Regulating Lethal and Harmful Autonomy: Drafting a Protocol VI of the Convention on Certain Conventional Weapons

Sean Welsh

Department of Philosophy, University of Canterbury, Christchurch, New Zealand
sean.welsh@pg.canterbury.ac.nz

Abstract

This short paper provides two partial drafts for a Protocol VI that might be added to the existing five Protocols of the Convention on Certain Conventional Weapons (CCW) to regulate "lethal autonomous weapons systems" (LAWS). Draft A sets the line of tolerance at a "human in the loop" between the critical functions of select and engage. Draft B sets the line of tolerance at a human in the "wider loop" that includes the critical function of defining target classes as well as select and engage. Draft A represents an interpretation of what NGOs such as the Campaign to Stop Killer Robots are seeking to get enacted. Draft B is a more cautious draft based on the Dutch concept of "meaningful human control in the wider loop" that does not seek to ban any system that currently exists. Such a draft may be more likely to achieve the consensus required by the UN CCW process. A list of weapons banned by both drafts is provided along with the rationale for each draft. The drafts are intended to stimulate debate on the precise form a binding instrument on LAWS would take and on what LAWS (if any) should be banned and why.

Introduction

After six years of UN debate on lethal autonomous weapons systems (LAWS) it seems timely to propose some treaty wording. Here two drafts of key clauses that might be included in a Protocol VI of the Convention on Certain Conventional Weapons (CCW) are provided.

Draft A is based on the notion that a "human in the loop" between select and engage reviewing and approving targeting decisions in real time is the normative requirement. This is based on positions articulated by NGOs.

Draft B is based on the Dutch concept of "meaningful human control in the wider loop" which as it bans nothing that exists today may be more likely to achieve the consensus required for a Protocol VI within the current CCW process. Wording is modelled on Protocols II and IV of the CCW.

These drafts are obviously tentative and incomplete. However, I hope they express some if not all of the essential points nations might see fit to include in a binding treaty instrument.

It is assumed the reader has some familiarity with the moral and legal arguments regarding LAWS and International Humanitarian Law (IHL). These are not addressed in detail here. ICRC (2018) provides a summary.

Rationale

With reference to the three approaches to definition presented in UNIDIR (2017), the rationale for the draft wording is as follows. A "technology-centric" definition of "autonomous" is not attempted. This is because technology is a moving target that changes every calendar quarter. Instead, the "human-centred" and "task/functions" definitional approaches described by UNIDIR are favoured and combined. It is held that regardless of the evolution of future technology the critical functions of targeting (defining, selecting and engaging targets) and the ability of humans to take part in critical functions and thus exercise control are clear and definable today.

Besides the define, select and engage critical functions, "meaningful human control" of AWS can be exercised by assigning responsibility to those who perform the function of activation and by ensuring that AWS can be deactivated by monitoring humans.

Definitions should capture non-lethal as well as lethal systems hence AWS is preferred to LAWS.

Also, definitions should capture existing systems not just future ones.

The definition of "autonomy" used here is based on that in Bekey (2005). Bekey defines autonomy as the ability to operate without a human operator for a protracted period of time.

This “no human operator” concept of autonomy is coupled with the ICRC “critical functions” approach to defining autonomy in an AWS. Three critical functions of targeting are defined: define, select and engage.

The concept of the “wider” loop defined in AIV/CAVV (2015) is used to draw the line of tolerance in Draft B.

The more commonly debated “narrower” loop of select and engage is used to draw the line of tolerance in Draft A.

With respect to weapons, a critical normative question is this: does fielding the weapon involve delegating a real-time lethal or harmful decision to a mechanical or computational device?

If the real-time decision to select (i.e. find, track, classify and prioritize a target) and engage (i.e. apply kinetic force and do harm to a target) does not involve a human and there is no possibility of human intervention, then the AWS is autonomous in the critical functions of select and engage and prohibited on the Draft A wording.

If in addition to autonomy in select and engage the AWS can define its own target classes (perhaps by machine learning from data sensed “in the wild”) and can go on to select and engage these self-defined targets without human review or approval of the defined rules and without human intervention in the select and engage functions then the AWS is prohibited on the Draft B wording.

While both drafts define “fully autonomous” in the same way, they prohibit different AWS. Draft A also has some “grandfather clauses” to exclude close-in weapons systems (CIWS), naval mines and anti-tank mines from the prohibition on weapons that have autonomy in the select and engage functions. Extra clauses might be added to Draft A to cater for other “defensive” systems.

