MEHUCO

Autonomous weapon systems: Conditions and limits of meaningful human control

With funding from the:

Federal Ministry
of Research, Technology
and Space

MEHUCO

Contents

Autonomous weapon systems: Conditions and limits of meaningful human control

Autonomous weapon systems (AWS) are a major challenge for international security and peace. The term "autonomous" refers to the high degree of independence these systems possess. Drones, tanks or submarines navigate in their environments; the identification, selection and attack of targets can be performed without humans. These systems are not just in development; they are widely used by militaries around the globe.

Proponents of AWS associate these technologies with the promise of being able to carry out military operations faster (than the enemy) without unnecessarily endangering soldiers' lives. They also claim that AI-enabled weapons and targeting systems deliver higher precision and reduce the risks for civilians. Critics, on the other hand, emphasise that a higher degree of autonomy will lead to a loss of human control. Policymakers, NGOs and researchers have paired this criticism with calls for regulation and effective, "meaningful" human control.

Organised in a research network titled "Meaningful Human Control – Autonomous Weapon Systems between Regulation and Reflection" (MEHUCO), projects at five research institutions in Germany are investigating the conditions necessary to ensure meaningful control of AWS as well as the consequences of machine autonomy for questions of legal and moral responsibility. The research network aims to contribute to political and public debates, offering a realistic understanding of the possibilities and limitations of AWS and their regulation.

This dossier aims to shed light on the complex challenges regarding international AWS regulation, the question of accountability and the social and political consequences that arise from these new types of technology. First, it introduces the three core concepts: autonomy, human control and meaningfulness. It then highlights answers to these questions reflecting the various disciplinary perspectives represented in the research network. Each of the five participating projects presents their unique research approach, coming from different academic backgrounds rooted in computer science, law, media studies, science & technology studies and sociology.

1. What does it mean that a (weapon) system is autonomous?

From a technical perspective, an autonomous computer system is often simply defined as one that operates independently of human input and can solve a predefined task. Autonomy is generally understood as a development that technologically proceeds from automation: While *automatic* systems react to a simple threshold value (for example, at a defined temperature the heating switches on), *automated* systems are more complex but still follow determined steps. *Autonomous* systems, however, are capable of adapting in ways that have not been preprogrammed in detail. Humans still define a goal (for example, a destination on a map), but the path to get there (the quickest route) is "found" by the system itself.

In military contexts, this functionality enables technical systems to navigate complex environments, perform tasks or decide on particular courses of action. More specifically, AWS are defined as being able to select and apply force to (or, euphemistically, "engage") targets without human intervention. Military models such as the "kill chain" describe the individual steps necessary for detecting, tracking, and destroying a target. In some cases, it has already become common practice to delegate all of these tasks to AI-based systems: many of the drones used in Russia's war on Ukraine can identify and destroy objects without human intervention. *AWS are a military reality.*

Despite this reality, prevailing expectations often remain disconnected from the actual operational capabilities of such systems. AWS are not rogue systems that are fully independent of humans. Technical autonomy is relational, which means that these systems depend on external factors such as infrastructure (energy supply, communication networks, data), maintenance or programming done by humans. No system is ever fully autonomous; developers distinguish between two scales, with varying degrees of self-directedness and self-sufficiency. These degrees, in turn, open up pathways for efficient and "appropriate" control, human intervention and regulation.

2. How are autonomous weapon systems designed to enable human control?

Human control can be broadly understood as the capacity of human actors to influence the autonomous processes of an AWS and the outcome of the decision. In order to determine more clearly what specific role humans should play, human control is generally modelled as taking two forms: either humans act as the central decision-maker ("human-in-the-loop") or as a supervisory authority ("human-on-the-loop.").

In a "human-in-the-loop" model, human control can range from analysing potential targets to merely deciding on weapon release. This could be implemented by humans specifying a particular target for the weapon system or being presented with a list of suggested targets from which they then select the desired target. In this scenario, the weapon system merely performs a supporting and subordinate role.

If a weapon system is granted a higher degree of autonomy, humans could assume the role of "human-on-the-loop". In this scenario, the weapon system independently processes the individual tasks associated with decision-making (such as tracking, targeting or engaging) and submits the selected target to humans for final review. This review also comes with different degrees of human involvement. The operator may be required to actively authorise the attack ("pressing the kill button"). Where there is a lesser degree of human involvement, human operators may only be granted the option to veto a machine-generated proposal ("pressing the cancel button"). In such cases, the weapon system would have the technical capability to eliminate a target independently without requiring explicit permission.

The scenarios described illustrate that human control decreases with increasing automation, and thus humans have less influence on the outcome of decisions. However, the specified degree of human control says little about how effective such human control actually is. Therefore, the human control in question has to be of a certain minimum quality level to ensure meaningful and responsible decision-making. This means that control must not only be present but also effective and substantial. The "meaningful human control" requirement is intended to solve this problem.

