REVIEW OF THE 2012 US POLICY ON AUTONOMY IN WEAPONS SYSTEMS

Human Rights Watch and Harvard Law School International Human Rights Clinic

April 2013

Introduction

On November 21, 2012, the US Department of Defense issued its first public policy on autonomy in weapons systems. Directive Number 3000.09 (the Directive) lays out guidelines for the development and use of autonomous and semi-autonomous weapon systems by the Department of Defense.[1] The Directive also represents the first policy announcement by any country on fully autonomous weapons, which do not yet exist but would be designed to select and engage targets without human intervention.

The Directive came out two days after Human Rights Watch and the Harvard Law School International Human Rights Clinic (IHRC) released their report *Losing Humanity: The Case against Killer Robots.*[2] The report calls for a preemptive ban on the development, production and use of fully autonomous weapons at the international and national levels.[3]

The Directive does not put in place such a preemptive ban. For a period of up to ten years, however, it allows the Department of Defense to develop or use only fully autonomous systems that deliver non-lethal force, unless department officials waive the policy at a high level. Importantly, the Directive also recognizes some of the dangers to civilians of fully autonomous weapons and the need for prohibitions or controls, including the basic requirement that a human being be "in the loop" when decisions are made to use lethal

[1] "Autonomy in Weapons Systems," US Department of Defense, Directive Number 3000.09, November 21, 2012, http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf (accessed March 4, 2013).

[2] Human Rights Watch, *Losing Humanity: The Case against Killer Robots*, November 2012, http://www.hrw.org/sites/default/files/reports/arms1112ForUpload_0_0.pdf.

[3] A prohibition on the development of fully autonomous weapons would not represent a ban on the development of all fully autonomous robotics technology or all autonomous weapons.

force. The Directive is in effect a moratorium on fully autonomous weapons with the possibility for certain waivers. It also establishes guidelines for other types of autonomous and semi-autonomous systems.

While a positive step, the Directive does not resolve the moral, legal, and practical problems posed by the potential development of fully autonomous systems. As noted, it is initially valid for a period of only five to ten years, and may be overridden by high level Pentagon officials. It establishes testing requirements that may be unfeasible, fails to address all technological concerns, and uses ambiguous terms. It also appears to allow for transfer of fully autonomous systems to other nations and does not apply to other parts of the US government, such as the Central Intelligence Agency (CIA). Finally, it lays out a policy of voluntary self-restraint that may not be sustainable if other countries begin to deploy fully autonomous weapons systems, and the United States feels pressure to follow suit.

The issuance of the Directive reflects that the United States is turning toward increasingly autonomous weapons systems and highlights the need to draw lines between different ones. At the same time, the elaboration by the Department of Defense of some of the dangers of fully autonomous weapons, and its stated intention not to pursue them in the near future, are important developments that other countries should take note of. Human Rights Watch and IHRC believe that this policy should lay the basis for the United States to embrace a permanent, comprehensive, preemptive ban on fully autonomous weapons in the coming years. The ultimate objective should be a legally binding global ban, so that all nations abide by the same restraints and avoid a robotic arms race.

Terminology

The Department of Defense Directive and *Losing Humanity* use different terminology to describe related technology. Like many other sources, *Losing Humanity* calls a system that requires a human command in order to select targets and deliver force, such as today's armed drones, a "human-in-the-loop weapon." The Directive uses the term "semi-autonomous weapon system" to describe the lowest level of autonomous systems it covers: weapon systems that engage specific targets or types of targets selected by a human operator. *Losing Humanity* describes as a "human-on-the-loop weapon" a system that can select targets and deliver force under the oversight of a human operator who has

the power to intervene and terminate engagements, while the Directive's counterpart is "human-supervised autonomous weapon system." *Losing Humanity* calls a system that is capable of selecting targets and delivering force without any human input or interaction a "human-out-of-the-loop weapon," and the Directive refers to it as an "autonomous weapon system," although the Directive's latter term also encompasses human-supervised autonomous weapon systems. For ease of discussion, the report by Human Rights Watch and IHRC as well as this briefing paper adopt the term "fully autonomous weapons" to refer to both human-out-of-the-loop weapons and weapons that technically have a human on the loop but are effectively out-of-the-loop because human supervision is so limited.

