Article36

Written submission by Article 36 to the Science and Technology Committee inquiry into robotics and artificial intelligence.

The need to ensure meaningful human control in the face of the weaponisation of robotics and artificial intelligence.

27 April 2016

This submission is made by Article 36 (www.article36.org), a UK based non-governmental organisation working to prevent the unintended, unnecessary or unacceptable harm caused by certain weapons.

Contact:

Richard Moyes,
Managing Partner, Article 36
richard@article36.org
www.article36.org

Summary

The UK's engagement to date in multilateral discussions on the implications of increased autonomy in weapons systems, facilitated by robotics and AI, is not adequate to the broad societal implications of the subject matter. How the relationship between human and machine decision-making is managed on issues of life and death is of fundamental importance to how society's relationship with computers and AI will develop in the future. In that context the UK's approach to policy making on autonomous weapons so far lacks foundations in a vision of the role of AI in society in the future. It fails to engage with the key questions of immediate relevance yet seeks to avoid movement towards multilateral agreement on the nature and form of human control that should be considered necessary in making decisions over how force is applied. UK policy making in this area should be subject to a broad review to ensure that a policy driven by defence interests also reflects the position the UK wishes to take on the wider roles of AI and computer autonomy in society in the future.

Introduction

Developments in robotics and artificial intelligence (AI) have the capacity to significantly transform weapons and the use of force. The potential for robotic or AI 'decision making' over critical functions in the use of force – such as who or what is targeted - raise urgent social, legal and ethical concerns. These issues are already the subject of debate in the media, in policy literature and in multilateral discussions under the United Nations Convention on Certain Conventional Weapons (CCW) which has been convening an informal Group of Experts on Lethal Autonomous Weapons Systems since 2014. There have been calls from senior UN officials that "autonomous weapons systems that require no meaningful human control should be prohibited"[1] – calls that are echoed by civil society[2] and leading artificial intelligence

[1] UN Human Rights Council, 'Joint report of the Special Rapporteur on the rights of freedom of

experts.[3] There is an opportunity for the UK to play a leading role in the development of international policy and legal responses to these issues, but to date the UK has not laid out an approach that suggests a coherent vision for policy making in the face of rapidly changing technological developments.

The UK Ministry of Defence and BAE Systems are investing in the development of their own autonomous system, the Taranis, which has been testing autonomous capabilities including target location and engagement.[4] Against this background of investment it is all the more urgent that the UK develop a coherent policy approach.

Introduction to UK policy on 'lethal autonomous weapons systems' to date

The UK's approach to policy and legal considerations around growing autonomy and AI in the development of weapons systems shows a number of features:

- Using reassuring language to argue that 'human control' will be retained.
- Unilaterally defining the subject of international discussions ('lethal autonomous weapons systems') in a highly futuristic way whilst seeking to avoid discussion of more realistic and relevant technological developments.
- Not providing more substantive explanation of the key terms that it uses – terms such as 'human control'.
- Asserting that no new international law is needed on the issue.

The combination of these features means that the UK is at once working to avoid the development of international law on this issue whilst creating substantial policy and legal space within which new weapons systems can be developed, including with autonomy in the critical functions of identifying, selecting and applying force to targets. Such an approach may be considered by some to serve the UK's best interests but it presents grounds for concern if it is not rooted in a wider conception of the role of robotics and AI in society and how that role might need to be constrained, in particular where it involves computer systems making 'decisions' over life and death.

The features identified above are evidenced and explained in the sections below:

Using reassuring language to argue that 'human control' will be retained

The UK has asserted that "the operation of weapons systems will always be under human oversight and control" [5], and that "systems will always have a human operator involved in all targeting decisions."[6]

Whilst such assertions seem on the surface to be reassuring, there has been no

peaceful assembly and of association, and the Special Rapporteur on extrajudicial, summary or arbitrary executions on the proper management of assemblies', 4 February 2016, A/HRC/31/66

[2] Campaign to Stop Killer Robots, 'Call to Action', available at: http://www.stopkillerrobots.org/call-to-action/

[3] See Future of Life Institute (2015), 'Autonomous Weapons: An Open Letter From AI & Robotics Researchers', available at: http://futureoflife.org/open-letter-autonomous-weapons/

[4] "Anglo-French UCAV Study Begins To Take Shape", Aviation Week, 4 February 2016
, http://aiationweek.com/defense/anglo-french-ucav-study-begins-to-take-shape

[5] Owen Bowcott, 'UK opposes international ban on developing "killer robots"', The Guardian, 13 April 2015, available at: http://www.theguardian.com/politics/2015/apr/13/uk-opposes-international-ban-on-developing-killer-robots

[6] UK CCW statement 12 April 2016

further explanation from UK officials as to what this means. The UK has vocally opposed the suggestion that there should be "meaningful human control" over weapons systems as being too subjective[7], yet asserts that there will be "human oversight and control" or "a human operator involved" as if that is self-explanatory.

