DISCUSSION PAPER | AUG 2019

Article36

Target profiles

An initial consideration of 'target profiles' as a basis for rule-making in the context of discussions on autonomy in weapons systems.

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

Summary

This paper argues that consideration of 'target profiles' can provide a basis for developing rules to preserve ethical standards and enable meaningful human control over systems that apply force on the basis of sensors. All systems that, during their use, independently identify and then strike targets must use target profiles, which are **the set of conditions under which such a system will apply force**. They are a technical necessity for the functioning of such systems, but they are also important as a basis upon which operators understand the utility and capabilities of a system. By applying rules regarding the parameters of target profiles, and obligations on the use of such systems, a framework can be constructed that has broad utility, is flexible in the face of future technological developments and can address the primary concerns raised around 'autonomy' in weapons systems.

This paper introduces a basic model of systems under consideration: systems where, during a period of time, sensor inputs are analysed by machine and, on the basis of that analysis, force is applied at a certain time and place. We situate that model, of systems that apply force to a 'what', in comparison to systems where a commander has stipulated specifically 'where and when' force should occur. The paper recognises target profiles as fundamental to the functioning of systems in the former category and then identifies key tensions that arise with respect to how a 'thing' is encoded as a target and how sufficient control is enacted where a commander does not know specifically where and when force will occur. From that we go on to consider the types of rules that might be applied to address such concerns. We also suggest questions for states which would be informative as to their orientations to this approach, noting that existing international law does not, by our reading, provide straightforward answers to such questions.

This paper concludes by highlighting the following structure as a basis for regulation, though it also suggests other rules and identifies alternative approaches in some areas:

- x The use of certain systems should be legally prohibited - such as systems that could change the conditions under which they will apply force, those where such conditions cannot effectively be understood, and those that would target people.
- x For all systems there need to be positive legal obligations to ensure that users understand and are responsible for the systems they use, and that the time, duration and location of any use is controlled such that the wider rules can be implemented effectively.

Whilst some such rules would provide a hard line against certain technological capabilities others provide obligations regarding how permissible systems can be used. This reflects a sense that this issue cannot be reduced to a problem presented simply by one set of technological 'objects'. Sufficient human control will always require some component of adequate human behaviour.

However, the focus of the paper is not so much on the assertion of specific rules as it is on framing the subject matter through a structure that enables concrete boundary questions to be asked. Such concrete questions can provide a basis for cutting through the current debate and developing a coherence of policy orientations amongst those that are committed to preserving the existing legal framework and avoiding the further automation of violence.

* This paper was written by Richard Moyes.

Introduction

International policy discussions on 'autonomy' in weapons systems have developed substantially since the first informal discussions of the issue were convened under the UN Convention on Certain Conventional Weapons (CCW) in 2014.[1] The discussions, now taken forward by a Group of Governmental Experts (GGE), continue to show a developing intellectual engagement with the subject matter and they have achieved broad agreement on certain key points. These include recognising that law is addressed at people, and so it is people not machines that must apply legal rules, and that human control (in some form) is necessary for the preservation of our ethical and legal frameworks.[2] However, international discussions on this issue are still complicated by divergent orientations to terminology and differing opinions on what is the real 'object' under consideration. The policy conversation requires simplification (which is, ultimately, the very role of policy making) but simplification is necessarily political in so far as it requires choices to be made and meanings to be stabilised.

Against that background, this paper suggests an approach to the subject matter that frames things in a new way, whilst intersecting with all of the key threads of the policy conversation to date. The approach suggested is in line with Article 36's previous policy writings on 'meaningful human control',[3] but here we approach the issue with a focus on the technology of the subject matter under discussion; not the physical structures of the technology (which may be diverse) but the technological process that is fundamental to the issue of concern. This paper is essentially an elaboration of Article 36's statement to the March 2019 session of the CCW GGE.[4] **The argument is broadly that the concept of the 'target profile' can provide a building block for definition of the subject matter and serve as a locus for the policy formulations or rules that should be applied to it**. The term 'target profile' has already been adopted within the CCW discussion, although it hasn't been discussed in any depth.[5] It is also used within the defence sector.[6]

At the political level this paper is somewhat at odds with the process of work now being pursued within the CCW. At the time of writing, that forum is preoccupied with developing consensus on further 'guiding principles,' whilst avoiding the sorts of political choices necessary for a more authoritative treatment of the subject matter. Despite discussions spanning some five years now, there has not yet been the coherence of policy thinking necessary to develop a treaty process amongst those that have some determination to protect shared societal interests in the future. It is anticipated that at some point such a group - of states, international organisations and civil society partners - will emerge. It is hoped that this paper might provide 'food for thought' towards the intellectual structuring of what such a group might do.

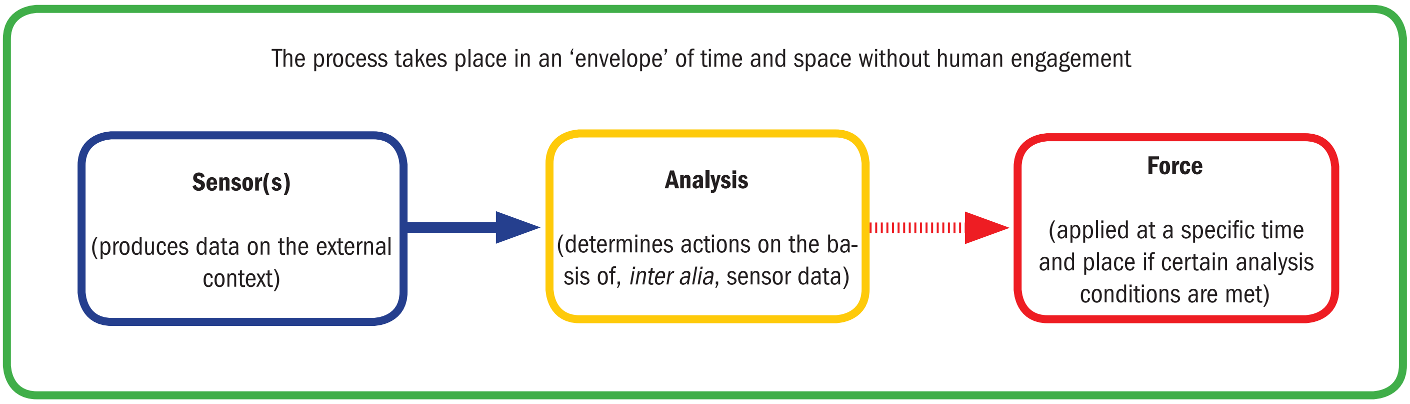
The issue under consideration – 'sensor-analysis-force'

Discussions on autonomy in weapons systems are challenging, in part, because different participants in the discussion have different starting assumptions and imaginings about what is being discussed. This has been spoken to explicitly in the CCW debate, with some stakeholders arguing that the proper subject of discussion is only very complex future systems whilst others maintain that certain existing systems would fall within the general object of study.[7] These differences often present themselves in different orientations to the term 'lethal autonomous weapons systems' (LAWS) – which is a central organising term in the international debate, but which is undefined and in many ways unhelpful.

This creates a familiar 'classification problem,' where political anxieties as well as intellectual analysis play a part in shaping different actors' orientations.[8] Certain states are likely wary of casting the net too wide for fear that some of their current weapons systems might be caught in any subsequent legal product.[9] Others are concerned that a narrower, futuristic orientation is difficult to approach intellectually without the building blocks that come from consideration of technologies in the present - and that such an approach may shield from critique technologies that it excludes from consideration. Article 36 has preferred the broader starting orientation and this paper continues that approach.

All conceptualisations of the 'issue of concern' in the international discussion relate to systems where, during a period of time, sensor inputs are analysed by a machine and, on the basis of that analysis, force is applied at a certain time and place. This sequence of technological transmission from sensor, through analysis, to the application of force at a certain time and place, without human action or intervention, underpins all of the framings of the issue under consideration. All of the differing policy orientations relate back to this. As a shorthand, we will refer to it as the 'sensor-analysis-force process' – and it is the effective management of that process that is the policy challenge addressed in discussions on autonomous weapons.

Figure 1: The 'sensor-analysis-force process' takes place without human engagement.



Some states have argued that there is a fundamental distinction between 'automatic' and 'autonomous' systems – arguing that the former has a greater degree of predictability or certainty regarding the system's functioning.[10] In the CCW, this distinction presents itself in disagreements regarding the starting orientation to the question of 'what is the object of study', including in different orientations to the term 'LAWS'. However, both of those labels are applied, by proponents of that distinction, to systems that employ the sensor-analysis-force process. A problem for those advocating for a narrower focus is that they have sought to establish a line of division here without having framed or elaborated the broader category that is being subdivided. This paper does not embrace the terminological distinctions between 'automatic' and 'autonomous' (because they are unnecessary for the line of argument here) – but we do see some of the arguments of 'difference' on the basis of predictability as important.

For the purposes of this paper, the key point is that the sensor-analysis-force process is implicitly recognised as central in the international discussion and, when made explicit, other policy issues and concerns can be organised in relation to that. A challenge in the debate to date has been an inability to stabilise this as a basis for classification – made difficult primarily by those who are seeking to establish sub-classifications before this conceptual foundation has been put in place.