While *Losing Humanity* concentrates on fully autonomous weapons, the Department of Defense Directive deals with policy questions related to weapon systems of varying levels of autonomy. It applies in part to existing, semi-autonomous (i.e., human-in-the-loop) weapons like drones. This briefing paper, however, will focus on the Directive's policies regarding fully autonomous weapon systems.

Positive Elements

The Department of Defense Directive acknowledges from the outset that fully autonomous weapons could endanger civilians in many ways. It recognizes in section 1(b) that failures may occur and lead to "unintended engagements." An unintended engagement is defined as "[t]he use of force resulting in damage to persons or objects that human operators did not intend to be the targets of U.S. military operations." From a humanitarian perspective, the most significant type of damage is "unacceptable levels of collateral damage beyond those consistent with the law of war, ROE [rules of engagement], and commander's intent," in other words, civilian injury or death and harm to civilian objects.

In describing its purpose, the Directive highlights the need to control these weapons, specifically, as indicated in section 1(b), the need to establish guidelines to "minimize the probability and consequences of failures." The guidelines it lays out cover a wide range of activities including development, testing, legal review, international sales and transfers, and use.

In a key statement of policy, the Directive prohibits the use of lethal force by fully autonomous systems, though for a limited time period and with the possibility of a waiver.

Section 4(c)(3) restricts these weapon systems to the application of nonlethal, non-kinetic force and states that such force may be used only against materiel targets. The ban on the use of lethal force by fully autonomous systems is consistent with the one called for by Human Rights Watch and IHRC. If made permanent and comprehensive, such a ban would not only directly protect civilians from the threat of these weapons but also help prevent an arms race among high-tech militaries that could increase the potential humanitarian harm.

The Directive creates a different level of regulation for human-supervised autonomous weapon systems. Section 4(c)(2) limits the use of these systems to "local defense to intercept attempted time-critical or saturation attacks" and excludes the selection of humans as targets. When used in this way, human-supervised autonomous weapon systems are comparable to what *Losing Humanity* describes as automatic weapons defense systems: both are intended for defensive use against materiel targets. The report defines automatic weapons defense systems as systems "designed to sense an incoming munition, such as a missile or rocket, and to respond automatically to neutralize the threat."[4] Examples include the US MK 15 Phalanx Close-In Weapons System (CIWS) and the Counter Rocket, Artillery, and Mortar System (C-RAM), as well as Israel's Iron Dome.

The Department of Defense's lower level of regulation for human-supervised autonomous weapons parallels the position of Human Rights Watch and IHRC, which do not envision them falling under the definition and thus the prohibition of fully autonomous weapons. Automatic weapons defense systems can better be classified as automatic than autonomous because they carry out only "a pre-programmed sequence of operations or moves in a structured environment."[5] As a result, they do not appear to pose the same concerns as fully autonomous weapons, which are harder to control because they operate more freely and in more unstructured environments.

Waivers

The effectiveness of the Directive's restrictions on the development and use of fully autonomous weapons is reduced by two waiver provisions. Section 4(d) creates an

[4] Human Rights Watch and International Human Rights Clinic, *Losing Humanity*, p. 9.

[5] Ibid., p. 12 (quoting roboticist Noel Sharkey, "Automating Warfare: Lessons Learned from the Drones," *Journal of Law, Information & Science* (2011): EAP 2, accessed March 19, 2013, http://www.jlisjounral.org/abtracts/sharkey.21.2.html).

exception to the Directive's regulations. It states that "[a]utonomous or semi-autonomous weapon systems intended to be used in a manner that falls outside the policies in subparagraphs 4.c.(1) through 4.c.(3)" must be approved by two under secretaries of defense and by the chairman of the Joint Chiefs of Staff "before formal development and again before fielding." While overriding the default rule requires high level approval, the provision represents a waiver to the prohibition on the use of lethal force by fully autonomous systems.

