Article36

BACKGROUND PAPER | APRIL 2016

Key elements of meaningful human control

Background paper to comments prepared by Richard Moyes, Managing Partner, Article 36, for the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)

Geneva, 11-15 April 2016

This paper draws on thinking around 'meaningful human control' developed in collaboration with Dr. Heather Roff in the context of a grant awarded to Arizona State University in partnership with Article 36 by the Future of Life Institute (www.futureoflife.org)

Article 36 is a UK-based not-for-profit organisation working to promote public scrutiny over the development and use of weapons.
www.article36.org
info@article36.org
@Article36

Article 36 is a founding member of the Campaign to Stop Killer Robots.
www.stopkillerrobots.org

Overview

The central area of concern regarding the development of autonomous weapons systems (AWS) is that they might lack the necessary human control in the critical functions of identifying, selecting and applying force to targets. Without the necessary human control, such systems might not allow the proper application of legal rules, or might produce interpretations of the legal framework that erode civilian protection, or lead to other negative outcomes relating to the morality of human interactions or the maintenance of peace and stability.

In this context, this paper argues that:

- Consideration of the form and nature of human control considered necessary is the most useful starting point for discussions on this issue.
- The existing legal framework of international humanitarian law provides a framework that should be understood as requiring human judgment and control over individual "attacks" as a unit of legal management and tactical action.
- That without recognizing a requirement for human control to be in some way substantial or meaningful, the existing legal framework does not ensure that human legal judgment will not be diluted to the point of being meaningless, as a result of the concept of "an attack" being construed more and more broadly.
- Against that background, delineation of the key elements of human control should be the primary focus of work by the international community.
- Towards such a process, the following key elements can be proposed:
  - Predictable, reliable and transparent technology.
  - Accurate information for the user on the outcome sought, the technology, and the context of use.
  - Timely human judgement and action, and a potential for timely intervention.
  - Accountability to a certain standard

- Whilst consideration of these key elements does not provide immediate answers regarding the form of control that should be considered sufficient or necessary, it provides a framework within which certain normative understandings should start to be articulated, which is vital to an effective response to the challenge posed by autonomous weapons systems.
- An approach to working definitions based on understanding 'lethal autonomous weapons systems' as weapons systems operating with elements of autonomy and without the necessary forms of human control would be the most straightforward way to structure discussion in a productive normative direction.

Introduction

"Meaningful human control over individual attacks" is a form of words that was coined by the NGO Article 36, to express the core element that is challenged by the movement towards greater autonomy in weapons systems. It is a policy formulation that has been picked up and used in different ways by different actors – in publications by various individuals and organisations, in state interventions at the UN Convention on Certain Conventional Weapons (CCW), in the open

letter from Artificial Intelligence practitioners organized by the Future of Life Institute. As used by Article 36 it has always been presented as an approach for structuring a productive debate rather then as providing a conclusion to that debate.

Asserting a need for meaningful human control is based on the idea that concerns regarding growing autonomy are rooted in the human aspect that autonomy removes, and therefore describing that human element is a necessary starting point if we are to evaluate whether current or future technologies challenge that. This is particularly important if a coherent policy conversation is to be had about diverse and often hypothetical future technologies. It is also a starting point for policy that is arguably more open to engagement from diverse stakeholders that might have different expectations of the advantages that may be afforded to them by future developments in autonomous weapons systems.

Considering the key elements necessary for human control to be meaningful does not preclude consideration of other more specific issues – but a structured analysis tends to find that those more specific issues fall within the key elements of human control. For example need for 'predictable' technology, the need for human 'judgment' to be applied in the use of force and the need for accountability all fall under the key elements of human control as laid out later in this paper. Furthermore, without a normative requirement regarding human control the legal framework itself is open to divergent and progressively broader interpretations that may render human legal application meaningless.

Recognizing the need for human control in some form

At its most basic level, the requirement for meaningful human control develops from two premises:

1. That a **machine** applying force and operating without any human control whatsoever is broadly considered unacceptable.
2. That a **human** simply pressing a 'fire' button in response to indications from a computer, without cognitive clarity or awareness, is not sufficient to be considered 'human control' in a substantive sense.

On this basis, some human control is required and it must be in some way substantial – we use the term 'meaningful' to express that threshold. From both of these premises, questions relating to what is required for human control to be 'meaningful' are open. Given that openness, meaningful human control represents a space for discussion and negotiation. The word 'meaningful' functions primarily as an indicator that the form or nature of human control necessary requires further definition in policy discourse.

