is that, once the AI has been successfully applied in one instance, it can be replicated easily and applied to hundreds of thousands of drones, making targeted mass killings within reach of terrorists or rogue nations or so the proponents of this campaign claim (Russell, 2022). Other arguments of those opposing 'killer robots' are accountability gaps (if no humans are involved, they cannot be held accountable), violation of International Humanitarian Law (IHL; principles of distinction and proportionality), and the dehumanization of warfare.

The general public is mostly concerned with the use of AI in AWS, also framed as 'killer robots'. However, AI – and certainly the recent developments in generative AI – may be applied across the military enterprise, ranging from human resource management systems to maintenance and logistics systems, from cyber defense systems to reconnaissance systems, and from decision support systems to joint protection and warfighting.

Any discussion of the use of AI in military systems, let alone the 'responsible' use of AI in such systems, needs to relate to these concerns and fears. It is very difficult to deal with these emotions on a rational basis, yet this is what those painting a more nuanced picture need to do. Presenting information and myth-busting are required to assuage those fears, without ignoring or diminishing in any way the real concerns many people have with the rapid development of AI.

One of the first arguments against the Slaughterbots movie is that it depicts a fictitious situation and that, as of yet, there are no killer robots enabled by AI. Although this is being debated, with some arguing that killer drones were employed in Libya in 2020 (in effect, the drone being used was the Kargu-2 rotary wing loitering munition, UN Security Council, 2021), the more general line of reasoning is familiar with anyone watching discussions between those who believe some form of autonomous technology is 'just around the corner' versus those who are more skeptical about those claims. The fact that the Slaughterbots movie depicts a fictitious situation is irrelevant to the first group, as they believe that, even though it may not be a reality today, it will be a reality in the foreseeable future (with 'foreseeable' being used as an ever-receding horizon if it does not materialize in time). Those who believe in this technology take the 'fake it till you make it' stance, which they claim is necessary or else there would be no funding for any new technological developments. This particular stance has been so ingrained in Silicon Valley culture that to criticize it is tantamount to criticizing progress in general (though with the funding drying up, faking it may be over, Griffith, 2023). Those warning about killer robots may not be the same as those Silicon Valley entrepreneurs warning about the dangers of AI, but they do share the same belief in technology that is always just around the corner. I am arguing here that this technology may never materialize.

An analogy may be drawn with the prophesies about self-driving ('autonomous') cars. In 2015, Musk predicted 'complete autonomy' by 2017. In 2016, Lyft co-founder and president Zimmer claimed that by 2025 car ownership in US cities would "all but end" (Sipe, 2023). General Motors in 2017 promised mass production of fully

autonomous vehicles in 2019. More than $100 billion has been invested in self-driving cars since 2010 (Chafkin, 2022). However, in 2020 and 2021, respectively, Uber and Lyft shut down their efforts. By the end of 2022, Volkswagen and Ford pulled the plug on their self-driving efforts (Sipe, 2023). According to Anthony Levandowski, a fervent believer turned apostate, "You'd be hard-pressed to find another industry that's invested so many dollars in R&D and that has delivered so little." (Chafkin, 2022). Most of the testing of self-driving cars is done in sunny California and Arizona, hardly representative of the rest of the US, let alone the world. What these cars have difficulties with, are what the engineers call 'edge cases'. Edge cases go far beyond the sunny weather these cars usually operate in: from broken traffic lights to a bicyclist crossing a street, the list is endless as the world is a messy place. Human drivers know what to expect of pigeons on the road and how to respond; self-driving cars have no such intelligence and will slam the brakes causing rear-end collisions. This example shows that the issues are not so much with object detection as such, but rather with interpretation of the information obtained in the context of driving.

The more general problem here is what Woods (2016, p.131) has stated as the gap between the demonstration and the real thing:

> Computer-based simulation and rapid prototyping tools are now broadly available and powerful enough that it is relatively easy to demonstrate almost anything, provided that conditions are made sufficiently idealized. However, the real world is typically far from idealized, and thus a system must have enough robustness in order to close the gap between demonstration and the real thing.
>
> *(Doyle/D. Alderson, personal communication, January 4, 2013)*

Demonstrating that one can drive a car autonomously under confined and idealized conditions, for a brief period of time, does not mean the car can be let loose in the real world, under less-than-ideal conditions. Similarly, as Lindsay (this Volume) stated about AI:

> AI relies on large-scale data and stable collective judgments. But these same conditions are elusive in war. AI, to put it glibly, is an economic product of peace. War destroys the conditions that make AI viable. The conditions that are conducive for AI are not conducive for war, and vice versa.

Hence, the second argument against a dystopian view on AI-enabled weapons is that it underestimates the differences between war and peace and that drones that are tested with AI in peacetime conditions are not robust enough to operate in wartime conditions. The difference may actually even be larger than the difference in traffic conditions for self-driving versus human-driven vehicles.

The analogy with self-driving cars may be taken even further to advance a third argument against dystopian views. Our focus on self-driving typically concerns removing the single human driver and replacing them with sensors. Considering the human task of driving a car an instance of perceiving objects in the environment and appropriately acting upon them, is an example of the 'reductive tendency' that humans are prone to (Feltovich, Hoffman, Woods, & Roesler, 2004). It means neglecting the

complexity and connectedness of driving and reducing it to a perceptual-motor task. Once we have completed this reduction in our minds, we can then proceed to automate the perceptual-motor task in order to achieve our goal of developing a self-driving car. But this reductive framing of driving misses very important aspects of the real-world driving task that cannot be easily automated. For instance, driving involves the constant awareness of others and their inferred intentions. If I see a bicyclist coming from the left, cycling at speed and avoiding eye contact, I decide to wait even though I have the right of way. If, however, the bicyclist looks at me and slows down, I may either slowly go ahead or wave to let the bicyclist pass anyway. This example may still be an oversimplification of many real-world traffic situations. The coordination between two interacting road users may be shaped by surrounding road users (Renner & Johansson, 2006). Nathanael and Papakostopoulos (2023) describe a range of coordination strategies employed by road users, such as the 'Pittsburgh left' in which a car is allowed to take a left turn at a two-lane intersection immediately after the traffic light turns green (as if there were a left turn signal) provided that the driver of the oncoming car is willing to cooperate. In general, a driver's exploitation of situational opportunities to gain priority is often contrary to regulatory provisions, but favoring overall traffic efficiency. These human coordination strategies pose the following design challenges for self-driving cars in mixed traffic: (i) distinguish these strategies from errant driving, (ii) recognize to whom a 'space-offering' is addressed, and (iii) assess the appropriateness or abusiveness of a particular strategy (Nathanael & Papakostopoulos, 2023). It is clear from actual observations of self-driving cars that these challenges are currently far beyond their capabilities (Brown & Laurier, 2017) and will frequently result in stalled traffic that requires human intervention (Metz, 2023).

Human coordination strategies are obviously of great importance in military operations as well. For instance, in drone warfare operations in Afghanistan, the focus has often been on the 'sharp end' of the drone warriors who operate these drones from a distance. And while these pilots often suffer from high workloads and moral vexations (Philipps, 2022), it is actually the 'blunt end' ('the customer') who designates the targets. Although targeting may be described as the practice of destroying enemy forces and equipment, it is more accurate and contemporary to describe it as a deliberate and methodical decision-making process to achieve the effects needed to meet strategic and operational campaign objectives (Ekelhof, 2018). This decision-making process may involve hundreds of people working over extended periods of time (at least months). Many automated tools already exist to assist in the targeting process and AI may certainly be used to further improve this process. However, this does not imply that AI will completely 'take over' the targeting process. It may change certain tasks, but the system should be designed such that ultimate control is still with the human. This is not to deny that drones could be used in autonomous mode against military objects and that AI increasingly plays a role in identifying these objects. AI-controlled military drones are reportedly being used in the war between Russia and Ukraine. These drones are able to independently identify and attack military objects. According to New Scientist, Ukraine is using the drone in autonomous attack mode: it is the first confirmed use of a "killer robot", the website says (Hambling, 2023). This brings us to the question of what it means for a weapon to be used in 'autonomous' mode. This will be taken up in the next section.

