HUMAN RIGHTS WATCH

Losing Humanity

The Case against Killer Robots

Summary

NOVEMBER 19, 2012

Ban 'Killer Robots' Before It's Too Late

Fully Autonomous Weapons Would Increase Danger to Civilians

OCTOBER 21, 2013

UN: Hold International Talks on 'Killer Robots'

Governments Should Address Fully Autonomous Weapons

With the rapid development and proliferation of robotic weapons, machines are starting to take the place of humans on the battlefield. Some military and robotics experts have predicted that "killer robots"—fully autonomous weapons that could select and engage targets without human intervention—could be developed within 20 to 30 years. At present, military officials generally say that humans will retain some level of supervision over decisions to use lethal force, but their statements often leave open the possibility that robots could one day have the ability to make such choices on their own power. Human Rights Watch and Harvard Law School's International Human Rights Clinic (IHRC) believe that such revolutionary weapons would not be consistent with international humanitarian law and would increase the risk of death or injury to civilians during armed conflict. A preemptive prohibition on their development and use is needed.

A relatively small community of specialists has hotly debated the benefits and dangers of fully autonomous weapons. Military personnel, scientists, ethicists, philosophers, and lawyers have contributed to the discussion. They have evaluated autonomous weapons from a range of perspectives, including military utility, cost, politics, and the ethics of delegating life-and-death decisions to a machine. According to Philip Alston, then UN special rapporteur on extrajudicial, summary or arbitrary executions, however, "the rapid growth of these technologies, especially those with lethal capacities and those with decreased levels of human control, raise serious concerns that have been almost entirely unexamined by human rights or humanitarian actors."[1] It is time for the broader public to consider the potential advantages and threats of fully autonomous weapons.

The primary concern of Human Rights Watch and IHRC is the impact fully autonomous weapons would have on the protection of civilians during times of war. This report analyzes whether the technology would comply with international humanitarian law and preserve other checks on the killing of civilians. It finds that fully autonomous weapons would not only be unable to meet legal standards but would also undermine essential non-legal safeguards for civilians. Our research and analysis strongly conclude that fully autonomous weapons should be banned and that governments should urgently pursue that end.

Definitions and Technology

Although experts debate the precise definition, robots are essentially machines that have the power to sense and act based on how they are programmed.[2] They all possess some degree of autonomy, which means the ability of a machine to operate without human supervision. The exact level of autonomy can vary greatly. Robotic weapons, which are unmanned, are often divided into three categories based on the amount of human involvement in their actions:

- **Human-*in*-the-Loop Weapons**: Robots that can select targets and deliver force only with a human command;
- **Human-*on*-the-Loop Weapons**: Robots that can select targets and deliver force under the oversight of a human operator who can override the robots' actions; and
- **Human-*out*-*of*-the-Loop Weapons**: Robots that are capable of selecting targets and delivering force without any human input or interaction.

In this report, the terms "robot" and "robotic weapons" encompass all three types of unmanned weapons, in other words everything from remote-controlled drones to weapons with complete autonomy. The term "fully autonomous weapon" refers to both out-of-the-loop weapons and those that allow a human on the loop, but that are effectively out-of-the-loop weapons because the supervision is so limited.[3] A range of other terms have been used to describe fully autonomous weapons, including "lethal autonomous robots" and "killer robots."[4]

Fully autonomous weapons, which are the focus of this report, do not yet exist, but technology is moving in the direction of their development and precursors are already in use. Many countries employ weapons defense systems that are programmed to respond automatically to threats from incoming munitions. Other precursors to fully autonomous weapons, either deployed or in development, have antipersonnel functions and are in some cases designed to be mobile and offensive weapons. Militaries value these weapons because they require less manpower, reduce the risks to their own soldiers, and can expedite response time. The examples described in this report show that a number of countries, most notably the United States, are coming close to producing the technology to make complete autonomy for robots a reality and have a strong interest in achieving this goal.

Safeguards for Civilian Protection

According to international law and best practices, states should evaluate new or modified weapons to ensure they do not violate the provisions of international humanitarian law, also called the laws of war.[5] States should conduct weapons reviews at the earliest stages of development and continue them up through any production decision. Given military plans to move toward increasing autonomy for robots, states should now undertake formal assessments of the impacts of proposed fully autonomous weapons and technology that could lead to them even if not yet weaponized.

As this report shows, robots with complete autonomy would be incapable of meeting international humanitarian law standards. The rules of distinction, proportionality, and military necessity are especially important tools for protecting civilians from the effects of war, and fully autonomous weapons would not be able to abide by those rules. Roboticists have proposed different mechanisms to promote autonomous weapons' compliance with these rules; options include developing an ability to process quantitative algorithms to analyze combat situations and "strong artificial intelligence (AI)," which would try to mimic human thought. But even with such compliance mechanisms, fully autonomous weapons would lack the human qualities necessary to meet the rules of international humanitarian law. These rules can be complex and entail subjective decision making, and their observance often requires human judgment. For example, distinguishing between a fearful civilian and a threatening enemy combatant requires a soldier to understand the intentions behind a human's actions, something a robot could not do. In addition, fully autonomous weapons would likely contravene the Martens Clause, which prohibits weapons that run counter to the "dictates of public conscience."

By eliminating human involvement in the decision to use lethal force in armed conflict, fully autonomous weapons would undermine other, non-legal protections for civilians. First, robots would not be restrained by human emotions and the capacity for compassion, which can provide an important check on the killing of civilians. Emotionless robots could, therefore, serve as tools of repressive dictators seeking to crack down on their own people without fear their troops would turn on them. While proponents argue robots would be less apt to harm civilians as a result of fear or anger, emotions do not always lead to irrational killing. In fact, a person who identifies and empathizes with another human being, something a robot cannot do, will be more reluctant to harm that individual. Second, although relying on machines to fight war would reduce military casualties—a laudable goal—it would also make it easier for political leaders to resort to force since their own troops would not face death or injury. The likelihood of armed conflict could thus increase, while the burden of war would shift from combatants to civilians caught in the crossfire.

Finally, the use of fully autonomous weapons raises serious questions of accountability, which would erode another established tool for civilian protection. Given that such a robot could identify a target and launch an attack on its own power, it is unclear who should be held responsible for any unlawful actions it commits. Options include the military commander that deployed it, the programmer, the manufacturer, and the robot itself, but all are unsatisfactory. It would be difficult and arguably unfair to hold the first three actors liable, and the actor that actually committed the crime—the robot—would not be punishable. As a result, these options for accountability would fail to deter violations of international humanitarian law and to provide victims meaningful retributive justice.

Related Content

NOVEMBER 19, 2012 | News Release

Ban ‘Killer Robots’ Before It’s Too Late

Recommendations

Based on the threats fully autonomous weapons would pose to civilians, Human Rights Watch and IHRC make the following recommendations, which are expanded on at the end of this report:

To All States

- Prohibit the development, production, and use of fully autonomous weapons through an international legally binding instrument.
- Adopt national laws and policies to prohibit the development, production, and use of fully autonomous weapons.
- Commence reviews of technologies and components that could lead to fully autonomous weapons. These reviews should take place at the very beginning of the development process and continue throughout the development and testing phases.

To Roboticists and Others Involved in the Development of Robotic Weapons

- Establish a professional code of conduct governing the research and development of autonomous robotic weapons, especially those capable of becoming fully autonomous, in order to ensure that legal and ethical concerns about their use in armed conflict are adequately considered at all stages of technological development.

I. Unmanned Robots and the Evolution toward Fully Autonomous Weapons

Robots are not new to the battlefield, but their expanding role encroaches upon traditional human responsibilities more than ever before. Most visibly, the use of US Predator, Reaper, and other drones in Afghanistan and elsewhere has provided an early sign of the distancing of human soldiers from their targets. Often piloted from halfway around the globe, these robotic aerial vehicles provide surveillance and identify targets before a human decides to pull the trigger, commanding the drone to deliver lethal force.

