Article36

www.article36.org

Submission to the UN Secretary-General by Article 36
Considerations on the development of an international legal instrument on autonomous weapons.

8 May 2024

Article 36 has written and worked extensively on the issue of autonomous weapons – including framing the requirement for meaningful human control and promoting the need for a structure of international legal regulation that includes both prohibitions and positive obligations.

Our basic position is that an international legal instrument is urgently needed. It should contain a prohibition on systems that would target people directly and a combination of prohibition and positive obligations that work together to ensure meaningful human control in the use of force.

This submission does not seek to restate all of our thinking on this issue but to highlight a number of key points that we consider to be particularly significant at this stage of the political and policy process.

- **We should recognise autonomous weapons systems as referring to 'systems' or 'processes', rather than 'objects'.**

  Discussions of this issue often talk about 'autonomous weapons' as concrete, unified physical 'objects' – that is to say, as physical objects that share a recognisable set of characteristics (akin to anti-personnel landmines, for example). However, the defining characteristics of autonomous weapons systems are the tied to the relationship of human users to processes of decision-making.

  Autonomous weapons systems may function through distinct and widely dispersed physical assets, all of which *might also* function in ways that would not constitute an autonomous weapons system. For example, an armed drone might direct force against a specific target under the control of a human operator, or that same hardware asset might direct force against a target based on instructions from a separate external computer and sensor system. In the latter configuration the armed drone is part of an autonomous weapons system but in the former it is not. The defining aspect is in the relationship of human decision-making to the process, not necessarily in the technological objects in themselves.

  This has important implications for how we write rules on this issue. Rules need to be focused on human understanding and control over individual attacks and on how such systems are used. There will still be unified physical systems that need to be subject to these rules (including systems that would be prohibited in such a framework), but the starting point should be to regulate human understanding and control of the 'process'.

- **AI is <u>not a necessary characteristic</u> of autonomous weapons, but it raises distinct challenges.**

  Artificial intelligence (AI) is one of the technical drivers that is making the issue of autonomous weapons systems particularly pressing. However, building on the point above, it is the relationship of human operator(s) to certain decision-making processes that is the defining characteristic, not the technology that is involved. It is also important to recognise that 'AI' is a broad umbrella term for a wide variety of computational and statistical processes.

  So, it is possible to have autonomous weapons systems that do not employ AI and we should not define the boundaries of this issue in relation to AI.
  However, AI does provide distinct additional challenges. For example, AI can make it more difficult for the users of systems to have a practical understanding of how their systems work and so to adequately predict outcomes from their use. In certain roles, AI might serve to embed bias from training datasets or algorithmic assumptions into the functioning of weapon systems – which is a particular challenge in relation to weapon systems that would target people, or specific groups of people. This is one of the reasons that systems targeting people should be prohibited.

- **A new legal instrument could be short - establishing the key overarching rules that provide a structure for shaping and evaluating technological developments in the future.**

  A legal instrument on this issue should focus on the key general rules that promote human dignity and meaningful human control. Rules should include:

  - A prohibition on using AWS to directly target people (anti-personnel systems).
  - Positive obligations to ensure meaningful human control, including requirements that:
    - Users sufficiently understand AWS they intend to use, including the conditions that would trigger an application of force by the system;
    - Users sufficiently evaluate the context where the system would be used; and
    - Users sufficiently limit the duration and area of system functioning in order to meaningfully apply existing legal rules.
  - A prohibition on systems that cannot be used in accordance with these positive obligations, and so are likely unpredictable and incompatible with the necessary human control.

  Such a legal structure can then provide a framework under which specific cases can be addressed.

  We have already noted that certain AI mechanisms in certain roles may be incompatible with the legal rules that should be adopted. However, we need to establish the rules that we need first and then evaluate specific technologies against those rules subsequently.

For example, we might broadly suggest (as above) a legal rule that 'users need to have a sufficient understanding of any autonomous weapons system that they are considering to use' and that 'it is prohibited to use systems that do not allow such an understanding'. Whether or not a particular form of machine learning used to develop a target profile of an enemy tank is compatible with that rule is something that would need to be evaluated a) in the context of the agreed rule, b) with an understanding of how the specific machine learning function works, and c) with an understanding of the implications of that function and its outputs for the operation of the system.

