ICRAC *Working Paper #2* May 2013

ICRAC
International Committee for Robot Arms Control

Compliance Measures for an Autonomous Weapons Convention

Principal Authors [1] Mark Gubrud[2], Jürgen Altmann[3]

General Considerations

Agreements to limit or prohibit certain types of arms – either in the context of arms control or of international humanitarian law[4] – always raise the concern that a party that violates the terms may gain an advantage, in armed conflict, over one that does not. Therefore many such agreements include measures for promoting, implementing and verifying compliance.

The types and extent of compliance measures may depend on many factors, including the military significance of the controlled weapons or actions, the difficulty of distinguishing systems and activities that are prohibited from those that are allowed, preexisting norms and levels of transparency, and the costs and acceptability of various measures. In the history of international arms limitations, the compliance measures agreed upon have ranged from leaving each state to monitor its own and others' compliance independently, to establishing international organizations with sophisticated technical inspection and monitoring systems.

Several arms control and international humanitarian law agreements and obligations lack any compliance measures, yet are regularly respected by states. Examples include the bans on "dumdum" bullets,[5] x-ray invisible fragments,[6] and blinding lasers,[7] as well as many other rules and principles of international humanitarian law, embodied in the Geneva Conventions, their Additional Protocols, and other documents, which govern both permissible weapons and conduct in war. Some of these have gained the status of customary international law,[8] and hence are incumbent even upon states that have not formally acceded to them; rules have been established in customary IHL for promoting compliance and prosecuting war crimes.[9]

Other agreements, such as those banning anti-personnel landmines and cluster munitions, set forth their own provisions for inquiry and investigation of suspected or alleged noncompliance. In addition, these agreements require state parties to enact their own national implementing measures which set penalties for banned activities, to report the numbers, type and status of banned weapons they are in the process of eliminating, and to participate in consultations and review conferences. These and similar measures set standards of implementation, promote transparency and build confidence, and make noncompliance more difficult to conceal. Non-governmental organizations (NGOs) can also help; in particular, the Landmine and Cluster Munitions Monitor (LCMM) plays as strong role as the *de facto* independent and respected verification mechanism of the treaty.

A higher level of verification is provided by official monitoring of declared facilities and weapon systems to ensure that their characteristics and uses fall within prescribed limits. Such measures, for a multilateral treaty, are typically implemented by a treaty implementing organization (TIO).[10] Technical measures include tamper-proof monitoring and tagging

devices as well as on-site inspections and forensic analysis by expert personnel. Evidence for the existence of undeclared activities and systems may also be collected and evaluated by the TIO when so mandated.

Considerations for Autonomous Weapons

The past decade has witnessed the advent and rapid growth in the development and use, especially by the United States, of weaponized "drones" and, more generally, air, land and water vehicles, large and small, that carry arms and have no on-board crew. A complete prohibition of all such uninhabited armed vehicles would be straightforward to verify through on-site inspections of military sites and other forms of monitoring. Most such vehicles would lack any accommodation for human crew and so would be easily distinguished from piloted and crewed vehicles.

A treaty that prohibits autonomous fire decision but allows remotely controlled and "semi-autonomous" weapons presents a more complex set of challenges. If a "semi-autonomous weapon system" may have capabilities to autonomously acquire, track, identify, group and prioritize targets, and to control their engagement once a "go" signal is given,[11] conversion to full lethal autonomy could be as simple as throwing a (software) switch. Given continued trends in technology, the addition of such capabilities to remotely controlled armed vehicles already equipped with sophisticated sensors and general purpose computers might also reduce to a matter of installing new software. Given the potentially high military importance of some kinds of fully autonomous weapons, especially those designed to attack major weapon systems (perhaps in swarms), there would be a significant risk of fully autonomous options being secretly prepared for systems officially declared to be under human control.

However, militarily potent fully autonomous weapons systems will likely require extensive development and testing while being operated under full autonomous control (though perhaps under human supervision). It would be difficult to conceal the large-scale activities that would be involved in such programs, especially if they are made clear violations of accepted norms and of a binding treaty.

By starting with a declaratory undertaking to forgo the development, testing, production and use of fully autonomous weapons, the international community would establish a normative goal and buy time to avoid a runaway arms race. As our understanding of the forms and capabilities of possible autonomous weapons deepens, more detailed limits may be established and clarified, with particular attention to blocking the development and deployment of those systems which pose the greatest threats. Provisions for such further clarifications, and a process for making them, should be incorporated in the treaty.

Since verification of the non-existence of an autonomous option in software is virtually impossible, and would be deemed far too intrusive, a tamper-proof system will be needed that can verify, after the fact, that an attack in question was under direct control of a human being ("in the loop," not "on the loop"). This could be achieved by keeping the records of each engagement and making the records of specific engagements available to a Treaty Implementing Organization, on request, when sufficient evidence exists to support suspicions of illegal autonomous operation.

