Meaningful human control

Presentation by Maya Brehm, Researcher, Geneva Academy of International Humanitarian Law and Human Rights (ADH), to the informal meeting of experts on lethal autonomous weapons systems of the Convention on Certain Conventional Weapons (CCW), Geneva, 14 April 2015

I. Introduction

The notion of '*meaningful human control*' has gained a lot of traction in discussions on autonomous weapon systems (AWS). The expression is used by commentators and some states for different purposes and can have slightly different connotations. In this presentation, I will look at four different facets of 'meaningful human control', and suggest how 'meaningful human control' can assist policy makers in effectively addressing the challenges raised by the weaponization of autonomous systems.

II. Facets of meaningful human control

There is general agreement that the use of armed force, weapon systems, weapons and weapon effects must not be 'uncontrollable' or 'out of control'. Control, exercised by human beings, in the use of weapons and over the consequences of weapon use has long been an important, implicit requirement for the moral acceptability, the socio-political legitimacy and the legality of organised armed violence.

Both, proponents and opponents of increasing autonomy in weapon systems acknowledge that human control is central to the acceptability of weapon systems that can detect, select and apply force without human intervention. Yet, the development of such systems may change what we have come to expect in terms of the *form* or *nature* of human control, and the *manner* in which and the *process* by which human beings exercise control over weapons.

Control is the ability to influence behaviors, circumstances, persons, and outcomes (such as achieving a goal or avoiding an undesired outcome). Control entails an assertion of agency, power and authority over someone or something. Control can be exercised by applying cognitive skills, managing resources, and participating in decision-making. Control involves some level of mastery of one's environment, and, thus, requires awareness of one's environment.

This control does not need to be absolute. In relation to the use of armed force, and to the use of weapons or weapon systems, specifically, it is generally expected in present practice that human beings exercise some degree of control over:

- *Who* or *what* is harmed, including:
    - harm to persons or objects against whom force is used (the target), and
    - harm to persons or objects that are incidentally affected by the use of armed force;
- *When* force is applied / harm is experienced (moment, duration);
- *Where* force is applied / harm is experienced (space, location);

- *Why* someone or something is targeted / harmed (rationale) and *how* armed force is used (process)

What makes human control in relation to these various aspects 'meaningful'?

The notion of meaningful human control has been proposed, in particular by the civil society organization, Article 36, as a way of identifying parameters of human control in present practice, with a view to identifying where normative boundaries should be drawn against unacceptable weapons or practices.

One way of approaching the question is by establishing what is *not* acceptable in terms of human control. Consider the following examples or scenarios:

1. Use of a weapon whose effects cannot be controlled in time or space.
   > Use of such a weapon is clearly unacceptable. There is an expectation that the effects of weapons are sufficiently controlled and limited in space and time. International legal bans on biological and chemical weapons, and on the use of balloon-borne bombs have, at least in part, been based on this consideration.
2. Use of a weapon system with mobile components that roam freely and apply force in multiple locations without reporting back to a person over an extended period of time.
   > Most would consider such a scenario unacceptable, which suggests that there is an expectation that human control is exercised in some form over the location and the moment of force application.
3. Use of a weapon system where human control consists solely of a person pushing a button every time a light comes on, without having any other information.
   > This too would be deemed unacceptable by most. This scenario suggests that to be acceptable, human control cannot be purely formulaic*. Not every form of human control is sufficient or adequate or meaningful.*

The second and third scenarios highlight, in particular, that for human control to be adequate or meaningful, *information* is required – information that allows those persons responsible for the use of force to *anticipate* the reasonably foreseeable consequences of force application. Only if these persons can anticipate these consequences, can they make the required *legal assessments* about the use of force.

Concretely, information is required about

- the kinetic or other effects produced by a weapon in different environments;
- the location and time of force application;
- who or what is being targeted and who or what else risks being harmed.

Only if the geographic area and the time of independent operation [NB: 'operation' refers to independent detection, selection and firing] of a weapon system is sufficiently bounded and the target parameters are sufficiently narrow can those responsible for the use of force *understand* who or what is

at risk of being made the object of attack (identified as a target) or of being harmed in different operational environments.

> In the specific case of use of force connected to the *conduct of hostilities*, information about these aspects is required for *every individual attack*. Only if this information is available can 'those who plan or decide upon an attack' apply and comply with the rules of IHL. An 'attack' (as a whole) can entail several instances of force application, but it is limited in space and time.

The *risk* that persons or objects are harmed who should be protected against the effects of armed force, increases:

- the more mobile a weapon system's components are,
- the wider the area within which a weapon system operates independently,
- the longer the time of independent operation,
- the more complex the operational environment,
- and the broader the target parameters.