In keeping with the escalating use of aerial drones, government planning documents and spending figures indicate that the military of the future will be increasingly unmanned. In recent years, for example, the US Department of Defense has spent approximately $6 billion annually on the research and development, procurement, operations, and maintenance of unmanned systems for war, and that figure is likely to increase rapidly.[6] Drones are seen as just the beginning of a technological revolution. As robotic warfare expert Peter W. Singer suggests, “Predators are merely the first generation—the equivalent of the Model T Ford or the Wright Brothers’ Flyer.”[7]

Unmanned technology possesses at least some level of autonomy, which refers to the ability of a machine to operate without human supervision.[8] At lower levels, autonomy can consist simply of the ability to return to base in case of a malfunction. If a weapon were fully autonomous, it would “identify targets and … trigger itself.”[9] Today’s robotic weapons still have a human being in the decision-making loop, requiring human intervention before the weapons take any lethal action. The aerial drones currently in operation, for instance, depend on a person to make the final decision whether to fire on a target. As this chapter illustrates, however, the autonomy of weapons that have been deployed or are under development is growing quickly. If this trend continues, humans could start to fade out of the decision-making loop, retaining a limited oversight role—or perhaps no role at all.

Plans for Autonomy

Military policy documents, especially from the United States, reflect clear plans to increase the autonomy of weapons systems. In its *Unmanned Systems Integrated Roadmap FY2011-2036*, the US Department of Defense wrote that it "envisions unmanned systems seamlessly operating with manned systems while gradually reducing the degree of human control and decision making required for the unmanned portion of the force structure."[10] The US plans cover developments in ground, air, and underwater systems. A US roadmap specifically for ground systems stated, "There is an ongoing push to increase UGV [unmanned ground vehicle] autonomy, with a current goal of 'supervised autonomy,' but with an ultimate goal of full autonomy."[11] According to the US Air Force, "[i]ncreasingly humans will no longer be 'in the loop' but rather 'on the loop'—monitoring the execution of certain decisions. Simultaneously, advances in AI will enable systems to make combat decisions and act within legal and policy constraints without necessarily requiring human input."[12] A 2004 US Navy planning document on unmanned undersea vehicles (UUVs) stated, "While admittedly futuristic in vision, one can conceive of scenarios where UUVs sense, track, identify, target, and destroy an enemy—all autonomously."[13] Other countries are also devoting attention and money to unmanned systems. [14]

While emphasizing the desirability of increased autonomy, many of these military documents also stress that human supervision over the use of deadly force will remain, at least in the immediate future. According to the US Department of Defense, "[f]or the foreseeable future, decisions over the use of force and the choice of which individual targets to engage with lethal force will be retained under human control in unmanned systems."[15] The UK Ministry of Defence stated in 2011 that it "currently has no intention to develop systems that operate without human intervention in the weapon command and control chain."[16] Such statements are laudable but do not preclude a change in that policy as the capacity for autonomy evolves.

Although the timeline for that evolution is debated, some military experts argue that the technology for fully autonomous weapons could be achieved within decades. The US Air Force predicted that "by 2030 machine capabilities will have increased to the point that humans will have become the weakest component in a wide array of systems and processes."[17] The UK Ministry of Defence estimated in 2011 that artificial intelligence "as opposed to complex and clever automated systems" could be achieved in 5 to 15 years and that fully autonomous swarms could be available in 2025.[18] Other experts have quoted similar estimates while cautioning that intelligence for weapons that equals that of a human is much further off, and many experts believe it is impossible.[19]

The next two sections examine the development of increasingly autonomous weapons. They are not the focus of this report, which instead highlights the risks posed by fully autonomous weapons. They show, however, that autonomous technology already exists and is evolving rapidly. An analysis of these weapons also leads to the conclusion that development of greater autonomy should proceed cautiously, if at all.

Automatic Weapons Defense Systems

Automatic weapons defense systems represent one step on the road to autonomy. These systems are designed to sense an incoming munition, such as a missile or rocket, and to respond automatically to neutralize the threat. Human involvement, when it exists at all, is limited to accepting or overriding the computer's plan of action in a matter of seconds.

The US Navy's MK 15 Phalanx Close-In Weapons System is designed to identify and fire at incoming missiles or threatening aircraft. This automatic weapons defense system, shown during a live fire exercise, is one step on the road to full autonomy. Photograph by Chief Fire Controlman Brian Kirkwood, US Navy.

The United States has several such systems. The US Navy's MK 15 Phalanx Close-In Weapons System, the earliest model, was first installed on a ship in 1980, and modified versions are still widely used by the United States and its allies.[20] It is designed to sense approaching anti-ship missiles or threatening aircraft and respond with fire from two 20mm guns with six rotating barrels.[21] The guns each fire 3,000 to 4,500 rounds per minute.[22] More recent models aim to defend against small gunboats, artillery, and helicopters.[23] The Navy describes the Phalanx as "the only deployed close-in weapon system capable of autonomously performing its own search, detect, evaluation, track, engage and kill assessment functions."[24]

The Counter Rocket, Artillery, and Mortar System (C-RAM) is a US land-based version of the Phalanx. The United States first deployed it at forward operating bases in Iraq in 2005. Twenty-two systems reportedly had more than 100 successful intercepts of rockets, artillery, and mortars [25] and provided more than 2,000 warnings to troops.[26] Like the Phalanx, it can fire 20mm rounds from a six-barrel gun at an incoming munition.[27] According to one US Army publication, after the C-RAM detects a threat, "a human operator certif[ies] the target,"[28] but that would have to happen almost instantaneously in order for the C-RAM to destroy the incoming munition in time.

Other countries have developed comparable weapons defense systems. Israel has deployed its Iron Dome near the border with Gaza and in Eilat, near the Sinai Peninsula.[29] It uses radar to identify short-range rockets and 155mm artillery shells up to 70 kilometers away.[30] It is armed with 20 Tamir interceptor missiles to respond to such threats, and extended range versions of those missiles are reportedly scheduled to be available in early 2013.[31] Israel has received financial support for the Iron Dome from the United States, and the US Department of Defense stated that the system, which reportedly has more than an 80 percent success rate, "has shot down scores of missiles that would have killed Israeli civilians since it was fielded in April 2011."[32] In a split second after detecting an incoming threat, the Iron Dome sends a recommended response to the threat to an operator. The operator must decide immediately whether or not to give the command to fire in order for the Iron Dome to be effective.[33]

An Iron Dome, an Israeli automatic weapons defense system, fires a missile from the city of Ashdod in response to a rocket launch from the nearby Gaza Strip on March 11, 2012. The Iron Dome sends warnings of incoming threats to an operator who must decide almost instantly whether to give the command to fire. © 2012 Jack Guez, AFP/Getty Images

Another example of an automatic weapons defense system is the NBS Mantis, which Germany designed to protect its forward operating bases in Afghanistan. The "short-range force protection system will detect, track and shoot the projectiles within a close range of the target base." Within 4.5 seconds after detecting targets about three kilometers away, it can fire six 35mm automatic guns at 1,000 rounds per minute.[34] The system has a “very high degree of automation, including automatic target detection and engagement processes which the operator only has to monitor.”[35] Sources were unclear whether “monitoring” also allowed the operator to override the process.