A broad classification of weapons

There are many ways of classifying weapons – and classification processes all serve different purposes. Recognising that the 'sensor-analysis-force process' provides the central characteristic of systems under consideration in discussions regarding 'autonomy' we can consider where that model fits within a broader terrain of weapons systems, simply for comparative purposes. At a macro level, a broad categorical differentiation of weapon systems is sketched out below:[11]

- **Applying force to a 'where and when'**

  Systems by which force is applied at a time and spatial location that a human commander evaluates as being co-located with an intended target.
  This is broadly how traditional guns, artillery shells and aircraft bombs function. In this approach, the 'thing' to be attacked is encoded by a person directly as a time and spatial location and requires no other formal modelling for the process to function.[12] The 'thing' to be attacked may be anything – so long as it can be situated in time and space. In practice, there may be errors or uncertainties: faulty intelligence could result in a location being targeted erroneously, or mechanical failures could result in force being applied in the wrong place; systemic factors can also mean that it is difficult to apply force as accurately as might be desired. Computers may assist in the process, doing the 'aiming' for example, but fundamentally there is a human authority that determines that force should be applied 'there and then.'

- **Applying force to a 'what'**

  Systems by which force is applied at a time and spatial location derived from an analysis of sensor inputs, without human assessment of those sensor inputs and whilst operating within an envelope of space and time.
  This is the broad category within which the issue of autonomy in weapons systems is situated. Of course, numerous systems in this category already exist – including naval mines, landmines, certain guided missiles, missile-defence systems and sensor-fused anti-armour weapons. These existing systems have different constraints on their movement or duration of operation, but a common feature across all of them is that a human commander does not stipulate the specific time and place where force will actually occur. Rather the weapons use sensors and apply force when the input of those sensors corresponds to a pre-encoded profile of a target. This pre-encoded profile is fundamental to how all of these systems function and is the primary focus of this paper. The policy issue of 'autonomous weapons' responds to a recognition that technological developments open up new possibilities for 'how' a thing (whether a person, materiel or a phenomenon such as another weapon being fired) is encoded as a target, and greater scope for 'where and when' this process of sensor-driven force application may occur.[13]

Our general area of concern is weapons that apply force to a 'what' – such that force is applied when and where a system receives certain sensor inputs from its environment, and without human analysis of those sensor inputs. We have noted already that many such systems already exist. Some have been associated with significant concerns or have had constraints placed on them from an orientation of precaution: landmines for example are subject to various legal prohibitions, regulations and political policy constraints; naval mines are subject to specific legal obligations; systems that use 'sensor-fuzed' submunitions are subject, inter alia, to constraints on their numbers and obligations to curtail their active life.[14] Yet others, such as certain anti-missile systems, have not been considered excessively problematic (even if certain problematic incidents have occurred) and are clearly militarily valued by those who deploy them.

The policy movement to constrain autonomy in weapons systems responds to certain concerns present in society, and these are felt and formulated differently by different stakeholders. These concerns are generally precipitated by a sense that technological change and development may open up moral and practical hazards within this category of systems that use the sensor-analysis-force process. However, recognising these issues as an emerging concern does not imply that hazards arise only from futuristic capabilities that are technologically distant. Rather, hazards flow from the structures that characterise the category – how can things reasonably be encoded as targets, and how should we manage systems where the specific time and place that force will occur is not known?

So, it is not that all systems that employ the sensor-analysis-force process have always been a concern (though, significantly, many have), but consideration of that category is necessary in order to identify and address the concerns that current technological developments are evoking.

This allows us to move on to the central theme of the paper. Whilst recognising that systems that fall within this category can (and already do) take many forms, our simple sensor-analysis-force process gives us a basis for saying more about how such systems work. This is where the concept of 'target profiles' comes to the fore.

Target profiles

Systems in our category of consideration all use 'target profiles'.[15] **Such profiles are a set of conditions which result in a specific application of force being undertaken by the system**. Some if not all of those conditions must be fulfilled by inputs from sensors that produce data on the external environment, resulting in force being applied at a particular place and time.

In systems we are considering here, the target profile is encoded (either physically or electronically) in a way that allows for an application of force to occur without any human evaluation of the source of the sensor input. Such a method of functioning confers various perceived military advantages, notably: an ability to apply force to objects when their precise location at a particular time is not known to a human commander; an ability to apply force *if* an object is present in, or enters into, an area; to target objects at greater speed; to extend the potential for force to be applied over time; and to strike objects more precisely, or with a lesser amount of force than might be required by other systems.

If we consider the basic sensor-analysis-force process, target profiles are fundamental to the 'analysis' phase. Where sensor inputs match or fall within a target profile, and if other system conditions allow it, the system applies force – where sensor inputs are not matching a target profile, force is not being applied.

Target profiles can be considered a pattern of sensor data that is taken to represent a target. It is this mechanism that gives a system the capacity to apply force to a certain type of 'thing' within the time and space envelope that the process is afforded. Examples from existing systems include profiles based on weight (many landmines), heat-shape (certain 'sensor-fuzed' anti-armour systems), acoustic signature (certain torpedoes) and radar signature (certain counter-missile systems).

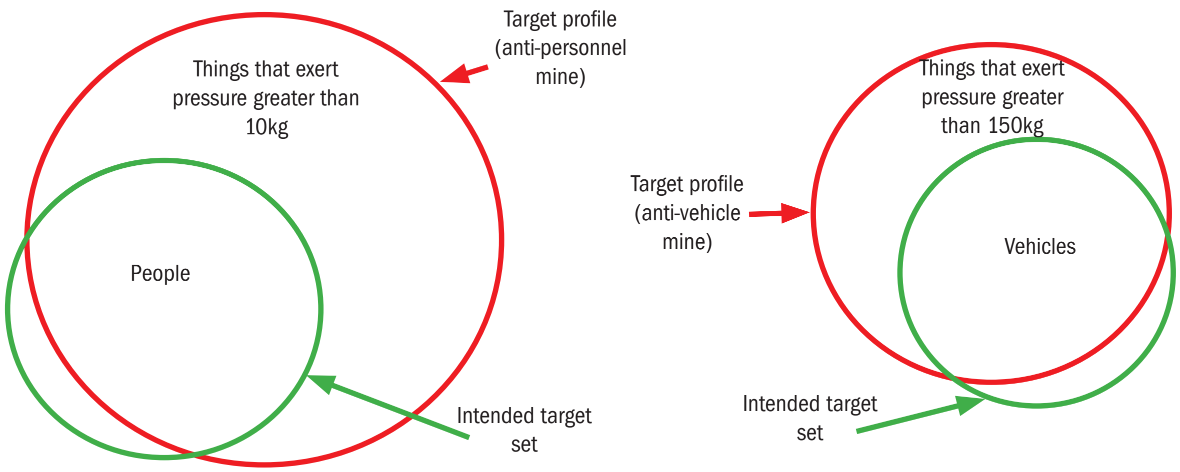
It is possible, in the future, that target profiles could be more complex – for example, factoring in previous sensor data, modelling certain contextual considerations, or actioning one type of force rather than another. There might be a number of layers of processing between the raw sensor signals and the assessment of different inputs against a profile. Conceivably, profiles could be constructed within the system itself rather than being pre-encoded with any fixity. This would seem to be in line with the futuristic definition proposed by Germany, where systems they are concerned with "evaluate the circumstances of a changing situation without reference to a set of pre-defined goals [and]...reason and decide on the most suited approach towards their realization."[16] This is proposed as a definitional boundary by Germany – effectively saying 'we should only be concerned with systems that define their own target profiles' (at least in so far as that could be assumed to a be a corollary of a weapon system setting its own goals.) Whilst that may be unsatisfactory as a starting point for discussion, it is notable that this can fit relatively easily into our conceptual model and, as discussed further below, it could and arguably should serve as one line of regulation.

However, any system within our broad category will need to utilise certain structures (whether pre-defined or developed within the system) that allow sensor inputs to result in an application of force. Without some such structure there is no way to move from 'sensing' to 'force application'.

The target profile functions as a technical mechanism for determining when sensor inputs should result in an application of force, but it also allows human operators to have certain expectations about the role and functioning of the system. It serves in part to translate the designed 'intent' behind a system into concrete technical form. The pressure activated anti-personnel mine provides a useful example.

As defined in the CCW, and later the Mine Ban Treaty, an anti-personnel mine is 'designed to be exploded by the presence, proximity or contact of a person...'. The legal definition does not elaborate on *how* that 'designed to be' is enacted and different antipersonnel mines enact it in different ways. A common target profile might be based on pressure, evaluated by compression of a spring, that will result in an application of force when a weight greater than 10kg bears upon the sensor area of the mine. Different models of mine will have slightly different specific pressure sensitivities, and others have utilised trip wires which will require a target profile based on the pulling force applied to a fuse. In these examples, the 'sensor'

Figure 2. Illustration of target profiles (red) and intended target sets (green) – anti-personnel mine and anti-vehicle mine – indicative only.



Note: the target profile of an anti-vehicle mine is 'smaller,' or more restrictive, than that of an anti-personnel mine because fewer things with which such systems might interact would exert a pressure greater than 150kg than would exert a pressure greater than 10kg.

is a pressure plate or tripwire, the 'analysis' is a simple evaluation of mechanical force bearing upon that sensor, and the 'application of force' takes the form of an explosive detonation. The target profile - in the first example - is easily understood as 'weight greater than 10kg'.

