Article36

DISCUSSION PAPER | NOV 2017

Autonomous weapon systems: Evaluating the capacity for 'meaningful human control' in weapon review processes

Discussion paper for the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts meeting on Lethal Autonomous Weapons Systems (LAWS)

Geneva, 13-17 November 2017

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

This paper considers the potential for national level weapon review processes to be strengthened to provide one implementation mechanism for a future international legal commitment to ensure adequate human control in the context of increasing autonomy in weapons systems.

It argues that under a future international instrument on autonomous weapons evaluations will need to be made of diverse technologies to ensure that they allow a sufficient level of human control. Despite recognised challenges, national level weapon review processes will be important to the effective implementation of any 'future orientated' international legal commitment. As such, the paper argues that in the development of such a legal commitment the role and parameters of the weapon review process should be further considered and elaborated.

A new legal instrument should provide additional guidance as to how new weapons, means or methods of warfare ought to be assessed in order to ensure that they allow for the necessary human control. Through such an approach a legal instrument could be established that is broad enough, and flexible enough, to be effective in the context of diverse future scenarios.

This is distinctly not an argument that national weapon review processes alone are a sufficient mechanism for addressing the challenges presented by growing autonomy. Rather, the paper argues that a specific additional legal obligation to ensure human control is necessary in order to require national weapon review processes to evaluate new weapons, means and methods of warfare effectively.

Introduction

The central area of concern regarding the development of autonomous weapons systems (AWS) is the depletion of human control over the critical functions of identifying, selecting and applying force to targets. Without the necessary capacity for human control there may be a moral deficit in the use of force, such systems might not allow the proper application of legal rules or might drive interpretations of the legal framework that erode the protection of civilians and combatants, and could further jeapordise international stability.

In the context of discussions in the Convention on Certain Conventional Weapons (CCW) a significant number of states have acknowledged that some form of human judgement and control must be retained as technological changes allow for greater autonomy in various weapon system functions.

In this context, Article 36 has published papers describing 'key elements' of what has been termed 'meaningful human control'. In those papers, and in this analysis, the term 'meaningful human control' is used as a placeholder for a recognition that some capacity for human control is necessary and that the parameters of that control need further elaboration. Other terms, such as sufficient-, adequate-, necessary-human control, could be chosen instead of the term 'meaningful'.[1] The choice of wording has certain subtle implications – for example, the term 'meaningful' arguably draws in broader concerns regarding the right to dignity, whereas a term like 'sufficient' implies a minimal requirement. However, whatever term is chosen it still leaves an open question as to what the parameters of that control must be.

This paper suggests an approach to articulating those parameters by describing them in the form of guidance to the assessment of human control in the context of a process to review new weapons, means and methods of warfare.[2] Such guidance would need to be provided in conjunction with an additional legal obligation to ensure meaningful human control, and to prohibit the use of

systems without that human control. Such an approach allows for a degree of flexibility in response to the possible diversity and complexity of future technological developments.

An obligation for human control vs. national level weapon reviews?

In the context of international discussions on autonomous weapons systems at the CCW, two distinct approaches have been proposed that have so far remained largely separate and have sometimes been presented as oppositional:

- One of these approaches argues for states to adopt a positive obligation for 'meaningful' or 'appropriate' human control in the operation of such systems. This orientation argues that a positive obligation is necessary in order to ensure that legal obligations can be upheld in the use of force and that the legal framework is not eroded by movements towards greater autonomy.
- Another approach argues that national level weapon review processes (for example as required under article 36 of Additional Protocol 1 to the Geneva Conventions) will provide a sufficient basis for avoiding legal challenges associated with autonomy. Such an approach tends to assert that the existing legal framework is sufficient and specific international rules are not required to respond to the challenges presented by increasing autonomy.

Whilst these two approaches have been presented as distinct, and have tended to proceed from different starting assessments regarding the sufficiency of the existing legal framework, this paper suggests an integration of these two approaches. This could provide a productive and practical way forwards: an international obligation to ensure that systems allow the necessary level of human control will require weapon review mechanisms for its effective implementation. In addition, the establishment of an international obligation for human control, even if broadly framed, will help to develop consistency in the implementation of these national weapon reviews as well as helping to extend weapons reviews in practice.