AUTONOMY

The concept of 'autonomy' has recently been used frequently in the discussion on AWS, in particular as a means of banning these weapon systems. It is not my intention in this section to arrive at a definition in order to ban autonomous weapons. Rather, it is my intention to shed light on various definitions of 'autonomy' and its relation with AI. The relation between autonomy and AI has not always been made sufficiently clear. It has frequently and implicitly been assumed that AWS need AI to function properly or even to be able to exist at all, but this largely depends on one's definition of AWS and to a lesser extent on one's definition of AI. For starters, systems such as the Phalanx or Goalkeeper behave more or less 'autonomously' without any use of AI. These systems were developed in the 1970s using linear programming methods. These are so-called close-in weapon systems to defend naval vessels automatically against incoming threats. According to Scharre (2018), these automated defensive systems are autonomous weapons, but they have been used to date in very narrow ways. These systems do need to be activated by humans, but, once activated, will search for, detect, track, and engage targets that match predetermined profiles (based on, for instance, angle of approach and speed – the minimum and maximum limits are set by operators). These systems are used as a last resort for self-defense and as such also destroy friendly targets that match the predetermined criteria. They are used to target objects, not people. Humans supervise these systems in real-time and could physically disable them if they stopped responding to their commands. This category of weapon systems is mentioned here to illustrate some of the conceptual difficulties that arise when using the term 'autonomous weapon system'. If one does not consider these systems to be fully 'autonomous', then what criteria do systems need to fulfill to be considered 'autonomous'? Would it be sufficient to equip these systems with AI, so they can, for instance, distinguish between friend and foe? Or would these systems have to be able to switch themselves on and engage targets without any human intervention? The former addition would make the systems more 'intelligent' perhaps, but no so much more 'autonomous'. The latter addition would make the systems definitely more 'autonomous' (by whatever definition of 'autonomy' one might entertain), but also more unpredictable and less trusted. Hence, it is necessary to discuss the concept of 'autonomy' in some more detail in order to be able to classify particular classes of weapon systems as being 'autonomous' or not.

One definition of 'autonomy' that has been cited frequently in this Volume by various authors is the one provided by the US DoD in its Directive 3000.09 (U.S. Department of Defense, 2023). Although this Directive is called "Autonomy in weapon systems", it does not provide a definition of 'autonomy'. It does, however, provide a definition of an 'autonomous weapon system':

> A weapon system that, once activated, can select and engage targets without further intervention by an operator. This includes, but is not limited to, operator-supervised autonomous weapon systems that are designed to allow operators to override operation of the weapon system, but can select and engage targets without further operator input after activation.

From this definition we can distill a definition of 'autonomy', in the context of weapon systems, as follows: 'the ability to select and engage targets without further intervention by an operator'. This definition is ambiguous as to what 'select target' actually means, but the Directive also provides a definition for target selection, as follows: "the identification of an individual target or a specific group of targets for engagement". What is crucial here, as stressed by Scharre and Horowitz (2015), is that an AWS does not engage a specific target but rather engages targets of a particular class within a broad geographic area without any human involvement. A loitering munition such as the Harpy would be an example of this. According to Scharre and Horowitz (2015, p.13), loitering munitions are "launched into a general area where they will loiter, flying a search pattern looking for targets within a general class, such as enemy radars, ships or tanks. Then, upon finding a target that meets its parameters, the weapon will fly into the target and destroy it". While this is a critical distinction with guided munitions, where the human controller must know the specific target to be engaged, Scharre and Horowitz do not mention the fact that in both cases targets must meet pre-set parameters and that the parameters in loitering munitions are also programmed by human operators. It is clear why Scharre and Horowitz emphasize the difference between specific targets and general classes of targets because otherwise the entire discussion on AWS "would be a lot of fuss for nothing" (Scharre & Horowitz, 2015, p.16), as guided munitions, which have been around for 75 years or more, would then also have to be classified as AWS. This "almost certainly misses the mark about what is novel about potential future autonomy in weapons", according to Scharre and Horowitz (2015, p. 17). It should be noted that the 'general area' and the 'general class of targets' are not without their own problems, such as the increased risk of collateral damage in populated areas and the fact that 'general targets' increase the risk of false positives (e.g., when using a face recognition algorithm).

Two other phrases are also noteworthy. The first is the phrase 'once activated'. The phrase 'after activation' occurs again at the end of the definition. This implies that an AWS first needs to be activated, presumably by a (human) operator, in order for it to identify and engage targets. The second is the phrase 'further intervention' or 'further operator input' which both strongly suggest that the activation of the weapon system is carried out by an operator. An AWS, in this definition, does not switch itself on to go on a killing spree. In this respect, close-in weapon systems fall under the general category of 'autonomous weapon systems', as they also need to be activated and are capable of selecting/identifying and engaging targets without further intervention by an operator. Loitering munitions also need to be activated by human operators, even though the time scale at which close-in weapon systems and loitering munitions are activated may differ.

Finally, it is noteworthy that in this definition of AWS, the operator can override the operation of the weapon system. This does not make the weapon system semi-autonomous. Hence, the defining difference between autonomous and semi-AWS is not in the possibility of humans to override the system, but rather in the capability of autonomous systems to select targets on their own (which should be interpreted as the capability to select classes of targets rather than specific targets). Hence, there is still the possibility of exerting human control over AWS, once they are activated, even though it is not a necessary condition for such systems to be called 'autonomous'. Once again, even close-in weapon systems may conform to this definition as they allow for a mode of

control in which the operator may override the operation of the weapon system. This does not make these systems less autonomous, in this definition.

The US DoD Directive does not mention the use of AI in its definition of AWS, and rightly so. AI is merely a technology to accomplish certain functions, in this case, the ability to select and engage targets. Given that close-in weapon systems do not use AI and conform to the definition of AWS, one may conclude that AI is not required for AWS to exist or even function properly. However, neither does the definition exclude AI for future use in AWS. It is foreseeable that AWS may function better using AI, but this does not make these systems more autonomous than they would have been without AI.

NATO has also provided a definition of autonomy that is noticeably different from the one provided by the US DoD. The Official NATO Terminology Database (NATOTerm, 2023) provides the following definition of 'autonomy':

> A system's ability to function, within parameters established by programming and without outside intervention, in accordance with desired goals, based on acquired knowledge and an evolving situational awareness.

If we parse this, it first and foremost states that autonomy is a system's ability to function in accordance with desired goals. This ability to function is bounded by parameters established by programming, hence the system is not capable of setting its own parameters. These boundaries are presumably set by humans, although the definition does not make this clear (the parameters could also be set by software outside of the system, but this would lead to an infinite regress). Furthermore, the system functions without outside intervention. This is a difference with the US DoD definition, where operators could override the operation of the system, and the system would still be called 'autonomous.' Finally, the definition says something about how the system is capable of functioning in accordance with desired goals: this is based on 'acquired knowledge' and an 'evolving situational awareness (SA)'.

This definition does not make clear who sets the desired goals, who acquires the knowledge, and how SA can evolve, or even what it means for a system to have SA. It is interesting, furthermore, that the NATO Database of terminology does not contain a definition of 'autonomous weapon system'. If it would have contained such a definition, and by applying the definition of 'autonomy' to a 'weapon system', we would have to arrive at the conclusion that such weapon systems would be able to operate without outside intervention, within parameters established by programming, that they would achieve desired goals based on acquired knowledge, and that they would possess an evolving SA. It is clear that such a definition would bring us much closer to AI than the definition stated in the US DoD Directive 3000.09. The NATO terminology is largely derived from the fields of cognitive science, cognitive engineering, and artificial intelligence. It sets a rather high bar for systems to be called 'autonomous', or so it seems. At first sight, one would have to exclude close-in weapon systems. However, if one equates 'acquired knowledge' with 'pre-programmed specifications' and 'evolving situational awareness' with 'dynamic model of target features', then these systems could still be considered 'autonomous' under this NATO definition.

Taddeo and Blanchard (2022, p. 15) have critically reviewed 12 definitions of 'autonomous weapon systems' and, based on cognitive systems engineering, arrived at the following definition themselves:

> An artificial agent which, at the very minimum, is able to change its own internal states to achieve a given goal, or set of goals, within its dynamic operating environment and without the direct intervention of another agent and may also be endowed with some abilities for changing its own transition rules without the intervention of another agent, and which is deployed with the purpose of exerting kinetic force against a physical entity (whether an object or a human being) and to this end is able to identify, select or attack the target without the intervention of another agent is an AWS.