Remaining on the example of landmines, it is noticeable that whilst anti-personnel mines and anti-vehicle mines have a different designed intent, they are distinguished in practice by different target profiles (a greater weight is required to activate an anti-vehicle mine) and a different capacity for the application of force (a greater quantity of explosives). See figure 2 for an illustration of these target profiles in relation to the intended target set of these different weapons. Both the 1997 Mine Ban Treaty and the discussions on mines other than anti-personnel mines (MOTAPM) in the CCW have engaged with efforts to police the boundary between the different target profiles of these systems – the former in relation to anti-vehicle mines with 'anti-handling devices' and the latter in policy concerns over anti-vehicle mines with 'sensitive fuzes'.[17] Concerns regarding both 'anti-handling devices' and 'sensitive fuzes' revolved around mechanisms that might, in different ways, make the target profile of an anti-vehicle mine too close to that of an anti-personnel mine. **The point being made here is that international legal discussions have already accepted the relevance of how target profiles function for understanding the boundaries between different sub-classifications and as a locus of regulation.**

The example of the anti-personnel mine also brings to the fore an important corollary of how such target profiles function. It is immediately obvious that whilst anti-personnel landmines might be 'designed to be exploded by the presence, proximity or contact of a person' the actual target profiles described here will detonate under a variety of circumstances that are not in line with this designed intent. We will elaborate on this further in the section below.

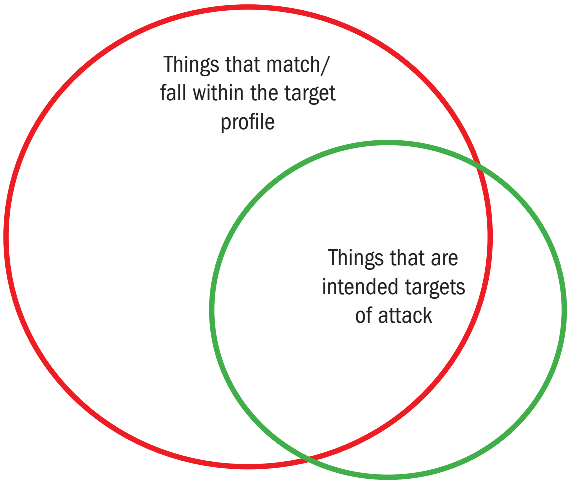
The intended and the unintended - what falls within the target profile?

Target profiles are constrained by the sensor 'language' that a system uses. For a system that only has a pressure sensor – pressure is the only way of encoding external objects as targets. In the case of the anti-personnel mine, this means that it is not only enemy combatants that will fall within the profile, but also other people, animals and vehicles that exert sufficient force on the mine.

Thus, from a human perspective, the target profile captures two sets of 'things' – those that may be intended objects of attack and those that are not. For different systems the relationship between these two sets may be very different. For example, a system programmed to recognise the specific acoustic signatures of individual enemy ships might be very finely tuned to the intended objects of attack, with few if any other objects likely to fall within the target profile.

The system itself, of course, does not make any distinction between these two sets of things – they all produce conditions that will result in an application of force. The distinction here arises from a difference between human intent and actual system performance. The target profile is always only an approximation of intent. Over time, and over a greater number of interactions with diverse objects in the environment, the difference between human intent and actual system performance becomes more and more likely to result in an unintended outcome.[18]

Figure 3. A relationship between a system's target profile (red) and intended target set (green).



Note: the degree of overlap between these two sets will vary between different systems. In some cases the profile might be matched very closely to a specific target object or phenomenon (which would result in these circles being very closely aligned), in others there are significant discrepancies – as we have seen with the example of anti-personnel mines. Managing the implications of such discrepancies is always one of the key challenges presented by systems that use target profiles.

In terms of the design and utility of weapons, there are potential tensions in this area between different interests and practicalities. Very narrow target profiles might only have a narrow set of applications, or might be vulnerable to false-negatives as a result of sensor data from an intended target object not aligning precisely enough with the target profile. Cruder profiles have a greater risk applying force to things that are not the intended target and may require more stringent controls on the context within which the system will function if that attack is not to be problematic or illegal.

For a human commander using any such systems it would seem to be important for them to understand what will fall within the target profile in order for them to evaluate the likely outcomes from the use of that system in a specific context; recognising that 'what falls within' includes both intended and unintended objects of attack.

For example, if a system will target the heat-output of vehicle engines in general, and a commander intends to attack enemy armoured vehicles, then the use of that system in an area where civilian and military vehicles are co-located could cause significant civilian harm. Being aware that the civilian vehicles fall within the target profile is a necessity if a meaningful evaluation of proportionality is to be undertaken. It is also necessary in order to evaluate if, in this circumstance, the weapon would be of a nature to strike military objects and civilian objects without distinction.[19]

It is notable, in this example, that for systems undertaking a single application of force, the probability of striking a military vehicle (and thus realising some military advantage) may be directly related to the number of civilian vehicles within the area where the system is functioning.

Whilst further legal analysis on this point should be considered, the key conclusion here is that **a commander would need to understand how the target profile relates to objects or phenomena within the area where the system will function, based on an understanding of what will result in that system producing an application of force.** It would seem wholly unsatisfactory for a commander to argue that they had destroyed the civilian vehicles because they did not know that civilian vehicles fell within the target profile of the weapon system they were using. This conclusion seems quite straightforward, but it reinforces the central argument of this paper – that the concept of the target profile is something around which expectations, obligations or rules can be organised.

The challenges presented by target profiles

Many of the challenges presented by systems that use target profiles as part of a sensor-analysis-force process we have covered in our previous writings on meaningful human control.[20] However, approaching the issue with a focus on target profiles brings certain issues more prominently to the fore. The section below will summarise the main sets of challenges, some of which relate to the **morality** of processing people through such a system and many of which relate to the well-established problems of '**control**'.

It is important to note that, approaching the issue from this starting point, all of the key challenges can be seen to derive from the way in which target profiles function. It is this reliance on an external, codified representation of a target that creates both the military opportunities and the humanitarian tensions associated with this policy concern. We consider first *how* something is encoded as a target – are all ways of modelling targets acceptable? – and what does a commander need to understand about that process? Building on this, we turn to the suggestion that systems might not only model targets, but also things to which force should not be applied. We then consider the duration and space over which the utilisation of a sensor-analysis-force process can be considered reasonable. Finally, we look at how the number of applications of force, and the nature of force, also has significant implications for managing the challenges of systems that use target profiles.

After this, in conclusion, we will sketch out the type of policy rules that present themselves as options from this form of analysis – and frame questions for states, that would help to build understanding of national orientations to those policy options.

How a target is encoded as a profile

The process of translating 'intent' into a target profile requires certain sensor-identifiable characteristics of the intended target to be taken as a proxy for the target or target type in question. Typically, the intended target for the purposes of weapons design is a 'type of thing' rather than a specific object – a category, or set of categories. The target profile then serves as a classification tool, dividing the world between 'targets' and 'not-targets'. In basic form, this process involves a reduction of the human concept of a target into a pattern of sensor signals. The weapon system does not then attack things as they would be understood in the human sense, it simply attacks the source of certain patterns of sensor data.

This is significant when it comes to considering the use of such systems to target people. A **system that targets people *per se*** represents a mechanism by which people are reduced to certain data points and then killed or injured, not as 'people' as a human would understand them, but simply as producers or possessors of certain sensor-identifiable characteristics. As a moral concern and from an orientation of societal precaution, continuing to allow systems that apply force to people on the basis of such a reduction to data-points risks endorsing a further dehumanisation of conflict and other violence. Such concerns are relevant to civilians and soldiers alike. The ICRC have noted that "it matters not just *if* a person is killed or injured but *how* they are killed or injured, including the process by which these decisions are made", noting that (absent a sufficiency of human control) such processes may "[undermine] the human dignity of those combatants targeted, and of civilians that are put at risk as a consequence of legitimate attacks on military targets".[21] It could be noted here that our mode of arguing against target profiles intended to identify people *per se* is not the same as arguing against harm to people as an outcome of a systems' functioning.

In practice, any system that would target people *per se* might only be permissible under existing rules relating to armed conflict in situations where civilians had been effectively excluded from the area where that system would function and for the duration of its functioning (though this is not explicitly agreed amongst states). Otherwise the system could be considered to be of a nature as to strike civilian and military people without distinction. Controlling the context of use then becomes the only mechanism by which the rules of distinction can be applied by a human commander (we return to this further in a separate section below). Controlling the context of use does not, however, address concerns regarding the reduction of people (including combatants) to objects within the system's process, nor would it address the hazards posed to soldiers rendered *hors de combat*, who could not legally be targeted if a human were making a decision at that time.

Concerns regarding the targeting of people arise in a different form if it is proposed to distinguish between groups of people on the basis of some sort of **collective identity markers**. The most obvious theoretical proposals in this area involve asserting that systems might target 'enemy combatants' rather than people in general. Again, this does not alleviate any moral concerns about reducing soldiers to objects as a basis for killing or wounding them. It also presents significant problems regarding the relationship between the target profile and actual intended targets – as discussed in the previous section. Given that 'combatant' identity is mutable it is hard to see how any markers relating to that identity would be reliably valid or wouldn't present very severe risks of false positives from the appearance or behaviour of civilians, or of combatants who cannot lawfully be targeted. It would likely be an extreme, if not impossible challenge to assert the sufficiency of any such process in a situation where civilians might also be present. **Adopting such a system would seem to displace too much of the burden of implementing the existing legal obligations (which bear on humans) onto human assumptions about the functioning of a technological system.**

Maya Brehm has highlighted that constructing profiles on the basis of perceived patterns of behaviour is already fraught with risks. She notes that, "the categorization of individuals based on correlations and inferences always entails a certain error rate. The very presupposition that 'relevant circumstances can be rendered algorithmically, and still adequately, as 'patterns of life'' is of course questionable. The approach can also be criticized for its dehumanizing quality in that 'it tends to reduce the person to the profile generated by automated processes which are liable to be used as a basis for decision-making'".[22] Notably here, it is not just the reduction of people to a profile that is problematic, but also the reliability of an identity attribution based on patterns of behaviour and the process by which such profiles might be generated.

