CAAI | Center for Autonomy and AI | www.cna.org/CAAI

REDEFINING HUMAN CONTROL

Lessons from the Battlefield for Autonomous Weapons

Larry Lewis

March 2018

3003 Washington Boulevard | Arlington, Virginia 22201 | USA

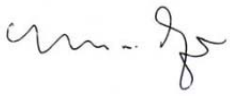
CNA's Occasional Paper series is published by CNA, but the opinions expressed are those of the author(s) and do not necessarily reflect the views of CNA or the Department of the Navy.

Distribution

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

Other requests for this document shall be referred to CNA Document Center at inquiries@cna.org.

Approved by: **March 2018**



Mr. Mark B. Geis
Executive Vice President
Center for Naval Analyses

This work was performed under Federal Government Contract No. N00014-16-D-5003.

Abstract

This report examines the issue of human control with regard to lethal autonomy, an issue of significant interest in United Nations discussions in the Convention on Certain Conventional Weapons (CCW) forum. We analyze this issue in light of lessons and best practices from recent U.S. operations. Based on this analysis, we make the case for a wider framework for the application of human control over the use of force. This report recommends that CCW discussions currently focusing on process considerations, such as human control, should instead focus on outcome—namely, mitigation of inadvertent engagements. This allows consideration of a more complete set of benefits and risks of lethal autonomy and better management of risks. The report also describes best practices that can collectively serve as a safety net for the use of lethal autonomous weapons. It concludes with concrete recommendations for how the international community can more effectively address the risk of inadvertent engagements from lethal autonomy.

This page intentionally left blank.

Executive Summary

With the impending development of lethal autonomous weapon systems (LAWS), observers agree on one thing: autonomy will revolutionize warfighting. Beyond that, the various parties disagree on how to address the risks of inadvertent engagement by weapons that can make their own decisions. Nongovernmental organizations are particularly worried about civilian casualties, while military leaders harbor additional concerns about friendly fire.

In four years of meetings on this topic by the Convention on Certain Conventional Weapons (CCW), two patterns have emerged. One is an approach to analyzing risk through a largely theoretical framework, perhaps because the novelty of autonomous weapons makes it difficult to imagine how real-world lessons from conventional warfare would apply. The second is a widespread belief that the panacea to the risks of autonomy rests in human control, particularly over the final engagement decision—the trigger pull.

Drawing upon CNA's long history of analyzing military operations, we employ empirical analysis of recent battlefield experiences to examine patterns and yield overall insights into protecting civilians from LAWS. And those lessons suggest that a narrow, trigger-pull approach will fail to adequately shield civilians. Fortunately, these empirical insights point the way toward a more comprehensive solution: a safety net woven from best practices in targeting, policy and testing, with the consideration of operational context. Instead of focusing on process considerations such as human control, this broader approach focuses on outcome, namely the mitigation of inadvertent engagements.

Mitigating human fallibility

One clear lesson from recent military experience is that human judgment during the trigger-pull decision is not perfect. Misidentifications were the reason for about half of all U.S.-caused civilian casualties in Afghanistan, with specific examples painfully abundant. In an area of daily attacks on coalition forces, girls using sickles to cut grass were misidentified as men with weapons. A sniper in the aftermath of a firefight mistook a farmer in a ditch for a combatant. And a helicopter crew thought it was preventing an expected attack when it took aim at a convoy carrying women and children. Soldiers themselves were also likely to be victims of misidentification:

for example, in major combat operations in Iraq in 2003, 17 percent of U.S. casualties were from fratricide. The fallibility of human judgment in real-world operations suggests that requiring a human in the loop for trigger-pull decisions will not eliminate the risks to civilians.

A more successful approach has been to reduce the number of decisions operators have to make in the heat of the moment, by front-loading some critical tasks earlier in the wider targeting process. From 2009, the International Security Assistance Force in Afghanistan modified its policies and procedures to help reduce the risk to civilians. Planning of operations began to consider risk factors for civilian casualties more effectively. One example was a focus on pattern-of-life determinations, in which forces used intelligence and reconnaissance data to establish a baseline of what was normal civilian activity. An analysis of available data suggests that these mitigation efforts were a win-win, reducing civilian casualties with no apparent cost to mission effectiveness. In a similar vein, the CCW will find that the most effective exercise of human control with autonomous weapons will take place over the entire targeting process. At the same time, the CCW should also consider the role of autonomous technologies that contribute to targeting without making the final engagement decision.