The Directive has a loophole within this loophole. Enclosure 3 lays out a list of requirements that must be followed if the above wavier is granted. Many of these testing and training requirements may be waived, however, under Section 2 of that enclosure, which states that the under secretaries of defense for policy and acquisition, technology, and logistics "may request a Deputy Secretary of Defense waiver … in cases of urgent military operational need." In this situation, development and fielding of fully autonomous weapons could proceed without certain testing and design safeguards. Only the legal review would remain mandatory. The Directive also does not define "urgent military operational need," a phrase that is vulnerable to broad interpretation. These loopholes open the door to the development and use of fully autonomous weapons that could apply lethal force and thus have the potential to endanger civilians in armed conflict.

Testing

As discussed above, the Directive establishes requirements, such as testing, that must be met before a waiver is granted for the development or use of fully autonomous weapons that deliver lethal force. While an important and well-intended prerequisite, effective testing could prove very challenging to implement, according to some experts. Enclosure 3 requires testing "under realistic conditions, including possible adversary actions." Elsewhere the Directive says that testing should "ensure that autonomous and semi-autonomous weapon systems: (a) Function as anticipated in realistic operational environments against adaptive adversaries." Recreating a realistic operational environment is difficult, and the problem is exacerbated by the expected presence of adaptive adversaries. Changing behavior in response to an opponent or other conditions is especially common behavior in contemporary warfare and could be difficult to predict. It would thus be very difficult, if not impossible, to prove through testing in advance of

development or use that a fully autonomous weapon would protect civilians in accordance with the laws of war in all circumstances.

Technical Concerns

The Directive itself highlights the range of weaknesses of fully autonomous weapons. Its glossary contains an extensive list of the possible causes of failures in these and other autonomous weapons: human error, human-machine interaction failures, malfunctions, communications degradation, software coding errors, enemy cyber attacks or infiltration into the industrial supply chain, jamming, spoofing, decoys, other enemy countermeasures or actions, or unanticipated situations on the battlefield.

Implicitly noting further potential risks, the Directive discusses the need to ensure that the weapons systems:

- "function as anticipated in realistic operational environments against adaptive adversaries";
- are able to "complete engagements in a timeframe consistent with commander and operator intentions and, if unable to do so, terminate engagements or seek additional human operator input before continuing the engagement"; and
- "minimize failures that could lead to unintended engagements or to loss of control of the system to unauthorized parties."

While not a flaw in the Directive itself, these lists serve as reminders of the dangers of fully autonomous weapons.

There are additional technological concerns that the Directive does not address. For example, while fully autonomous weapons would be limited under the Directive to attacking anti-materiel targets, robot-on-robot engagements are inherently unpredictable and could create unforeseeable harm to civilians.[6]

[6] Noel Sharkey, "America's mindless killer robots must be stopped," *The Guardian*, December 3, 2012, http://www.guardian.co.uk/commentisfree/2012/dec/03/mindless-killer-robots (accessed March 4, 2013). Sharkey is a professor of artificial intelligence and robotics at the University of Sheffield. Sharkey also notes that the failures recognized in the Directive's definition lie largely outside of developers' control and thus "show the weakness of the whole enterprise."

Ambiguity

In two key places, ambiguity interferes with the Directive's clarity. Section 4(a) of the Directive mandates that weapon systems, including fully autonomous ones, "be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force." This language is repeated in Enclosure 3, which as discussed above sets guidelines for development and fielding of weapons that are excused from the policy's default rules. The language wisely recognizes the value of human judgment, which is essential for distinguishing between soldier and civilian under international humanitarian law. It is unclear, however, what an "appropriate level" of such judgment is. For example, could management of a dozen on-the-loop weapon systems by a single operator meet that standard? Human Rights Watch and IHRC recognize the difficulty at this stage of concretely defining the "appropriate" level of human judgment, but it will be essential for the Department of Defense to do so if it pursues ever greater autonomy in its weapons systems.

