POLICY NOTE| NOV 2019

Article36

Autonomy in weapons systems: mapping a structure for regulation through specific policy questions

An outline for regulation and national-level questions for analysing policy orientations

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

* This paper was written by Richard Moyes.

Introduction

This policy note sketches a structure for the legal regulation of autonomy in the 'critical functions' of weapons systems. The focus is on building a shared conceptual structure for such regulation, rather than on the specifics of terminology. Having outlined such a structure, we suggest questions that could be asked at the national level in order to assess government orientations to the key policy questions that this structure raises.

Outline of a structure for regulation

The approach outlined here draws on Article 36's paper on target profiles, and on previous writings on 'meaningful human control'.[1] The structure is based on three key elements:

- As a starting point, a broad range of systems should be subject to regulation.
- Within that broad structure, certain system configurations should be prohibited.
- Other configurations, those that are not prohibited, should be subject to obligations regarding their development and use to ensure that they can be used in accordance with established legal obligations.

A broad range of systems should be subject to regulation
A legal instrument should be addressed at a broad range of systems. It should cover the wide category of systems where force is applied on the basis of sensor data, without human evaluation of that data, and without a human setting the time and place of that application of force. A formal definition of that broad category would need to set this boundary carefully, but all international discussions currently recognise that the primary issues of concern regarding autonomy in weapons systems fall *within* this boundary. Some would like to set such a boundary more narrowly, but their proposed approaches *still fall within* this category. Addressing a broad range of systems is necessary in order to respond to systems that do not need to be subject to outright prohibition, but still must be subject to certain constraints if we are to retain meaningful human control over attacks in armed conflict.

Certain system configurations should be prohibited
Within that broad category of concern, certain system configurations should be prohibited. Article 36 suggests two main prohibitions based on different sets of objections.

In our paper '**Targeting people**'[2] we have elaborated a set of significant concerns about using sensors to apply force to people. In particular we note that, the prospect of using AWS to direct force against human beings raises a set of interrelated moral and legal issues, including:

- the risk that the 'wrong people' (e.g. civilians or combatants *hors de combat*) may be targeted;
- questions about how measures adopted to manage the risk of undesired consequences might affect the normative protection of people in the long run;
- procedural concerns about the *process* of targeting, that is, how and why a person is made the object of attack or harmed.

Within the parameters of the broad definition noted in the section above, Article 36 supports a prohibition on all systems that use sensors to identify people as targets for the application on force.

Secondly, there should be a prohibition, or set of prohibitions, on systems that, in basic terms, are so **complex in their functioning** that they cannot be meaningfully controlled. Factors producing an unacceptable level of complexity and unpredictability could include:

- systems where target profiles are built on the basis of machine learning such that commanders do not know the actual characteristics of conditions under which they will apply force (i.e. they are not able to assess the presence of conditions that will produce 'false positives');
- systems where target profiles, the conditions under which force will be applied, change within the system after it has been put into use and are not approved by a human commander.

Prohibitions in these areas would prevent the adoption of systems where a commander cannot effectively evaluate the specific risks a system presents to civilians and civilian objects. They would address concerns regarding systems 'setting their own goals' as well as underpinning concerns that target profiles may be based on opaque characteristics, subject to dataset bias, and not amenable to an evaluation of their wider implications.

In our paper on 'Target profiles' we have noted that consideration of target profiles, as the conditions under which force will be applied, may provide a useful conceptual tool for formulating such prohibitions.

The remaining systems should be subject to obligations regarding their development and use

All systems within our broad starting category present challenges because they involve the application of force at a specific time and specific location that has not been set by a human commander. This means that there is a window of uncertainty regarding the actual effects that will occur. That window of uncertainty necessarily produces additional challenges for efforts to assess, mitigate and apply legal judgments regarding anticipated harms to civilians and civilian objects.

Furthermore, all of these systems use some form of target profile(s), i.e. a preprogrammed set of conditions under which force will be applied, typically based on some proxy indicators of an 'intended' target. Such proxy indicators, in existing systems, are already built on the basis of sensor-identifiable characteristics, such as weight, heat-shape, acoustic signature, radar profiles etc. The relationship between target profiles and actual circumstances in the operating environment produces another mode of uncertainty - what will actually trigger an application of force in practice?

Because of these two sets of uncertainties - regarding the location of force in time and space, and regarding the matching of target profiles to intended targets - it is necessary to apply specific obligations to all such systems to ensure that they are used in accordance with established legal rules and principles. Such obligations will likely need to be drafted in broad terms, but such obligations are vital if such systems are not to be used over such wide areas or for such long durations of time that they erode the structure of the law, which requires judgements to be applied to specific circumstances. Obligations should also ensure that commanders understand the systems that they are using, including the implications of the target profiles that systems use and their relationship to things that are not the intended targets.

Questions for states

The structure for regulation sketched out above includes both prohibitions and positive obligations. How definitions and rules are drafted would require careful work, but that can be taken forward once a sufficient group of stakeholders is working within the same conceptual structure for how a legal instrument should be framed.

In order to help to evaluate, at a national level, where states are positioned in relation to this structure, the questions below orientate towards some of the key lines along which prohibitions and positive obligations could be developed. These are suggested here as a resource that can be drawn upon in support of policy engagement at a national level.

There are two sets of questions, both comprised of specific policy points framed under a general question:

For systems that process sensor inputs to determine where and when to apply force (without further human involvement) would the following be acceptable or unacceptable?