Certain strictly defensive systems, where human safety is at stake and where human reactions are too slow for an effective response, may be exempted from the prohibition, provided they are operated under human supervision. Cases which meet these criteria may include missile and artillery interception systems which defend human-inhabited vehicles or locations. A strict criterion of necessity should be applied; in cases where human reaction is possible, the system should delay engagement to allow a human decision until imperative safety reasons compel an automatic response. In no case should autonomous engagement of human targets be permitted. Such allowances will complicate the terms of an agreement, but if they are narrowly restricted and clearly defined they do not pose particularly difficult challenges for verification.

Specific Proposals

Given the challenges in drawing a clear line across a complicated space of possibilities, and of holding that line when it is easily crossed and there are potential military advantages from doing so, prohibition of autonomous weapons requires a strong set of compliance measures. Perhaps the most fundamental is global recognition of the dangers of an open-ended robot arms race, and, responding to this, state commitment to forgoing autonomous weapons, and to establishing and sustaining a regime of preventive arms control. Entwined with this is the establishment of an unequivocal, universal norm demanding a human decision for each single use of violent force, and the implementation of measures to verify human control and to enforce accountability in each instance.

As a philosophical and legal foundation, the principles that the use of violent force must always be under human control, that decision in the use of force is a human responsibility, and that it is a human right not to be subjected to violent force or coercion on the decision of a machine, should be asserted as primary, and added to the canons of just war theory, ethics and international humanitarian law, especially as taught to military officers and personnel. Together with specific legal terms of prohibition and its implementation, these can be embodied in an Autonomous Weapons Convention (AWC).

The central obligations of state parties to an AWC will be: not to develop, test, produce, stockpile, deploy, transfer, broker transfer, or use weapon systems capable of autonomous target selection and engagement; not to permit autonomous target selection and weapons engagement by any machines under its jurisdiction; and to ensure that for each use of force against any target by means of any robotic weapon under its jurisdiction or control (whether lethal or nonlethal), the selection of the target, and decision to engage, are made by a human being who is responsible and accountable for that decision.

National implementing legislation should prohibit and impose penalties for any activities contrary to these obligations, and make it the responsibility of soldiers and citizens to refuse participation in and to report violations. State parties should be required to declare any pre-existing weapon systems that will be destroyed and programs that will be terminated when treaty comes into force. There should be provisions for consultations and procedures for requesting consultations in case of compliance issues arising. A treaty implementing organization (TIO) should be established to facilitate consultations, implement technical

safeguards, and conduct inquiries and investigations when so mandated. It should also be charged to develop a body of technical expertise on autonomous weapons and verification of their non-use. An NGO body like the LCMM should also independently monitor compliance and address gaps in national and TIO monitoring and reporting.

ICRAC scientists are developing proposals for technical safeguards which could verify that a responsible human operator has selected each target and initiated each engagement of a weapon system, under the authority of a responsible commander (which might be the same person), based on human, not machine judgment. Some initial ideas are presented below.

A compliance model based on transparency and confidence-building measures, inspections, technical safeguards, and forensic investigation of suspicious incidents, together with verification of human control and enforcement of accountability in the use of violent force, particularly by means of remotely-operated weapons and uninhabited vehicles, is sufficient for effective verification of a ban on fully autonomous weapons designed to engage personnel and nonstrategic military targets.

For issues of strategic concern, stronger and more specific measures may need to be developed, nationally and through the TIO, and could be added to the treaty regime as protocols or amendments. National technical means of verification will also be important resources.[12]

Definitions

Careful and explicit definitions will need to be given for each of the terms used; for example, "autonomous" is generally understood, in this context, to mean functioning independently of human action, though possibly under human supervision and with the possibility of human intervention. Here a distinction must be made with the word "automatic." The general sense is that "autonomous" implies a higher level of complexity in a system's ability to collect and process relevant information and in the relationship between that information and behavior; in other words, a higher level of (artificial) intelligence. It is possible to give a technical definition of "autonomy" in this sense which permits us to distinguish "autonomous" from "automatic" quantitatively, on the criterion of a measure of complexity.

As an alternative, it may be sufficient to define an "autonomous weapon" (AW),[13] as any system that acts independently of human action in "engagement-related functions" such as the acquisition, tracking, identification, grouping, selection, prioritization and engagement of targets.[14] Each step in this so-called "kill chain" or "loop" involves functions which the weapon system might fulfill autonomously. If any of these functions are autonomous, the weapon system may be classified as an AW, and if all of them are autonomous, the system is a fully autonomous weapon (FAW).