Critical responses to this policy formulation tend to fixate on the term 'meaningful' because it is undefined or might be argued to be vague – responses that may also be motivated by state representative anxieties at policy formulations not initiated by states. Such responses, however, miss the point. There are other words that could be used instead of 'meaningful', for example: appropriate, effective, sufficient, necessary. Any one of these terms leaves open the same key question: how will the international community delineate the key elements of human control needed to meet these criteria? Any one of these would also be vague until the necessary form of human control is further defined, giving the chosen adjective some further calibration.

The term 'meaningful' can be argued to be preferable because it is broad, it is general rather than context specific (e.g. appropriate), derives from an overarching principle rather being outcome driven (e.g. effective, sufficient), and it implies human meaning rather than something administrative, technical or bureaucratic.

That said, fixating on which adjective is most appropriate should not stand as a barrier to the next step required of the international community, which is to begin to delineate the elements of human control that should be considered necessary in the use of force.

Situating human control in the legal framework

Article 36 has called on states, in the context of discussions on autonomous weapons systems in armed conflict, to recognise the need for 'meaningful human control over individual attacks.' In its use of the term 'attacks', this formulation situates the issue of human control within the legal framework of international humanitarian law (IHL).

It is important to recognize that IHL is not the only legal framework relevant to AWS, nor are legal frameworks the only basis for assessing whether further development of such technologies is appropriate or advisable. However, the relationship between human control, AWS and IHL are given particular focus in this paper.

Human beings as addressees of the law

When discussing AWS, however complex, Article 36 orientates to these systems as 'machines'. Discussion on this issue is prone to a slippage towards treating these machines as 'agents' and in particular as 'legal agents'. It is common for diplomats and 'experts' to refer to concerns about whether AWS will 'be able to apply legal rules', or 'to follow the law'. Machines don't apply legal rules. They may undertake functions that are in some ways analogous to the legal rules (for example being programmed to apply force to certain heat patterns common to armoured fighting vehicles) but in doing so they are not 'applying the law' – they are simply implementing a process that a human commander anticipates in their assessment of the legality of a planned attack. Prof. Marco Sassoli in his presentation to the 2014 ICRC expert meeting on autonomous weapons stated that, "only human beings are addressees of international humanitarian law."

Human judgment in relation to 'attacks' – part of the structure of IHL

Given that human beings are the addressees of the law, whether collectively or individually, then there are certain boundaries of machine operation that the law implies in relation to humans. The terms 'attacks' in IHL provides a unit of military action and it is over individual 'attacks' that certain legal judgments must be applied. So attacks are part of the structure of the law, in that they represent units of military action and of human legal application.

For example, Article 57 of Additional Protocol I, provides rules on precautions to be taken in attack. Where it refers to "those who plan or decide upon an attack" it is referring to humans. It is therefore humans that shall apply these legal rules – including verifying the objective, choosing the means and method of attack, and refraining from or cancelling an attack in certain circumstances.

We know that an attack must be directed at a specific military objective otherwise it is indiscriminate (Article 51. 4 a). We also know that a military objective must be of a sort (nature, location etc.) to offer military advantage at the time (Article 52. 2), and that in the application of the legal rules the concrete and direct military advantage must be assessed by the humans that plan and decide upon an attack (Article 51. 5 b and Article 57. 2 a.i&iii). Therefore humans must make a legal determination about an attack on a specific military objective based on the circumstances at the time. There should also be a capacity to cancel or suspend an attack (Article 57. 2 b).

These rules imply that a machine cannot identify and attack a military objective without human legal judgment and control being applied in relation to an attack on that specific military objective at that time (control being necessary in some form to act on the legal judgment that is required). Arguing that this capacity can be programmed into the machine is an abrogation of human legal agency - breaching the 'case-by-case' approach that forms the structure of these legal rules.

This line of argument is not dependent upon claims regarding the technical capacity of complex future AWS to do this or that, but is based on the law as a framework that applies to humans and that is structured to require human legal judgements at certain points.

However, this is not to argue that the law straightforwardly implies a very narrow constraint on what an AWS might do under its existing terms. Nor is it suggesting that existing law alone represents a sufficient basis for managing AWS. It is simply to point out that the existing legal structure (human judgement being required over 'attacks') implies certain boundaries to independent machine operation and that this is separate from arguments about how a machine might perform in relation to the implementation of individual legal rules (for example, the rule of proportionality).