According to Taddeo and Blanchard (2022), once deployed, an AWS can be operated with or without some forms of human control (in, on or out the loop). In this regard, they are in agreement with the US DoD Directive 3000.09. A lethal AWS is specific subset of an AWS with the goal of exerting kinetic force against human beings.

Although one may critique the term 'artificial agent' for legal purposes (Seixas Nunes, SJ, this Volume), given that agency is inherently linked to the notion of liability, I take it that Taddeo and Blanchard are referring to software agents, or software in brief. This allows the system to identify, select, or attack a target, which again is in agreement with the US DoD. The definition mentions no less than three times 'without the intervention of another agent', which is in agreement with the NATO definition that states 'without outside intervention'. However, Taddeo and Blanchard go further than either the US DoD or NATO, particularly when they stress that an AWS is capable of 'changing its own internal states' and 'changing its own transition rules' in order to achieve a set of goals in a dynamic operating environment. Regardless of whether this can be accomplished with rule-based AI or with machine learning using neural networks, this requirement specifies some kind of learning system. As such, it would in all likelihood be unacceptable for military commanders, as it would render the system practically unpredictable and therefore untrustworthy. What makes it all the more questionable is that this kind of self-learning is accomplished without the intervention of another agent. Hence, there is no level of human control whatsoever over this kind of AWS. Although Taddeo and Blanchard claim that their definition is 'value-neutral', it sets the bar so high that any system that potentially meets their criteria will in all likelihood be unacceptable for military commanders. It should also be clear from this definition that close-in weapon systems do not fall within this categorization of AWS, as close-in weapon systems are unable to change their own internal states or transition rules.

Finally, Kaber (2018) presented a conceptual framework of autonomous and automated agents. Although not geared to weapon systems, it is nevertheless an interesting perspective on autonomy, as it contrasts this concept with the concept of automation. Kaber makes clear that the Levels of Automation approach should not be evaluated from a 'lens' of autonomy, as the concepts are quite distinct. Kaber conceptualizes autonomy as a multifaceted construct including: (1) viability of an agent in an environment; (2) agent independence or capacity for function/performance without assistance from other agents; and (3) agent 'self-governance' or freedom to define goals and formulate an operational strategy. An agent could be a software agent, but also a 'thing' in an environment with sensing and effecting capabilities. Viability is the capability of an agent/human to sustain the basic functions necessary for survival in context. Self-governance requires cognitive abilities such as learning and strategizing so the agent can formulate goals and initiate tasks. According to Kaber, this is beyond the capabilities of present

advanced computerized and mechanical systems (although AI is capable of formulating subgoals, it is not capable of formulating the top-level goal).

The facet of self-governance is critical to differentiating autonomy from automation. Loss of capacity for self-governance relegates an autonomous agent to an automated agent. According to Kaber (2018, p. 413): "In general, when autonomous agents are pushed past the boundaries of their intended design context, they become forms of automated or functional agents.". In Kaber's conceptual framework of autonomy, one can only speak of autonomy when an agent scores high on all three facets of autonomy (i.e., viability, independence, self-governance) for a particular context. There are no levels of autonomy and therefore it is a 'misnomer' to speak of 'semi-autonomous systems' (Kaber, 2018, p. 417).

Using this framework, we can establish whether a system is autonomous as a result of the absence of specific characteristics. For instance, close-in weapon systems are hardened to their environment (i.e., they are able to operate in the specific naval context for which they were designed), they do not require monitoring or intervention by humans in a defined operating context (even though they are designed for, and might require, monitoring in different contexts), but they are not 'self-governing' (i.e., they are not responsible for mission goals or control of resources as they do not have the capacity to learn or strategize). The absence of the latter facet of autonomy means these systems are not autonomous, according to Kaber's (2018) framework. In fact, according to Kaber, there are currently no autonomous systems beyond known and static environments. This reinforces what was stated above when the difference between the simulation and the 'real thing' was discussed in the context of self-driving cars, or when the context of the use of AI (peace versus wartime) was discussed.

It is important to note that while autonomous agents pose low demands on humans for supervision or management (whereas automation requires human supervision), this does not imply that they cannot serve as partners for humans in achieving a broader mission. Also, many application environments or work systems require humans to support autonomous agents and vice versa. This dictates additional agent design requirements, particularly from a coordination perspective, as already stated above when discussing the targeting process or the sophisticated coordination strategies employed by humans in traffic. Finally, there may also be a dynamic shifting of functions back and forth between humans and autonomous agents, particularly when environmental conditions change beyond the capacities of an autonomous agent. This may be the case, for instance, when road and weather conditions force self-driving cars to enlist the driver's assistance, fully recognizing that human drivers may also experience difficulties under these circumstances. It is well-known in the human factors literature that such sudden transitions of control may lead to 'automation surprises' (Sarter et al., 1997). It is not sufficient to state that humans should be able to exercise 'appropriate levels of human judgment' (U.S. Department of Defense, 2023) or 'meaningful human control' (Ekelhof, 2019) in these cases of shifting control. Even if, in the far future, there will be AWS that are able to deal with unknown and dynamic environments, dynamic shifts in control will occur and humans will have been out of the loop for so long that they either lack the skills to regain control ('deskilling'; Bainbridge, 1983) or are confronted with an 'automation conundrum' (Endsley, 2017). The latter reflects the fact that the better

the automation, the less likely humans are able to take over manual control when needed (Endsley deliberately uses automation and autonomy interchangeably).

Comparing Kaber's (2018) definition of 'autonomy' with the other definitions discussed, we may note some similarities with Taddeo and Blanchard's (2022) definition. Both definitions stress agent independence and self-governance. However, Kaber additionally stresses the viability of an agent in an environment, an aspect that other definitions have overlooked or have ignored. Viability is not absolute, obviously. Just as humans, who are generally considered to be 'autonomous' creatures, display limits to their viability across different contexts (Kaber suggests relocating a human outside the Earth's surface atmosphere to reveal the limits of any autonomy), so other agents' viability is always relative to a particular context. Self-driving cars may be viable under Sunny State contexts, but not viable in harsh winter weather. What this means for AWS is not immediately clear. At the very least, one would, when accepting Kaber's framework, have to add viability to the definition provided by Taddeo and Blanchard. This would imply that AWS would have to be able to sustain their operations across at least a range of contexts (imposed by, e.g., weather, terrain, enemy operations, available time) and be able to adapt themselves, through rule modification, to these various contexts. It may be that this is what Taddeo and Blanchard meant by achieving goals 'within its dynamic operating environment'. In that case, their definition of autonomy meets Kaber's viability characteristic.

In summary, we have seen a wide variety of definitions of 'autonomy', as used in 'autonomous weapon systems'. We are left with a choice between definitions that set a high bar for AWS, insofar they are required to possess learning (Taddeo & Blanchard, 2022) or self-governing capabilities (Kaber, 2018), versus definitions that set a lower bar and that include close-in weapon systems and loitering munitions as a class of AWS (US DoD Directive 3000.09). Systems that learn without outside intervention and that are therefore 'self-governing' may not be acceptable to military commanders as they are essentially unpredictable, may not conform to Rules of Engagement, and can therefore not be trusted. Accepting this definition would in effect mean that any use of the term 'autonomous weapon system' would be inappropriate, at least when describing current weapon systems and possible weapon systems for the foreseeable future. It does not preclude that there will ever be weapon systems that conform to this definition, yet, if they are developed, they will in all likelihood be unacceptable for responsible military use. Finally, there is also a very pragmatic reason not to adopt this 'high bar' definition, which is that the current usage of the term 'autonomous weapon system' is much more in line with the US DoD Directive 3000.09 definition. From an academic point of view, the high bar definitions might be preferable, but they leave us mostly empty-handed: we would have to exclude all current weapon systems from this definition, as well as most to come, and, in the unlikely case there will be a future weapon system conforming to these definitions, we would have to ban it from being used, as there will be no guarantee that it will conform to Rules of Engagement. From an ethical point of view, this could be precisely what is desirable, and it could be the entire point of advancing this definition. But then we are left with countless current weapon systems that are 'highly automated' rather than 'autonomous' and whose effects are just as lethal.