Categories and definitions in the current CCW debate

The CCW has been discussing this subject matter under the rubric of "lethal autonomous weapons systems" – abbreviated to LAWS. Whilst there is developing consensus that the central issues of concern regarding LAWS are related to autonomy in the 'critical functions' of selecting and applying force to targets,[3] there is significant uncertainty in other areas of the debate, including around the approach taken to terminology and definition.

There is no working definition of LAWS and there is uncertainty and controversy about the scope of the discussion in the CCW. This is common in future-oriented debates about the regulation or control of technologies given that the choice of organising terms and principles may have implications for the scope of future constraints. There are divergent opinions on whether the term LAWS should refer:

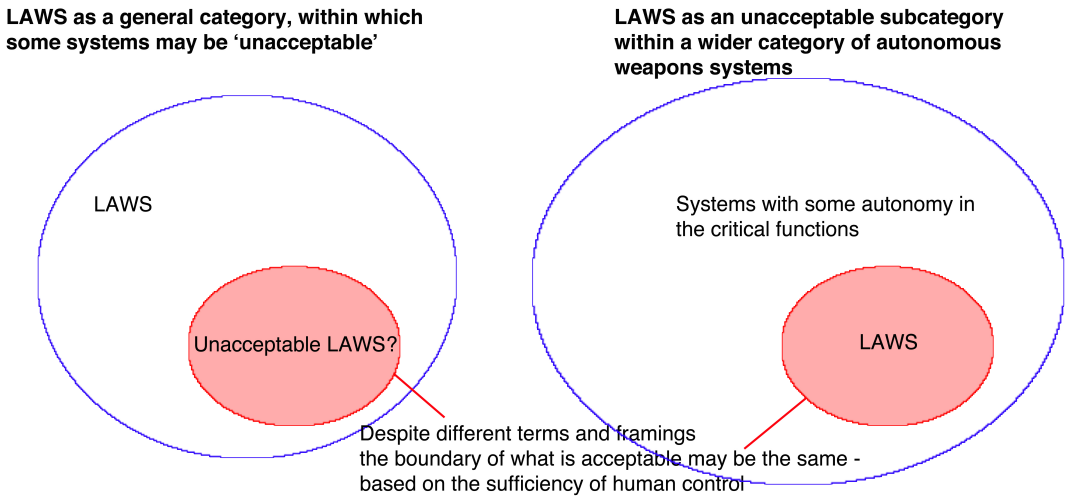
(1) to a broad category of technologies within which certain systems may be deemed particularly problematic or unacceptable (and which might be termed 'fully autonomous weapons'), or
(2) whether the term LAWS refers to particularly problematic/unacceptable systems within a wider category of systems with some autonomy in the critical functions.

In the positions articulated by states there are various further subtleties to how these different categories are described or bounded – but some version of this structure is apparent in most position papers of substance.[4] Such divergent starting points add a layer of complexity to the discussion in the CCW which it will be necessary to get beyond in order to have a productive debate. For example, any discussion of 'working definitions' will need to make choices about the basic hieracharchy of terminology and the categories to which such terms apply.

The term 'LAWS' is a neologism coined by the CCW specifically to structure the CCW's work. There is, thus, no 'right answer' to the question of which definitional starting assumption should be adopted. Rather, this is a political question, with different political and policy implications for how the subsequent conversation is likely to proceed.

Whatever terminological approach is adopted, it is possible that the boundaries between what is considered particularly problematic/unacceptable and what is not might fall in the same place. In the illustration below it is suggested that bounding the 'unaccpetable' category in relation to the sufficiency of human control produces a category of the same dimensions, regardless of the terminology or the wider field that it falls within. More fundamental, therefore, than the choice of terminology are the principles by which the different boundaries are configured. That said, achieving the establishment of a boundary in one place or another might be advantaged or disadvantaged by the choice of terms that is use. It is for this reason that terminology and definitions is primarily a political rather than a technical question at this stage of the debate.