3. What makes human control “meaningful”?

The term “meaningful human control” originates in civil society interventions undertaken against the unrestricted use of weapons by militaries. It gained major relevance in debates on how to regulate military AI innovation, including AWS, and was later appropriated in military discourses, not least to showcase responsible and appropriate use to the general public.

The term, therefore, is not only vague in nature but also contested and widely politicised. To grasp the range of meanings that “meaningful” can adopt in different discourses, it is helpful to utilise the multiple perspectives represented in the research network.

The different disciplines may emphasise computer system architecture, interface design, organisational norms, legal frameworks and human-machine interactions. The higher the degree of human authority over technical and institutional choices, the more meaningful it may be.

From the perspective of science & technology studies, meaningful human control is a highly demanding concept. It presupposes the possibility of a kind of human-machine interaction in which the human operator has a high degree of independence in terms of his situational awareness and decision-making. Meaningful control is difficult to achieve, particularly given the complexity of AWS and the unpredictability of their behaviour.

From a sociological perspective, the focus is on the sociocultural aspects of AWS, such as human systems of meaning, social practices, and cultural notions of accountability, agency and control. The emphasis is on the interactive relationships between humans, algorithms and weapon systems in knowledge generation and decision-making processes. This approach considers how actors understand the term “meaningful” in the context of AWS and military decisions. The design, production and use of AWS are shaped by notions of law and ethics, as well as by concepts of enemies and future wars.

Media studies approaches analyse the forms, relations and contents that media create and enable. These interfacing effects fundamentally shape the relations between humans and machines, which include interactions between operators and weapon systems. The design of interfaces, such as a screen or a controller, frames human perceptions and determines the ways humans can intervene – a major factor in making control effective and meaningful.

In computer science, meaningful control typically is addressed through system architectures, algorithmic transparency and real-time feedback. Ideally, interfaces should ensure that operators receive clear information and can intervene decisively. A secure recording of all activities enables human oversight over technological decision-making.

From a criminal law perspective, meaningful human control must be analysed using a human-centred and case-by-case approach in order to ensure a fair attribution of responsibility. This could be achieved by employing a criteria-based method that includes the following factors: the operator's technical and situational awareness, adequate time for evaluating machine suggestions and proper training. AWS should also be reliable and predictable so that the operator can trust the machine's suggestions.

Researching autonomous weapon systems

Agency in complex human-machine interaction and the limits of human control

Science & Technology Studies, Paderborn University

Research into interactions between humans and machines in different social contexts has established that neither humans nor machines possess autonomy as a given attribute. Rather, in these contexts, a capacity for action ("agency") arises only situationally and under the conditions set by either the machine or the human. The agency of humans and machines is hence the temporary result of mutual enabling and framing. The claim that a complex machine acts autonomously obscures, among other things, the necessity of design decisions, construction processes and appropriation strategies by users. Conversely, the assumption that a human being can act autonomously when using or operating a complex machine obscures the fact that their specific agency is only made possible by technology and is at the same time determined by its material conditions.

Furthermore, in the age of opaque, self-learning algorithms, it is unclear how this machine influence on human agency can be made transparent and understandable for users and affected third parties, and how human responsibility and accountability can be conceived against this backdrop. The rapid technical development of AI-based weapon systems thus undermines the concept of autonomy that underlies both the term "AWS" and that of "meaningful human control".

Any approach to regulating AWS has to account for these limitations. There are numerous approaches that attempt to maintain human control over the use of complex weapon systems. For instance, there is an emerging international trend towards transferring the concept of responsible, human-centred, trustworthy and ethical AI that originated in the regulation of civilian AI applications to the military sector. However, none of these approaches have been able to do justice to the complex processes and effects of human-machine interaction.

Understanding how AI systems interpret human behaviour in military decision-making

Sociology, University of Hamburg

The autonomy of a weapon system is usually achieved through algorithms, primarily through various forms of artificial intelligence. Although advanced artificial intelligence systems can pass basic intelligence tests, they fundamentally lack human capacities such as critical judgement and moral understanding. This raises serious concerns in the context of military operations where lives are at stake. AI systems work by drawing connections between different data points and generating patterns based on large amounts of data in order to predict human behaviour and military operations. While they can make suggestions and decisions based on probabilities, they do not truly understand the complexity of human actions or the moral and legal implications of military decisions.

In an attempt to address these limitations, developers are trying to program human values, such as dignity and respect for life, into these systems. Human operators are also required to maintain control over critical decisions, particularly when selecting targets. Furthermore, these systems receive feedback from various stakeholders to improve their decision-making capabilities over time.

Despite these safeguards, significant problems remain. Research has revealed accountability gaps – when something goes wrong, it is often unclear who is legally responsible. The programmers who create these systems have their own biases, which affect how AI identifies legitimate targets, interprets enemy behaviour and determines who is considered an enemy. Furthermore, the data used to train these systems can produce distortions that lead to flawed conclusions.

These challenges highlight the difficulty of converting complex human values and legal principles into computer code, particularly in situations where upholding international humanitarian law and human rights standards is indispensable.