Proliferation

The Directive makes possible the proliferation of fully autonomous weapons, which could present additional humanitarian risks. Section 4(e) of the Directive allows for "international sales and transfers." Transfers must be "approved in accordance with existing technology security and foreign disclosure requirements and processes," but once these weapons leave the country, the United States would lose exclusive control over them. If a waiver were granted for development and use of fully autonomous weapons that deliver lethal force, the United States might also be able to export them. Once in circulation, such weapons could end up in the hands of rogue actors either through direct acquisition or reengineering of technology these actors gained access to through trade or capture. Abusive leaders could then employ fully autonomous weapons as tools of repression without the fear of mutiny presented by asking human soldiers to fire on their own people.

Expiration Date

The power of the Directive's ban on the use of fully autonomous weapons to deliver lethal force is undermined in the long run by the policy's temporary nature. Section 7 states that the Directive "[m]ust be reissued, cancelled, or certified current within 5 years of its publication" or will expire November 21, 2022 (10 years after it took effect). The

Department of Defense may be reluctant to renew the policy within the five-year period or to replace it with something equivalent after it expires. The United States might develop new technology under the waiver described above that would pave the way to development and production of fully autonomous weapons that could deliver lethal force, and having invested in the technology, the United States might find it difficult to give up. In addition, other nations not applying similar voluntary restraints could race ahead in efforts to deploy fully autonomous weapons systems, causing US officials to believe they have no choice but to deploy them as well. Other states, such as China and Russia, are known to be investing in unmanned systems already.[7]

By its very nature, a directive such as this one cannot guarantee long-term protections for civilians from the threat of fully autonomous weapons. Only an international regime binding on all nations could hope to do that.

Applicability

Finally, while a Department of Defense directive can by definition apply only to defense and military entities, that limitation leaves the possibility of unrestricted development, production, and use by other agencies. Paragraph 2(a)(2) specifies that the Directive is applicable only to the Office of the Secretary of Defense, the Military Departments, the Office of the Chairman of the Joint Chiefs of Staff and the Joint Staff, the Combatant Commands, the Office of the Inspector General of the Department of Defense, the Defense Agencies, the Department of Defense Field Activities, and all other organizational entities within the Department of Defense. This list notably excludes the Central Intelligence Agency (CIA). The CIA has played an active role in the use of drones, and its strikes have been controversial because of a lack of transparency as to the rules of engagement and its responsibility for civilian casualties.[8] It is unclear if the CIA could fund development and production of fully autonomous weapons. If it could, however, the same concerns about transparency and accountability could arise. The Directive should therefore be only a first step on the road to a broader US government policy.

[7] Human Rights Watch and IHRC, *Losing Humanity*, p. 7, n. 14.

[8] "US: Transfer CIA Drone Strikes to Military," Human Rights Watch news release, April 20, 2012, http://www.hrw.org/news/2012/04/20/us-transfer-cia-drone-strikes-military.

Conclusion

The Department of Defense Directive on autonomy in weapon systems has several positive elements that could have humanitarian benefits. It establishes that fully autonomous weapons are an important and pressing issue deserving of serious concern by the United States as well as other nations. It makes clear that fully autonomous weapons could pose grave dangers and are in need of restrictions or prohibitions. It is only valid for a limited time period of five to ten years, however, and contains a number of provisions that could weaken its intended effect considerably. The Directive's restrictions regarding development and use can be waived under certain circumstances. In addition, the Directive highlights the challenges of designing adequate testing and technology, is subject to certain ambiguity, opens the door to proliferation, and applies only to the Department of Defense.

Human Rights Watch and IHRC call on the United States to strengthen its national policy and absolutely prohibit the development, production, and use of fully autonomous weapons. They also urge it to lead other states in the creation of an international legally binding instrument that would create a comprehensive ban.