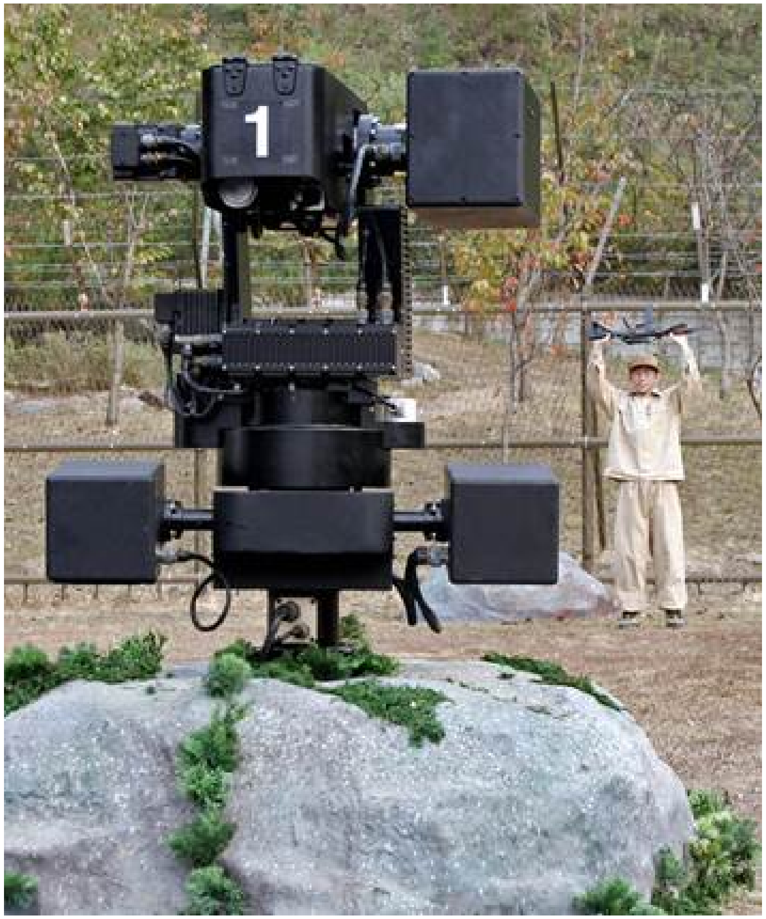
These weapon defense systems have a significant degree of autonomy because they can sense and attack targets with minimal human input. Technically, they fall short of being fully autonomous and can better be classified as automatic. Robotics professor Noel Sharkey defines an automatic robot as one that “carries out a pre-programmed sequence of operations or moves in a structured environment. A good example is a robot arm painting a car.” An autonomous robot, he continues, “is similar to an automatic machine except that it operates in open and unstructured environments. The robot is still controlled by a program but now receives information from its sensors that enable it to adjust the speed and direction of its motors (and actuators) as specified by the program.”[36] Nevertheless, while not the focus of this report, these automatic defense systems can be seen as a step toward greater autonomy in weapons.

As weapons that operate with limited intervention from humans, automatic weapons defense systems warrant further study. On the one hand, they seem to present less danger to civilians because they are stationary and defensive weapons that are designed to destroy munitions, not launch offensive attacks.[37] On the other hand, commentators have questioned the effectiveness of the human supervision in the C-RAM and other automatic weapons defense systems. Writing about the C-RAM, Singer notes, “The human is certainly part of the decision making but mainly in the initial programming of the robot. During the actual operation of the machine, the operator really only exercises veto power, and a decision to override a robot’s decision must be made in only half a second, with few willing to challenge what they view as the better judgment of the machine.”[38] When faced with such a situation, people often experience “automation bias,” which is “the tendency to trust an automated system, in spite of evidence that the system is unreliable, or wrong in a particular case.”[39] In addition, automatic weapons defense systems have the potential to endanger civilians when used in populated areas. For example, even the successful destruction of an incoming threat can produce shrapnel that causes civilian casualties.[40] Thus these systems raise concerns about the protection of civilians that full autonomy would only magnify.

Other Precursors to Fully Autonomous Weapons

Other unmanned weapons systems that currently retain humans in or on the loop are also potential precursors to fully autonomous weapons. Militaries have already deployed ground robots, and air models are under development.[41] If their potential for full autonomy in the use of lethal force were realized, these systems would pose a greater threat to civilians than automatic weapons defense systems. As currently designed, the systems discussed below would all have the capability to target humans. In addition, the increased mobility and offensive nature of the air systems in particular would give them more range and make them harder to control than weapons like the Phalanx.

South Korea and Israel have developed and started to use sentry robots that operate on the ground. South Korea installed SGR-1s, costing $200,000 each, along the Demilitarized Zone (DMZ) for testing in 2010.[42] These stationary robots can sense people in the DMZ with heat and motion sensors and send warnings back to a command center.



The South Korean SGR-1 sentry robot, a precursor to a fully autonomous weapon, can detect people in the Demilitarized Zone and, if a human grants the command, fire its weapons. The robot is shown here during a test with a surrendering enemy soldier. © 2007 Kim Dong-Joo/AFP/Getty Images

From there, human soldiers can communicate with the individual identified and decide whether to fire the robot sentry's 5.5mm machine gun or 40mm automatic grenade launcher.[43] The SGR-1's sensors can detect people two miles away during the day and one mile away at night. Its guns can hit targets two miles away.[44] At present, the sentry has autonomous surveillance capabilities, but it cannot fire without a human command.

[45] The journal of the Institute of Electrical and Electronics Engineers reported, however, "[T]he robot does have an automatic mode, in which it can make the decision."[46]

The Israel Defense Forces (IDF) has deployed Sentry Tech systems along Israel's 60 kilometer border with Gaza. The sentry detects movement and sends signals to a facility "at a distant location."[47] Soldiers then evaluate the data and decide whether to fire at the target. The Sentry Tech currently has a 12.7mm or a .50 caliber machine gun with a kill zone of about 1 to 1.5 kilometers.[48] To increase its range to several kilometers, the IDF is considering adding anti-armor missiles.[49] Sentry Tech is reportedly designed to defend against people trying to cross the border as well as sniper and rocket attacks.[50] In 2007, *Jane's Defence Weekly* described Sentry Tech as "revolutionary" because it could not only detect threats but also engage them.[51] While the system is currently operated by remote control, an IDF division commander told *Defense News*: "[A]t least in the initial phases of deployment, we're going to have to keep the man in the loop."[52] The commander thus implied that human involvement may not always be the case.

Israel has also deployed the Guardium, "a semi-autonomous unmanned ground system," which is reportedly used for patrolling Israel's border with Gaza. It can carry lethal or non-lethal payloads. According to the manufacturer G-NIUS's brochure, "[t]he Guardium UGV™ was designed to perform routine missions, such as programmed patrols along border routes, but also to autonomously react to unscheduled events, in line with a set of guidelines specifically programmed for the site characteristics and security doctrine."[53] While the brochure implies there is some level of human oversight because it refers to stationary, mobile, and portable control terminals, it also notes that the Guardium can have "autonomous mission execution."[54]

Unmanned aircraft are moving beyond existing drones to have greater autonomy. The US Navy has commissioned the X-47B, which will be able to take off from and land on an aircraft carrier and refuel on its own power.[55] It was tested multiple times in 2012 and is scheduled in 2013 to do a trial landing on a carrier, "one of aviation's most difficult maneuvers."[56] Although as a prototype it will not carry weapons, it has reportedly been designed for eventual "combat purposes,"[57] and it has two weapons bays with a total payload capacity of 4,500 pounds.[58] Humans remain on the loop for the time being, but their role in the flight of the X-47B is limited. Northrop Grumman described it as a system that "takes off, flies a preprogrammed mission, and then returns to base in response to mouse clicks from its mission operator. The mission operator monitors the X-47B air vehicle's operation, but does not actively 'fly' it via remote control as is the case for other unmanned systems currently in operation."[59] The *Los Angeles Times* called it "a paradigm shift in warfare, one that is likely to have far-reaching consequences. With the drone's ability to be flown autonomously by onboard computers, it could usher in an era when death and destruction can be dealt by machines operating semi-independently."[60]

The US Navy’s X-47B, currently undergoing testing, has been commissioned to fly with greater autonomy than existing drones. While the prototype will not carry weapons, it has two weapons bays that could make later models serve a combat function. Photograph by DARPA.