On the common wording in both drafts, crude and simple AWS existed in the American Civil War. A land mine or naval mine is an AWS on these definitions. Such mines have autonomy in select and engage but do not have the ability to define their own target classes. Humans do this.

To the best of my knowledge, there are no existing AWS that are “fully autonomous” as defined here. Such AWS would be able to define, select and engage targets without any human involvement beyond setting up the original machine learning. AWS like this are theoretically possible but do not yet exist.

No distinction is made between autonomous and automated. Machines are machines and humans are humans. You are either delegating a critical function of targeting that can kill or harm humans to a machine in combat or you are not.

However, I accept there is a critical distinction between a rule-following system and a rule-initiating system. This, I think, brings out a critical point about control that the automated/autonomous distinction is trying to express. While I come at this though human review and approval of defining targeting classes (which could in theory be “autonomously” generated by an AI in a format readily comprehensible to humans), there is an assumption here that such human review and approval will constitute acceptance of targeting rules or targeting behaviour. Such rules or behaviour may emerge from a machine learning process or be keyed in by humans setting up a more traditional AI expert system where behaviour follows from explicitly defined rules. From the point of view of affirming “meaningful human control” there is a case for treating rule-following systems differently from rule-initiating systems.

A rule-following system follows human-defined targeting rules and accepts human-defined normative constraints to achieve human-defined goals. On current technology it seems to me that an autonomous system such as Aegis can adequately express the will of the ship’s commander as described in Scharre (2018). If such systems are to be classified as automated not autonomous, then it should be recognized that some people want to ban offensive “automated” weapons systems as well as offensive “autonomous” ones on moral grounds such as the “dignitarian” argument presented in Heyns (2016). This claims that delegating lethal decisions to machinery violates a fundamental human right to dignity even more basic than the right to life.

A rule-initiating system might discover targeting rules and normative constraints in training data. More ominously, a rule-initiating system might choose non-human goals based on deep reinforcement learning or some other form of machine learning and define its own goals that may be hostile to humans. Also such a system may create or discover its own rules or select action on the basis of oscillations in neural networks that may be inscrutable (as much machine learning currently is) to humans.

There is a case to stigmatize such systems as being “beyond” any form of “meaningful human control.” A system that can define its own targeting policy and execute it without any human review or approval is clearly unacceptable. It is hard to see how command responsibility could work with such a system. Arguably, such a system is already unlawful under current IHL.

It is also hard to see what interest any state has in building a system that might decide on the basis of an evolving “value function” or “genetic algorithm” that the world is better off without that state or the humans in it or indeed that the entire world is better off without any states or any humans in it at all.

A system has to be able to demonstrate to those fielding it that it selects action in accordance with IHL. From a systems architecture point of view, it should be possible to design a machine learning system that can learn new tactics and yet abide by normative constraints. The Alpha Go and Alpha Go Zero systems were both capable of superhuman performance in choosing tactical moves but neither ignored the normative rules of Go.

The reasoning an AWS uses should be auditable by humans. Its targeting policy should be comprehensible to humans prior to activation so they can approve it and accept responsibility for the actions of the AWS. This latter requirement poses deep challenges for “inscrutable” machine learning systems. However future research may solve these problems.

Obviously, regardless of the system architecture, Article 36 review (ICRC 1977) is critical in verifying that an AWS can be operated in compliance with IHL before fielding.

The fundamental ideas of Auditable Reasoning and a Responsible Officer assuming responsibility for AWS configuration and operation derive from the Responsibility Advisor in Arkin (2009). The phrase “Auditable Reasoning” comes from statements by the NZ delegation to the CCW (New Zealand 2018).

Banned Weapons

Table 1 provides examples of weapons banned by the two drafts.

Weapon	**Autonomy in**	**Tactical Role**	**Draft A**	**Draft B**
Naval & anti-tank mines	Select & Engage	Defensive	Permit	Permit
‘Fire & forget’ torpedo	Engage	Offensive	Permit	Permit
‘Fire & forget’ loitering missile (e.g. Harpy)	Select & Engage	Offensive	Permit	Permit
CIWS (e.g. Phalanx/Aegis)	Select & Engage	Defensive	Permit	Permit
Arkin Drone	Select & Engage	Offensive	Ban	Permit
Kalashnikov Autonomous Tank	Select & Engage	Offensive	Ban	Permit
Taranis with onboard autonomy	Select & Engage	Offensive	Ban	Permit
Future rule-following system	Select & Engage	Offensive	Ban	Permit
Future rule-initiating system	Define, Select & Engage	Offensive	Ban	Ban

Table 1: Examples of Weapons Banned in Drafts A and B

A future rule-following system might take the form of a stealthy radio-silent offensive UCAV with onboard autonomy. Its targeting policy (rules of engagement) might be generated by a “strategic AI”. Even so, human review of its inspectable (not “inscrutable”) rules of engagement that would include IHL could be possible either using expert systems or by developing “explainable AI” that provides an “explanation” for machine learned behaviour. Such a system would have a “human in the wider loop” between define and select but no human between select and engage. I suspect many would bitterly oppose such a system. However states seeking to maintain “top tier” status in air power may insist such systems are not banned.