Building the safety net

But even with a broader approach including the wider targeting process, the CCW risks missing important elements of a safety net against civilian casualties. Discussions regarding LAWS in the CCW often have not addressed the question of how weapons are used. This context, consisting of both the operational environment and the mission, needs to be a key part of the evaluation process for the development and use of LAWS. For example, there are some environments where civilians will be rarely encountered or can easily be identified—including the underwater and air-defense domains. Another element of context is self-defense: should LAWS be handled differently if they are acting in defense of humans? One possibility is to take a crawl-walk-run approach, in which lethal autonomy is first pursued in less demanding missions and environments, where civilian casualties are less likely.

Many members of the CCW have also focused their attention on International Humanitarian Law (IHL), and the requirement that any use of autonomous weapons meet the requirements of IHL. While this should be considered necessary, a cautionary tale on the risks of elevating rules above outcomes can be found in the early results from self-driving cars. A recent study found that self-driving cars were five times more likely to be in a crash than conventional vehicles. Yet the autonomous vehicles were never at fault; they strictly followed the rules of the road. What they did not do was to anticipate the bad driving of humans. Just as the

necessary outcome for self-driving cars is not to follow the rules but to reduce crashes, a key desired outcome for autonomous weapons is to avoid inadvertent casualties. Designing systems to follow a set of rules—specifically, International Humanitarian Law—is necessary but not sufficient in itself for addressing this risk through a comprehensive safety net.

To complete the safety net, IHL must be supplemented with best practices. These include practices for avoiding inadvertent engagement already discussed, such as careful attention to possible misidentification of civilians. They also include best practices in policy. States can establish policies intended to strengthen oversight of procurement and use of LAWS. Implementing guidance can be developed and shared among states. Finally, testing and evaluation best practices will have to be developed and refined to accommodate the unique challenges posed by autonomous systems.

Lessons from the battlefield indicate that human judgment remains essential for autonomous weapons. But this is not requiring a finger on a trigger; rather, it is building a comprehensive safety net consisting of best practices in targeting, policy and testing, and focusing on the mitigation of inadvertent engagements.

Recommendations

- CCW discussions currently focusing on such considerations as human control should instead focus on mitigation of inadvertent engagements. This allows consideration of a more complete set of benefits and risks of lethal autonomy and more effective management of risks.
    - As part of this new focus, human control should be addressed not just as part of the final engagement decision but as a broader set of military doctrinal functions distributed across the wider targeting cycle.
    - Another implication of considering autonomy in this broader context is the need to consider the role of autonomous technologies that contribute to targeting without making the final engagement decision. For example, the use of artificial intelligence in target identification.
- Ways to comprehensively address risks posed by LAWS should be prominent features in future CCW discussions. States should not only ensure that LAWS follow the rules for IHL compliance but also incorporate best practices for avoiding inadvertent engagements, such as civilian casualties or fratricide, creating an overall safety net for LAWS. These best practices include:

- **Specific measures to help avoid inadvertent engagements.** Such measures should be gleaned from military operations and maintained and updated to learn lessons from operational experiences with LAWS.

- **Policy.** In addition to legal reviews mandated by Article 36 of Additional Protocol I, States can establish policy intended to strengthen oversight of procurement and use of LAWS. Implementing guidance can be developed and shared among States.

- **Testing.** It can be expected that traditional Test and Evaluation (T&E) processes will be unsuited for autonomous systems that can change and adapt over time, and not be predictable in a deterministic sense. New approaches to T&E will need to be developed to ensure reliability and predictability of these systems.

- The issue of context for LAWS should be a prominent feature in future CCW discussions. One possible approach to LAWS is to take a crawl-walk-run approach where lethal autonomy is first pursued in easier environments and missions where discrimination is more straightforward. The next steps are to learn applicable lessons and evaluate the suitability of LAWS in more challenging contexts.

- States should pursue applications of artificial intelligence to help avoid negative second order effects from the use of force. For example, applying machine learning to improve pattern of life analysis, obtain more accurate collateral damage estimates, and better anticipate reverberating effects (e.g., effects on civilian infrastructure).

- The moral dilemma concerning LAWS hinges in part on the relative risk posed by LAWS to civilians compared to that from human combatants. To evaluate this in future applications of LAWS, States should examine relative rates of inadvertent engagements by LAWS versus non-autonomous weapons. If the rate is lower for LAWS, it may also make the use of LAWS a "feasible precaution" per Additional Protocol I Article 57.