The United Kingdom unveiled a prototype of its Taranis combat aircraft in 2010. Designers described it as “an autonomous and stealthy unmanned aircraft” that aims to strike “targets with real precision at long range, even in another continent.”[61] Because Taranis is only a prototype, it is not armed, but it includes two weapons bays and could eventually carry bombs or missiles.[62] Similar to existing drones, Taranis would presumably be designed to launch attacks against persons as well as materiel. It would also be able to defend itself from enemy aircraft.[63] At this point, the Taranis is expected to retain a human in the loop. The UK Ministry of Defence stated, “Should such systems enter into service, they will at all times be under the control of highly trained military crews on the ground.”[64] Asked if the Taranis would one day choose its own targets, Royal Air Force Air Chief Marshal Simon Bryant responded, “This is a very sensitive area we are paying a lot of attention to.”[65] He thus left the door open to the possibility of greater autonomy in the future.[66]

The United Kingdom's Taranis combat aircraft, whose prototype was unveiled in 2010, is designed to strike distant targets, "even in another continent." While the Ministry of Defence has stated that humans will remain in the loop, the Taranis exemplifies the move toward increased autonomy. © 2010 Associated Press

Other countries have also developed or procured unmanned aircraft that are precursors to fully autonomous weapons. The Israeli Harpy, for example, has been described as a combination of an unmanned aerial vehicle and a cruise missile. It is designed to fly "autonomously to the patrol area." Once there, it seeks to detect hostile radar signals and then destroy a target with a high explosive warhead.[67]

The US military's SWARMS technology would also involve autonomous aircraft, but in this case, many such aircraft would navigate in a synchronized way with a human controller directing them as a group "swarm" rather than individually.[68] While initially designed to gather intelligence,[69] SWARMS could one day undertake offensive operations. For example, their numbers, designed to be a force multiplier, could overwhelm an air defense system.[70] At least at this point, designers envision that SWARMS would have a human on the loop. Tests done in August 2012 showed that a single "operator on the ground, using only a laptop and a military radio, can command an unmanned aerial vehicle (UAV) 'swarm.'"[71] The ability of a single operator to have effective oversight of dozens or even hundreds of aircraft seems implausible to many experts.[72] As a result, a swarm could operate as a de facto out-of-the-loop weapon.

Because humans still retain control over the decisions to use lethal force, the above weapons systems are not, at least yet, fully autonomous, but they are moving rapidly in that direction. Human oversight is minimal, especially in the case of SWARMS. At the same time, technology is developing that allows weapons to identify targets and travel to and around a battle zone on their own power. Proponents tout military advantages, such as a reduction in the number of human troops required for military operations, the availability of sentries not influenced by laziness or fear, and faster response time.[73] Technological developments combined with these advantages of autonomy create incentives for states to develop weapons with greater autonomy.

Critics have two major concerns, however. First, they question the effectiveness of the existing limited human oversight.[74] Second, they worry that the next step will be to grant these systems control over launching attacks. Speaking of Taranis, for example, Sharkey, a computer scientist and vocal critic of fully autonomous weapons, said, “But warning bells ring for me when they talk about Taranis being ‘a fully autonomous intelligent system’ together with applications in ‘deep missions’ and having a ‘deep target attack’ capability.... We need to know if this means the robot planes will choose their own targets and destroy them—because they certainly will not have the intelligence to discriminate between civilians and combatants.”[75] Control systems specialist Nick Jennings did not object to SWARMS technology as a surveillance tool, but he warned, “We don’t want UAVs selecting targets and working out how best to carry out an attack.”[76] Full autonomy would give weapons this power to decide when to fire.

Given that some believe that full autonomy could become a reality within 20 or 30 years, it is essential to consider the implications of the technology as soon as possible. Both supporters and skeptics have agreed on this point.[77] The UK Ministry of Defence wrote, “[I]f we wish to allow systems to make independent decisions without human intervention, some considerable work will be required to show how such systems will operate legally.”[78] Philip Alston, when serving as the UN special rapporteur on extrajudicial, summary or arbitrary executions, warned that “[u]rgent consideration needs to be given to the legal, ethical and moral implications of the development and use of robot technologies, especially but not limited to uses for warfare.”[79] The rest of this report will explore these implications, particularly as they relate to the protection of civilians during times of armed conflict.

II. Article 36 and the Requirement to Review New Weapons

States should review new and modified weapons for their compliance with international law. This rule is codified in Article 36 of Additional Protocol I to the Geneva Conventions, which states:

> In the study, development, acquisition or adoption of a new weapon, means or method of war, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.[80]

Whether considered new types of weapons or modifications of ones that have greater human supervision, autonomous weapons should be subject to such reviews. In fact, the International Committee of the Red Cross (ICRC) specifically highlighted autonomous weapons as an area of concern in its authoritative commentary on Article 36. The ICRC wrote, "The use of long distance, remote control weapons, or weapons connected to sensors positioned in the field, leads to the automation of the battlefield in which the soldier plays an increasingly less important role.... [A]ll predictions agree that if man does not master technology, but allows it to master him, he will be destroyed by technology."[81] This statement from 1987 raised alarms about the risks of partly autonomous weapons. The warning is even more apt for fully autonomous models.

All states, whether or not they are party to Protocol I, should conduct weapons reviews. Some experts contend that Article 36 is customary international law binding on all states, while others see it as a best practice.[82] The ICRC argues that the obligation applies to all states because "the faithful and responsible application of its international law obligations would require a State to ensure that the new weapons, means and methods of warfare it develops or acquires will not violate these obligations."[83] Regardless of their opinion of Article 36's legal status, many weapons producing states have accepted the obligation to review. The list includesthe United States, which is not party to Protocol I but has been a leader in robot research.[84]

The review of weapons, including robotic ones, should take place at the earliest stage possible and continue through any development that proceeds. For a producing state, “reviews should take place at the stage of the conception/design of the weapon, and thereafter at the stages of its technological development (development of prototypes and testing), and in any case before entering into the production contract.”[85] Evaluations of weapons being modified should similarly be started early in the process.[86] States must also review weapons that they plan to acquire rather than produce themselves. Given that certain states, such as the United States, are putting large amounts of money into research and development of autonomous weapons, the time to begin reviews is now. In addition to being required, an early assessment is in a state’s interest because weapons development can be expensive and it makes sense to avoid costs that may produce only an unlawful weapon.[87]

When determining if a review is necessary, states should define “weapon” broadly to encompass major components and final products. There has been some concern that the United States dodged a review of at least one new unmanned system by arguing that evaluations of its individual components were sufficient. The Judge Advocate General’s office reportedly said that it did not need to review a newly weaponized Predator drone because both the Predator, when used for surveillance, and the Hellfire missile with which it was to be armed, had previously passed reviews when considered separately.[88] An ICRC guide to Article 36, however, says reviews should cover “an existing weapon that is modified in a way that alters its function, or a weapon that has already passed a legal review but that is subsequently modified.”[89] This rule is especially important for robots because they are complex systems that often combine a multitude of components that work differently in different combinations.

While international legal standards encourage states to review new weapon systems for their compliance with international law, the United States said it did not need to review the newly weaponized Predator drone, shown here firing a Hellfire missile. It argued that the drone and missile had already received approval separately, at a time when the Predator was used for surveillance. © Pan-African News

Reviews should also be sensitive to the fact that some robotic technology, while not inherently harmful, has the potential one day to be weaponized. As soon as such robots are weaponized, states should initiate their regular, rigorous review process. It would be even better to review this technology before weaponization to ensure that robots do not get to that stage, especially since states are more reluctant to give up weapons later in the development process.[90] Such reviews would be designed to preempt fully autonomous weapons that are inconsistent with international humanitarian law, not to block all work in robotics.

The purpose of a weapons review is to determine if the new or modified weapon would be prohibited by international law. First, states should consider prohibitions under existing weapons treaties.[91] While it is possible that fully autonomous weapons could include components banned or regulated by such treaties, there is no existing treaty that prohibits them as a class.[92] States must then evaluate whether a weapon runs counter to other treaties or customary law. Particularly significant for this discussion are the rules of distinction, proportionality, and military necessity, the cornerstones of international humanitarian law, all of which are accepted as customary.[93] The Martens Clause, which prohibits weapons that run counter to the "dictates of public conscience," may also be relevant.[94]

The requirement of distinction is arguably the bedrock principle of international humanitarian law. According to customary international law, articulated in Protocol I to the Geneva Conventions, combatants must "distinguish between the civilian population and combatants."[95] Attacks that fail to distinguish are indiscriminate and unlawful. Indiscriminate attacks include those that do not target a specific military objective, "employ a method or means of combat which cannot be directed at a specific military objective," or "employ a method or means of combat the effects of which cannot be limited."[96]

International humanitarian law also prohibits disproportionate attacks, in which civilian harm outweighs military benefits. Protocol I defines a disproportionate attack as one that "may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."[97] Determination of proportionality requires a subjective balancing of the military and civilian impacts of an attack as anticipated before it is launched.