Other collective identity markers could relate to physiological or social indicators of **age, race, gender or other social identities**. It might seem straightforward to many that a weapon system should not identify people as targets on the basis of indicators that pattern these characteristics. However, the recent discourse relating to casualty recording – which saw the suggestion that 'military-age' men could be assumed to be combatants in evaluations of the impact of armed drone strikes in the practice of the USA[23] – illustrates how close certain bureaucratic structures already are to accepting markers such as these as sufficient to categorise someone as a target. After all, if people could be categorised as combatants on this basis when they present as casualties (resulting from their proximity to a weapon being used) why not also as targets by a weapon that is, necessarily, proximate and in use? It is also notable that international humanitarian law does not make explicit that factors such as age, race, gender or other social identities cannot be used as a basis for identifying people as combatants in international armed conflict.[24] That said, using such identity markers alone would appear to be insufficient to meet legal obligations. However, it is not implausible that some would argue that certain social identity markers, in a certain context, provide a sufficient basis for assuming people to be combatants.

Even if identifying people as targets explicitly on the basis of such markers may be unacceptable to many, there is a further challenge relating to the use of **data sets and machine learning** in any construction of target profiles. Given an established recognition of the prevalence of **biases in algorithmic processing**,[25] care would need to be taken to ensure that target profiles have not been constructed in a way that results in people being classified as targets, or being more likely to be classified as targets, on the basis of such collective identities. Developing overarching rules to avoid that might require us to assert, as a starting point, the impermissibility of explicitly targeting people on this basis, in any context. This also brings us on to the set of challenges resulting from the use of target profiles that we will look at below, the capacity for commanders to understand how they work.

However, before turning to that, we should note again that practical challenges regarding attributing identities to people, and challenges of bias derived from collective identities, are both layered on top of a process of reducing people to data points as a basis for killing or injuring them. Moral concerns evoked by such a process may be reinforced by appeals to the public conscience. This in turn can be supported by an argument that it is simply not in the best interest of our shared society to allow machines to target people. It is not the purpose of this paper fully to elaborate the lines of argument in this area, but a straighforward prohibition on profiling people as targets would have a compelling simplicity and significant social utility.

Understanding the target profile

As we have noted already, in the section on 'the intended and the unintended,' it seems necessary that a commander understands what will fall within a target profile, where 'what falls within' includes both intended and unintended objects of attack. It might be conceded here that an operational commander would not necessarily be expected to know the specific values of different sensors that constitute the conditions whereupon the system will proceed to apply force. The primary requirement is that they understand the practical implications of those technical processes – and, vital to this, that the testing and evaluation process behind the system has done a sufficient job of identifying and communicating the conditions that will result in an application of force.[26]

Returning to the example we used previously, of a system that identifies the heat-shape of vehicle engines in general – it would not seem appropriate to test that mechanism of target identification only on certain armoured vehicles and then to convey to an operational commander that the system only targets armoured vehicles, without undertaking any evaluation of what else might fall within the target profile. So, drawing on adequate testing and evaluation processes, an operational commander should have a sufficient understanding of all of the things that can be expected to trigger an application of force, whether they are intended objects of attack or not.

As was suggested in the section above regarding bias in datasets, the process by which a target profile is technologically constructed can also have a bearing on how readily it is understandable. If 'machine learning' processes are used as a basis for object recognition, it may not be possible for a person to understand or interrogate the specific characteristics or features of the object that are providing the triggers for recognition. This would make it very difficult to determine what situations might produce false-positives and so to have any meaningful understanding of what things might fall within the target profile but which are not intended objects of attack. It is notable from such machine learning processes that the sensor-identified characteristics that do produce false positives might not present any humanly perceivable 'similarity' to target characteristics that we might consider significant. The challenge here is that target profiles need to be sufficiently transparent in their functioning if an operational commander is going to understand their implications in an operational context.[27]

A similar challenge of transparency would arise for systems that develop or change their target profiles after their operation has been initiated. For example, if a target profile were developed within a system after it has been put into operation, based on various abstractions and assumptions from its operating environment, it would be hard to argue that the operational commander was in a position to predict the outcomes of the system's operation and so to form a reasonable legal judgement beforehand. Similarly, if a system were to develop or refine pre-encoded profiles after it had been put into operational use, any prior legal judgement by a commander would not have been based on the actual operational parameters of the system. This would also seem to be both practically and legally problematic. States within the CCW have generally argued that a 'weapon review' process would need to be undertaken if a system was changed in a way that was significant in terms of its operational functioning.[28] The development or change of electronically encoded target profiles, undertaken by and within a system where the potential for such internal change had been accepted and agreed by a prior weapon review, could be argued not to be a change of the system itself. However, accepting such a capability within a weapon system and then arguing in this way would be highly erosive of the intent and purpose of such review processes. After all, what is to be classified as a target, and how, would seem to be critical to a practical and legal understanding of the system.

This is not to say, of course, that the target profiles used in a particular system could not be changed *between* uses, so long as those target profiles had been approved for use in that system and were understood by the operational commander.

In the international discussion on autonomy in weapons systems a significant number of states have asserted the importance of weapons review processes, as required under article 36 of Additional Protocol I of the Geneva Conventions.[29] Such reviews would be a very important component of any policy or legal responses seeking to regulate on this particular set of challenges. If it were required that weapons systems only use target profiles that are sufficiently understandable, for example, this would likely need to be evaluated through national weapon review mechanisms. It is notable, however, that the legal obligation to undertake weapon review processes only requires an evaluation of weapons, means and methods of warfare in relation to established legal obligations. Many of the central issues outlined in the section above, including the example of a system developing its own target profiles, are not subject to explicit legal obligations. With weapon reviews a purely national level mechanism, such a mechanism cannot be considered sufficient to address concerns on these points, or to build shared normative understandings, without further legal articulations.

Profiles of ‘what is not to be targeted’

An additional set of issues, related to target profiles themselves, is the suggestion that such profile structures might be used also to identify things that mean that **force should not be applied in a particular situation**. The suggestion of such a capability has arisen a number of times in the international discussion, generally as a hypothetical positive potential for technological development in this area. Such a capacity would require a system to detect multiple object types in its environment - though technically those multiple objects, in machine analysis terms, could be physically co-located in the real world (to draw on an example we will explore further below, such objects could be 1. a vehicle heat-shape and 2. a distinctive medical emblem). Based on an analysis of these objects, the system would make an evaluation of whether something falls within a target profile or not. Some of the more futuristic visions of autonomous weapons systems build upon this sort of concept of operation.

At a basic level the process could be represented as:

> Target profile is x, unless the proximity of y or z...
>
> Where x, y and z are all sensor-identifiable profiles of external objects or phenomenon and where ‘proximity’ also needs to be encoded.

The suggestion has been that such a mechanism could serve to reduce civilian harm by enabling weapons to identify certain ‘things’ which mean that force should not be applied in a particular location – ‘things’ such as civilians or protected objects. Such proposals present certain challenges. As noted above, a common, crude suggestion is the notion of a weapon that self-deactivates if it detects an ambulance in its target zone, and we will return to that in our comments below.

Firstly, such a structure requires formalising and encoding of what is to be considered a civilian, a civilian object or a specially protected object. This means highlighting only certain, sensor-identifiable things as being protected when the established legal obligation is that *things in general* should be assumed to be civilian (and therefore should not be subjected to attack) unless there is evidence to the contrary.[30] This shift erodes the fundamental presumption of civilian status and moves the burden onto civilians to identify themselves as such, whilst moving the burden of responsibility off human commanders. In this example, such a process may also put at risk the status of distinctive emblems, such as the red cross or the red crescent, if it becomes known that such markings might automatically prevent certain weapons from striking in an area.

Constructing profiles of protected objects presents the same challenges that we have noted previously regarding target profiles. Some things might fall within such profile that were not intended to result in force being withheld – reducing military effectiveness. Other things that a profile was intended to capture could sometimes fall outside it, for example if they were not presenting the signatures that the profile needs to match – producing unintended harms. After all, many types of vehicles might be used as an ambulance, with different visibility of markings in different conditions.

A commander would need to understand these multiple profiles in their assessments about the use of a system – in order to understand the likelihood of achieving their military effect if nothing else. Operating in conjunction, the two sets of profiles would create multiple layers of unpredictability regarding what a system could be practically expected to do in a specific operational context. At anything other than a most basic structure such an approach would quickly become similar to a target profile that changes during the process of use.

If the process was intended to identify protected objects in proximity to an intended target (not just to refine the identification of targets themselves) then this would also imply an effort to diminish collateral damage risks. This, in turn, implies that not only is some aspect of the human responsibility to distinguish a military target being displaced into assumptions about system performance, but also aspects of the obligation of assess and minimise incidental harm. So as well as eroding the presumption of civilian status, such an approach also begins to erode a wider set of human legal responsibilities.