The acceptability of weapons systems, therefore, hinges on *limitations* being placed on their independent operation in technical, policy and legal terms, to reduce that risk. It is limitations along the parameters of space/time/target identification that enable the exercise of meaningful human control over the use of force.

Can we guarantee that AWS will do what we want them to do?

A second, related facet of 'meaningful human control' also relates to our ability to predict the outcome produced by the use of AWS: It concerns the '*controllability', including the predictability,* of artificial intelligence systems.

The concern here is with avoiding system *failures*, and with the development of *robust, reliable systems*, that perform exactly as desired, so that it can be *guaranteed* that they do what we want them to do.

[An important challenge is that as the flexibility of a system increases, its predictability decreases.]

Why is armed force used and how is it used?

A third facet of meaningful human control relates to why and how armed force is used. In this connection, the question has been raised: is it acceptable to apply force by means of a weapon system in the knowledge that, based on its algorithms, on average, the system will make a certain percentage of persons/objects that are not legitimate targets, the object of attack? Is it acceptable if, in statistical terms, this percentage is relatively low?

[NB: this concern about 'kill decisions' being based on probabilistic matching algorithms is an overarching moral and legal concern that is distinct from the concern about system failures.]

Many would agree that the *process* by which decisions are made to cause harm matters. It matters why something is done, especially when decision have grave implications for an individual's life. The notion of 'meaningful human control' in this connection draws attention to human activities of *meaning-making*. Meaning-making is a social practice by which human beings interact with each other to make *common sense* of the world. Meaning includes moral understandings of right and wrong, cognitive understandings of true and false, perceptual understandings of like and unlike.

In present practice, organized violence entails that decisions to use armed force are the result of a *deliberative process* of human interaction. It involves moral agents making moral judgments involving shared values.

When weapon systems detect, select and apply force without human intervention, this deliberative process of human interaction is replaced by algorithmic calculations. However, whether the human abilities to sense, think, decide or act can accurately be ascribed to a machine is a matter of *significant contention*. After all, machines lack intentionality.

Some, including the UN Special Rapporteur on extrajudicial killings, have warned that when machine processes are treated as equivalent to social/human interaction, deliberation, and judgment, critical aspects of how we make sense of violence and suffering are lost. Many find this dehumanizing. They warn that it diminishes the value of life itself. That it is an affront to *human dignity*. That it is 'inherently arbitrary', and that it offends our sense of *humanity* and *justice*.

Who is responsible?

Finally, exercising meaningful human control over the use of armed force, it has been argued, also entails that *someone somewhere is responsible and can be held accountable* for harm done and for wrongdoing. The ability to designate a responsible moral agent, both, for the outcome of armed force and the processes involved, is critical to the legitimacy and the legality of organized armed violence.

Some appear to suggest that human beings, will in any case, always be involved, in the development, building and programming of AWS. In this connection, the notion of 'meaningful human control' serves to highlight that if the locus of human decision-making and of moral and legal responsibility becomes too far removed from the locus where harm is experienced; if the connection between the two becomes too remote or diffuse or distributed, human control ceases to be meaningful (filled with meaning). *Meaningless or senseless violence is not legitimate violence*.

III. The contribution of 'meaningful human control' to the policy debate on AWS

States parties to the CCW recognize 'the need to continue the codification and *progressive* development of the rules of international law applicable in armed conflict' (CCW Preamble). In discussions to this end, focusing on 'meaningful human control' can have advantages over more technology-centred approaches.

Most importantly, focusing on how we ensure meaningful human control over present weapon systems allows us to *elaborate policies and legal norms on the basis of what we know*, rather than relying on conjectures about the hypothetical capabilities of potential future technological developments.

It enables us to *preventively* address humanitarian, rule of law and security/arms control concerns around weapon systems whose technical characteristics, capabilities and applications remain highly uncertain.

The notion of 'meaningful human control' can contribute to the policy debate on AWS by:

- Helping to structure the debate: We already see a number of commentators and state representatives adopting this language to articulate their arguments and suggest next steps for policy makers.
- Guiding technological developments: It has been recognized that meaningful human control is important to ensuring that artificial intelligence systems are beneficial to humanity. The notion can help guide research and development in AI and other areas.
- Indicating where normative boundaries should be drawn: Some have said that a requirement of meaningful human control is already implicit in international law governing the use of force. Several commentators have called for the articulation of *an explicit legal requirement* that there must always be meaningful human control over every individual attack. Weapon systems that operate without meaningful human control would, thus, be explicitly prohibited. On the basis of this principle, more specific technical, policy and legal limitations on the weaponization of autonomous systems can then be elaborated.

It would, of course, be very helpful in this regard if states that possess or are developing weapon systems that can apply force without direct human intervention could explain what limitations, in policy, legal or technical terms, they have put in place to ensure meaningful human control in respect of the different considerations laid out.