Contents

Glossary

AI	Artificial intelligence
CCW	Convention on Certain Conventional Weapons
CIVCAS	Civilian casualties
COMISAF	Commander, International Security Assistance Force
DODD	DOD Directive
EOF	Escalation of Force
GGE	Group of Government Experts
ICRC	International Committee of the Red Cross
IED	Improvised explosive device
IHL	International humanitarian law
ISAF	International Security Assistance Force
ISIS	Islamic State of Iraq and Syria
LAWS	Lethal Autonomous Weapon Systems
MOAB	Mother of All Bombs
NATO	North Atlantic Treaty Organization
NGO	Nongovernmental Organization
T&E	Test and Evaluation
UN	United Nations
UNODA	UN Office for Disarmament Affairs

This page intentionally left blank.

Introduction

Background

States have always aimed to harness technological advances to obtain a military edge. Examples include the development of bows and arrows; gunpowder-powered projectile weapons; rockets, jet aircraft, and nuclear warheads in World War II; and stealth, unmanned vehicles, and precision-guided munitions in recent decades. Many have noted how States are working to leverage the rapid and advancing progress in artificial intelligence (AI). Dominated by the commercial industry and its innovations, the past two years have seen dramatic advances in which machines have been able to complete complex tasks and match or exceed human performance. This trend is expected to continue and probably accelerate. One prominent feature of these AI and related information technology developments is the attribute of autonomy, where systems can make decisions and actions without the need for human intervention.

A number of States have already announced that these technological developments are central to their national security plans. For example, the U.S. response to this new security environment is the "Third Offset" strategy, an asymmetric approach that aims to "exploit all the advances in artificial intelligence and autonomy…to achieve a step increase in performance that the department [U.S. Department of Defense] believes will strengthen conventional deterrence."[1] China, similarly, is pursuing AI to achieve the "intelligentization" of warfare, declaring its intent to "seize the advantage in military competition and the initiative in future warfare."[2] And Russian President Vladimir Putin recently remarked that "the one who becomes the leader in this sphere [AI] will be the ruler of the world," while Russia's former Chief of Air and Space Forces underscored the emerging importance of autonomous and AI-driven systems, stating that "the day is nearing when vehicles will get artificial intelligence. So, why not entrust aviation or air defense to them?"[3]

[1] Larry Lewis, *Insights for the Third Offset: Opportunities and Challenges for Autonomy and AI,* CNA, DRM-2017-U-016281-Final, September 2017.

[2] Elsa B. Kania, *Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power,* Center for a New American Security, November 2017.

[3] "Putin: Leader in Artificial Intelligence Will Rule the World," Associated Press, September 1, 2017; correspondence with Samuel Bendett, expert on Russian autonomy and unmanned systems, CNA, December 19, 2017.

The most controversial aspect of this new technology has been whether weapon systems operating autonomously (without a human operator) should be allowed to use lethal force. The United Nations (UN) Convention on Certain Conventional Weapons (CCW) has spent four years discussing ethical, legal, and operational considerations of lethal autonomous weapon systems (LAWS).[4] There are varying national positions on LAWS as a future weapon system, some nations seeing potential benefits and regarding their operation as being sufficiently addressed under international humanitarian law (IHL), others believing that additional steps should be taken to ensure oversight and safety of these systems, and about two dozen countries calling for a preemptive ban of LAWS. While there are several areas of potential concern about the use of LAWS, the predominant concern with LAWS in discussions to date has been the risk of civilian casualties.[5] While this paper will examine a range of potential concerns (including compliance with IHL), a chief focus will be how to effectively address the issue of civilian protection with LAWS.[6]

In international discussions, parties have consistently expressed support for some kind of human control over LAWS to mitigate potential risks. But the exact form of this control is debated: for example, some add such terms as "meaningful" or "effective" to modify their conception of human control, while the U.S. uses the alternate term, "appropriate levels of human judgment."[7] Note that the intended scope of this human control is not always clear. Some argue that having adequate human control necessitates that a human pull the trigger—or otherwise activate the weapon system for a particular engagement—and that "systems operating outside of that [form of control] should not be considered acceptable."[8] In this argument, having a human in the loop is a panacea for civilian protection and related concerns with lethal autonomy.

[4] The CCW is properly referred to as the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects.

[5] Professor Stuart Russell, All-Party Parliamentary Group on Drones Inquiry Evidence Session: "The Use of Armed Drones: Working with Partners and Emerging Technologies," UK Parliament, October 31, 2017. "Much of the debate around LAWS has centered on the risk of non-combatant deaths."

[6] IHL is also commonly referred to as the Law of Armed Conflict or the Law of War.

[7] *Perspectives on Lethal Autonomous Weapon Systems,* UNODA Occasional Papers, No. 30, November 2017.

[8] *Key Elements of Meaningful Human Control,* Background Paper, Article 36, April 2016.