Although not clearly articulated in a single treaty, the concept of military necessity "infuses" international humanitarian law.[98] It provides that lethal force may only be used for the explicit purpose of defeating an adversary, it must be proportionate, and it "cannot justify violation of the other rules of [international humanitarian law]."[99] As one scholar described it, "Military necessity dictates that military force should only be used against the enemy to the extent necessary for winning the war."[100] For example, attacking surrendering or wounded troops would be unlawful because it is not essential for victory and is expressly prohibited by the Geneva Conventions.[101]

Finally, reviews should assess a weapon under the Martens Clause.[102] The clause dates back to the 1899 and 1907 Hague Conventions and was codified more recently in Article 1(2) of Protocol 1, which states:

> In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from dictates of public conscience.[103]

In other words, even if a means of war does not violate an existing treaty or customary law, it can still be found unlawful if it contravenes the principles of humanity or the dictates of public conscience. The International Court of Justice, which found the rule to be customary international law, noted that it "had proved to be an effective means of addressing rapid evolution of military technology."[104] The clause is a useful tool for evaluating and governing emerging weapons because they often develop faster than international law.[105]

States interested in developing or acquiring fully autonomous weapons should initiate detailed legal reviews of any existing or proposed technology that could lead to such robots. These reviews should begin in the early stages of development, address all configurations of the weapons, and consider such key principles of international humanitarian law as distinction, proportionality, and military necessity. They should also take into account the Martens Clause. States should then cease development of any weapons that fail to meet legal requirements before they become so invested in the technology that they will be reluctant to give it up.

III. International Humanitarian Law Compliance Mechanisms

Proponents of fully autonomous weapons have recognized that such new robots would have to comply with international humanitarian law. Supporters have therefore proposed a variety of compliance mechanisms, two of which will be discussed below, that seek to prevent any violations of the laws of war.[106]

Arkin's "Ethical Governor"

Ronald Arkin, a roboticist at the Georgia Institute of Technology, has articulated the "most comprehensive architecture" for a compliance mechanism.[107] Recognizing the importance of new weapons meeting legal standards, Arkin writes, "The application of lethal force as a response *must* be constrained by the LOW [law of war] and ROE [rules of engagement] before it can be employed by the autonomous system."[108] He argues that such constraints can be achieved through an "ethical governor."

The ethical governor is a complex proposal that would essentially require robots to follow a two-step process before firing. First, a fully autonomous weapon with this mechanism must evaluate the information it senses and determine whether an attack is prohibited under international humanitarian law and the rules of engagement. If an attack violates a constraint, such as the requirement that an attack must distinguish between combatant and noncombatant, it cannot go forward. If it does not violate a constraint, it can still only proceed if attacking the target is required under operational orders.[109] The evaluation at this stage consists of binary yes-or-no answers.

Under the second step, the autonomous robot must assess the attack under the proportionality test.[110] The ethical governor quantifies a variety of criteria, such as the likelihood of a militarily effective strike and the possibility of damage to civilians or civilian objects, based on technical data. Then it uses an algorithm that combines statistical data with "incoming perceptual information" to evaluate the proposed strike "in a utilitarian manner."[111] The robot can fire only if it finds the attack "satisfies all ethical constraints and minimizes collateral damage in relation to the military necessity of the target."[112]

Arkin argues that with the ethical governor, fully autonomous weapons would be able to comply with international humanitarian law better than humans. For example, they would be able to sense more information and process it faster than humans could. They would not be inhibited by the desire for self-preservation. They would not be influenced by emotions such as anger or fear. They could also monitor the ethical behavior of their human counterparts.[113] While optimistic, Arkin recognizes that it is premature to determine whether effective compliance with this mechanism is feasible.[114]

"Strong AI"

Another, even more ambitious approach strives to "match and possibly exceed human intelligence" in engineering international humanitarian law-compliant autonomous robots.[115] The UK Ministry of Defence has recognized that "some form of artificial intelligence [AI]" will be necessary to ensure autonomous weapons fully comply with principles of international humanitarian law.[116] It defines a machine with "true artificial intelligence" as having "a similar or greater capacity to think like a human" and distinguishes that intelligence from "complex and clever automated systems."[117] John McGinnis, a Northwestern University law professor, advocates for the development of robotic weapons with "strong AI," which he defines as the "creation of machines with the general human capacity for abstract thought and problem solving."[118] McGinnis argues that "AI-driven robots on the battlefield may actually lead to less destruction, becoming a civilizing force in wars as well as an aid to civilization in its fight against terrorism."[119]

Such a system presumes that computing power will approach the cognitive power of the human brain, but many experts believe this assumption may be more of an aspiration than a reality. Whether and when scientists could develop strong AI is "still very much disputed."[120] While some scientists have argued that strong AI could be developed in the twenty-first century, so far it has been "the Holy Grail in AI research: highly desirable, but still unattainable."[121] Even if the development of fully autonomous weapons with human-like cognition became feasible, they would lack certain human qualities, such as emotion, compassion, and the ability to understand humans. As a result, the widespread adoption of such weapons would still raise troubling legal concerns and pose other threats to civilians. As detailed in the following sections, Human Rights Watch and IHRC believe human oversight of robotic weapons is necessary to ensure adequate protection of civilians in armed conflict.

IV. Challenges to Compliance with International Humanitarian Law

An initial evaluation of fully autonomous weapons shows that even with the proposed compliance mechanisms, such robots would appear to be incapable of abiding by the key principles of international humanitarian law. They would be unable to follow the rules of distinction, proportionality, and military necessity and might contravene the Martens Clause. Even strong proponents of fully autonomous weapons have acknowledged that finding ways to meet those rules of international humanitarian law are "outstanding issues" and that the challenge of distinguishing a soldier from a civilian is one of several "daunting problems."[122] Full autonomy would strip civilians of protections from the effects of war that are guaranteed under the law.

Distinction

The rule of distinction, which requires armed forces to distinguish between combatants and noncombatants, poses one of the greatest obstacles to fully autonomous weapons complying with international humanitarian law. Fully autonomous weapons would not have the ability to sense or interpret the difference between soldiers and civilians, especially in contemporary combat environments.

Changes in the character of armed conflict over the past several decades, from state-to-state warfare to asymmetric conflicts characterized by urban battles fought among civilian populations, have made distinguishing between legitimate targets and noncombatants increasingly difficult. States likely to field autonomous weapons first—the United States, Israel, and European countries—have been fighting predominately counterinsurgency and unconventional wars in recent years. In these conflicts, combatants often do not wear uniforms or insignia. Instead they seek to blend in with the civilian population and are frequently identified by their conduct, or their "direct participation in hostilities." Although there is no consensus on the definition of direct participation in hostilities, it can be summarized as engaging in or directly supporting military operations.[123] Armed forces may attack individuals directly participating in hostilities, but they must spare noncombatants.[124]

It would seem that a question with a binary answer, such as "is an individual a combatant?" would be easy for a robot to answer, but in fact, fully autonomous weapons would not be able to make such a determination when combatants are not identifiable by physical markings. First, this kind of robot might not have adequate sensors. Krishnan writes, "Distinguishing between a harmless civilian and an armed insurgent could be beyond anything machine perception could possibly do. In any case, it would be easy for terrorists or insurgents to trick these robots by concealing weapons or by exploiting their sensual and behavioral limitations."[125]

An even more serious problem is that fully autonomous weapons would not possess human qualities necessary to assess an individual's intentions, an assessment that is key to distinguishing targets. According to philosopher Marcello Guarini and computer scientist Paul Bello, "[i]n a context where we cannot assume that everyone present is a combatant, then we have to figure out who is a combatant and who is not. This frequently requires the attribution of intention."[126] One way to determine intention is to understand an individual's emotional state, something that can only be done if the soldier has emotions. Guarini and Bello continue, "A system without emotion … could not predict the emotions or action of others based on its own states because it has no emotional states."[127] Roboticist Noel Sharkey echoes this argument: "Humans understand one another in a way that machines cannot. Cues can be very subtle, and there are an infinite number of circumstances where lethal force is inappropriate."[128] For example, a frightened mother may run after her two children and yell at them to stop playing with toy guns near a soldier. A human soldier could identify with the mother's fear and the children's game and thus recognize their intentions as harmless, while a fully autonomous weapon might see only a person running toward it and two armed individuals.[129] The former would hold fire, and the latter might launch an attack. Technological fixes could not give fully autonomous weapons the ability to relate to and understand humans that is needed to pick up on such cues.