In the end, what is important is the level of human control that can be exerted over the weapon systems. This brings us to the discussion of 'meaningful human control',

which will be taken up in the next section. For now, it suffices to say that accepting the US DoD Directive 3000.09 definition of 'autonomous weapon system' explicitly includes that level of human control. First, by requiring that the system has to be activated by a human operator. Second, by noting that an operator can override the operation of the weapon system, although not necessarily so. Third, by making clear that although the weapon system may 'select' targets on its own, this in fact boils down to selecting classes of targets rather than specific targets. The latter point is not immediately obvious when first reading this definition. The discussion is also muddled by the US DoD's use of the term 'semi-autonomous', which is reserved for systems that only engage targets but do not select them. However, there are two uses of the term 'target selection'. One is target selection once the system has been activated. This is the sense in which the US DoD uses the term. If, once activated, an operator does the target selection, the system is called 'semi-autonomous'; if the system does the target selection (meaning it does not select a specific target but rather a class of targets over a wide geographic area), it is called 'autonomous'. However, a second use of the term 'target selection' applies to the programming and development phase of the weapon system: 'target selection' here means the specification, frequently in software, of the target parameters. This is clearly under human control and no AWS that meets the definition of the US DoD Directive 3000.09 can do without such target parameters programmed into the weapon system. Parameters in loitering munitions are also programmed by human programmers. This leaves us with the uneasy conclusion that even loitering munitions, considered by Scharre and Horowitz (2015) to be the only examples of autonomous systems (apart from close-in weapon systems), are in fact under human control and cannot be considered 'autonomous'. This would mean that the distinction between a specific target (chosen by a human in, for instance, guided munitions) and a general class of targets (programmed by a human in loitering munitions) would not be as large as Scharre and Horowitz (2015) claimed, making the discussion on the definition of AWS indeed a lot of fuss for nothing.

Accepting the high bar definitions leads to the conclusion that there are no AWS yet, and they are not likely to be developed in the near future. Accepting what Scharre and Horowitz (2015) called a 'common sense definition' yields a superficial distinction between widely used guided missile systems that engage specific targets versus AWS that engage a class of targets on their own. This distinction is superficial in that both classes of weapon systems are ultimately under human control. We therefore seem to be caught between a rock and a hard place and will need to expand our view on autonomy and AWS.

BOUNDED AUTONOMY

In this section, I will introduce the concept of 'bounded autonomy' in order to arrive at a broader and more acceptable definition of autonomy. Introducing this concept will also serve the purpose to gain more clarity on the issue of 'meaningful human control' or being able to 'exercise appropriate levels of human judgment', as the US DoD states.

The concept of 'bounded autonomy' is introduced here by analogy to the concept of 'bounded rationality', as introduced by Simon (1947, 1955). According to Simon (1957, p. 198):

> The capacity of the human mind for formulating and solving complex problems is very small compared with the size of the problems whose solution is required for objectively rational behavior in the real world – or even for a reasonable approximation to such objective rationality.

As a result, the human actor must "construct a simplified model of the real situation in order to deal with it" (Simon, 1957, p. 199). Humans behave rationally with regard to these simplified models, but such behavior does not even approximate objective rationality. Rational choice exists and is meaningful, but it is severely bounded. Our knowledge is necessarily always imperfect, because of fundamental limitations to our information-processing systems (e.g., limits on working memory capacity) and because of fundamental limitations to the attention we can pay to the external world. When confronted with choices of any complexity, we necessarily have to 'satisfice' rather than optimize. Rational behavior is as much determined by the "inner environments" of people's minds, both their memory contents and their processes, as by the "outer environment" of the world on which they act (cf. Simon, 2000).

In terms of economics theory, we could reformulate this as a relaxation of one or more of the assumptions underlying Subjective Expected Utility (SEU) theory underlying neoclassical economics:

> Instead of assuming a fixed set of alternatives among which the decision-maker chooses, we may postulate a process for generating alternatives. Instead of assuming known probability distributions of outcomes, we may introduce estimating procedures for them, or we may look for strategies for dealing with uncertainty that do not assume knowledge of probabilities. Instead of assuming the maximization of a utility function, we may postulate a satisfying strategy. The particular deviations from the SEU assumptions of global maximization introduced by behaviorally oriented economists are derived from what is known, empirically, about human thought and choice processes, and especially what is known about the limits of human cognitive capacity for discovering alternatives, computing their consequences under certainty or uncertainty, and making comparisons among them (Simon, 1990, p. 15).

If we take Kaber's (2018) framework for autonomy as the equivalent of SEU theory, we can reformulate our proposed concept of 'bounded autonomy' as a relaxation of one or more of Kaber's three facets of autonomy (i.e., viability, independence, and self-governance). Rather than relegating the agent to the domain of automation, when the facet of self-governance is lacking, we may view the agent as displaying bounded autonomy. Instead of assuming the viability of basic functions in particular contexts, we may postulate the viability of a limited set of functions for a shorter duration. Instead of assuming independence, we may postulate dependence on parameters established during a preceding targeting process or dependence on mission and task constraints. Instead of assuming self-governance, we may postulate performance in accordance with desired mission goals, based on knowledge acquired during controlled training sessions and a continuously updated model of the environment.

By analogy to bounded rationality, I will now put forth the following description of bounded autonomy:

> The capacity of a system to display viability, independence and self-governance (i.e. 'autonomy') is very limited compared with the variety of the environments to which adaptation is required for objectively autonomous behavior in the real world – or even for a reasonable approximation to such objective autonomy.

This is in accordance with Kaber (2018) who notes that autonomous systems are currently restricted to known and static environments. As a result, the system is dependent on a simplified model of the real situation in order to deal with it (cf. Woods, 2016). Systems behave autonomously with regard to these simplified models, but such behavior does not even approximate objective autonomy, that is, autonomy that fully meets all three facets of viability, independence, and self-governance.

Applying this general concept of bounded autonomy to AWS, brings us to the following maxim:

> The capacity of an autonomous weapon system to select and engage targets on its own ('platform autonomy') is highly dependent on parameters established during the preceding targeting process as well as constraints imposed by legal, ethical, and societal considerations ('mission autonomy').

According to this maxim, a distinction has to be made between platform autonomy and mission autonomy. Mission autonomy concerns what an autonomous system should do and within which constraints. This is the domain of the human. This might be restricted to a single commander, but this is frequently an oversimplification and usually involves hundreds of people (cf. Ekelhof, 2018), not merely in the targeting process, but more generally in the governance and design loops (Heijnen et al., this Volume). This involves the entire design and development process preceding the deployment of an AWS, including testing, evaluation, validation and verification, training with humans in the loop, as well as post-deployment evaluation processes. Mission autonomy is what makes the platform boundedly autonomous. Focusing exclusively on the selection and engagement of targets ('platform autonomy') misses the point. Humans are in control of assessing the necessity and applicability of autonomy. They set the boundaries within which a platform can then operate.

Autonomy is the possibility of an unmanned system to execute an ordered task. The military commander, assisted by countless others, orders the task for the unmanned systems to execute, the AI can help in developing the plan, and the plan is presented to the military commander, he or she can adjust it or he or she can approve it, and then the plan is transferred to the platform (e.g., robot or drone). Hence, the platform is specifically ordered what to do in terms of tasks, conditions, and constraints. Only then do we have controllable, that is, bounded, autonomy.

In the previous section, we were caught between a rock and a hard place in terms of what definition to choose for 'autonomous weapon system' and how to apply that definition to a range of existing and future weapon systems. Given the discussion above, it is now clear that there currently are only 'boundedly autonomous' weapon systems. Moreover, each of these weapon systems contains varying degrees of platform and

mission autonomy. There are no hard and clear-cut distinctions to be made between various weapon systems in terms of their 'autonomy'. Loitering munitions may be considered to exhibit less bounded autonomy than guided missiles, as their mission autonomy allows them to select and engage a wider range of targets than guided missiles. The mission autonomy for guided missiles is fairly restrictive, in that these weapon systems have generally been programmed to attack a single target. That this constraint has been relaxed somewhat in the case of loitering munitions (as well as in the case of close-in weapon systems), does not make these systems qualitatively different from guided missile systems. All current systems display bounded autonomy and the discussion would be more fruitful if we focused on the various ways platform autonomy and mission autonomy are achieved than on whether these weapon systems belong to qualitatively different categories.