Articles Common to Drafts A and B

The first three articles in both drafts are identical. They relate to scope, definitions and an “auditable reasoning” or “explicability” requirement (Floridi, Cowls et al. 2018).

Article 1: Scope of Application

1. This protocol relates to the use of autonomous weapons systems, defined herein, on land, sea and air.
2. This protocol applies only to autonomy in the critical functions of targeting as defined herein and to the functions of activation and deactivation as defined herein.
3. Autonomy in non-targeting functions such as navigation and refuelling is not regulated by this protocol.

Article 2: Definitions

1. “Autonomy” and “autonomous” refer to systems that are capable of operating in a real-world environment without external human control for a protracted period of time.
2. The “critical functions of targeting” are 1) defining targets, 2) selecting targets, and, 3) engaging targets.
3. “Autonomous Weapons System” (AWS) means a weapons system that has autonomy in one or more of the critical functions of targeting.
4. “Defining targets” means defining what classes of objects the autonomous weapon selects and engages.
5. “Selecting targets” means sensing and confirming objects meets the defined targeting criteria and aiming at them.
6. “Engaging targets” means firing on or using force against the selected targets.
7. “Activation” means turning on the AWS and sending it into offensive combat or enabling its defensive combat function.
8. “Deactivation” means withdrawing the AWS from combat and turning it off.
9. "Responsible Officer(s)" means the human or humans who assume responsibility for the configuration, fitness for purpose and state of repair of the AWS and who can be held accountable for its actions between activation and deactivation.
10. "Fully autonomous" means an AWS that has no humans involved in any of the critical functions of targeting. A fully autonomous AWS defines, selects and engages targets with no external human control.

Article 3: Auditable Reasoning

1. Logs containing timestamped data used by the AWS to make targeting decisions must be kept in an auditable form. It must be possible to inspect the logs and audit the reasoning used by the AWS to engage targets.

Draft A Articles

Draft A requires a "human in the loop" between the critical functions of select and engage. A human is required to assume responsibility prior to activating the AWS. Humans must define targeting criteria. Some "grandfather clauses" provide exemptions from the "human in the loop" requirement for existing widely fielded AWS.

Article 4: Prohibitions

1. AWS that are fully autonomous as defined in Article 2.10 are prohibited.
2. Autonomy in the function of defining targeting criteria is prohibited. The Responsible Officer must understand what targets the AWS will attack and be satisfied that the AWS can conform to IHL in such attacks.
3. Autonomy in the function of engaging targets is prohibited. Except as provided for in Article 6, a Responsible Officer must confirm the decisions of the AWS to engage selected targets with a positive act.
4. Autonomy in the activation function is not permitted. A human must activate an AWS after a Responsible Officer has assumed responsibility for its configuration, fitness for purpose and state of repair.

Article 5: Grandfather Clauses

1. This protocol does not apply to anti-tank and anti-ship mines.
2. This protocol does not apply to 'fire and forget' acoustic torpedoes and anti-radiation missiles.
3. Close-in weapons systems that due to military necessity must operate at an operational tempo too rapid for effective human control are permitted to be designed so that a Responsible Officer can monitor the select decisions of the AWS and abort engagements or deactivate the AWS in real time. Such AWS are permitted to fire autonomously if the Responsible Officer does not intervene to abort the engage decision.

Draft B Article

Draft B is similar to Draft A in that it prohibits autonomy in the critical function of defining targets and requires a human to assume responsibility for the configuration and state of repair of the AWS at activation. Unlike Draft A, it does not require a "human in the loop" between the critical functions of select and engage. A human in the "wider loop" between the critical functions of define and select is deemed sufficient to implement "meaningful human control."