Finally, such a mode of functioning offers the prospect of the machine taking up the burden of dealing with uncertainties or absences of information. It may make the use of force in situations where information about the context of use is lacking seem more permissible. Rather than reducing the number of iterations of force, the perceived permissibility of using a system in more situations might increase the number of attacks that are undertaken. Thus, what might on first sight appear a positive potential for increasing civilian protection, presents significant risks of undermining civilian protections more generally, increasing unpredictability, eroding human legal responsibilities and enabling a wider use of force.

However, other hypothetical examples could be less contentious. For example, a ship-mounted missile defence system that was able to automatically stop firing if it identified that the radar signature of an aircraft would pass across its line of fire may present many of the challenges noted here, but in a manageable form. So long as this capacity were not taken as a justification to keep the system operating continuously or with diminished human oversight, the comparatively controlled context, with limited and identifiable ways in which generally protected objects are likely to present themselves, might be amenable to effective control.

Beyond the added layer of complexity and unpredictability, much in our evaluation of these examples hinges on whether the structures of human analysis, judgement and oversight remain the same or are diminished in response to a technological capability being adopted, and whether force is now considered to be acceptable in an expanded set of circumstances, over a wider area or over a longer period of time. As a general orientation, in order to ensure such mechanisms function as a constraint on force rather than as an enabler of the wider use of force, it could be argued that operators should not assume the effective functioning of such profiles in their evaluations of the incidental loss of civilian life, injury to civilians, or damage to civilian objects that might be expected from the use of a system.

Controlling the space, duration and time of the 'sensor-analysis-force process'

Whilst the previous sections have highlighted challenges that relate to the types of target profiles used, this section summarises themes that arise with all systems that use target profiles in a sensor-analysis-force process – themes that are increasingly accepted within the international discussion.

It is widely recognised now that a key requirement for meaningful human control is that systems operating a sensor-analysis-force process do so within a physical space and over a duration of time that is in some way set by a human commander.[31] The CCW Chair's draft report on the 2019 GGE session noted that "human operators and commanders need to understand, inter alia, the operational environment, since the use of force is contextual, and how the weapon system is likely to interact with the operating environment, in order to be able to ensure their use of force is consistent with applicable international law."[32] Given that a commander needs to make legal judgements based on an anticipation of the interaction of a system with its operational context, there needs to be some bounding of that context in space and time in order for such judgements to be substantive. The wider the physical area, and the longer the duration of operation, the less detailed the information a commander will likely have regarding that area, and the less predictable that system's use will be. It is also recognised that different operational contexts and different system configurations may allow a system to function more or less predictably - and so the area and duration of sufficiently predictability may be different for different systems in different contexts.

Article 36 has previously tied this requirement for some bounding of the time and space of operation to the concept of 'an attack' in the structure of international humanitarian law. If human legal judgements need to be applied to each attack – as the law is written – then an attack must have some spatial, temporal and conceptual boundaries if the structure of the law is not to be undermined.[33] Similarly, if legal judgements about the use of force are widely understood to be made 'contextually', on a 'case-by-case basis', then there needs to be a specific context, and a specific 'case'. Systems should not be allowed to proceed from attack to attack without the application of distinct human legal decisions, and multiple, distinct military objectives should not be amalgamated into one single objective.

For the sake of clarity, we can note that mobility or 'range' of a system are factors in generating the physical area in which a system operates. Similarly, self-destruct or self-deactivation mechanisms, as well as any capacity for an operator to intervene to stop a system functioning, or to re-task it, are all mechanisms that control or constrain the duration of operation (and through that potentially also its range). Conceptually, it is worth nothing that what we are concerned with here is the space and duration over which the sensor-analysis-force process operates. This is not necessarily the same as the range or duration of a physical system (which might be instructed to apply that process only for certain portions of its operational functioning).

A similar and compelling line of argument is that the human legal judgement needs to be made sufficiently proximate in time to the sensor-analysis-force process of the system for that judgment to be relevant and, again, substantive. The ICRC noted in March 2019 that, "assessments of distinction, proportionality and precautions made by combatants must be reasonably proximate in time to the attack (or 'strike'). Where these assessments form part of planning assumptions, these assumptions must have continuing validity until the execution of the attack in order to comply with IHL."[34] If the judgement is made too far in advance then circumstances within the operational area may have changed significantly by the time the sensor-analysis-force process is put into effect.

Ultimately, and as is implied in the ICRC statement noted above, the sensor-analysis-force process and the target profiles employed within that, represent a set of assumptions about how force can be applied. The time at which that process begins to operate, its duration and the area over which it extends need to be known, understood, and bounded such that those assumptions can be reasonably expected to remain valid. The more outdated the information upon which human judgements are based, the wider the area and the longer the duration of the sensor-analysis-force process, the more complex the context, the greater number of actual applications of force, and the greater the destructive power of those applications of force, the more chance there is of prior assumptions not being valid and of unintended consequences occurring. Controlling the space, duration and time of any such process are effectively mechanisms to allow humans to have adequate information about the context that these parameters enclose – such constraints do not, of themselves, ensure adequate information which must still be driven by people taking adequate operational and legal responsibility.

Controlling conditions in the context of use

Mechanisms for controlling the space, duration and time of a sensor-analysis-force process' operation are tools by which informed judgements can be made about the circumstances prevailing in the context within which that operation will occur. An additional approach to that challenge is to apply certain obligations to control the actual conditions in that context of use.

We have noted previously that certain systems that use the sensor-analysis-force process are already subject to specific legal regulation. Anti-personnel mines, specifically, are prohibited in the 1997 Mine Ban Treaty. However, another form of regulation in the CCW requires certain landmines only to be used in "a perimeter-marked area which is monitored by military personnel and protected by fencing or other means, to ensure the effective exclusion of civilians from the area."[35] Such an approach was also reaffirmed later in a "Declaration on Anti-Vehicle Mines" tabled by 25 states in the CCW in November 2006.[36] This opens up a different mechanism for managing the challenges presented by sensor-analysis-force processes, which is an exclusion of those who are at risk of being mistakenly targeted from the area in which those processes are operating. This again fits with the conceptual structure outlined in this paper – whereby measures to ensure the effective exclusion of civilians are a further means by which the circumstances prevailing in the context of operation can be controlled (so as to manage the tension between the actual target profile and the intended targets of attack).

Historically, various practical challenges with marking and monitoring have been identified, particularly in the context of systems that continue to function if that marking and fencing becomes degraded or military monitors are forced to flee. More broadly, there is a danger that such approaches contribute to shifting of the burden onto civilians to be absent from the area where weapons are to be used.

However, controlling who or what is within the context of operation though active management efforts, not just through situating and limiting the context of operation to a certain area and duration of time, could provide an additional locus for regulation for certain systems. After all, systems that use profiles that will capture both civilian and military objects, used in a context where such objects are intermingled, could apply force to those objects without distinction.

The 'quantity' of force

Finally in this section on the challenges presented by target profiles we should note that the quantity of force to be applied also has a significant bearing. Where target profiles present some risk of generating false positives – i.e. applying force to things that are not the intended object of attack – then multiple iterations of that process increase the risk of unintended effects being produced.[37] Simply, where a system presents a degree of unpredictability, empowering such a system to apply force numerous times simply increases the risk of serious negative effects.

As well as the number of iterations of force being a factor, the scale of each such iteration is also significant. An explosive warhead that projects blast and fragmentation across a wide area will typically present a level of direct harm to surrounding people and objects, and of longer-term, indirect harms, that is significantly greater than a single 'explosively formed projectile' or a bullet. This scale of force has a bearing on the commander's responsibility to understand the conditions in the area immediately surrounding an application of force, which is made much more difficult given that, in the systems under consideration here, the specific time and location at which force will occur cannot be fully known. The United Kingdom noted at the March 2019 GGE of the CCW that "emerging technology offers opportunities to address some of [the humanitarian concerns regarding the use of explosive weapons in populated areas] by improving the fidelity of military decision making, improving situational awareness and offering lower-yield and higher-precision weaponeering options for operations in an urban or access-denied environment."[38] That may be the case, but there are also no clear boundaries as to the scale of explosive force that states might consider permissible in the use of sensor-targeted systems, in populated areas or elsewhere.

Both of these characteristics have the potential to result in an attack being indiscriminate under existing legal rules. Constraining the number applications of force, and the scale of force, that systems may undertake under an individual attack, that is subject to human legal approval, is difficult without recourse to metrics that are perhaps inevitably arbitrary. There is potential to draw clear lines on such a basis, but given the broad range of current and future technologies that fall into this category it may be better, at this stage, to emphasise general principles. As with space and duration, the number of applications of force and the scale of force that can be applied should be understood and controlled by a human commander such that they can meet their legal obligations.

What types of rules might be applied around target profiles?

As we have noted, target profiles can provide a structure around which certain rules might be articulated. The utility of this concept as a locus for regulation comes from the fact that it operates at a general technical level and from its role in mediating between the designed purpose of a system and its technical functioning.

Following the broad structure of the preceding sections, we suggest below types of rules that could be adopted to address the concerns raised in these areas. We recognise, of course, that different actors might be more or less enthusiastic about specific rules, or about the prospect of any rules at all. In the section below we suggest different types of rules that could be applied without being too specific about how they should be formulated. The point is to illustrate that target profiles are amenable to rule-setting, and from that to create space for a more grounded dialogue about what those specific rules should be. We then go on to suggest some questions that could be asked of states to assess their orientation to these sorts of rules.