One could, of course, deliberately restrict one's definition to the phase after activation, but this would be an oversimplification of a complicated process of decision-making finally leading to weapons release. It is akin to blaming a nurse for a medication error that clearly is the result of an entire work organization or design issue. 'Human error', then, is a symptom of a system that needs to be redesigned, not a symptom of a human that needs to be retrained or fired. The emphasis, in system safety, has changed from preventing failures to enforcing constraints on system behavior (Leveson, 2011). In the same way, the emphasis in discussions on AWS needs to change from preventing failures that occur after such weapon systems have been activated to enforcing constraints, through mission autonomy, on such weapon systems. Eggert (this Volume) counters this conclusion by stating: "The fact that humans are in control *before* a weapon system is activated hardly shows that no human control is required *after* a weapon system is activated". This is true and is called 'human on the loop' or the ability to intervene if the weapon system fails or malfunctions. There are many systems, particularly defensive systems that target objects, not people, where such control is possible after the weapon system is activated. Systems where there is no human control after they are activated are rare, however (Scharre, 2018). Loitering munitions are the primary example and the Harpy (and possibly the Saker Scout drone; Hambling, 2023) are currently likely the only operational examples where the human is 'out of the loop' (other circumstances can be imagined in which autonomy without real-time human control could be desirable, for instance in underwater or silent operations). The Harpy dive-bombs into a radar and self-destructs (as well as destroying the radar) after it detects any radar that meets its criteria. The Saker Scout drone is said to be able to identify 64 types of military objects. It is not clear why Eggert would want human control over the Harpy *after* activation, unless she is suspicious about the criteria that the Harpy has been provided with *before* it is activated, and she is distrustful of the entire testing, evaluation, validation, and verification cycle that the Harpy presumably has undergone. Granted, there could be collateral damage if the Harpy destroys a radar, just as there will be collateral damage if the Saker Scout drone destroys a tank. However, human operators can also make mistakes in drone warfare and cause collateral damage (Philipps, 2022). Moreover, in these cases the Harpy and Saker Scout would still comply with key principles of IHL, notably those of distinction and proportionality (Eggert would counter this by stating that complying with IHL is not the same as complying with morality). The real fear, of course, is with loitering munitions that are programmed with face-recognition capabilities and

dive-bomb unto humans meeting the face-recognition criteria. But again, this is an issue of human control *before* a weapon system is activated. This brings us to a further discussion of the limits of human control in the next section.

MEANINGFUL HUMAN CONTROL

As noted by Eggert (this Volume), a common claim is that AWS must remain under 'meaningful human control' (see also Ekelhof, 2019, and NATO STO, 2023, for a critical discussion). As noted by Ekelhof (2019, p. 344): "the concept of MHC or a similar concept, is one of the few things that states agree could be part of such a CCW outcome". And Vignard (this Volume) notes: "While there is not a common understanding of the phrase, nor is there international consensus agreement on its utility, there are groups of States who have embraced the concept as a principle to guide AI development in the military domain". Numerous states have explicitly declared their support for the idea that all weaponry should be subject to MHC (Crootof, 2016).

There are several conceptual distinctions to be made when defining 'meaningful human control'. The first is to ask: "control over what"? Do we mean informed human approval of each possible action of a weapon system ('human in the loop'), the ability to intervene if a weapon system fails or malfunctions ('human on the loop'), or do we mean control over the entire distributed targeting process (Cummings, 2019; Ekelhof, 2019)? Given the increased speed of modern warfare, the need for rapid self-defense in some situations, inherent human limitations in time-critical targeting scenarios, and the effects of high workload, fatigue, and stress on human decision-making, it is a far cry from 'meaningful human control' to put a human 'in' or even 'on the loop'. In this sense, it could be more in accordance with IHL to use a precision-strike Tomahawk guided missile that is not under direct operator control, but is under control by the programmers who enter targeting information or the hundreds of people involved in the targeting process. Obviously, there is no guarantee that programmers or others involved in the targeting process do not make any mistakes, but at least the likelihood of such mistakes is much smaller compared to "putting military operators into a crucible of time pressure, overwhelming volumes of information, and life and death decisions in the fog of war (…)" (Cummings, 2019, p. 24).

A second question to ask about MHC is: "control by whom"? This is usually interpreted as control exercised by an individual operator or commander. However, as argued by Ekelhof (2018; 2019) and Cummings (2019), targeting is frequently, though by no means always, a distributed process involving many people (exceptions may be found in urban warfare where individuals, assisted by robots, need to make split-second decisions). This does not make the line of accountability less clear, it just means we need to shift our attention away from the individual operator at the sharp end to the organization at the blunt end. At the very least we should make a distinction between what Cummings (2019) refers to as the 'strategic layer' and the 'design layer', and what I have referred to above as the distinction between 'mission autonomy' and 'platform autonomy'. Human control is exercised at the strategic layer when targets are designated in accordance with

mission objectives and abiding by principles of proportion and distinction within the Law of Armed Conflict framework. At the design layer, it depends on whether the target is static, well-mapped, and within the rules of engagement, whether human control can or needs to be exercised. When, in the minority of cases, targets meet these criteria, particular weapon systems may be used without human control at the design layer (e.g., the deployment of a Tomahawk missile against a particular building). When targets do not meet these criteria, and are dynamic and emerging, Cummings (2019, p. 25) argues that there needs to be human certification rather than human control. By this, she means that the use of weapon systems for such targets "should be proven through objective and rigorous testing, and should demonstrate an ability to perform better than humans would in similar circumstances, with safeguards against cybersecurity attacks".

A third question to ask is: "meaningful human control to what end"? Do we want to involve human beings in the decision-making process regarding the use of force? Do we want to ensure compliance with existing legal obligations? Do we want to establish a higher legal or ethical standard? Do we want to improve military effectiveness? As argued by Crootof (2016), MHC will usefully augment the humanitarian norms of proportionality and distinction. However, an overly strict interpretation of what constitutes MHC may actually undermine fundamental humanitarian norms governing targeting. For instance, if MHC is taken to mean that the human commander or operator has full contextual or SA of the target area and the means for the rapid suspension or abortion of the attack, then this would rule out the use of entire classes of precision-guided munitions or close-in defensive weapon systems. This may actually increase collateral damage and the killing of non-combatants. Crootof (2016) argues that the distinction, proportionality, and feasible precaution requirements should serve as an interpretive floor for a definition of MHC. This means that if we augment existing humanitarian norms governing targeting, for instance by strengthening the certification process as suggested by Cummings (2019), or by paying more attention to the distributed nature of the targeting process as suggested by Ekelhof (2019), then the notion of MHC, however imprecise, can fruitfully advance the conversation regarding the appropriate regulation of AWS.

Eggert (this Volume) challenges this widespread faith in MHC and discusses three main problems with AWS: the compliance problem, the dignity problem, and the responsibility problem. She argues that MHC does considerably less to address these problems than typically assumed. The compliance problem was already discussed above, in that legal compliance is not the same as moral compliance. Making AWS comply with moral principles would make them effectively not autonomous. Eggert raises the important question of whether control over autonomous weapons is not just an *apparent* but a *real* contradiction, meaning that we must ultimately choose between autonomy in machines and control in human hands. Can we argue for MHC while at the same time arguing for AWS? I think the distinction introduced above between platform autonomy and mission autonomy largely answers this question. We can both have platform autonomy while at the same time having mission autonomy as well: we have human control via mission autonomy *before* a weapon system is activated, while we have platform autonomy *within the constraints set by mission autonomy after* a weapon system is activated. Platform autonomy is never absolute and is always bounded by higher-order constraints. Advocates of MHC do not need to call for a ban on AWS, provided they make the

distinction between mission autonomy and platform autonomy and hence adopt the concept of bounded autonomy.