Conceptualising 'an attack'

Whilst seeing in the structure of the law an assumption of human legal judgement in relation to individual attacks, it is also recognised that 'an attack' is not necessarily a single application of kinetic force to a single target object. In practice an attack may involve multiple kinetic events against multiple specific target objects. However, there has to be some spatial, temporal, or conceptual boundaries to an attack if the law is to function. This is linked to the different layers at which military action is often conceptualised – from the local tactical level, through the operational to the broad strategic level. If 'attacks' were not conceptualised and subject to legal judgement at the tactical level, but only say the broad strategic level, then a large operation may be determined to be permissible (on the basis of broad anticipated outcomes) whilst containing multiple individual actions that would in themselves be legal violations. Clearly for the law to function meaningfully there needs to be legal judgments and accountability over actions at the most local level.

Recognition that human legal engagement must occur over each attack means that a machine cannot proceed from one attack to another, to another, without human legal judgment being applied in each case, and without capacity for the results of that legal judgment to be acted upon in a timely manner – i.e. through some form of control system. Given that an attack is undertaken, in the law, towards a specific military objective that has been subject to human assessment in the circumstances at the time, it follows that a machine cannot set its own military objective without human authorization based on a human legal judgment.

Preventing an expansion of the concept of 'an attack'

Our starting point in this paper was concern that greater autonomy in weapons systems may result in human control not being meaningful. Based on the analysis above regarding the relationship of autonomy to the legal framework, we can see that this concern is linked to a risk that autonomy in certain critical functions of weapons systems might produce an expansion of the concept of 'an attack' away from the granularity of the tactical level, towards the operational and strategic. That is to say, AWS being used in 'attacks' which in their spatial, temporal or conceptual boundaries go significantly beyond the units of military action over which specific legal judgement would currently be expected to be applied.

Greater specificity of legal assessment - by this we mean a legal assessment that is evaluating specific events expected to occur over a shorter period of time, and within a narrower area - allows for specific risks to the civilian population to be more accurately assessed, and therefore for civilian protection to be better protected. Furthermore, allowing greater autonomy to facilitate progressive broader interpretations of what constitutes an attack would have a corrosive function upon the legal framework as a whole. This raises a key objection to assertions that national weapon review processes would be a sufficient response to the concerns posed by autonomous weapons. If the very tests that are applied to determine permissibility of a weapon system are being undermined by the development of that weapon system itself, how can the review process remain meaningful?

By asserting the need for meaningful human control over attacks in the context of autonomous weapons systems, states would be asserting a principle intended to protect the structure of the law, as a framework for application of wider moral principles. Moving the debate on to delineate the elements needed for human control to be meaningful would start to develop a normative understanding that should pull towards greater granularity and specificity of legal assessment, rather than greater generalisation.

Key elements of human control

So, as framed by the previous section, a meaningful form of human control is necessary both to allow for legal application and to protect the structure of the law from progressive erosion. In that context the section below sketches out 'key elements' through which human control can be understood to be applied in the use of weapons systems. These elements are not simply about technological characteristics but recognise that human control is necessarily part of a wider system that allows a specific technology to be controlled in a specific context of use.

Predictable, reliable and transparent technology

Starting with the technology itself, human control is facilitated where the technology is:

- predictable - it can be expected to respond in certain ways;
- reliable - it is not prone to failure, and is designed to fail without causing outcomes that should be avoided;
- transparent – practical users can understand how it works.

However the technology is to be used, there are certain characteristics that may be designed and manufactured into the technology that have a bearing upon the subsequent capacity for human control. A technology that is by design unpredictable, unreliable and un-transparent is necessarily more difficult for a human to control in a given situation of use.

Accurate information for the user on the outcome sought, the technology, and the context of use.

Human control in the use of a technology is then based upon those planning and deciding upon an attack having certain information. Control in the use of a weapons system can be understood as a mechanism for achieving commander's 'intent'. So information on the objective that is sought is an important starting point – including information on the unintended consequences that a commander wishes to avoid. This information is necessary for a human commander to assess the validity of a specific military objective at the time of an attack, and to evaluate a proposed attack in the context of the legal rules.

Such assessments also require an understanding of the technology. For example, we need to know what types of object a weapons system will identify as a target object – target 'profiles' – whether these are the commander's intended targets or not. We need to know how kinetic force will be applied – it makes a difference if the force will be a heavy explosive weapon with large blast and fragmentation radius, or if it will apply force quite narrowly, such as with an explosively formed projectile with no fragmentation effects.

'Predictability' is an important concept in that it provides a link between commander's intent and the likelihood of outcomes that match that intent. Predictability is partly a characteristic of the technology, but more fundamentally it is a characteristic of the interaction between that technology and the specific environment within which it will operate. As a result, information on context of use is very significant. We should have some understanding of the environment in which the technology will operate, including the presence of civilians and civilian objects for example.

