WRITTEN BY RICHARD MOYES

SYSTEMS THAT CANNOT BE EFFECTIVELY CONTROLLED

Article36

WWW.ARTICLE36.ORG
INFO@ARTICLE36.ORG
@ARTICLE36

Article 36 is a specialist non-profit organisation, focused on reducing harm from weapons.

NOTE: This paper is a short consideration of issues around the concept of 'systems that cannot be effectively controlled'. It is based on a background paper for a meeting. It does not consider the issue of systems designed or used to target people. Article 36 calls for a prohibition on the use of autonomous systems to target people.[1]

We argue that 'meaningful human control' is needed in order to ensure sufficient human moral engagement with the use of force, to meet existing legal obligations regarding the use of a weapons system in an attack, and to ensure that responsibility and accountability are situated appropriately.

From this it follows that systems that 'don't allow' or 'cannot be used with' meaningful human control should be considered unacceptable and that their development and employment should be prevented. Whilst this proposition seems reasonably straightforward, it is a more complex question to assert what technical structures of system configuration would necessarily produce systems that cannot be sufficiently controlled.

All systems that use sensors to determine automatically where and when force will be applied present some degree of uncertainty regarding the effects that will occur - and these are the systems that we are concerned with in discussions regarding 'autonomy'. Even in other weapon systems, where a human operator does select the specific time and location of force, there will often be some margin of error – for example, as a product of weapon system 'inaccuracy'. But where the operator is not setting the specific time and location for the application of force, uncertainties regarding actual effects are greatly increased.

Our basic model of control in practice involves 'management' of the following key elements:

- The technical functioning of the system – including its sensors, target profiles, the nature of force it applies and the number of iterations of force it is capable of.
- The 'context' in which those technical functions will occur – recognising that complexity of context will make prediction of specific effects more difficult and that complexity of context increases with greater spatial area and greater duration of operation over time.

We tend to highlight the importance of 'context' in our recommendations regarding operational obligations (positive obligations) to control systems in their use. For example, we argue that a system user must be able to limit the area and duration of a system's functioning sufficiently for them to meet their existing legal obligations pertaining to 'an attack' in a meaningful way - i.e. to understand the specific context enough to make an informed judgement about what will happen when a particular system is used.

Synthesising this requirement with our earlier recognition that systems that do not allow meaningful human control should be considered unacceptable, we can create a rule such that:

- It is not acceptable to use systems where the location and duration of their functioning cannot be appropriately limited.[2]

It might be argued of course that militaries would not wish to develop such systems - though in practice there is the example of landmines, which have caused significant humanitarian problems as a result of the lack of limits on their duration of functioning (coupled with a loss of control over their location).[3] It might also be argued that existing law and existing weapon reviews would (in the interpretation of an individual state) not allow such systems anyway. This may be the case for a specific state making this assertion, but that does not really provide an argument against making this an explicit international rule.

Beyond ensuring that the technical characteristics of a system allow the management of 'context' sufficiently, we should turn our attention to the other aspects of technical functioning that we highlighted earlier, notably:

- system sensors
- target profiles
- the nature of force a system applies
- the number of iterations of force a system is capable of.

A system's sensors and target profiles work together to determine if external conditions at any particular point in time are such that force should be applied. These are the technical components of the system that produce an application of force at a specific place and at a specific point in time.

It follows that if a human commander is to make meaningful judgements about the implications of using a system in a particular context they must have a sufficient understanding of how the system will determine, within that context, where and when to apply force. They would need to understand what characteristics in the environment would produce an application of force, whether those characteristics

were associated with a military objective or not. This could be formulated as a rule:

- It is not acceptable to use systems where the external conditions and circumstances that will trigger a specific application of force cannot be appropriately understood.

More specific consideration could be given to technical characteristics that are likely to fall foul of a rule such as this. We have noted elsewhere that systems where the target profiles can change without human authorisation would present a barrier to a user making an appropriate judgement about the use of that system. Likewise, we have highlighted the possibility of systems where target profiles are constructed through forms of current machine learning – producing a situation where the actual parameters of those profiles cannot be interrogated or understood, and so the conditions and circumstances that would trigger force cannot be described in terms other than similarity to the abstract objects or situations that are the intended targets of attack.[4] Whilst these formulations might need further elaboration, they both serve as specific examples that might run afoul of the general rule here.

The rule formulation above applies to the technical characteristics of a machine 'system'. Such a rule has a possible corollary in terms of an obligation on human users of systems where an appropriate level of understanding is possible but may, nevertheless, not be straightforward:

- The users of a system must be able to provide a meaningful explanation of how a system functions, including meaningful information on the external conditions and characteristics that will trigger an application of force.