Definitional

Target profiles and the sensor-analysis-force process could be used as a basis for bounding a set of technologies to which specific rules are being applied. Because the target profile is part of the functioning of such systems (as the conditions that result in an application of force), rules applied within that definitional boundary would relate to the functioning of systems, or to their use, rather than to those systems as physical 'things.' A definitional boundary based on system functioning could relate both to systems that are unified in a single physical entity, or dispersed – with sensing, analysis and force application being based on different platforms. It would provide an approach to definition that is 'neutral' in the sense that it is describing factual features of how such systems work. Article 36 continues to consider that all systems that, during their use, independently identify and apply force to objects, should be subject to legal rules that preserve our ethical standards and enable meaningful human control.

Obligations relating to the target profiles themselves

Within the space provided by an initial definitional boundary, certain rules could be applied to how target profiles function. Because target profiles function as both a mechanical feature of such systems (as the system conditions that result in force being applied) and as a tool for designers and operators to use systems towards a certain 'intent', then such rules can be articulated along a number of different lines. However, it should be noted that the options here are formulated towards controlling such systems in the context of hostilities in armed conflict. International human rights law may not be compatible with the leeway afforded by some of the options here.

- It could be prohibited to use target profiles that represent people in any form;
- Alternatively, and presenting significant weaknesses, it could be required that for systems that use target profiles that capture people in any form that this only occurs where measures have been taken to ensure that civilians are effectively excluded from the area where such use would occur.

A clear prohibition on targeting people within a sensor-analysis-force process would best be argued on the basis of the morality of reducing people to data-points and from an orientation of societal precaution – i.e. regardless of what may be argued regarding the military utility of such systems, expanding the potential for systems to kill people based on sensors would be very negative for our shared future society, and for soldiers and civilians alike. Such arguments can be coupled, in turn, with appeals to the public conscience and a recognition of significant practical risks associated with partial responses.

- It could be prohibited to use target profiles that categorise targets solely or partly on the basis of characteristics associated with people's age, race, gender or other social identities; and prohibited to use profiles that are developed in a way that could inadvertently have this effect.

This would be redundant in the context of a prohibition on the use of such systems to target people. However, if states are not prepared to adopt such a prohibition then it would be hard to argue against the adoption of a rule such as this, which is also not explicitly stated in existing rules regarding international armed conflict and which is important in the context bias in data sets and machine learning.

- Where target profiles may capture both civilian and military objects, it could be prohibited to use such systems where such objects are intermingled or it could be required that measures have been taken to ensure that civilians are effectively excluded from the area where use would occur.

Where civilian and military objects both fall within the target profile, the use of a system where such objects are intermingled presents a significant risk of striking such objects without distinction. Specific rules could be established to recognise and prevent this. This risk may apply to systems currently in use by certain states – but in the CCW context, at least, they have not offered their legal reading of these points.

- It could be prohibited to use target profiles that can change during the process of operational use, or that are developed or defined within a system during such a process of operational use.

This would follow from the requirement of a commander to understand the system that they are using such that they can make a substantive legal judgement about its use in a given context. A rule such as this would prohibit the sorts of advanced systems that Germany has expressed concerns about (such that the German Foreign Minister has called on states to prohibited them).[40]

- There could be a positive obligation for a user to understand the target profiles of the systems that they use, including objects that might fall within the target profile but which are not military objectives.

A rule such as this would reinforce what might be considered implicit in existing law – that parties need to understand the weapons that they use. However, made explicit it has implications regarding how target profiles might be constructed, such as through machine learning. This could be reinforced through an explicit reference for such an understanding to be established and communicated internally through weapon review processes.

- There could be a positive obligation for users to ensure, in the use of a system, that legal judgements do not assume a system will, through its technical functioning, identify and avoid incidental loss of civilian life, injury to civilians, or damage to civilian objects.

Formulated in this way such a rule would ensure the burden remains on human commanders to make legal determinations and to fully assess the risks to civilians from the use of a system. This is important in relation to the obligations of article 57 of Additional Protocol one which requires those "who plan and decide upon an attack" to "do everything feasible to verify that the objectives to be attacked are neither civilians nor civilian objects and are not subject to special protection." Assuming that a machine will subsequently distinguish between what is combatant and civilian would be an erosion of human legal responsibility.

Whilst we have generally rejected anxieties that regulation on this issue risks some unintended constraint on civilian technological development, it could be noted that by using target profiles as the locus of the rules above there is no risk of them having any such implications.

Obligations to control the space, duration and timeliness in the use of any systems that are not prohibited

Beyond rules regarding the target profiles themselves, certain requirements would still be necessary regarding the use of any systems that confirm with those rules. As formulated below, there is a recognition that different systems and operational contexts may have different implications, but a system's functioning always needs to be contained such that human commanders can apply the law.

- There could be a positive obligation to ensure that the spatial area and duration over which a sensor-analysis-force process can occur is set or controlled such that a human commander can fulfil their legal obligations in relation to an attack.
- There could be a positive obligation to ensure that the time at which a sensor-analysis-force process occurs will be sufficiently proximate to the application of human legal judgement for that legal judgement to be relevant to the circumstances prevailing at the time of an attack.

Both of these rules are quite open in their terms, but they again ensure that the burden of responsibility falls on those that use such systems to control the functioning of these processes such that legal judgements are substantive and relevant. We emphasis the applicability of obligations to 'an attack' in order to bring to bear the structure of international humanitarian law, which requires human legal determinations in each attack.

Obligations to control the quantity of force

Linked to the obligations suggested above, it should be recognised that the number of applications of force a system can undertake, and the scale of that force, can significantly expand the risks associated with the sensor-analysis-force process.

- There could be a positive obligation to ensure that the number of applications of force a system may undertake within an individual attack are set or controlled such that a human commander can fulfil their legal obligations.
- There could be a positive obligation to ensure that a commander understands the actual force effects that a system may apply in a specific attack, and makes their legal judgements based on that understanding.

As with the suggestions regarding space and duration, these rules are open and broad in their terms – emphasising human responsibilities rather than establishing specific proscriptive limits.

The purpose of the section above has been to highlight the sorts of rules that could be developed based on the approach taken in this paper. There are no doubt other rules that could be proposed, both more restrictive and more permissive. However, as suggested here, a relatively small number of distinct obligations can create a framework for maintaining meaningful human control over weapons systems in the future. Such an approach would need to be supported by a commitment to national level weapon reviews, and should not close off consideration of further obligations in response to particular challenges that might arise. As with any future oriented instrument, such rules would need to be brought to bear socially and politically over time in order to have normative effect.

Questions for states

A significant number of states in the CCW have asserted that existing international law is sufficient, as it stands, to ensure the acceptability of future weapons systems. Yet despite this shared confidence in the sufficiency of existing law, it is not clear that these states share the same understanding of what the law might require. Other states do not believe that existing law provides clear answers to all of the questions raised by technological developments in this area.The questions below are suggested as tools for national-level dialogue and are based on the sort of rule formulations we considered in the previous section.

For systems that use target profiles to determine where and when to apply force, would the following be acceptable or unacceptable?

- Target profiles that are designed to identify humans on the basis of human biometrics (i.e. on the basis of the target simply being human)?
- Target profiles that identify different groups of people on the basis of perceived racial, gender or age characteristics?
- Target profiles that change or develop, within the system, after it has been activated and without being specifically certified by a human?
- Target profiles where a human commander understands what it is *intended* to target but does not know the actual physical/emission characteristics of objects that the profile will match against – such as a profile built through current neural-network/machine learning?
- Target profiles where a human commander does not have an understanding of what things, other than intended targets, might fall within the profiles?

For systems that use target profiles to determine where and when to apply force, are the following assertions reasonable?

- Human commanders should be fully responsible for verifying the risk to civilians from the 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 area over which a sensor-targeting function can occur should be controlled such that a human commander can fulfil their legal obligations;
- The duration over which a sensor-targeting function can occur should be controlled such that a human commander 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 a human commander;
- Human commanders need to understand the actual weapon effects that such systems will create.

Conclusions

There are diverse challenges posed by the use of new technologies in weapon systems. Some of these challenges require states to set aside ambitions for national military advantage in order to protect the shared interests of our international society. A wider reduction of people to objects, and a wider application of force by machines operating at a level of abstraction from increasingly remote human judgements, threatens a further dehumanisation of our relationships with each other. Conflict and violence are always a product of certain human social failings. We should resist suggestions that such failings can be 'cleaned up' by handing greater authority over to machines to determine when, where and against what force is applied.

Retaining meaningful human control over the use of force requires us to attend to the morality of how things are encoded as targets, for the users of such systems to understand those encodings and their implications, and to ensure that they are only used in areas and for durations such that those implications can be anticipated and acceptably managed. Allowing systems to undertake applications of force across wider areas and longer durations risks eroding the notion of 'an attack' as the unit around which human legal application is demanded in armed conflict. Those that assert that international humanitarian law is already sufficient to manage such implications should consider the questions posed at the end of this paper. Failure to effectively codify our collective orientation to such questions risks allowing human legal authority to be progressively subsumed within technological and bureacratic structures from which scant real humanity may subsequently emerge.