Under this paradigm, the treaty definition may simply exclude certain very simple systems, to be considered as merely "automatic" and not as AW. These exclusions, such as proximity fuses, mines, and heat-seeking missiles, can be enumerated and described in detail, either as an exhaustive list or as a set of typical examples. General technical criteria can also be given, including weapons type and complexity.

In addition, definitions will need to be given for those high-complexity FAW which are to be permitted as exceptions, principally those systems which are purely defensive against incoming projectiles which must be engaged in a time too short for human decision and response. The conditions under which such systems are permissible need to be spelled out; potential requirements include that they must be defending a human-inhabited location or vehicle, that they must be operated under accountable human supervision, and that to the greatest extent feasible they must give the human operator adequate information and maximum time and opportunity to abort or intervene in an erroneous engagement.

Standards

A problem related to definitions is the setting of criteria for human control and responsibility in the decision to use violent force. The difficulty and importance of this is indicated by the language of the US Department of Defense's Directive on Autonomy in Weapon Systems, which refers repeatedly to "selection" of targets by a "human operator" as the crucial step that distinguishes "semi-autonomous weapons" (SAW) from fully autonomous weapons; a SAW may "cue" its operator to "potential targets," but the operator must "select" them. Yet the definition offered for "target selection" – "The determination that an individual target or a specific group of targets is to be engaged." – fails to clarify what this means in practice. Does the operator need to move a cursor over the potential target's image, or if there is only one potential "target group" in play, can the operator just say "Go"? If the operator is using some type of brain-computer interface, can "determination" be as little as a conscious decision?

We believe that in order for any level of "autonomy in engagement-related functions" to be acceptable under an AWC, clear requirements must be stated and met. Each engagement decision must be taken under the authority of an accountable commander, and the weapon system itself must be under the control of an accountable operator (who may be the same person). The commander must have sufficient information, without relying on machine assessment, target recognition or preprocessing of raw data, to distinguish combatants from noncombatants, to determine that the military objectives outweigh harm or risks to noncombatants and civilian objects, and to respect all other applicable rules of international humanitarian law. If the system, and other resources, do not provide sufficient information to make these determinations, the commander's obligation is to hold fire. The operator must have positive control of target selection and engagement, so that unintended engagements are nearly impossible. If the system does not provide such positive control, the operator's obligation is to refuse use of the system. Neither the commander nor the operator may evade responsibility as a result of technical limitations of the system.

Additionally, the AWC may set forth standards for the operator's interface. An unmistakable, undeniable, affirmative action of the operator may be required both for "selection" when there is any degree of ambiguity, such as when multiple "potential targets" or "target groups" are indicated, and again to initiate engagement. Some kind of "handshaking" between the operator and system may be required for confirmation. Control by brain-computer interface may be prohibited.[15]

Technical Safeguards and Verification of Human Control and Responsibility

If remotely-operated weapons (ROW), including armed uninhabited vehicles, and semi-autonomous weapons as described by the US Department of Defense, will continue to be used and permitted under an AWC, in order to hold the line and prevent its crossing into prohibited fully autonomous weapons, technical safeguards and verification measures should be implemented to verify that each engagement of a weapon falling into one of these categories, as well as the operation of permitted FAW for terminal defense, is carried out under the authority of an accountable commander and control of an accountable operator. ICRAC scientists have begun work toward proposing such technical measures.

ICRAC's assumption is that state parties to the AWC will be willing to accept on-site inspections, sharing of some data, requirements for more extensive private data recording and preservation, and the installation of monitoring and reporting devices with known, open-source functions, provided that the information revealed by such procedures is strictly circumscribed and costs are not excessive. The benefit to participating state parties is to provide evidence of their compliance and thereby promote the compliance of other states as well as refuting spurious allegations of noncompliance.

Proving that the command to select and to engage a particular target was the action of a particular person is difficult, but an evidence trail that such a command was given can be generated and made difficult to forge or tamper with. Such evidence could include video and other data that was presented to the operator, plus a video view that includes the operator's hands on controls and the operator's view of the console. The data record would also include the commands as received by the console and codes for the identities of the accountable operator and accountable commander, which might be encrypted in physical keys which they are personally accountable for controlling at all times, and which are needed in order to operate the weapon.

A time slice of the data stream immediately prior to and including the selection and engagement commands could be designated as the primary record of the engagement. This record would be held by the state party, but a cryptographic code called a "hash" of the record would be recorded by a "glass box" (not "black" because its hardware and software would be known and open) together with a time stamp of the moment the engagement command was issued. The hash would serve as a digital seal of the engagement record; if even a single bit of the record were later altered, the hash would not match. The hash and the time stamp, recorded together, could be referred to as a "use of force identifier" (UFI). The UFIs would be periodically downloaded during on-site inspections by the TIO, which would also verify that the glass boxes were functional and properly installed. The UFIs would be held in a repository by the TIO. While the TIO would make every effort to ensure security of the UFI database, its compromise would not reveal any useful intelligence, but only strings of gibberish.