The second problem, the dignity problem, essentially states that MHC is no guarantee that human dignity is respected. Humans are capable of extreme cruelty and they violate human dignity all the time, according to Eggert. MHC does nothing to prevent this. Indeed, the moral constraints that we impose upon a physical platform may, in the hands of a dictator, have the opposite effect of what we intended with MHC. The dictator will have a different set of moral values and the AWS may be programmed in such a way as to kill many innocent civilians. There will be human control, but it will not be 'meaningful' in the sense in which democratic countries use this word. This argument does not imply, however, that democratic countries would have to abstain from MHC altogether. Quite the opposite, particularly given that MHC is a distributed process that avoids concentration of power in any one malevolent individual. MHC is thus a form of democratic control and this leads to questions on who should be involved and why. Even though Eggert claims that militaries are hardly democratic institutions, at least in some countries, they do act in accordance with democratically elected governments and can be held accountable (at least in democracies). Furthermore, there are currently efforts underway to establish Ethics Advisory Boards and processes to consult with all kinds of stakeholders. This makes MHC even more of a distributed process, as well as a more ethically inspired process.

Finally, the responsibility problem states that if humans are in control, they can be held accountable if something goes wrong. According to Eggert, the most promising sense in which MHC addresses concerns about responsibility is by allowing us to categorize harms wrongfully inflicted by AWS as *human omissions to prevent such harms*. These are cases in which humans omit to act; instances of allowing harm to be caused by an AWS which human agents should prevent. However, the kind of control necessary for ascribing responsibility may altogether rule out autonomy in weapon systems, as humans should always be able to intervene in order to prevent harms wrongfully inflicted by AWS. The conclusion should be that MHC is incompatible with AWS. Again, I think this conclusion is incorrect. First, as noted by Eggert herself, when time is of the essence, for instance in self-defense, humans cannot reasonably be asked to intervene. AWS, such as close-in weapon systems, are used against objects, not people, so the risk of harm wrongfully inflicted by an AWS is minimal anyway. Secondly, in footnote 8 Eggert states:

> Humans might program an AWS to target enemy combatants at t1, but it is still up to the machine, at t2, to select individual targets, and it may not always be clear why it targeted one person rather than another. The fact that a human person, at t1, programmed an AWS to behave a certain way doesn't mean that the AWS is under human control at t2.

If this would be the case, then such an AWS would not have been tested properly and it would not be trusted and accepted by commanders. If there is a disconnect between mission autonomy and platform autonomy, then something is wrong with the constraints put upon the platform (or weapon system) and this should have become apparent during the testing and evaluation phase (I will deal with some of the challenges with

Testing, Evaluation, Validation, Verification (TEVV) later). If the AWS displays signs of self-governance, setting its own goals, and behaving in ways it was not programmed to do, then this is a reason not to use the AWS.

THE RESPONSIBLE USE OF AI IN MILITARY SYSTEMS

So far, little discussion has been devoted to AI, and even less to the responsible use of AI in military systems. This is partly due to the fact that military systems are broader than weapon systems, but even when we restrict our discussion to weapon systems, we have seen that AI is not always necessary (e.g., close-in weapon systems and loitering munitions can fulfill their intended functions perfectly without AI). The notion of developing AI in a responsible fashion has not emerged primarily within the context of AWS. It was primarily driven by concerns over the use of AI in general commercial applications, and efforts to formulate policies and guidelines for the responsible use of AI have emerged first of all in the commercial sector. Nearly every respectable company these days has a responsible AI (RAI) framework, approach, or toolkit. It is beyond the scope of this chapter and this book to discuss these. Instead, I will be focusing on the responsible use of AI in (autonomous) weapon systems.

Any standard for RAI needs to cover the entire AI system lifecycle, from initial design to the decommissioning of the system. This is because decisions are made along the entire lifecycle that can and will impact later decisions. RAI may then be described as a dynamic approach to the design, development, deployment, and use of AI-enabled capabilities that ensure the safety of these capabilities and their ethical employment. By 'ethical employment', I mean conformance to AI principles and guidelines, such as bias mitigation, explainability, traceability, governability, and reliability. Exactly how the process of developing AI-enabled capabilities should conform to these principles is a matter of ongoing research.

In the following sections, three important challenges will be discussed: (i) testing, evaluation, validation, and verification; (ii) human–AI teaming; (iii) transparency and explainability. Although the challenges may seem to arise primarily from an AI perspective, there are important HFE aspects to these challenges as well. These will be explicitly discussed at the end of each section, particularly since these HFE aspects have received insufficient attention in the other chapters in this Volume.

Testing, Evaluation, Validation, and Verification

Reliability is an important ethical value in AI development. Reliability means that a system will provide the same output given the same input. With self-learning AI systems, that could be an issue, as these systems constantly improve themselves, in ways not always transparent to human users. But it is also an issue with AI systems that

encounter slightly different situations than they have been trained for (e.g., the camera angle is slightly different; the object is slightly different than the one encountered during training). In practice, this means that a lot of effort needs to go in the validation and verification process in all imaginable contexts. Hence, there should be more focus on TEVV rather than on regulation afterward. A comprehensive approach is crucial during the entire lifecycle: from design to maintenance, from training to doctrine development, and ethics (see Dunnmon, Goodman, Kirechu, Smith, & Van Deusen, 2021 for a first attempt at operationalizing ethical principles for AI within the lifecycle process). Algorithm auditing to ensure explainability, robustness, fairness, and privacy will become an important part of the technical lifecycle of an AI system (Koshiyama et al., 2022). Validating, verifying, and testing AI in all kinds of scenarios and use cases is very important because context is an important factor, also with regard to values such as responsibility and accountability.

All of this is easier said than done. Consider the following example. In a military conflict, a coalition commander has to decide whether or not to employ an AI-enabled system that was developed in a particular country within the coalition. The commander asks whether the system has been thoroughly tested and evaluated (ideally, of course, he or she should be aware of that, but in this case the system was developed in a different country), and receives a confirmatory response by AI experts. There can be two problems here. One is the classical issue that the conditions under which a system has been tested are not representative of the conditions under which the system will be employed. Surprise is continuous and ever-present. There is always the need to close the gap between the demonstration and the real thing (Woods, 2016). These are called 'AI blind spots', or conditions for which the system is not robust. The other issue is that the system may not be sufficiently explainable at this point, and under these time-pressed circumstances, for the commander to have sufficient trust in this system. Given that the commander is currently ultimately responsible and accountable for the use of this system on the battlefield, should the commander trust his or her experts and their V&V process? There is no answer to this question without considering the exact context in which the system will be deployed. If situations are not completely routine, the commander will have to make a decision based on fundamentally incomplete knowledge and taking into account principles of IHL. For this reason, an 'ethical risk analysis' should be developed integrally, together with the technical and human factors aspects of system development. In addition to Technology Readiness Levels, a Moral Organization AI Readiness Level (MORAL) could be developed. MORAL would describe how well AI has been evaluated and tested in terms of ethical risks. Yet it could also be argued that the burden of proof of performance and safety should fall on the shoulders of the industry as well as the military branches who buy their weapons (Cummings, 2019), rather than on the individual commander.

For the HFE community, there are important knowledge gaps and future developments to address the emerging challenges in the TEVV field. Validating, verifying, and testing AI in all kinds of scenarios and use cases leads to the questions "how representative are the scenarios and use cases" and "how many are sufficient"? Given that most AI applications will always work together with humans (see the Human–AI teaming section below), we need to ask whether the AI is intended to support only routine cases

or also 'edge cases' where the AI is likely to fail. Thus, HFE professionals need to look at the cognitive support objectives to understand the range of situations where the AI is intended to provide effective support and sample broadly from that range of situations (e.g., Roth et al., 2021). The 'how many are sufficient' question is a familiar one in the usability evaluation literature (e.g., Hwang & Salvendy, 2010), yet the answer may be quite different depending on the criticality of the device to be tested (Schmettow et al., 2013). Measuring and controlling the effectiveness of formative evaluation, usability testing, in particular, is crucial for risk reduction in the development of AI. The number of scenarios that need to be taken into account should not be set to a fixed (or even 'magical') number, but rather should be dependent on the nature of the risk involved when applying the AI (see also Panwar, this Volume). In general, the higher the risk, the larger and the more representative the scenarios to be included need to be. This 'late control strategy' (Schmettow et al., 2013) was developed to determine the number of users required to test a particular device. This strategy may be useful to determine the number of scenarios required as well.