Proportionality

The requirement that an attack be proportionate, one of the most complex rules of international humanitarian law, requires human judgment that a fully autonomous weapon would not have. The proportionality test prohibits attacks if the expected civilian harm of an attack outweighs its anticipated military advantage.[130] Michael Schmitt, professor at the US Naval War College, writes, "While the rule is easily stated, there is no question that proportionality is among the most difficult of LOIAC [law of international armed conflict] norms to apply."[131] Peter Asaro, who has written extensively on military robotics, describes it as "abstract, not easily quantified, and highly relative to specific contexts and subjective estimates of value."[132]

Determining the proportionality of a military operation depends heavily on context. The legally compliant response in one situation could change considerably by slightly altering the facts. According to the US Air Force, "[p]roportionality in attack is an inherently subjective determination that will be resolved on a case-by-case basis."[133] It is highly unlikely that a robot could be pre-programmed to handle the infinite number of scenarios it might face so it would have to interpret a situation in real time. Sharkey contends that "the number of such circumstances that could occur simultaneously in military encounters is vast and could cause chaotic robot behavior with deadly consequences."[134] Others argue that the "frame problem," or the autonomous robot's incomplete understanding of its external environment resulting from software limitations, would inevitably lead to "faulty behavior."[135] According to such experts, the robot's problems with analyzing so many situations would interfere with its ability to comply with the proportionality test.

Those who interpret international humanitarian law in complicated and shifting scenarios consistently invoke human judgment, rather than the automatic decision making characteristic of a computer. The authoritative ICRC commentary states that the proportionality test is subjective, allows for a "fairly broad margin of judgment," and "must above all be a question of common sense and good faith for military commanders."[136] International courts, armed forces, and others have adopted a "reasonable military commander" standard.[137] The International Criminal Tribunal for the Former Yugoslavia, for example, wrote, "In determining whether an attack was proportionate it is necessary to examine whether a reasonably well-informed person in the circumstances of the actual perpetrator, making reasonable use of the information available to him or her, could have expected excessive civilian casualties to result from the attack."[138] The test requires more than a balancing of quantitative data, and a robot could not be programmed to duplicate the psychological processes in human judgment that are necessary to assess proportionality.

A scenario in which a fully autonomous aircraft identifies an emerging leadership target exemplifies the challenges such robots would face in applying the proportionality test. The aircraft might correctly locate an enemy leader in a populated area, but then it would have to assess whether it was lawful to fire. This assessment could pose two problems. First, if the target were in a city, the situation would be constantly changing and thus potentially overwhelming; civilian cars would drive to and fro and a school bus might even enter the scene. As discussed above, experts have questioned whether a fully autonomous aircraft could be designed to take into account every movement and adapt to an ever-evolving proportionality calculus. Second, the aircraft would also need to weigh the anticipated advantages of attacking the leader against the number of civilians expected to be killed. Each leader might carry a different weight and that weight could change depending on the moment in the conflict. Furthermore, humans are better suited to make such value judgments, which cannot be boiled down to a simple algorithm.[139]

Proponents might argue that fully autonomous weapons with strong AI would have the capacity to apply reason to questions of proportionality. Such claims assume the technology is possible, but that is in dispute as discussed above. There is also the threat that the development of robotic technology would almost certainly outpace that of artificial intelligence. As a result, there is a strong likelihood that advanced militaries would introduce fully autonomous weapons to the battlefield before the robotics industry knew whether it could produce strong AI capabilities. Finally, even if a robot could reach the required level of reason, it would fail to have other characteristics—such as the ability to understand humans and the ability to show mercy—that are necessary to make wise legal and ethical choices beyond the proportionality test.

Military Necessity

Like proportionality, military necessity requires a subjective analysis of a situation. It allows "military forces in planning military actions … to take into account the practical requirements of a military situation at any given moment and the imperatives of winning," but those factors are limited by the requirement of "humanity."[140] One scholar described military necessity as "a context-dependent, value-based judgment of a commander (within certain reasonableness restraints).”[141] Identifying whether an enemy soldier has become *hors de combat*, for example, demands human judgment.[142] A fully autonomous robot sentry would find it difficult to determine whether an intruder it shot once was merely knocked to the ground by the blast, faking an injury, slightly wounded but able to be detained with quick action, or wounded seriously enough to no longer pose a threat. It might therefore unnecessarily shoot the individual a second time. Fully autonomous weapons are unlikely to be any better at establishing military necessity than they are proportionality.

Military necessity is also relevant to this discussion because proponents could argue that, if fully autonomous weapons were developed, their use itself could become a military necessity in certain circumstances. Krishnan warns that the development of “[t]echnology can largely affect the calculation of military necessity.”[143] He writes: “Once [autonomous weapons] are widely introduced, it becomes a matter of military necessity to use them, as they could prove far superior to any other type of weapon.”[144] He argues such a situation could lead to armed conflict dominated by machines, which he believes could have “disastrous consequences.” Therefore, “it might be necessary to restrict, or maybe even prohibit [autonomous weapons] from the beginning in order to prevent a dynamics that will lead to the complete automation of war that is justified by the principle of necessity.”[145] The consequences of applying the principle of military necessity to the use of fully autonomous weapons could be so dire that a preemptive restriction on their use is justified.

Martens Clause

Fully autonomous weapons also raise serious concerns under the Martens Clause. The clause, which encompasses rules beyond those found in treaties, requires that means of warfare be evaluated according to the “principles of humanity” and the “dictates of public conscience.”[146] Both experts and laypeople have an expressed a range of strong opinions about whether or not fully autonomous machines should be given the power to deliver lethal force without human supervision. While there is no consensus, there is certainly a large number for whom the idea is shocking and unacceptable. States should take their perspective into account when determining the dictates of public conscience.

Ronald Arkin, who supports the development of fully autonomous weapons, helped conduct a survey that offers a glimpse into people’s thoughts about the technology. The survey sought opinions from the public, researchers, policymakers, and military personnel, and given the sample size it should be viewed more as descriptive than quantitative, as Arkin noted.[147] The results indicated that people believed that the less an autonomous weapon was controlled by humans, the less acceptable it was.[148] In particular, the survey determined that “[t]aking life by an autonomous robot in both open warfare and covert operations is unacceptable to more than half of the participants.”[149] Arkin concluded, “People are clearly concerned about the potential use of lethal autonomous robots. Despite the perceived ability to save soldiers' lives, there is clear concern for collateral damage, in particular civilian loss of life."[150] Even if such anecdotal evidence does not create binding law, any review of fully autonomous weapons should recognize that for many people these weapons are unacceptable under the principles laid out in the Martens Clause.

Conclusion

To comply with international humanitarian law, fully autonomous weapons would need human qualities that they inherently lack. In particular, such robots would not have the ability to relate to other humans and understand their intentions. They could find it difficult to process complex and evolving situations effectively and could not apply human judgment to deal with subjective tests. In addition, for many the thought of machines making life-and-death decisions previously in the hands of humans shocks the conscience. This inability to meet the core principles of international humanitarian law would erode legal protections and lead fully autonomous weapons to endanger civilians during armed conflict. The development of autonomous technology should be halted before it reaches the point where humans fall completely out of the loop.

V. Other Threats to Civilian Protection

In addition to being unable to meet international humanitarian law standards, fully autonomous weapons would threaten other safeguards against civilian deaths and injuries. Two characteristics touted by proponents as making these robots superior to human soldiers—their lack of emotion and their ability to reduce military casualties—can in fact undermine civilian protection. First, delegating to machines the decision of when to fire on a target would eliminate the influence of human empathy, an important check on killing. Second, assigning combat functions to robots minimizes military casualties but risks making it easier to engage in armed conflict and shifts the burden of war onto the civilian population. Humans should therefore retain control over the choice to use deadly force. Eliminating human intervention in the choice to use deadly force could increase civilian casualties in armed conflict.