END NOTES

1 For background and documents relating to that stream of discussion, from the first informal discussions to meetings of the Group of Governmental Experts in 2017-2019, see https://www.unog.ch/80256EE600585943/(httpPages)/8FA3C2562A60FF81C1257CE600393DF6?OpenDocument

2 Whilst not a consensus document, the GGE Chair's draft *Conclusions and Recommendations* dated 14 June 2019 provide a summary of sorts. The latest consensus document is CCW/GGE.1/2018/3, *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, dated 23 October 2018, online https://undocs.org/en/CCW/GGE.1/2018/3

3 See for example *Key elements of meaningful human control*, Article 36, April 2016, http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf and Richard Moyes and Heather Roff, *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons*, Article 36, April 2016, http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf also, notably, *Killing by Machine: Key issues for understanding meaningful human control*, Article 36, April 2015, http://www.article36.org/wp-content/uploads/2013/06/KILLING_BY_MACHINE_6.4.15.pdf. Many of these papers use the term 'proxy data' in a way which is synonymous with the term 'target profile' in this paper.

4 Article 36 statement to the CCW, March 2019, on agenda item 5c, Further consideration of the human element, http://www.article36.org/other-issues/definitions/article-36-statement-on-human-control-to-the-un-discussions-on-autonomous-weapons/

5 In their statement March 2019 statement to the GGE, on agenda item 5(d), the United Kingdom noted: "Specific constraints or parameters that might be placed on a machine's freedom of action might include limiting the target sets or profiles – defined by appropriate decision makers during the targeting cycle – which a machine can prosecute without additional human input" (https://www.unog.ch/80256EDD006B8954/(httpAssets)/85A4AA89AFCFD316C12583D3003EAB3E/$file/20190318-5(d)_HMI_Statement.pdf). The term appeared subsequently in the Chair's draft *Conclusions and Recommendations* where in the 14 June 2019 version, at C3(g), it stated that, "Setting operational constraints on systems, regarding tasks, target profiles, time-frame of operation, and scope of movement over an area and operating environment, could increase predictability and thereby assist with compliance with IHL, but, in general, would not alone be sufficient to ensure IHL compliance."

6 For example, a recent article, 'U.S. Army to receive remote weapons stations with automated targeting system', *Defence Blog*, 26 July 2019, noted that, "The active target detection component actively searches the field of view for known target profiles and highlights them for instant evaluation by the shooter", online at: https://defence-blog.com/army/u-s-army-to-receive-remote-weapons-stations-with-automated-targeting-system.html

7 For example, in April 2018, a joint statement by France and Germany stated that, "With regards to the defining [lethal autonomous weapons systems] LAWS France and Germany would like to underline understanding that the object of our discussions is limited to fully autonomous weapons, which do not exist to date" (https://www.unog.ch/80256EDD006B8954/(httpAssets)/895931D082ECE219C12582720056F12F/$file/2018_LAWSGeneralExchange_Germany-France.pdf). By contrast, as the culmination of a sophisticated line of argument, Switzerland argued in August 2018 that, "... technical sophistication doesn't seem to our delegation to be a defining characteristic of whether a weapon is autonomous. A weapon could be very simple and 'unintelligent' in its design, but highly autonomous in its critical functions" (https://www.unog.ch/80256EDD006B8954/(httpAssets)/2AFEC18F31D3930DC12583BB0044FC56/$file/2018.08.27+GGE+LAWS_Switzerland_Item+6a.pdf).

8 For an introduction to thinking around the politics of classification see Geoffrey Bowker and Susan Leigh Star, 1999, *Sorting Things Out: Classification and its Consequences*, MIT Press.

9 France in particular have consistently pressed the perspective that the object of study is only things that do not exist. The first point of their August 2018 *Working Paper on Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, stated that "It must first be recalled that although autonomous technologies are growing quickly, LAWS are still very much a long-term initiative: there are currently no fully-autonomous lethal weapons systems." https://www.unog.ch/80256EDD006B8954/(httpAssets)/2E16E59C0AB73F2FC12582F30055113C/$file/2018_GGE+LAWS_August_Working+Paper_France.pdf

10 Germany's CCW statement of 25 March 2019 (on agenda item 5.d) elaborated characteristics of 'autonomy' that are very close to the sensor-analysis-force process set out here, but with the caveat that the analysis phase requires the system to, "evaluate the circumstances of a changing situation without reference to a set of pre-defined goals" and "to reason and decide on the most suited approach towards their realization", https://www.unog.ch/80256EDD006B8954/(httpAssets)/D6C11A0A68D81F4AC12583CB0036650B/$file/20190325+Statement2+Germany+GGE+LAWS.pdf. That position built upon their statement from April 2018 on a 'Working Definition of LAWS' that noted that, "a proper distinction of automatisation and autonomy is key for an improved understanding and an informed debate on the issue within this GGE as well as in the wider public ... The characteristics of autonomous systems will primarily be the result of the application of information processing technologies based on algorithms", https://www.unog.ch/80256EDD006B8954/(httpAssets)/2440CD1922B86091C12582720057898F/$file/2018_LAWS6a_Germany.pdf.

11 Aspects of this typology were inspired by comments by Dr. Heather Roff at an informal Meeting of Experts of the CCW in April 2016 where she noted the distinction between systems that apply force to a 'what' by comparison with those that apply force at a 'where and when'.

12 It could be noted that some of these systems also use a sensor-analysis-force process at the immediate point of detonation – however, in these systems the sensor input in question is rendered inevitable by the ballistic trajectory of the weapon.

13 We should also acknowledge an additional category that sits alongside these: **Weapons by which force is applied at a time and spatial location derived from an analysis of sensor inputs, based on some human assessment of those sensor inputs and whilst operating within an envelope of space and time**. Such systems use sensors to guide a weapon into contact with the object to be attacked, but a human operator has identified the specific object in question (or at least a specific sensor-representation of that object). Some forms of missile guidance work in this way – with a human selecting certain radar signals (for example) as objects to attack. As with the category above, the specific time and spatial location where force will be applied is not set by a person and, again, there is typically an envelope of space and time within which it could occur. The key point of difference from the category above is the role of a human in specifying the sensor-input to be subsequently targeted.

14 See 1997 Mine Ban Treaty, CCW Amended Protocol II, 1907 Hague VIII Convention, 2008 Convention on Cluster Munitions.

15 We noted previously the use of this term in the United Kingdom's March 2019 statement to the GGE on agenda item 5(d) and in the Chair's 14 June 2019 draft *Conclusions and Recommendations*.

16 Statement to the CCW GGE by Germany on agenda item 5(d), Geneva, 25 March 2019, https://www.unog.ch/80256EDD006B8954/(httpAssets)/D6C11A0A68D81F4AC12583CB0036650B/$file/20190325+Statement2+Germany+GGE+LAWS.pdf

17 See 1997 Mine Ban Treaty, Art 2(1 & 3) for the conclusion of considerations regarding anti-handling devices in that context. On sensitive fuzes see, amongst others, CCW Working Paper of Germany, Nov 2003, *Sensitive Fuzes for Anti-Vehicle Mines*, CCW/GGE/VI/WG.2/WP.3 and CCW Working Paper of the Russian Federation, Nov 2003, *Considerations relating to the sensitivity of the fuses of mines other than anti-personnel mines*, CCW/GGE/VI/WG.2/WP.7, which notes that "...modern MOTAPMs are equipped with fuses which can classify the object of destruction with a probability exceeding 0.95." The UK NGO Landmine Action also undertook research in 2001 to see if anti-vehicle mines could anyway be detonated by common human actions (Landmine Action, 2001, *Civilian Footsteps*). As reported in the *New Statesman*, "it was up to the Sports Biomechanics Research Group at Loughborough University to show, in its Landmine Action report, that simple human activities such as jumping from the back of a light goods vehicle can exert enough pressure to detonate an AVM. More alarmingly, the researchers found that 'a child can quite easily, while skipping for instance, exert downward forces in excess of a common AVM initiation pressure of 150kg", *New Statesman*, 7 January 2002, 'The Ministry of Defence has anti-vehicle mines, but has not bothered to find out that a skipping child could easily explode them', https://www.newstatesman.com/node/154741

18 For elaboration, see Richard Moyes, 2016, 'Meaningful human control over individual attacks', in ICRC, *Autonomous Weapon Systems: Implications of Increasing*

Autonomy in the Critical Functions, expert meeting report, 2016, online at: https://shop.icrc.org/autonomous-weapon-systems.html?___store=default

19 See Gro Nystuen and Stuart Casey-Maslen [eds], 2010, *The Convention on Cluster Munitions: A Commentary*, Oxford Commentaries on International Law, Oxford University Press, p.200 which notes that: "Munitions such as Smart 155 and BONUS, and the 'skeet' explosive submunitions of the BLU-108 do not distinguish between military objectives and civilian objects but rather are programmes to identify target objects that fit profiles of heat and shape associated with certain military objects (typically armoured fighting vehicles). It is not known what analysis has been done of how such profiles compare with the profiles of objects commonly in civilian use ... or how susceptible such submunitions are to identifying objects commonly in civilian use as targets...Any limitations of such systems in distinguishing between objects that may be in civilian use and objectives that may be in military use might limit the circumstances in which they can be lawfully used. For example, the likelihood of an attack being lawful in a desert situation where a party is attacking an area full of military vehicles is greater than for an attack in a populated area where military vehicles are interspersed with vehicles in civilian use."