Example 1 here arguably creates space to raise critical questions about existing systems with some autonomy in relevant functions



under the term LAWS. However, some states wish to exclude such systems from any consideration and have sought to embed the mantra that 'LAWS do not exist yet' into their working definitions despite using LAWS as a more general organising term. The approach in example 1 might require a further term to be adopted to describe those types of systems that might be considered unacceptable and so subject to prohibition (because, for example, they do not allow 'sufficient human control').

Example 2 arguably confers certain advantages to those that wish to see a prohibition of a named category – i.e. lethal autonomous weapons systems. It allows the key term of the debate in the CCW to be argued to be directly congruent with the objects that such actors wish to see prohibited – e.g. systems that don't allow 'meaningful human control', 'fully auonomous weapons', 'killer robots'. Creating equivalence between those terms and the term LAWS simplifies the conversation and makes it clear that a process of defining will be a process of defining 'that which is unacceptable' and therefore must be the subject of a prohibition.

From the perspective of ensuring meaningful human control, the key issue in both orientations is to foster recognition that the smaller, inner boundary delineates systems that do not allow, or that are operating without, sufficient human control. Any state that has acknowelged that a certain degree of human control needs to be retained in the use of force should be able to acknowledge that a boundary exists based on that requirement.

Defining the necessary human control through practical guidance

Article 36's previous policy papers on this theme have focused primarily on promoting a positive obligation to ensure that there is 'meaningful human control' in the way an 'attack' is undertaken in conflict. In this approach 'meaningful human control' is the positive characteristic that is lost when autonomy in the critical functions of weapons systems is extended too far. This sets up a principle for definition – based on describing the form of human control that is considered necessary.

This is a very different challenge to that of defining a technology. The description of 'meaningful human control' (or whatever form of words is chosen to express this) is likely to be framed in terms of a number of further tests of sufficiency, an approach which is explored later in this paper. It is therefore unlikely to present, in itself, a single hard boundary. It may, however, serve to establish a set of reference points against which technologies can be considered and certain boundaries established, including possible determinations that certain configurations of technology are always unacceptable.

The framework presented below suggests that rather than seeking to define 'meaningful', 'appropriate' or 'necessary' human control in detailed and specific terms such a concept should be understood as a broad principle.[5] In order for that principle to be given practical utility it should be augmented by a set of understandings regarding the key areas through which human control may be enacted, in different ways, in different systems.

In line with Article 36's previous writings on this issue, it is recognised that the necessary form of human control does not equate to some form of 'absolute control'. For example, already in the use of weapons there are substantial levels of uncertainty over exactly what will happen when a weapon is employed. Additionally, the emphasis on 'control' in Article 36's writings on this issue is not a rejection of the need for human 'judgement' in the enacting of legal requirements. However, control is seen as the mechanism by which human judgement is transmitted into the functioning of a technology. Furthermore, an analysis of 'control' allows consideration of how that transmission takes place (which is directly linked to the human-machine interface), and of the ability to understand the context where force will be applied, as well as prior and subsequent aspects of technological development and management. On this basis we see 'control' to be the primary issue for consideration.

Weapon reviews background

Under article 36 of Additional Protocol I states are under a legal obligation to review new weapons, means and methods of warfare to determine if their use would be prohibited in some or all circumstances. It is generally recognised that weapon reviews should be undertaken through the process from study and development to the adoption of a weapons system into operational use.

The primary purpose of weapon reviews, as articulated in law and in the interpretation of most states that have presented opinions on the issue, is to evaluate proposed weapons systems in relation to the existing legal framework. Weapon reviews, as currently understood, are not therefore a mechanism by which new legal orientations are brought about, but are fundamentally a tool to facilitate the implementation of existing, explicit legal obligations.