To strengthen the evidence trail, glass boxes could also be installed at the receiving end, on armed uninhabited vehicles and other ROW. All ROW would need to be registered with the TIO, and the glass boxes would need to be periodically inspected and their data downloaded.

The glass box on the weapon would be capable of detecting the launch of a missile, firing of a gun, or other engagement action of the weapon, either independently, with a signal provided by the weapon, or both. It would record the time of the event, plus a hash of data generated by the weapon system, which the system would retain until downloaded to custody of the state party. The UFI would also be transmitted, through the weapon system's communications links, from the glass box on the console to the glass box on the weapon, immediately following the engagement, and would be recorded by the glass box on the weapon. The time stamp of the UFI's issuance at the console would have to be prior to the time stamp recorded for the weapon firing, in order for the firing to have been caused by a command from the safeguarded console. The presence of the UFI in the glass box on the weapon would also show that the particular weapon was in communication with the particular safeguarded console at the time of the engagement.

This conceptual sketch is intended as representative of initial thinking about technical verification measures for an AWC, not the final word. The basic approach, though, seems plausible. Tactical information about engagements and technical details of weapon system hardware and software would not be disclosed, but the UFI hash codes would serve to prevent tampering with the records kept by the state party. In the event of a question about whether the weapon involved in a particular use of force was operating autonomously or under accountable human control, the state operating the weapon could be asked to produce the records which it kept of that use of force, perhaps in an encrypted form but tamper protected by the hash code held by the TIO. The state party could then selectively reveal verified details of the use of force event to an orderly process of inquiry conducted by the TIO.

---

[1] The principal authors gratefully acknowledge the contributions of other ICRAC members to this paper, including valuable ideas, suggestions and critical review.

[2] Program on Science and Global Security, Princeton University, USA

[3] Experimentelle Physik III, Technische Universität Dortmund, Germany

[4] This paper addresses compliance with an AWC as an international problem, so we have assumed that the AWC is framed as a security measure, as an extension of international humanitarian law, or both. It is also possible that the AWC would reference human rights law, as the core principles (discussed herein) fit also within a human rights framework. HRL is a vitally important source setting restraints on arms transfers and use. In view of the non-derogable right to life and the customary prohibitions against torture and genocide, arms transfers ought to take into account the risks resulting from contravening the *erga omnes* obligations in the context of HRL. In addition, the torture and genocide conventions rose to the level of peremptory or *ius cogens* obligations from which their absolute character cannot be derogated. D. Garcia 2011, Disarmament Diplomacy and Human Security - Norms, Regimes, and Moral Progress in International Relations, Routledge 2011.

[5] 1899 Hague Convention, Declaration III; see http://www.icrc.org/eng/assets/files/other/irrc_849_coupland_et_loye.pdf.

[6] Convention on Certain Conventional Weapons, Protocol I.

[7] *Ibid*, Protocol IV.

[8] See e.g. ICRC, Customary International Humanitarian Law, http://www.icrc.org/customary-ihl/eng/docs/home.

[9] http://www.icrc.org/customary-ihl/eng/docs/v1_cha, Ch. 40-44.

[10] Examples include the Organization for Prohibition of Chemical Weapons, Comprehensive Test Ban Treaty Organization, and International Atomic Energy Agency; the latter has a broader mandate than just arms control.

[11] Autonomy in Weapon Systems, Department of Defense Directive No. 3000.09, November 21, 2012, http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf (DoDD 3000.09).

[12] In case of a serious concern about strategically threatening noncompliance, a state party may seek private diplomatic conversation with the suspected perpetrator, formal consultations or inquiries through the treaty-mandated consultative bodies, or bring the matter to the attention of the UN Security Council. In the last resort, a state party may withdraw from the treaty, although this would be a highly provocative and destabilizing move.

[13] There is in general no need to distinguish between "autonomous weapon" and "autonomous weapon system" or to enforce the use of one term or the other. In other contexts, "weapon" may be considered as referring specifically to that part of a larger system which acts directly on the target, but in this context it is only the fact of being part of an autonomous system that can possibly make a weapon autonomous (it makes no sense to speak, for example, of an autonomous knife blade). Sometimes it may be desired to emphasize the view of an AWS as a system which includes many parts besides the directly harmful part; in other cases it may be desired to emphasize the view of the entire system as a weapon. In some cases, a subsystem of an AWS may itself constitute an AWS, which we may call an AW.

[14] See definition of "semi-autonomous weapon system" in DoDD 3000.09.

[15] The potential tactical advantages of such hypothetical interfaces have been greatly exaggerated.