The concept of a meaningful explanation and 'meaningful information' here draws upon the language of European data-protection legislation regarding the rights of people subject to automated decision-making.[5] Being able to provide a meaningful explanation of how a system will interact with its environment, in terms of the critical question of 'what will result in force being applied' is arguably an essential element for meeting moral and legal obligations, and to ensuring appropriate responsibility and accountability. Meeting a rule such as this is not simply a function of a technological system but would require appropriate considerations in the development of a system, review processes and training.

The other technical considerations we highlighted above were the nature of force that a system can apply, and the number of iterations of force that a system is capable of. The 'nature of force' in many ways returns to basic moral and legal considerations regarding the use of weapons – that there must be an understanding of the immediate and longer-term physical effects that weapons will cause, and human judgements made on that basis. This is simply to say that the human user of a system, in making judgements about the use of a system, must understand the form of force that will be applied and factor that into their assessments. It would also follow that if a particular weapon type is prohibited for use directly by a human operator it is prohibited also for that operator to use that weapon type through the intervening medium of a sensor-targeting system. These points should be wholly uncontentious.

By 'number of iterations of force' we are referring to the number of times the system might cycle through the process of 'sensor input—calculation—force application'. Where we recognise that target profiles are a simplified representation of an intended target type, in the language of a system's sensors, then we recognise also a potential for false positives (that is, circumstances that trigger an application of force to things not of the intended target type). While this is always present as a risk, that risk necessarily grows each time the system cycles through the sensor-targeting process. Therefore, limiting the number of iterations of force that a system is capable of presents an additional mechanism by which its effects can be rendered incrementally more predictable.

All systems will have some limit to the number of iterations of force that they can apply. Some existing systems are self-expending – with the sensor-calculation unit being destroyed in the process of applying force (as in a sensor-fused munition, for example). Other systems may be able to undertake numerous applications of force before they would run out of ammunition (as in the case of a Phalanx anti-missile system).

Rather than being amenable to set boundaries of technology, both the nature of force and the number of iterations of force could be subject to obligations relating to the use of systems:

- The users of a system must understand the nature and extent of force that a system will exert in any application of force.
- The users of a system must limit the number of applications of force that a system can undertake within the context of an attack, such that they can make appropriate judgements about the use of that system in that attack.

CONCLUSIONS

The sections above have highlighted possible rules that might flow from a particular mode of analysis.

It is important to note that:

1) We are regulating systems that use sensors to determine where and when force will occur, and that automatically apply force, without that specific place and time being set by a person.
2) Within that category, we are proposing a prohibition on systems where people are identified as objects to be attacked - but that is not the subject of this paper.
3) We argue that systems that cannot be effectively controlled should be prohibited (and we have been elaborating this line further in this paper).
4) We argue that systems within this scope (1) and not prohibited (under 2 & 3) should be subject to positive obligations regarding their development and use.

Whilst we are not suggesting that these rules sketched in this paper provide a comprehensive / sufficient set of obligations, they serve to illustrate that broad obligations can be created in this space.

Bringing them together our example rules read as follows:

Towards prohibitions:

- It is not acceptable to use systems where the location and duration of their functioning cannot be appropriately limited.
- It is not acceptable to use systems where the external conditions and circumstances that will trigger a specific application of force cannot be appropriately understood.
- The users of a system must be able to provide a meaningful explanation of how a system functions, including meaningful information on the external conditions and characteristics that will trigger an application of force

Towards additional positive obligations:

- The users of a system must understand the nature and extent of force that a system will exert in any application of force.
- The users of a system must limit the number of applications of force that a system can undertake within the context of an attack, such that they can make appropriate judgements about the use of that system in that attack.

ENDNOTES

1 For a summary of Article 36's overall policy approach see, Article 36, 2020, Regulating autonomy in weapons systems, https://article36.org/updates/treaty-structure-leaflet/

For more analysis on a rejection of targeting people, see Article 36, 2019, Targeting people, https://article36.org/wp-content/uploads/2019/11/targeting-people.pdf

2 Such a rule would echo the Additional Protocol 1 Art 51 (4.c) prohibition on employment of means and methods the effects of which cannot be limited as required – though it is not necessary to argue that it flows from that established obligation.

3 Whilst anti-personnel mines have been prohibited it is notable also that the political Declaration on Anti-Vehicle Mines adopted by a number of states within the CCW (CCW/CONF.III/WP.16, 2006) is primarily focused on measures to curtail the duration of functioning of these systems (through allowing their retrieval, or self-destruction etc.).

4 See Article 36, 2019, Target Profiles, https://article36.org/wp-content/uploads/2019/08/Target-profiles.pdf - the notion that the conditions or circumstances cannot be described other than in terms of abstract descriptions of the 'intended target' links to the requirement for a 'meaningful explanation' in the subsequent rule proposed here.

5. For more, see Article 36, 2020, 'Explicability' as a way to secure accountability, https://article36.org/wp-content/uploads/2020/12/Explicability-and-accountability.pdf