The Lack of Human Emotion

Proponents of fully autonomous weapons suggest that the absence of human emotions is a key advantage, yet they fail adequately to consider the downsides. Proponents emphasize, for example, that robots are immune from emotional factors, such as fear and rage, that can cloud judgment, distract humans from their military missions, or lead to attacks on civilians. They also note that robots can be programmed to act without concern for their own survival and thus can sacrifice themselves for a mission without reservations.[151] Such observations have some merit, and these characteristics accrue to both a robot's military utility and its humanitarian benefits.

Human emotions, however, also provide one of the best safeguards against killing civilians, and a lack of emotion can make killing easier. In training their troops to kill enemy forces, armed forces often attempt "to produce something close to a 'robot psychology,' in which what would otherwise seem horrifying acts can be carried out coldly."[152] This desensitizing process may be necessary to help soldiers carry out combat operations and cope with the horrors of war, yet it illustrates that robots are held up as the ultimate killing machines.

Whatever their military training, human soldiers retain the possibility of emotionally identifying with civilians, "an important part of the empathy that is central to compassion."[153] Robots cannot identify with humans, which means that they are unable to show compassion, a powerful check on the willingness to kill. For example, a robot in a combat zone might shoot a child pointing a gun at it, which might be a lawful response but not necessarily the most ethical one. By contrast, even if not required under the law to do so, a human soldier might remember his or her children, hold fire, and seek a more merciful solution to the situation, such as trying to capture the child or advance in a different direction. Thus militaries that generally seek to minimize civilian casualties would find it more difficult to achieve that goal if they relied on emotionless robotic warriors.

Fully autonomous weapons would conversely be perfect tools of repression for autocrats seeking to strengthen or retain power. Even the most hardened troops can eventually turn on their leader if ordered to fire on their own people. A leader who resorted to fully autonomous weapons would be free of the fear that armed forces would rebel. Robots would not identify with their victims and would have to follow orders no matter how inhumane they were.

Several commentators have expressed concern about fully autonomous weapons' lack of emotion. Calling for preservation of the role of humans in decisions to use lethal force, a US colonel who worked on the US Future Combat Systems program recognized the value of human feelings.[154] He said, "We would be morally bereft if we abrogate our responsibility to make the life-and-death decisions required on a battlefield as leaders and soldiers with human compassion and understanding."[155] Krishnan writes:

> One of the greatest restraints for the cruelty in war has always been the natural inhibition of humans not to kill or hurt fellow human beings. The natural inhibition is, in fact, so strong that most people would rather die than kill somebody.... Taking away the inhibition to kill by using robots for the job could weaken the most powerful psychological and ethical restraint in war. War would be inhumanely efficient and would no longer be constrained by the natural urge of soldiers not to kill.[156]

Rather than being understood as irrational influences and obstacles to reason, emotions should instead be viewed as central to restraint in war.

Making War Easier and Shifting the Burden to Civilians

Advances in technology have enabled militaries to reduce significantly direct human involvement in fighting wars. The invention of the drone in particular has allowed the United States to conduct military operations in Afghanistan, Pakistan, Yemen, Libya, and elsewhere without fear of casualties to its own personnel. As Singer notes, "[M]ost of the focus on military robotics is to use robots as a replacement for human losses."[157] Despite this advantage, the development brings complications. The UK Ministry of Defence highlighted the urgency of more vigorous debate on the policy implications of the use of unmanned weapons to "ensure that we do not risk losing our controlling humanity and make war more likely."[158] Indeed, the gradual replacement of humans with fully autonomous weapons could make decisions to go to war easier and shift the burden of armed conflict from soldiers to civilians in battle zones.

While technological advances promising to reduce military casualties are laudable, removing humans from combat entirely could be a step too far. Warfare will inevitably result in human casualties, whether combatant or civilian. Evaluating the human cost of warfare should therefore be a calculation political leaders always make before resorting to the use of military force. Leaders might be less reluctant to go to war, however, if the threat to their own troops were decreased or eliminated. In that case, "states with roboticized forces might behave more aggressively.... [R]obotic weapons alter the political calculation for war."[159] The potential threat to the lives of enemy civilians might be devalued or even ignored in decisions about the use of force.[160]

The effect of drone warfare offers a hint of what weapons with even greater autonomy could lead to. Singer and other military experts contend that drones have already lowered the threshold for war, making it easier for political leaders to choose to use force.[161] Furthermore, the proliferation of unmanned systems, which according to Singer has a "profound effect on 'the impersonalization of battle,'"[162] may remove some of the instinctual objections to killing. Unmanned systems create both physical and emotional distance from the battlefield, which a number of scholars argue makes killing easier.[163] Indeed, some drone operators compare drone strikes to a video game because they feel emotionally detached from the act of killing.[164] As D. Keith Shurtleff, Army chaplain and ethics instructor for the Soldier Support Institute at Fort Jackson, pointed out, "[A]s war becomes safer and easier, as soldiers are removed from the horrors of war and see the enemy not as humans but as blips on a screen, there is a very real danger of losing the deterrent that such horrors provide."[165] Fully autonomous weapons raise the same concerns.

The prospect of fighting wars without military fatalities would remove one of the greatest deterrents to combat.[166] It would also shift the burden of armed conflict onto civilians in conflict zones because their lives could become more at risk than those of soldiers. Such a shift would be counter to the international community's growing concern for the protection of civilians.[167] While some advances in military technology can be credited with preventing war or saving lives, the development of fully autonomous weapons could make war more likely and lead to disproportionate civilian suffering. As a result, they should never be made available for use in the arsenals of armed forces.

VI. Problems of Accountability for Fully Autonomous Weapons

Given the challenges fully autonomous weapons present to adherence to international humanitarian law and the way they undermine other humanitarian protections, it is inevitable that they will at some point kill or injure civilians. When civilian casualties in armed conflict occur unlawfully, people want to see someone held accountable.[168] Accountability in such cases serves at least two functions: it deters future harm to civilians and provides victims a sense of retribution.[169] If the killing were done by a fully autonomous weapon, however, the question would become: whom to hold responsible. Options include the military commander, the programmer, the manufacturer, and even the robot itself, but none of these options is satisfactory. Since there is no fair and effective way to assign legal responsibility for unlawful acts committed by fully autonomous weapons, granting them complete control over targeting decisions would undermine yet another tool for promoting civilian protection.

The first option is to hold the military commanders who deploy such weapons responsible for the weapons' actions on the battlefield.[170] Given that soldiers are autonomous beings, commanders are not held legally responsible for the actions of their subordinates except in very particular circumstances. It seems equally unfair to impose liability on commanders for their fully autonomous weapons. These weapons' autonomy creates a "responsibility gap," and it is arguably unjust to hold people "responsible for actions of machines over which they *could not have* sufficient control."[171]

In certain situations, under the principle of "command responsibility," a commander may be held accountable for war crimes perpetrated by a subordinate. It applies if the commander knew or should have known that the individual planned to commit a crime yet he or she failed to take action to prevent it or did not punish the perpetrator after the fact.[172] While this principle seeks to curb international humanitarian law violations by strengthening commander oversight, the doctrine is ill suited for fully autonomous weapons. On the one hand, command responsibility would likely apply if a commander was aware in advance of the potential for unlawful actions against civilians and still recklessly deployed a fully autonomous weapon. This application would be legally appropriate. On the other hand, a commander might not be able to identify a threat pre-deployment because he or she had not programmed the robot. If the commander realized once a robot was in the field that it might commit a crime, the commander would be unable to reprogram it in real time to prevent the crime because it was designed to operate with complete autonomy. Furthermore, as will be discussed in greater detail below, a commander cannot effectively punish a robot after it commits a crime. Thus except in cases of reckless conduct, command responsibility would not apply, and the commander would not be held accountable for the actions of a fully autonomous weapon.