The initial legal instrument should not be expected to work through all such specific cases in advance. This must be a future orientated instrument against which new technological structures are evaluated as they are developed (including through weapon review processes). Sharing good practices in such processes, and in the understanding and evaluation of certain technologies, or on the necessary trainings needed to meet the positive obligations suggested above would all be valuable technical multilateral work streams once the legal instrument has been established.

- **Regulating autonomous weapons is an important opportunity to limit the negative potential of AI without curbing its wider positive potential**.

  Although we have noted that AI is not a defining characteristic of autonomous weapons systems, adopting this legal treaty should be recognised as a critical action to prevent negative effects from AI in the world. The legal treaty would establish guardrails that prevent the development and adoption of AI functions in some critical roles that undermine human control and human dignity in the use of force. As such, it points to one potential mode for regulating AI more broadly which is to limit its scope of use in specific roles and manifestations. The key to that regulatory mode is not to regulate the AI directly (which is too amorphous) but to establish the obligations for human understanding and action.

- **Current 'defensive systems' (missile defence systems etc) should not be prohibited, but *should* fall within a legal instrument and be used in accordance with its positive obligations (this is in line with current practice).**

  Some states have raised concerns that 'defensive' systems should not fall within the scope of consideration of discussions regarding autonomous weapons. For some states, this concern underpins a preference to work under the terminology of "lethal autonomous weapons systems" rather than "autonomous weapons systems" more generally. The types of 'defensive' systems driving these concerns are broadly 'anti-missile' systems that use sensors and computer-directed guns to detect and apply force to incoming weapons (such as missiles, rockets and mortars). Such systems include Iron Dome, CRAM, Aegis etc.

  'Anti-missile' systems such as those noted above fall within the understanding of an autonomous weapons system as described in the policies of many states as well as by organisations such as ICRC and Article 36.

  It is a fundamental aspect of the widely adopted 'two-tier' approach that many of the systems that fall under that approach are subject to regulations rather than

prohibitions. This two-tier approach is now the predominant orientation to autonomous weapons systems in international discussions.

These 'anti-missile' systems fall within the scope of the consideration because they use sensors to determine specifically where and when force will occur in response to matching data from the environment against a generalised target-profile. However, such systems would not be considered at risk of prohibition under a future instrument because:

A. they can be used with meaningful human control, appropriate human judgement etc. The users of such systems can have an effective understanding of how these systems function, including what will trigger an application of force by the system. Furthermore, the location and duration of system functioning can be specifically controlled by the human operator.

B. they do not target 'people' directly. Some states and organisations are calling for a prohibition on systems that target people directly. The 'anti-missile' systems discussed here clearly do not fall within this area of concern because the target weapons rather than people.

Given this analysis, we would suggest that states might move on from this line of concern or anxiety that anti-missile systems might be prohibited under the two-tier approach.

- **A prohibition on systems that would target people directly should be a critical moral and societal priority.**

The ethical and moral concerns with respect to autonomous weapons are most critical in relation to systems that would target people directly. Acknowledging this is akin to recognising the specific problems associated with anti-personnel mines by comparison with anti-vehicle mines. Allowing systems to be used to harm people as a result of machine processing is dehumanising and should be considered incompatible within requirements to protect human dignity. Such systems would also be fraught with legal risks.

If it were claimed that systems could somehow distinguish combatants from civilians this would be a transference to machine functioning of determinations that should be made by a human commander. Furthermore, such mechanisms would likely neglect the obligation to protect soldiers *hors de combat* and may be liable to problems of racial, age and gender bias if built on certain AI processes.

We have an opportunity to act now to prevent the adoption of autonomous systems that target people. The working presumption for future negotiations should be that systems targeting people are unacceptable.

- **A legal instrument should be developed through an inclusive multilateral process that is open to all states but that cannot be blocked by any one country.**

It is urgent to start negotiations on a legal instrument to address autonomous weapons systems. That process needs to start in a forum that can bring in the views of diverse stakeholders and that is open to all states to participate (if they wish) on equal terms.

However, it is not prudent to insist that such discussions should only take place in forums where certain militarised states (who are most invested in military technologies) are consistently allowed to prevent the majority from moving forwards. The CCW has provided a useful framework for building shared understandings of the parameters of this issue and it can continue to play an important role.

However, energising international humanitarian law and international commitment to protect civilians requires action in a framework that has the potential to reflect to will of the majority. This issue is too fundamentally important for society to continue to remain constrained by procedural exploitation.