Clearly, it is the form and extent of that human 'control' or 'involvement' that is the critical question in policy discussions around these issues. In the context of discussions of complex technologies terms such as 'human control' cannot be taken as understood, but need substantive explanation regarding their form and boundaries. Advocates of "meaningful human control" over weapons systems recognize this explicitly – proposing it as a principle that requires further collective definition as a process of collective policymaking.

The UK position uses terms that sound supportive of meaningful human control over weapons systems, but rejects that specific formulation and does not provide any explanation of what the terms it uses instead actually mean. Such an approach is primarily obstructive of the international policy conversation.

Unilaterally defining the subject of international discussions ('lethal autonomous weapons systems') in a highly futuristic way whilst seeking to avoid discussion of more realistic and relevant technological developments.

Against the background of international UN expert meetings on 'lethal autonomous weapons systems' (LAWS), the UK has stated that it will not develop LAWS. However, there is no internationally agreed definition of this term and the UK has unilaterally adopted a definition that is highly futuristic, to the point of being almost fantastical (this is evidenced by the UK's own assertions that the technologies it is referring to may never exist). The UK definition of "autonomous weapons systems", from Ministry of Defence (MOD) Joint Doctrine Note (2011) states:[8]

> An autonomous system is capable of understanding higher level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present.

This was elaborated further in a recent statement to the UN CCW in which the UK said:

> The UK understands such a system to be one which is capable of understanding, interpreting and applying higher level intent and direction based on a precise understanding and appreciation of what a commander intends to do and perhaps more importantly why. Critically, this understanding is focused on the overall effect the use of force is to have and the desired situation it aims to bring about. From this understanding, as well as a sophisticated perception of its environment and the context in which it is operating, such a system would decide to take - or abort - appropriate actions to bring about a desired end state,

[7] UK CCW statement 12 April 2016

[8] Ministry of Defence (2011), 'Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems' available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/33711/20110505JDN_211_UAS_v2U.pdf

> without human oversight, although a human may still be present. The output of such a system could, at times, be unpredictable - it would not merely follow a pattern of rules within defined parameters.[9]

And

> Based on these definitions we believe that lethal autonomous weapon systems do not exist, and may never exist.

And in this context

> …the UK Government does not possess fully autonomous lethal weapon systems and has no intention of developing them.

This approach to the definition of the subject matter of these international discussions is one that serves to jump over less complex systems wherein computers may undertake critical functions regarding the identification and application of force to targets, but where it would not be argued that the system has any capacity to have a "precise understanding of what a commander intends to do". By jumping so far ahead of current technological capacities in its definition of the subject the UK appears to be trying to avoid international scrutiny of developments in weapons systems that might have a fundamental bearing on military conduct and on the relationship between computers, AI and society. For example, a weapons system may not have a precise understanding of a commander's intent, but could still use sensors and algorithms to identify and attack certain categories of people within a pre-defined area. Whether such a system is acceptable or not is an open question, but the UK approach appears designed to take such questions off the table. Is a human decision on the predefined area for such an attack sufficient to meet their policy assertion that there will be 'human involvement'? How should categories of people be encoded into computer systems? Avoiding such questions amounts to avoiding critical issues regarding the role of AI and society.

However, the UK's approach to definitions appears to have little traction in the international policy debate. As a result it is the UK's relevance to the conversation, rather than the conversation itself, that is diminished. Thus, whilst our primary concern is that the UK's policy position is not anchored in a wider vision of the role and possible constraints of AI in society, we also note that their chosen approach does not seem likely to result in the achievement of their own goals within the framework of multilateral discussions.

Not providing more substantive explanation of key terms that it uses – terms such as 'human control'.

The UK has stated in multilateral discussions that:

> It is worth reiterating that as a matter of policy the UK Government is clear that the operation of its weapons will always be under human control as an absolute guarantee of human oversight, authority and accountability for weapons use.[10]

[9] UK statement to the CCW 12 April 2016

[10] Ministry of Defence (2011), 'Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems' available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/33711/20110505JDN_211_UAS_v2U.pdf

Similarly, in Parliament, the government has committed that the "operation of weapons systems will always remain under human control".[11]

As noted above, it is surprising that the UK vocally rejects calls for agreement that there be "meaningful human control" over weapons systems whilst in the same diplomatic statements makes an assertion that "human control" over the operation of weapons provides a "guarantee" of oversight, authority and accountability. In that context, questioning why calling for that control to be "meaningful" should be explicitly rejected can only result in the conclusion that either the UK wishes to promote a concept of human control that is actually extremely limited (mere "involvement" in their other formulation) or that they are diplomatically anxious that they don't have the political authority to guide a multilateral process asserting the need for "meaningful human control" to a conclusion that matches their interests. If, as in their formulation, human control provides a guarantee of human oversight, authority and accountability, then it seems incongruous to actively work against calls for "meaningful" human control in relation to such weapons systems.