A second emerging challenge for the HFE community lies in the role it may play in the determination of the Moral Organization AI Readiness Level (MORAL). Currently, it is difficult to identify relevant moral values (especially considering that they may change over time) and to ensure that the human–AI system continues to operate in accordance with these moral values and their context-dependencies. There is a role for the HFE community in applying requirements analysis to ethical, legal, and societal aspects of AI, as well as applying Value Sensitive Design methods (Friedman & Hendry, 2019). A necessary discussion is on who to involve with what responsibility to derive requirements from identified relevant moral values, ethics, and laws given the application that is considered. Methods are required that can facilitate this, on top of (existing) methods that shape the more general design process (see also Heijnen et al., this Volume). The IEEE 7000™-2021 standard is an important first step toward value-based engineering, as it is the only value-based engineering standard worldwide so far. The Value Lead, a new profession introduced by IEEE 7000, is trained in ethical and value theories, yet also fits into the system engineering process. HFE professionals are well-suited for fulfilling this profession of Value Lead, as they are used to working together with both end users and system engineers.

Human–AI Teaming

A second HFE challenge that has received a lot of attention recently is 'human–AI teaming', 'human-autonomy teaming', or 'human-machine teaming'. A comprehensive state-of-the-art report on human–AI teaming, including research needs, was published by the National Academies of Sciences, Engineering and Medicine (NAS) in December 2021 (National Academies of Sciences, 2021). Several journals have devoted special issues to this topic, reflecting the burgeoning field. However, neither the NAS report nor recent empirical work in this field is concerned with the moral, ethical, or legal aspects of military decision-making using AI (a notable exception is the NATO STO, 2023, report).

Embedding AI in a Human–AI Team, taken in its broadest organizational context, is an essential part of achieving responsible military AI, both for ethical and legal

reasons and to realize the 'multiplier effect' that comes from combining human cognition and inventiveness with machine-speed analytical capabilities. Research gaps are in what the human needs to know about the AI, but also what the AI needs to know about the world and the human. What human needs to know about AI has implications for the training and education of military personnel, and also for human–AI interfaces (see the section on display transparency below). What the AI needs to know about the world and the human has implications for real-time model updating as well as for human enhancement questions (e.g., operator state monitoring). Yet, the responsible use of AI in military systems goes beyond these well-known HFE challenges. An AI system could critique a human's moral reasoning in terms of presenting the moral acceptability of what-if scenarios. Research gaps are in what information is needed for moral SA and how to augment and support the commander's moral model. From an HFE perspective, the use of virtual or augmented reality technology could have the potential to improve moral SA. Sushereba et al. (2021) developed a framework for using augmented reality to train sensemaking skills in combat medics and civilian emergency management personnel. The four key elements of sensemaking that they list – perceptual skills, assessment skills, mental models, and generating/evaluating hypotheses – also appear relevant for training and supporting a commander's moral model.

A second venue for future research lies in the area of trust repair in human-autonomy teams. As artificial teammates may increasingly behave like human teammates, the question arises how human teammates respond when an artificial teammate violates their trust. Trust violations are an inevitable aspect of the cycle of trust and since repairing damaged trust proves to be more difficult than building trust initially, effective trust repair strategies are needed to ensure durable and successful team performance (Kox et al., 2021). A trust repair strategy could be an expression of regret that accompanies the apology, providing promises or explanations, or delaying the repair strategy until the next trust opportunity. Research shows that a single unethical behavior immediately worsened participants' perceptions of an autonomous teammate and that apologies and denials following unethical behaviors were insufficient in rebuilding trust in a military-based human–AI teaming context (Textor et al., 2022). Other research has shown that the intelligent agent was the most effective in its attempt of rebuilding trust when it provided an apology that was both affective and informational (Kox et al., 2021). Hence, future research should evaluate the efficacy of alternative trust repair strategies.

Transparency and Explainability

Transparency (also referred to as observability, and sometimes as traceability) can refer to many objects and processes, for instance, data, algorithms, decisions, or organizations. In the context of this chapter, I follow the definition used in the NAS report (2021, p. 33): "(…) 'the understandability and predictability of the system' (Endsley, Bolte, and Jones, 2003, p. 146), including the AI system's 'abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process' (Chen et al., 2014, p. 2)". Display transparency provides a real-time understanding of the actions of the AI system, whereas explainability provides information in a backward-looking manner. AI explainability is the "ability to provide satisfactory,

accurate, and efficient explanations of the results of an AI system" (National Academies of Sciences, 2021, p. 38).

Transparency is important because AI systems learn and change over time and may be applied in contexts and situations they were not initially trained for. In order for humans to keep up with these changes and maintain adequate situation awareness, system changes need to be presented in a transparent manner. Explainability is important due to the black-box nature of machine learning AI. One kind of explainability is post-hoc explanations that make a non-interpretable model understandable after an action has been executed. This is considered more active than it is in transparency because the system only gets understandable with the provided explanation (Arrieta et al., 2020).

As far as empirical evidence is concerned for the effects of transparency and explainability, a recent systematic literature review covering 17 experimental studies found a promising effect of automation transparency on situation awareness and operator performance, without the cost of added mental workload (Van de Merwe, Mallam, & Nazir, 2024). According to Endsley (2023), SA is best supported by display transparency that is current and prospective, whereas explainable AI is primarily retrospective and directed at building mental models. In a study directly comparing the effects of transparency and explainability on trust, situation awareness, and satisfaction in the context of an automated car, Schraagen et al. (2021) showed that transparency resulted in higher trust, higher satisfaction, and higher level 2 SA than explainability. Transparency also resulted in a higher level 2 SA than the combined (transparency+explainability) condition, but did not differ in terms of trust or satisfaction.

These results are promising and show how abstract ethical principles such as transparency and explainability can be operationalized in AI systems. According to Endsley (2023), future research is needed to design effective AI transparency techniques and to demonstrate that needed SA and trust at manageable workload levels are achieved in the real-world conditions where these systems will be used. It is particularly important to keep in mind the time-constrained conditions under which most military commanders operate. Care must be taken not to overload commanders with information. In some cases, there just is not enough time for a lengthy explanation. To safeguard these situations, we need to focus more on the development process up-front and make sure that things go right there. AI should be able to explain itself during this development process, not merely to system designers but also to end users. And once end users have obtained sufficient trust in these systems, they may be fielded in operational contexts, with high time pressure and high stakes. And even then, the military leader should always be aware of contextual limitations.

CONCLUSIONS

The responsible use of AI in military systems is a relatively recent development. Operationalizing ethical principles in the design and development process is an ongoing challenge. There are also many research challenges, not only in the software engineering

field but also in the human factors engineering field. The discussion on the interconnections between AI, AWS, and MHC is a complicated and politically charged one. This has sometimes led to oversimplifications and a tendency to reduce the inherent complexity and ambiguity of the subject matter.

The central thesis in this chapter has been that there is no such thing as absolute autonomy, only various gradations of 'bounded' autonomy. Autonomy is bounded by the mission control, or strategic control, exercised by humans over the platforms or weapon systems that they develop. This type of mission control should lead to extensive testing and evaluation, in order to verify that the systems being developed do what they are supposed to do. No military commander would want this otherwise, or else they would not trust the systems they are in control of. We should therefore shift some of our attention from the 'sharp end' of weapon system impact or use, to 'blunt end' weapon system design and development, without neglecting the human-machine interaction that is crucial for the soldier on the ground to interact effectively with AI. This is not to say that this is easy or without its own challenges. There are still, and for the foreseeable future, huge challenges in the certification process, the requirement for AI to be able to explain itself, certainly during the development process, and the way humans and AI need to be able to work together. Over the past couple of years, research has made considerable progress in these areas, yet, there is still a lot to learn.