An unlawful act committed by a fully autonomous weapon could be characterized as the result of a design flaw. The notion that a violation is a technical glitch points toward placing responsibility for the robot's actions on its programmer or manufacturer, but this solution is equally unfair and ineffective. While the individual programmer would certainly lay the foundation for the robot's future decisions, the weapon would still be autonomous. The programmer could not predict with complete certainty the decisions a fully autonomous robot might eventually make in a complex battlefield scenario.[173] As Robert Sparrow, a professor of political philosophy and applied ethics, writes, "[T]he possibility that an autonomous system will make choices other than those predicted and encouraged by its programmers is inherent in the claim that it is autonomous."[174] To hold the programmer accountable, therefore, "will only be fair if the situation described occurred as a result of negligence on the part of the design/programming team."[175] Furthermore, to be held criminally liable under international humanitarian law, the programmer would have had to cause the unlawful act intentionally.[176] Assuming any miscoding by the programmer was inadvertent or produced unforeseeable effects, there would be no option for accountability here.

Some have pointed to the product liability regime as a potential model for holding manufacturers responsible for international humanitarian law violations caused by fully autonomous weapons.[177] If manufacturers could be held strictly liable for flaws in these weapons, it would provide an incentive for those manufacturers to produce highly reliable weapons to avoid liability. Yet the product liability regime also falls short of an adequate solution. First, private weapons manufacturers are not typically punished for how their weapons are used, particularly if the manufacturers disclose the risks of malfunction to military purchasers up front.[178] It is highly unlikely that any company would produce and sell weapons, which are inherently dangerous, knowing the firm could be held strictly liable for any use that violates international humanitarian law. Second, product liability requires a civil suit, which puts the onus on victims. It is unrealistic to expect civilian victims of war, who are often poverty stricken and geographically displaced by conflict, to sue for relief against a manufacturer in a foreign court, even if legal rules would allow them to recover damages. Thus, the strict liability model would fail to create a credible deterrent for manufacturers or provide retribution for victims.

Holding accountable any of the actors described above—commanders, programmers, or manufacturers—is not only unlikely to be fair or effective, but it also does nothing to deter robots themselves from harming civilians through unlawful acts. Fully autonomous weapons operate, by definition, free of human supervision and so their actions are not dependent on human controllers.[179] Fully autonomous weapons also lack any emotion that might give them remorse if someone else were punished for their actions. Therefore, punishment of these other actors would do nothing to change robot behavior.

Looking into the future, some have argued that the remaining party—the fully autonomous weapon itself—might be held responsible for the unlawful killing of civilians. Krishnan writes, “At the moment, it would obviously be nonsensical to do this, as any robot that exists today, or that will be built in the next 10-20 years, is too dumb to possess anything like intentionality or a real capability for agency. However, this might change in a more distant future once robots become more sophisticated and intelligent.”[180] If a robot were truly autonomous, the robot might be punished by being destroyed or having its programming restricted in some way. Merely altering a robot’s software, however, is unlikely to satisfy victims seeking retribution.[181] Furthermore, unless the robot understood that it would be punished for violating the law, its decisions would not be influenced by the threat of accountability.[182]

These proposed methods would all fail to ensure accountability for the same reasons. They would neither effectively deter future violations of international humanitarian law nor provide victims with meaningful retributive justice. Taking human beings out of the loop of robotic decision making would remove the possibility for real accountability for unlawful harm to civilians, making it all the more important that fully autonomous weapons are never developed or used.

Conclusion

Fully autonomous weapons have the potential to increase harm to civilians during armed conflict. They would be unable to meet basic principles of international humanitarian law, they would undercut other, non-legal safeguards that protect civilians, and they would present obstacles to accountability for any casualties that occur. Although fully autonomous weapons do not exist yet, technology is rapidly moving in that direction. These types of weaponized robots could become feasible within decades, and militaries are becoming increasingly invested in their successful development. Before it becomes even more challenging to change course, therefore, states and scientists should take urgent steps to review and regulate the development of technology related to robot autonomy. In particular, states should prohibit the creation of weapons that have full autonomy to decide when to apply lethal force.

To achieve these goals, Human Rights Watch and IHRC recommend:

To All States

Prohibit the development, production, and use of fully autonomous weapons through an international legally binding instrument.

States should preemptively ban fully autonomous weapons because of the threat these kinds of robots would pose to civilians during times of war. A prohibition would ensure that firing decisions are made by humans, who possess the ability to interpret targets' actions more accurately, have better capacity for judging complex situations, and possess empathy that can lead to acts of mercy. Preserving human involvement in the decision-making loop would also make it easier to identify an individual to hold accountable for any unlawful acts that occur from the use of a robotic weapon, thus increasing deterrence and allowing for retribution.

This prohibition should apply to robotic weapons that can make the choice to use lethal force without human input or supervision. It should also apply to weapons with such limited human involvement in targeting decisions that humans are effectively out of the loop. For example, a human may not have enough time to override a computer's decision to fire on a target, or a single human operator may not be able to maintain adequate oversight of a swarm of dozens of unmanned aircraft. Some on-the-loop weapons could prove as dangerous to civilians as out-of-the-loop ones. Further study will be required to determine where to draw the line between acceptable and unacceptable autonomy for weaponized robots.

Adopt national laws and policies to prohibit the development, production, and use of fully autonomous weapons.

National measures could serve as means of prohibition before the creation of an international instrument. They could also raise awareness of the problems of fully autonomous weapons and help establish best practices on how to deal with them.

Commence reviews of technologies and components that could lead to fully autonomous weapons. These reviews should take place at the very beginning of the development process and continue throughout the development and testing phases.

Such early and ongoing reviews help ensure that states do not develop weapons, like fully autonomous weapons, that fail to comply with international humanitarian law. States should make public their determinations about a weapon's or technology's ability to meet legal standards because transparency can allow for monitoring and confidence building. In addition, transparency would allow reviews to facilitate public debate about the problems and potential solutions.

To Roboticists and Others Involved in the Development of Robotic Weapons

Establish a professional code of conduct governing the research and development of autonomous robotic weapons, especially those capable of becoming fully autonomous, in order to ensure that legal and ethical concerns about their use in armed conflict are adequately considered at all stages of technological development.

A code of conduct for those involved with developing robotic weapons could help ensure that such technology evolves in accordance with the legal and ethical frameworks that protect civilians in armed conflict. Academic and scientific associations could draft and distribute the code. Codes of conduct for military technological development already exist in the fields of synthetic biology and nanotechnology.[183] They serve to increase transparency in research agendas and encourage researchers to adopt socially responsible approaches to scientific development.

Acknowledgments

This report was researched and written by Bonnie Docherty, senior researcher in the Arms Division of Human Rights Watch and senior clinical instructor at the International Human Rights Clinic (IHRC) at Harvard Law School. Julia Fitzpatrick and Trevor Keck, students in IHRC, contributed to the research and writing. Steve Goose, director of the Arms Division, edited the report. Tom Malinowski, Washington director for Human Rights Watch, Dinah PoKempner, general counsel, and Tom Porteous, deputy program director, all reviewed the report.

Kate Castenson, coordinator in the Arms Division, provided research and production assistance. Arms Division interns Rachel Borrell and Denise Tugade provided additional research assistance. This report was prepared for publication by Kate Castenson, Anna Lopriore, photo editor, Grace Choi, publications director, and Fitzroy Hepkins, administrative manager. Russell Christian produced the cartoon for the report cover.

Human Rights Watch and IHRC would like to thank Noel Sharkey, professor of artificial intelligence and robotics at the University of Sheffield, for providing a technical review of the report. Human Rights Watch and IHRC would also like to thank Jody Williams, Nobel Peace Laureate and chair of the Nobel Women's Initiative, for encouraging us to undertake work on this issue and for helpful comments on the report itself.

Region / Country Asia, Europe/Central Asia, Middle East/North Africa, United States, Afghanistan, Pakistan, United Kingdom, Israel/Palestine, Yemen, US Foreign Policy

Topic Arms, Terrorism / Counterterrorism, Killer Robots