Of course we may not achieve complete predictability – already in the use of weapons we accept degrees of uncertainty about the actual effects that will occur, and we know that there may be limitations on the information available about the context. However, our ability to understand the context is directly linked to both the size of the area within which the technology will operate, and the duration over which it will operate. For any given environment, it follows logically that greater area and longer duration of independent operation by a technology result in reduced predictability and so reduced human control.

It is recognized that different environmental domains present different general characteristics – with land, air and sea presenting different levels of complexity. This may mean that a large area of operation in the sea may still facilitate better contextual understanding than a smaller area on land. However, for environments of equal complexity, greater area and greater time of operation necessarily mean reduced control. In relation to the duration of an attack, this might be because certain people or objects enter or leave an area over time in a way that could not be anticipated, or it could be because the commander's intent has changed from the point at which the attack was initiated.

And from an understanding of the technology, and an understanding of the context within which it will operate, a commander should able to assess likely outcomes, including the risk of civilian harm, which is the basis for the legal assessment. It is important to note that information on these different elements may be the product of wider human and technological systems, but at some point understanding of these three elements must coalesce to a point where an informed judgement can be made.

Timely human judgement and action, and a potential for timely intervention

Based on the information on the outcome sought, the technology and the context, we need humans to apply their judgment – as implied by the legal analysis earlier in this paper – and to choose to activate the technology. This point of human engagement ties together the systems of information upon which judgements are made, but also provides a primary point of reference for the framework of accountability within which these actions are taking place. Of course responsibility for negative outcomes may turn out to result from problems elsewhere in the system (e.g. malfunctioning technology or inaccurate information on the context of use), but human judgement and action at this point is likely to be the starting point from which any negative outcomes are investigated.

The timeliness of this process is also significant because the accuracy and relevance of the information upon which it is based, about context for example, also degrades over time. For a system that may operate over a longer period of time, some capacity for timely intervention (e.g. to stop the independent operation of a system) may be necessary if it is not to operate outside of the necessary human control.

A framework of accountability

Finally, this broad system requires structures of accountability. Such structures should encompass not just the commander responsible for a specific attack, but also the wider system that produces and maintains the technology, and that produces information on the outcomes being sought and the context of use.

Conclusion on the key elements of human control

All of these areas cumulatively contribute to the extent of human control that is being applied in a specific context of use. In all of

these areas there are tests of 'sufficiency' that would need to be met in order for the overall extent of human control to be assessed itself as sufficient. Where some have asserted that the existing legal framework provides the answers needed for evaluating autonomous weapons systems, these tests suggest that this is not straightforwardly the case.

For example, it is not clear what level of information about the context within which a weapon will be used is considered 'sufficient' to provide a basis for an informed legal judgement. If a weapons system were to apply force to the individual vehicles of a group of fighting vehicles this might be considered reasonable if the group were known to be in a reasonably bounded geographical area over which a commander had knowledge. However, if the area within which that group of vehicles was situated was spread over a wider area, about which the commander necessarily had a lesser and lesser understanding, at what point does that understanding become so diluted as to make a legal assessment unreasonable? In legal terms, this is a question about what can reasonably be considered a 'specific military objective' and about what can reasonably be considered 'an attack'. The law alone does not provide an answer to these questions that resolve the uncertainty here, yet such questions are fundamental to avoiding the erosion of the legal framework that can be envisaged should states choose to develop autonomous weapons systems.

Whilst consideration of the key elements of human control does not immediately provide the answers to such questions either, it would at least allow states to recognise that these questions are fundamental, and it provides a framework within which certain normative understandings should start to be articulated, which is vital to an effective response to the challenge posed by autonomous weapons systems.

Working definitions - facilitating discussion within the CCW

The most direct way in which to establish such a discussion within the CCW is to adopt an approach to working definitions that is based on a recognition that certain forms of human control are required over the use of force, and that systems operating outside of that should not be considered acceptable. That would most straightforwardly be facilitated by adopting a working definition of 'lethal autonomous weapons systems' that is based on these being 'weapons systems operating with elements of autonomy and without the necessary forms of human control'. In such an approach the concept of weapons systems operating with elements of autonomy then refers to a broad category of systems within which a certain subset (either by design or by their manner of use) is considered unacceptable. Such an approach then sets up delineation of the key elements of human control as a primary focus of work in order to understand where the boundaries of permissibility should lie.