Article 4: Prohibitions

1. AWS that are fully autonomous as defined in Article 2.10 are prohibited.
2. Autonomy in the function of defining targeting criteria is prohibited. The Responsible Officer(s) must understand what classes of targets the AWS will attack and be satisfied that the AWS can conform to IHL in such attacks.
3. An AWS may not be activated without at least one Responsible Officer assuming responsibility for its targeting configuration, fitness for purpose and state of repair.
4. Autonomy in the activation function is not permitted. A human must activate an AWS after a Responsible Officer has assumed responsibility for its configuration, fitness for purpose and state of repair.

Conclusion

It is reiterated that these drafts are tentative and incomplete. Only the most critical articles relating to scope, definitions and prohibitions are included. Some readers might think them suspiciously short, however Protocol VI of the CCW banning blinding lasers has only 4 articles (ICRC 1995) comprising 174 words. Protocol II, which regulated anti-personnel landmines, is considerably longer (14 articles) and also has a technical annex (ICRC 1996). Should nations see fit to add a Protocol VI to the CCW I imagine it would contain such an annex and be of similar length to Protocol II.

Similar to Protocol II, a Protocol VI would contain clauses relating to such matters as transfers, recordkeeping, compliance, technological cooperation and the like as well as the core clauses regarding scope, definitions and prohibitions presented here.

Additional articles could also cover such things as range and payload to address concerns about weapons of mass destruction (WMD) becoming autonomous. Matters such as distinction, proportionality, necessity and legal review of new AWS are already covered by existing IHL.

The regulation of weapons is and will continue to be complex. However, numerous delegations attending CCW meetings at the UN have pressed for discussions to move towards a tangible outcome. This short paper hopes to stimulate concrete and focused discussion as to the eventual form AWS regulation should take.

Acknowledgements

The author is much indebted to Mark Pearson of the New Zealand Department of Foreign Affairs and Trade, Reuben Steff of Waikato University, Jai Galliott of the Australian Defence Force Academy, Ron Arkin of Georgia Tech, Paul Scharre of the Centre for a New American Security and Mary Wareham, Noel Sharkey and Rob Sparrow of the Campaign to Stop Killer Robots for helpful discussions on LAWS and LAWS regulation.

References

AIV/CAVV. (2015). "Autonomous Weapons Systems: The Need for Meaningful Human Control." Advisory Council on International Affairs / Advisory Committee on Issues of Public International Law Retrieved 1st Nov, 2018, from https://aiv-advies.nl/download/606cb3b1-a800-4f8a-936f-af61ac991dd0.pdf.

Arkin, R. C. (2009). Governing Lethal Behaviour in Autonomous Robots. Boca Rouge, CRC Press.

Bekey, G. A. (2005). Autonomous robots: from biological inspiration to implementation and control, MIT press.

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke and E. Vayena (2018). "AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." Minds and Machines **28**(689).

Heyns, C. (2016). Autonomous Weapons Systems: Living a Dignified Life and Dying a Dignified Death. Autonomous Weapons Systems: Law, Ethics Policy. N. Bhuta, S. Beck, R. Geiss, H.-Y. Liu and C. Kress. Cambridge, Cambridge University Press**:** 3-20.

ICRC. (1977). "Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977. Article 36." Retrieved 24th Feb, 2015, from https://www.icrc.org/ihl/WebART/470-750045?OpenDocument.

ICRC. (1995). "Protocol on Blinding Laser Weapons (Protocol IV to the 1980 Convention), 13 October 1995." Retrieved 2nd Nov, 2018, from https://ihl-databases.icrc.org/ihl/INTRO/570.

ICRC. (1996). "Protocol on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices as amended on 3 May 1996 (Protocol II to the 1980 CCW Convention as amended on 3 May 1996)." Retrieved 6th Mar, 2015, from https://www.icrc.org/ihl/INTRO/575.

ICRC. (2018). "Ethics and Autonomous Weapons Systems: an ethical basis for human control?" International Committee of the Red Cross Retrieved 5th Nov, 2018, from https://www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf.

New Zealand. (2018). "Statement by Katy Donnelly Deputy Permanent Representative to the Conference on Disarmament, Geneva." CCW - Discussions on emerging technologies in the area of LAWS Retrieved 12th April, 2018, from https://www.unog.ch/80256EDD006B8954/(httpAssets)/4A54EE38E2F8223AC12582720057E761/$file/2018_LAWS6b_New+Zealand.pdf.

Scharre, P. (2018). Army of None. New York, WW Norton & Co.

UNIDIR. (2017). "The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches." UNIDIR Resources Retrieved 1st Nov, 2018, from http://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-concerns-characteristics-and-definitional-approaches-en-689.pdf.