The existing legal framework is insufficient to manage the implications of developing autonomy in the critical functions of weapons systems. This assessment is not based on claims around what future technologies will or will not be able to do, but rather on a recognition that autonomy in the critical functions will facilitate expanding interpretations of legal terms such as "attack" and erode the fundamental role of humans as the addressees of the law as written.[6] On this basis Article 36 has pressed states to adopt a positive legal obligation to ensure that there is 'meaningful human control' over individual attacks. Such an obligation would establish the idea of a line between systems with acceptable/unacceptable forms of autonomy and new systems would need to be evaluated against that obligation in future weapons review processes.

What is a system and what constitutes 'new'?

One of the challenges presented by movements towards autonomy in the critical functions of weapons systems is that these functions may be dispersed between different physical structures or locations. As a result, whilst review processes will need to be applied to new, physically unified weapons systems, they will also need to be applied where different components are brought together to function in a different system configuration. Thus, for example, where an armed drone might currently be configured to operate under direct human control the wider system would need to be reviewed as a whole if it was reconfigured to take targeting or weapon release instructions from a separate computer system. As such, it should be recognised that the separate 'nodes' that make up a system might be integrated into one physical platform or dispersed across several physical platforms. In this context the framing of the legal obligation to review not only new weapons but also new "means and methods" of warfare is important.

Linked to the considerations noted above, changes within a physically unified system that has previously been accepted for deployment should also precipitate a further review process. Given that the development of autonomy in weapons systems is largely driven by the interaction of sensors and computing technology, changes in the configuration of these systems will necessarily require further evaluation. Software changes could

make a fundamental difference to how human control is excercised over or through a system. This requirement would be complicated further by any use of machine learning which might result in unstable or uncertain analytical processes within a system.

Evaluating systems to ensure meaningful human control

In conjunction with adopting a legal obligation to ensure a sufficient form of human control, states should also develop complementary guidance on how the sufficiency of human control should be evaluated.

In previous papers circulated at the CCW, Article 36 has suggested the following broad areas through which human control is enacted:

- 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.

Some of these requirements are embedded to some extent in the technology itself, whereas others relate to the wider framework within which it will be used. This is significant because it suggests that in the review of a particular technology – or means of warfare – consideration will also need to be given to the operational structures within which its use will be considered. That said, the comparative breadth of this approach, and its comparative openness in terms of the lines that are drawn, does not preclude the possibility that certain sorts of technology might be held to not allow the necessary human control, regardless of the wider operational structures.

Key questions for evaluating human control

The considerations suggested below provide initial examples that could guide an evaluation of the sufficiency of human control that a system allows. It should be recognised that these questions may function at different levels, at different stages in the process of system development and testing, and in relation to specific contexts of use.

This framework does not indicate clear lines regarding where human control moves from being sufficient to insufficient, but rather suggests a set of subsidiary tests. And some of these tests themselves might interact with each other – such that, for example, a system that operates on the basis of comparatively broad 'target profiles' might need to be more constrained in its area and time of operation than a system using much narrower target profiles.

Predictable, reliable and transparent technology

- Consideration should be given to whether the technology itself is sufficiently predictable, reliable and transparent in relation to those components that identify, select and apply force to target objects.
    - These questions should be evaluated not only in relation to issues of performance against design specification, testing, technical validation and operator instruction and training, but also in relation to transparency regarding the process of target categorization.

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

- Consideration should be given to whether any 'targeting profiles' (the proxy data taken to indicate a target) are specific enough and sufficiently understood in relation to the objects intended to be targetted by the system.
    - Such an evaluation should consider what military objects will fall within the target profiles and what other objects (including persons) might fall within these profiles.
    - Consideration should be given to the form and sufficiency of testing, verification and validation of how target profiles match target objects and other objects, and it should consider what factors might influence the reliability of such matching in different operational contexts.
- The review process should consider whether it will be possible for a commander to have a sufficient understanding of the context where force will actually be applied.
    - This should include evaluation of whether the geographical space where force will be applied is or can be sufficiently constrained for a commander to evaluate the implications of specific applications of force that may be undertaken by the system.
    - A review should consider whether the time period within which force will be applied is or can be sufficiently constrained for a commander to evaluate the implications of specific applications of force that may be undertaken, and to evaluate whether external circumstances may change in a way that would make such an application of force ineffective, undesireable or illegal.
    - Consideration should be given to whether a commander can have sufficient information about the times and locations where force might be applied to make an informed assessment of the effectiveness, legality and desirability of initiating an attack given any target profiles employed by the system.
- The review process should also consider whether the physical effects of the weapons to be employed by a system are sufficiently understood and can be considered appropriate in the context of any uncertainty about the time and place where they may be used.