At the same time, current developments on the battlefield impose their own dynamics on the weapon systems being developed. Systems are being deployed that use AI to recognize targets. Some of these systems are said to operate in 'autonomous mode'. It is my hope that the reader of this book will be able to look critically at these developments and discussions. While the term 'autonomy' is often used loosely, such usage does conjure up images of 'killer robots' that are out on a killing spree on innocent citizens. This is not what modern warfare, with its highly distributed and deliberate targeting process is about. In democratic societies, we should hold on to accountability and conformance with IHL. The distinction between mission autonomy and platform autonomy makes it clear that weapon systems need to conform to ethical and legal standards, and that MHC is a way to impose those standards on the design, development, and deployment process (rather than on 'controlling' the weapon systems after activation). The fact that terrorists, rogue nations, or even democratic states, may use MHC for their own, undemocratic, authoritarian, or immoral ends, and may deploy weapon systems to such ends, violating humanitarian norms of proportionality and distinction, should not dissuade us from using AI responsibly in military systems.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779.

Brown, B., & Laurier, E. (2017). The trouble with autopilots: Assisted and autonomous driving on the social road. In *Proceedings of the CHI conference on human factors in computing systems* (pp. 416–429). Denver, CO.

Chafkin, M. (October 6, 2022). *Even after $100 billion, self-driving cars are going nowhere*. Bloomberg. Retrieved September 26, 2023.

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency*. Aberdeen Proving Ground, MD: Army Research Laboratory. https://apps.dtic.mil/sti/pdfs/ADA600351.pdf

Crootof, R. (2016). A meaningful floor for 'meaningful human control'. *Temple International and Comparative Law Journal, 30*(1), 53–62.

Cummings, M. L. (2019). Lethal autonomous weapons: Meaningful human control or meaningful human certification? *IEEE Technology and Society Magazine, 38*(4), 20–26.

Dunnmon, J., Goodman, B., Kirechu, P., Smith, C., & Van Deusen, A. (2021). *Responsible AI guidelines in practice: Lessons learned from the DIU portfolio*. Washington, DC: Defense Innovation Unit.

Ekelhof, M. A. C. (2018). Lifting the fog of targeting: "autonomous weapons" and human control through the lens of military targeting. *Naval War College Review, 71*(3), 61–94.

Ekelhof, M. A. C. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343–348.

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors, 59*(1), 5–27.

Endsley, M. R. (2023). Supporting human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior, 140*. https://doi.org/10.1016/j.chb.2022.107574

Endsley, M. R., Bolte, B., & Jones, D. G. (2003). *Designing for situation awareness: An approach to human-centered design*. London: Taylor and Francis.

Feltovich, P. J., Hoffman, R. R., Woods, D. D., & Roesler, A. (2004). Keeping it too simple: How the reductive tendency affects cognitive engineering. *IEEE Intelligent Systems, 19*(3), 90–94.

Friedman, B., & Hendry, D. G., (2019). *Value sensitive design: Shaping technology with moral imagination*. MIT Press.

Griffith, E. (April 15, 2023). The end of faking it in Silicon Valley. *The New York Times*. Retrieved November 1, 2023.

Hambling, D. (2023). Ukrainian AI attack drones may be killing without human oversight. *New Scientist*. Retrieved October 16, 2023.

Hwang, W., & Salvendy, G. (2010). Number of people required for usability evaluation: The 10+/-2 rule. *Communications of the ACM, 53*(5), 130–133.

Kaber, D. B. (2018). A conceptual framework of autonomous and automated agents. *Theoretical Issues in Ergonomics Science, 19*(4), 406–430.

Koshiyama, A., Kazim, E., & Treleaven, P. (2022). Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. *Computer, 55*(4), 40–50.

Kox, E. S., Kerstholt, J. H., Hueting, T. F., & De Vries, P. W. (2021). Trust repair in human-agent teams: The effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems, 35*(2), 1–20.

Leveson, N. G. (2011). *Engineering a safer world: Systems thinking applied to safety*. Cambridge, MA: The MIT Press.

Metz, C. (February 1, 2023). Self-driving car services want to expand in San Francisco despite recent hiccups. *The New York Times* (nytimes.com). Retrieved September 26, 2023.

Nathanael, D., & Papakostopoulos, V. (2023). Three player interactions in urban settings: design challenges for autonomous vehicles. *Journal of Cognitive Engineering and Decision Making, 17*(3), 236–255. https://doi.org/10.1177/15553434231155032

National Academies of Sciences, Engineering, and Medicine (2021). *Human-AI teaming: State of the art and research needs.* Washington, DC: The National Academies Press.

NATOTerm (2023). NATOTerm: The Official NATO Terminology Database. NATOTermOTAN - Home

NATO Science and Technology Organization (STO) (2023). Meaningful human control of AI-based systems workshop: Technical evaluation report, thematic perspectives and associated scenarios. In *STO Meeting Proceedings* (MP-HFM-322). Neuilly-sur-Seine Cedex: NATO STO.

Philipps, D. (April 15, 2022). The unseen scars of those who kill via remote control. *The New York Times* (nytimes.com)

Renner, L., & Johansson, B. (2006). Driver coordination in complex traffic environment. In *Proceedings of the 13th European conference on cognitive ergonomics (ECCE 2006): Trust and control in complex socio-technical system* (pp. 35–40). Zurich, Switzerland.

Roth, E. M., Bisantz, A. M., Wang, X., Kim, T., & Hettinger, A. Z. (2021). A work-centered approach to system user-evaluation. *Journal of Cognitive Engineering and Decision Making, 15*(4), 155–174.

Russell, S. (2022). Banning lethal autonomous weapons: An education. *Issues in Science and Technology, 38*(3), 60–65.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed.) (pp. 1926–1943). New York: Wiley.

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war.* New York: W.W. Norton & Company.

Scharre, P., & Horowitz, M. C. (2015). *An introduction to autonomy in weapon systems*. Center for a New American Security Working Paper.

Schmettow, M., Vos, W., & Schraagen, J.M. (2013). With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems. *Journal of Biomedical Informatics, 46*, 626–641.

Schraagen, J. M. C., Kerwien Lopez, S., Schneider, C., Schneider, V., Tonjes, S., & Wiechmann, E. (2021). The role of transparency and explainability in automated systems. In *Proceedings of the 2021 HFES 65th international annual meeting* (pp. 27–31). Santa Monica, CA: Human Factors and Ergonomics Society.

Simon, H. A. (1947). *Administrative behavior.* New York: Palgrave Macmillan.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Simon, H. A. (1957). *Models of man: Social and rational*. New York: John Wiley.

Simon, H. A. (1990). Bounded rationality. In J. Eatwell, M. Milgate, & P. Newman, (Eds). *Utility and probability* (pp. 15–18). London: Palgrave Macmillan.

Simon, H. A. (2000). Bounded rationality in social science: Today and tomorrow. *Mind & Society, 1*(1), 25–39.

Sipe, N. (March 23, 2023). We were told we'd be riding in self-driving cars by now. What happened to the promised revolution? (theconversation.com). Retrieved September 26, 2023.

Slaughterbots (2017). Future of Life Institute. Retrieved October 13, 2023, from https://youtu.be/9CO6M2HsoIA?feature=shared

Sushereba, C. E., Militello, L. G., Wolf, S., & Patterson, E. S. (2021). Use of augmented reality to train sensemaking in high-stakes medical environments. *Journal of Cognitive Engineering and Decision Making, 15*(2–3), 55–65.

Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons systems. *Science and Engineering Ethics, 28*, 37–59.

Textor, C., Zhang, R., Lopez, J., Schelble, B. G., McNeese, N. J., Freeman, G., Pak, R., Tossell, C., & de Visser, E. J. (2022). Exploring the relationship between ethics and trust in human-artificial intelligence teaming: A mixed methods approach. *Journal of Cognitive Engineering and Decision Making, 16*(4), 252–281.

United Nations Security Council (2021). Letter dated 8 March 2021 from the Panel of Experts on Libya established pursuant to resolution 1973 (2011) addressed to the President of the Security Council. N2103772.pdf (un.org). Retrieved October 13, 2023.

U.S. Department of Defense (2023). DoD Directive 3000.09. Autonomy in weapon systems. Washington, DC: Department of Defense. DoD Directive 3000.09, "Autonomy in Weapon Systems," January 25, 2023 (whs.mil).

Van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review. *Human Factors*, *66*(1), 180–208

Woods, D. D. (2016). The risks of autonomy: Doyle's catch. *Journal of Cognitive Engineering and Decision Making, 10*(2), 131–133.