Proponents of "meaningful human control" have repeatedly noted that the term "meaningful" simply serves to indicate that further definition of "human control" is required – because in detailed discussions regarding the functions that can be delegated to computers, 'human control' is not something that can be taken for granted but is a central component of the debate. If the UK is actively opposed to agreement that their should be "meaningful human control" over weapons systems, yet repeatedly considers asserting a commitment to "human control" to be politically useful, then the government should explain the key terms that it is using.

For example, when the UK asserts that the "operation of its weapons will always be under human control" what does that mean?

- Does it mean a human will always identify and affirm the specific object against which force will be applied?
- Does it mean that a human will sign off on a broad category of target objects against which a weapons system will subsequently 'choose' where force is applied?
- Does it mean that a human will authorize the activation of a weapon system that will then go out and identify and engage targets independently based on its own assessment of the contextual situation?

These examples are presented to illustrate the broad spectrum of technological functions that could be captured under different interpretations of 'human control'.

Similarly, the UK has stated that there will always be human control over 'weapons release.' In that context the UK should clarify whether, once a human has authorised the release of a weapon, the weapons system itself might:

- have the capacity to select a specific object to be struck from within a target area established by a human operator;
- have the capacity to determine for itself what target (military objective) it will strike; or
- whether it might permissibly have even more scope for action than this.

[11] Lord Astor of Hever (Parliamentary Under Secretary of State, Defence; Conservative), House of Lords debate, 26 March 2013, available at: http://www.publications.parliament.uk/pa/ld201213/ldhansrd/text/130326-0001.htm#st_14

The questions presented above are only a few examples of critical issues within the international debate regarding the role of computers and AI in weapons systems, and yet UK policy to date avoids these relevant issues in favour of a focus on technologies that it admits may never exist.

Asserting that no new international law is needed on the issue.

Despite not engaging with the central questions of the debate around realistic and relevant technological developments the UK asserts that no new international law is needed in this area. Whilst the subject of LAWS is the focus of international discussion at the UN CCW the UK appears to be the only state to have explicitly ruled out the development of new international law. Whilst the UK asserts that existing international humanitarian law is adequate to manage these new technological developments[12], its failure to provide answers to the sorts of questions presented above make it impossible to determine where its interpretation of the law lies.

The UK has sought to argue that implementing legal reviews of weapons as required by article 36 of the 1977 Additional Protocol I of the Geneva Conventions will be a sufficient response for the international community. The UK has not published analysis nor explained how assessments would be made in the face of the various highly complex implications of artificial intelligence and autonomy in weapons systems: or of which systems it considers acceptable and which are unacceptable (in the context of the sorts of questions presented above), nor the rationale for any such assertions.

The UK is effectively asserting that existing international humanitarian law is adequate to regulate these technologies without explaining how that law is to be interpreted or applied. The position amounts to an assertion that "if it is illegal it will be illegal" – but without any capacity to explain how it will be known where the threshold of acceptable human control/machine autonomy lies.

Conclusion and recommendations to the UK government

The UK's current policy orientation to the implications of AI and greater autonomy in weapons systems is not tackling the central issues regarding the nature and form of human control that should be considered necessary as computers are given a greater role in the making of life and death decisions. Given the broad societal implications of these questions, beyond just consideration of weapons, it is urgent that a wider governmental perspective is brought to bear on this matter that allows the UK's approach to be rooted in a coherent conception of the role of AI in society in the future.

The UK's futuristic definition of lethal autonomous weapons systems is out of step with technological developments that are happening now. Applying a focus to systems that are not, and may never be, technologically achievable neglects discussion over the systems that are currently on the cusp of development. The

[12] Statement of the UK to the CCW Meeting of High Contracting Parties, November 2015, available at: http://www.unog.ch/80256EDD006B8954/(httpAssets)/880AB56F1A934474C1257F170056A8F2/$file/2015_CCWMSP_LAWS_UnitedKingdom.pdf The UK stated: "Given the uncertainties in the current debate, the UK is not convinced of the value of creating additional guidelines or legislation. Instead, the UK continues to believe that international humanitarian law remains the appropriate legal basis and framework for the assessment of the use of all weapons systems in armed conflict".

UK’s failure to elaborate key terms such as ‘human control’ amounts to an avoidance of engagement in the real debate regarding the role of AI in weapons systems in the future.

The UK should engage in debate on this issue nationally and internationally, with the aim of setting a normative standard that prevents the development and use of autonomous weapons systems operating without meaningful human control. The UK should work to build a collective understanding of the form and nature of the human control that should be required.