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

- Consideration should be given to whether the system provides sufficient capacity for intervention in its operation once it has been activated.
    - Such an evaluation should consider the possible duration of system operation and its potential to engage in multiple applications of force.
    - It should consider what information the system can report back to a commander and whether this is sufficient for them to make an informed judgement about the continued operation of the system, in the context of other available contextual information.
    - It should consider whether a commander has the capacity to deactivate the system, or to change

operational parameters, once the system has been activated.

Accountability to a certain standard

- The review process should ensure that the system and the operating procedures around it allow for a sufficient level of accountability within the procedural context in which it will be used.
  - This should ensure that the system and operating procedures allow for command responsibility.
  - It should consider the policies, procedures and administrative structures necessary to ensure that, beyond command responsibility, there is accountability for the technical performance of the system and its components.

Informing such an evaluation process

Approaching the questions sketched out above would require consideration of the design purpose and technical specifications of a system, as well as performance data from testing and validation processes in sufficiently realistic conditions, and consideration of the operational and accountability structures within which the system is intended to be used.

Conclusions

This preliminary framework does not indicate clear lines regarding where human control moves from being sufficient to insufficient, but rather suggests a set of subsidiary tests – which themselves may need to be assessed in conjunction with each other in order to draw a conclusion. Such an approach allows for the diversity of possible future technologies to be managed appropriately, as well as giving some consideration to the way different operating environments may allow human control to be more or less tightly enacted.

The approach to evaluation sketched out above is specifically relevant to interrogating a legal obligation for there to be a sufficient level of human control in the operation of such systems. Without such a legal obligation there would be no specific basis from which to require such considerations within the review process. Clearly, these approaches to evaluation do not nullify states obligations also to evaluate the permissibility of systems in relation to the wider set of legal obligations that they are bound by.

Article36

www.article36.org

[1] See for example, the working paper of Germany and France CCW/GGE.1/2017/WP.4, November 2017, stating that "humans should ... continue to exert sufficient control"; statement of South Africa to the informal meeting of experts to the CCW, April 2016, "'necessary human control' is a requirement that my delegation is supportive of"; statement of Canada to the informal meeting of experts to the CCW, April 2016, "'meaninful human control' or appropriate human judgement ... may help us to develop norms of responsible behavior".

[2] Art. 36, 1977 Additional Protocol I to the Geneva Conventions.

[3] The ICRC's working paper to the 2016 informal meeting of experts of the CCW, April 2016, states that, "The ICRC has defined autonomous weapon systems as: 'Any weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention.'"

[4] For example, the working paper of the Netherlands CCW/GGE.1/2017/WP.2, October 2017, distinguishes between "autonomous weapons" and "fully autonomous weapons, without meaningful human control", where the latter category is "outright reject[ed]". The November 2017 working paper of Germany and France (CCW/GGE.1/2017/WP.4) however, argues that LAWS should refer to "fully autonomous lethal weapons systems" but then suggests that these could still remain under "sufficient [human] control". This presents a confusing approach to definition that does not help to clarify the current debate.

[5] The October 2017 working paper of the Netherlands (CCW/GGE.1/2017/WP.2) argues that 'meaningful human control' should be regarded as "a standard deriving from existing legislation and practices" and partly uses that as a justification for arguing that specific legal articulation is not necessary. However, given that autonomy threatens to erode understanding of established legal terms and categories the importance of this principle, even if implicit in the current law, should be made explicit.

[6] See Maya Brehm, 2017, "Defending the boundary: constraints and requirements on the use of autonomous weapon systems under international humanitarian and human rights law", Geneva Academy 2017, especially pp32-39.