20 See, for example, *Key elements of meaningful human control*, Article 36, April 2016, http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf and Richard Moyes and Heather Roff, *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons*, Article 36, April 2016, http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf also, notably, *Killing by Machine: Key issues for understanding meaningful human control*, Article 36, April 2015, http://www.article36.org/wp-content/uploads/2013/06/KILLING_BY_MACHINE_6.4.15.pdf. Many of these papers use the term 'proxy data' in a way which is synonymous with the term 'target profile' in this paper. Also very significantly on the challenges of 'categorical targeting' and the procedural implications of subjecting people to evaluation as potential targets under a target profile process, 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 of International Humanitarian Law and Human Rights, https://www.geneva-academy.ch/joomlatools-files/docman-files/Briefing9_interactif.pdf

21 ICRC Working Paper to the CCW, March 2018, *Ethics and autonomous weapon systems: An ethical basis for human control?*, CCW/GGE.1/2018/WP.5, https://www.unog.ch/80256EDD006B8954/(httpAssets)/42010361723DC854C1258264005C3A7D/$file/CCW_GGE.1_2018_WP.5+ICRC+final.pdf

22 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 of International Humanitarian Law and Human Rights, p.60, https://www.geneva-academy.ch/joomlatools-files/docman-files/Briefing9_interactif.pdf

23 See Article 36 and Reaching Critical Will, 2014, *Sex and drone strikes - Gender and identity in targeting and casualty analysis*, http://www.article36.org/wp-content/uploads/2014/10/sex-and-drone-strikes.pdf. This paper notes that: "According to a New York Times report from May 2012, in counting casualties from armed drone strikes, the US government reportedly records 'all military-age males in a strike zone as combatants ... unless there is explicit intelligence posthumously proving them innocent.' The non-governmental organisation Center for Civilians in Conflict (CIVIC) subsequently highlighted some attempted nuancing of this position by administration officials, reporting that an aide to the US President described the New York Times article's characterization as a 'wild oversimplification.' States have a legal obligation to provide due process to those suspected of crimes and cannot just kill them. See for example Article 14(1) of the International Covenant on Civil and Political Rights; A/68/382, Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christopher Heyns, September 2013. However, CIVIC also noted that administration officials did not deny that they presume unknown men killed in a strike to be militants. It also pointed to another administration official stating that the article was 'not wrong that if a group of fighting age males are in a home where we know they are constructing explosives or plotting an attack, it's assumed that all of them are in on that effort.' In a May 2013 release of 'policy standards and procedures for the use of force in counterterrorism operations outside the United States and areas of active hostilities,' the US government stated in a footnote that 'males of military age may be non-combatants; it is not the case that all military-aged males in the vicinity of a target are deemed to be combatants' [emphasis in original]. It is unclear where geographically this policy applies. The statement was only made in reference to operations conducted outside 'areas of active hostilities'. Given the lack of clarity from the administration about which areas are considered to represent hostilities, it may not include some of the areas where drones are most heavily used. It is also unclear whether this statement is referring to the classification of people for purposes of post-strike analysis (which was the subject of the New York Times report) or the process of targeting. Some US drone attacks are 'personality strikes,' in which there is intelligence about a specific identified individual. But the majority of US strikes are reported to be 'signature strikes,' in which people are attacked on the basis of observed characteristics with no substantial intelligence regarding actual identity or affiliations. 'Signature strikes' use patterns of behaviour assumed to indicate militancy as a basis for targeting. These patterns are determined by analysing information collected by drones that survey selected areas. Legal scholars, UN officials, and civil society groups have noted that the 2013 policy provides no clarity about what 'signatures' have been or are being used in targeting decisions."

24 However, 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 of International Humanitarian Law and Human Rights, https://www.geneva-academy.ch/joomlatools-files/docman-files/Briefing9_interactif.pdf pp.61-62, notes that: "Targeting security measures based on broad-brush racial, ethnic, religious and national origin stereotypes in order to identify potential threats and vulnerabilities has been harshly criticized in the counter-terrorism context. It is doubtful that the severe interference with the rights of everyone affected by algorithmic targeting can ever be justified as necessary and proportionate to safeguard democratic institutions. As noted previously, there is scope for 'categorical targeting' within a conduct of hostilities framework but the principle of non-discrimination continues to apply in armed conflict. Adverse distinction based on race, sex, religion, national origin or similar criteria is prohibited. Remarkably, under Additional Protocol II to the Geneva Conventions, applicable in certain NIACs, the prohibition on adverse distinction also applies to persons directly participating in hostilities." Additional Protocol I does identify certain bases for categorising people that should not lead to discriminatory treatment, noting that provisions of assistance shall be provided "without any adverse distinction founded on race, colour, sex, language, religion or belief, political or other opinion, national or social origin, wealth, birth or other status, or on any other similar criteria."

25 On this see, ICRC, 2019, *Artificial intelligence and machine learning in armed conflict: A human-centred approach*, Geneva, 6 June 2019, online at https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach

26 The UK, in a statement to the CCW, in March 2019, noted that "where a human operator is required to supervise or interact directly with an autonomous system it must be designed in a way which enables them to understand the situation and the system status to an appropriate level. Again, what constitutes appropriate will depend on the operational context and environment. Two important factors are the type of feedback available to the user before, during and after use, and the familiarity of the operators with the system – particularly regarding its capabilities and limitations." UK statement on agenda item 5(d), March 2019, https://www.unog.ch/80256EDD006B8954/(httpAssets)/85A4AA89AFCFD316C12583D3003EAB3E/$file/20190318-5(d)_HMI_Statement.pdf

27 See, ICRC, 2019, *Artificial intelligence and machine learning in armed conflict: A human-centred approach*, Geneva, 6 June 2019, online at https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach and previously Richard Moyes and Heather Roff, *Meaningful Human Control, Artificial Intelligence and Autonomous Weapons*, Article 36, April 2016, http://www.article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf

28 For example, the UK have noted that they "[conduct] Article 36 legal reviews at multiple procurement milestones and additional reviews if there is a change in system use [or] capability," UK Working Paper to CCW, August 2018, *Human Machine Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles*, CCW/GGE.2/2018/WP.1 https://www.unog.ch/80256EDD006B8954/(httpAssets)/050CF806D90934F5C12582E5002EB800/$file/2018_GGE+LAWS_August_Working+Paper_UK.pdf

29 See for example working papers and statements by the UK, Australia, Canada and others.

30 For example, Article 50(1) of the 1977 Additional Protocol I provides: "In case of doubt whether a person is a civilian, that person shall be considered to be a civilian," similarly Article 52(1), "Civilian objects shall not be the object of attack or of reprisals. Civilian objects are all objects which are not military objectives as defined in paragraph 2" and (3), "In case of doubt whether an object which is normally dedicated to civilian purposes, such as a place of worship, a house or other dwelling or a school, is being used to make an effective contribution to military action, it shall be presumed not to be so used."

31 For example, UK Statement to the CCW, March 2019, on Agenda item 5(d), noted that "specific constraints or parameters that might be placed on a machine's freedom of action might include ... limiting the range and task duration within which the system may be allowed to operate away from direct human involvement," https://www.unog.ch/80256EDD006B8954/(httpAssets)/85A4AA89AFCFD316C12583D3003EAB3E/$file/20190318-5(d)_HMI_Statement.pdf. This sort of formulation has also been taken up in the Chair's draft report from the 2019 session

32 *Draft Conclusions* - Chair's Non-Paper 19-05-07 v.2.2 https://www.unog.ch/80256EDD006B8954/(httpAssets)/D8C04EC71F502A77C12583F400476619/$file/Draft+Conclusions+(%D1%84).pdf

33 See Richard Moyes, 2016, 'Meaningful Human Control', in Robin Geiss ed. *Lethal Autonomous Weapons Systems: Technology, Definition, Ethics, Law and Security*, Federal Foreign Office of Germany, 2016 (https://www.unog.ch/80256EDD006B8954/(httpAssets)/5AC684827AD95693C12580A5002E5F97/$file/GER+MFA+LAWS-experts+publication_cover.pdf) and Richard Moyes, 2016, 'Meaningful human control over individual attacks', in ICRC, Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions, expert meeting report, 2016, online at: https://shop.icrc.org/autonomous-weapon-systems.html?___store=default

34 Statement of the ICRC to the CCW, March 2019, on agenda item 5(a), https://www.unog.ch/80256EDD006B8954/(httpAssets)/5C76B1301CEC4BE6C12583CC002F6A15/$file/CCW+GGE+LAWS+ICRC+statement+agenda+item+5a+26+03+2019.pdf

35 See CCW Amended Protocol II, Art.5(2.a) and, related, (6.b).

36 CCW/CONF.III/11 (Part III) Section X (2.i & ii) https://undocs.org/CCW/CONF.III/11%20(Part%20III)

37 It is notable that the 2008 Convention on Cluster Munitions, in its definition of a cluster munition, excludes from prohibition systems that apply force to individual target objects only, inter alia, on the basis that they have a limited number of submunitions – effectively containing the number of applications of force that a system could undertake with each human use of the weapon system.

38 UK statement to the CCW, March 2019, on agenda item 5(a): https://www.unog.ch/80256EDD006B8954/(httpAssets)/1ED3972D40AE53B5C12583D3003F8E5E/$file/20190318-5(a)_IHL_Statement.pdf

39 We noted earlier that, at a level of principle, such systems still reduce people to data-points as a basis for applying force to them and, at a practical level, would likely still target people that are hors de combat, and that that different mechanisms for ensuring the effective exclusion of civilians both serve to push the burden of avoiding harm onto the civilian population and have presented various shortcomings in practice.

40 Speech of German Foreign Minister, Heiko Maas, at the general debate of the 73rd General Assembly of the United Nations, 2018, online at: https://www.auswaertiges-amt.de/en/newsroom/news/maas-general-assembly-un/2142300

Article36

www.article36.org