- Systems that are designed to identify people as targets on the basis of human biometrics?
- Systems that identify different groups of people as targets on the basis of perceived racial, gender or age characteristics?
- Systems where the sensor-identifiable characteristics of possible targets can change or develop, within the system, after it has been activated and without being specifically certified by a human?
- Situations where the human users understand what the system is *intended* to target, but do not know the actual physical/emission characteristics that will be identified as a target – such as where target profiles have been built through current neural network/ machine learning?
- Situations where the human users do not have an understanding of what, other than intended targets, might also be identified as targets by the system?

For systems that process sensor inputs to determine where and when to apply force (without further human involvement) are the following assertions reasonable?

- Human users should be fully responsible for verifying the risk to civilians from their use of a system;
- Systems that will target both certain civilian objects and certain military objects should not be used in situations where those objects are intermingled;

- The geographic area over which a sensor-targeting function can occur should be controlled such that human users can fulfil their legal obligations;

- The duration over which a sensor-targeting function can occur should be controlled such that human users can fulfil their legal obligations;

- The time at which a sensor-targeting function may occur should be sufficiently proximate to the application of human legal judgement for that legal judgement to be relevant to the circumstances in which the function will occur;

- The number of applications of force that a system can undertake in an individual attack should be set by the human users;

- Human users need to understand the actual weapon effects (type of force) that such systems will create.

Conclusion

Developing a shared concept of how a regulatory approach should be structured in relation to autonomy in weapons systems is of paramount importance. Such a structure does not pre-judge the specific rules that might be formulated, but it is necessary in order to foster a constructive conversation about the best approach to any such rules. Whilst formal definitions and specific rules will require focused collective work, a group of actors working within the same conceptual structure is vital to establishing a context within which that work can be undertaken.

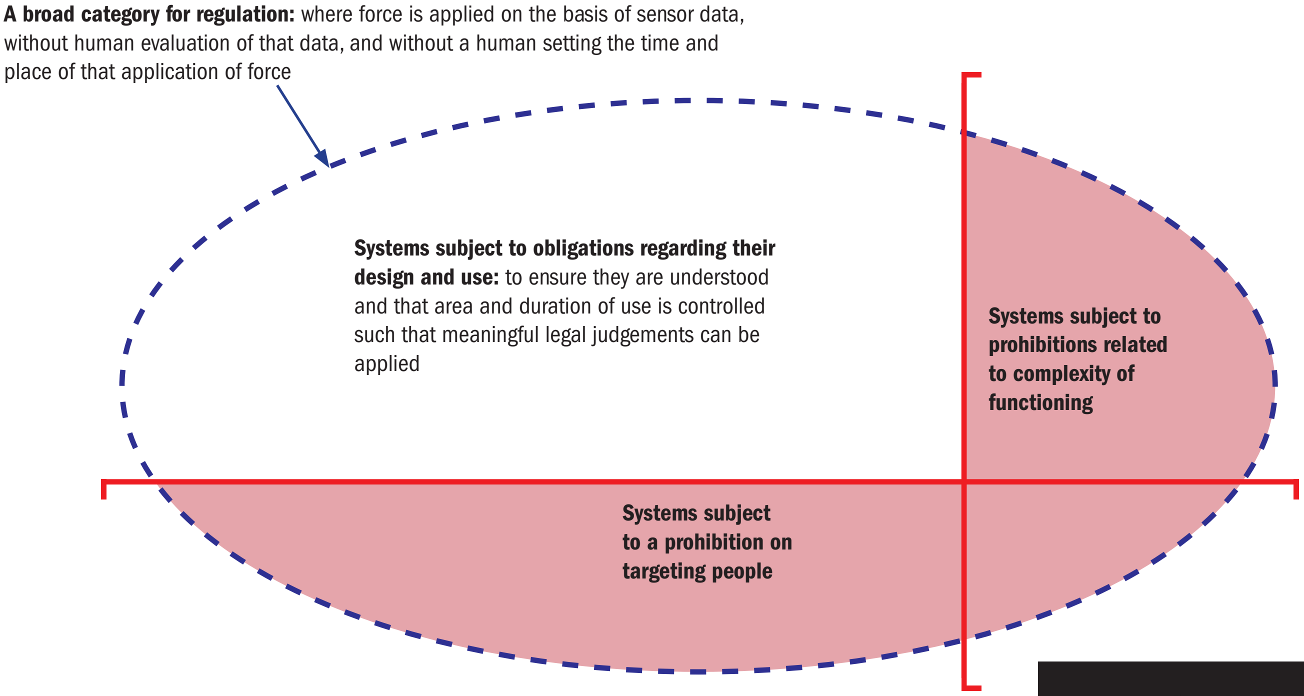
A number of states argue that existing international law is sufficient to provide adequate regulation of the issues posed by autonomy in weapons systems. We encourage those states to consider the questions suggested in this paper and to indicate what answers their reading of the law guides them towards.

NOTES

[1] 'Target profiles', Article 36, 2019, http://www.article36.org/wp-content/uploads/2019/08/Target-profiles.pdf and, for example, 'Key elements of meaningful human control', Article 36, 2016, http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf

[2] 'Targeting people: key issues in the regulation of autonomous weapons systems', Article 36, 2019, http://www.article36.org/wp-content/uploads/2019/11/targeting-people.pdf

A structure of regulation - prohibitions and positive obligations within a broad category



A broad category for regulation: where force is applied on the basis of sensor data, without human evaluation of that data, and without a human setting the time and place of that application of force

Systems subject to obligations regarding their design and use: to ensure they are understood and that area and duration of use is controlled such that meaningful legal judgements can be applied

Systems subject to prohibitions related to complexity of functioning

Systems subject to a prohibition on targeting people

Article36

www.article36.org