How do humans and AI interact? Weapons, human operators and the role of interfaces

Media Studies, University of Bonn

As AI has advanced, the autonomy of weapons has also risen steadily. At the same time, this makes it increasingly important to regulate or curtail the capabilities and forms of use of AWS. The strongly voiced need for "meaningful human control" or "appropriate levels of judgment and care" are intended to achieve this – or are meant to appease public concerns about these weapons. Conceptually, these considerations often start with the technical functionality of these systems: what can and should AI be able to do independently? Irrespective of these efforts, the meaning of attributes such as "meaningful" or "appropriate" often remains vague.

Against this background, the MEHUCO research project at the University of Bonn addresses these questions with a relational rather than normative approach. It focuses on the complex relationship between humans and machines, which goes far beyond the simple idea of "humans controlling AI weapons" implied by the "human in/on the loop" notion. Military decisions, such as the selection or engagement of targets, are not made before the weapon is deployed but in concordance with the AI system – there is a "teaming" of humans and machines, which collectively constitute a system and cannot be clearly distinguished as separate elements. The project therefore specifically investigates the interaction of humans with AI and the role that interfaces (for example, controllers, screens or sensors) play in this.

To this end, the project analyses technical solutions, research and development scenarios and publicly available materials on the operational use of weapons. At the same time, imaginations on AI-based warfare are examined in order to gain insight into possible future technological developments and military scenarios.

What role do humans play in autonomous system behaviour?

Computer Science, Ostfalia University of Applied Sciences

AWS are designed and programmed to "sense" task-specific information and to execute the task. Systems, when fully autonomous, are developed and implemented to handle all this without human intervention after deployment. Errors may occur due to various factors, for example, inadequate design, inaccurate representations of the environment or misleading predictions. This can have different impacts on the system. Various organisations such as the North Atlantic Treaty Organization (NATO), the International Organization for Standardization (ISO) or the Institute of Electrical and Electronics Engineers (IEEE), have developed standards for the design, development, testing, production and usage of systems. These standards are meant to prevent undesired effects.

The interface that is used to exercise control over the system must be designed to ensure that the human operator's intended course of action is clearly communicated to the system. It should also be intuitive to use and easy for the operator to understand. When interfaces are flawed or unsuitable, this leads to a higher risk of miscommunication of the intention.

AWS should be designed with meaningful human control in mind, so that a given decision is based on a sufficient understanding of the situation at hand and the system's output. This understanding, in turn, guides the decision-making process. For example, targets should be presented with a threshold value that is assigned to them so that the human operator is able to make an informed decision that reflects the probability that the system's output is correct.

Why the human-in-the-loop is not enough – the necessity of meaningful human control from a criminal law perspective

Law, University of Hannover

The increase in the autonomy of weapon systems raises many legal questions and concerns. For example, there is controversy over whether the fundamental principles of international humanitarian law are being upheld. These include the principle of distinction, which states that civilians must be protected as far as possible. In order for AWS to be allowed to eliminate targets independently, i.e. without human verification, they must be able to reliably distinguish which targets are permissible. It is difficult to determine with certainty what constitutes a civilian – as opposed to a member of the armed forces. Human control over AWS therefore remains necessary in order to uphold international law.

Furthermore, it is important to clarify who should be held criminally liable if such (partially) autonomous weapon systems eliminate the (de facto) wrong targets, in violation of international criminal law. In any event, only a human can be held criminally liable. To avoid accountability gaps, one could therefore consider always involving a "human-in-the-loop". However, there is a risk that the person will rely too heavily on the machine's suggestion or have too little information about the situation as well as a lack of time and training. This could lead to a situation where the person appears to be in control but in reality has no real opportunity to review the machine's suggestion. The person would then ultimately be a mere scapegoat.

Based on these considerations, the "human-in-the-loop" should be supplemented by a qualitative component of control: meaningful human control. The person interacting with the partially autonomous weapon system should be empowered to take control in order to bear accountability in an appropriate manner.

A future with autonomous weapon systems?

When discussing AWS, it is never sufficient to solely focus on a system's functionality; indeed, it is essential to consider the relations between machines and humans. As clearly highlighted by the projects of the research network, neither humans nor machines are fully autonomous entities: AWS strongly depend on external factors, such as infrastructure, maintenance and specific human programming. Human decision-making, on the other hand, is shaped by factors such as technical design or the specific requirements of the context of use. These interdependencies make it difficult to clearly assign accountability or responsibility, especially since AI systems' decision-making remains opaque.

From both an ethical and a legal perspective, it is therefore necessary to replace the idea of merely including the "human-in-the-loop" with the concept of a genuinely *meaningful* human control. This requires humans to have the knowledge, time and technical capability needed to understand and question machine decisions and to override them, if necessary. The principles of international law, particularly the protection of civilians, can only be upheld if this qualitative dimension of control is taken seriously. Responsible regulation of AWS must recognise the complex human-machine relationship and ensure that legal and ethical considerations are integral to technological development and political decision-making. If this ideal remains out of reach, there will be no alternative but to ban these systems.