Overview of this report

Given that civilian casualties are a key concern in the CCW, is having human control over the final engagement decision the right answer? Many reports have used law and logic to deduce implications for the permitted use of autonomous weapons, including support for the argument for human control over the final trigger-pull decision. This paper takes a different approach, looking at lessons from military operations and identifying insights and operational best practices that are relevant to the employment of lethal autonomy. Based on real-world examples, the benefit of human control over the final engagement decision is seen to fall short of what is hoped for; instead, military practices adopted in the last 16 years point to a more effective way of pursuing the humanitarian aim of civilian protection within the context of lethal autonomy. The report contains four sections, followed by conclusions and recommendations:

- **The Fallibility of Human Judgment.** Human soldiers do not always make perfect decisions; limitations in these decisions can result in inadvertent engagements.
- **Operational Adjustments in Recent U.S. Operations.** Operational adjustments to U.S. military operations front-loaded some critical tasks earlier in the targeting process, reducing the opportunity for mistakes in human judgment.
- **Human Control: A Broader Framework.** The CCW framework for human control should include not just the final engagement decision, but a broader set of military doctrinal functions inherent in control.
- **Other Considerations.** Operational context and the pursuit of steps to more comprehensively address risks posed by LAWS should be prominent features in future CCW discussions. Specifically, the past discussion of human control should shift to a more comprehensive discussion of how to avoid inadvertent engagements. This section also discusses autonomous targeting functions that are not weapons and a moral argument that could support the use of LAWS.

The fallibility of human judgment: operational examples

A central argument for having meaningful human control over lethal autonomous weapons is that it will promote good decisions regarding lethal force. This belief was seen, for example, in a discussion at the 2017 LAWS Group of Government Experts (GGE) where one observer raised concerns that LAWS would not be able to detect when a perceived combatant was, in fact, mentally ill. The unspoken assertion was that a human in the loop would surely be able to detect this and apply restraint. According to actual experiences in combat operations, however, this assertion is unfounded: for example, there were several instances in U.S. operations in Afghanistan where human soldiers, faced with hostile action or perceived hostile intent, did in fact (and in accordance with international law) kill people who seemed to pose a threat and were only later discovered to be mentally ill or impaired. This is one example of a reality of combat: human soldiers do not always make perfect decisions, and limitations in these decisions can result in civilian casualties—as well as fratricide—in armed conflict.

Civilian casualties

Civilian casualties are often assumed to be the result of collateral damage from the engagement of a valid military target; for example, small arms fire or a 500-pound bomb is aimed at and effective against the intended target, but unobserved civilians in the area are also casualties. But this is not necessarily the case. Analysis of civilian casualties from operations showed that another mechanism was just as common a contributor to civilian casualties: misidentification. This occurs when military forces target civilians in the mistaken belief that they are military targets. In analysis of Afghanistan operations, misidentifications were the cause of about half of all U.S.-caused civilian casualties in Afghanistan. These misidentifications often involved human judgment that the targeted individuals represented a threat to military forces. This could result from mistakes in associating intelligence with specific individuals, misinterpreting their behavior or appearance, or losing situational awareness. We provide three civilian casualty incidents (uses of force that resulted in inadvertent civilian casualties) as examples.

Example 1: Self-defense attack in Kunar

On May 15, 2011, U.S. soldiers at an observation post in a mountainous area in Kunar Province, Afghanistan, observed activity they believed to be suspicious. Four persons appeared to be digging and moving rocks, possibly improving their fighting position, on an adjacent ridge. The location was near a position they had been attacked from in the past; they had been attacked every day the previous week, so they expected another attack that day. None of the four were wearing headscarves that women traditionally wear, so they were believed to be men. The U.S. soldiers saw what appeared to be a weapon slung over the back of one person. The commander approved an attack against the group in self-defense, believing it to be an imminent threat. An hour later, the district police notified the unit of civilian casualties: one girl was killed and four wounded, all ages 6 to 17. They were gathering grass for their animals, carrying metal sickles to cut grass and using their headwear to carry the cut grass.[9]

Example 2: Sniper engages suspected combatant

On February 12 2010, an Afghan National Army vehicle was struck by an improvised explosive device (IED) and then received small arms fire from a nearby compound. U.S. snipers worked with the Afghan National Army to eliminate the combatants in the compound. A few minutes later, the snipers observed a person in a ditch with something in his hand that looked like a weapon. The man climbed out of the ditch and was still carrying the object, so he was shot and killed in the belief that he was a threat. The Afghan National Police, who were collocated with the sniper team, realized the man was a local farmer and that a civilian casualty incident had occurred.[10]

Example 3: Close air support in Uruzgan

Early on February 21, 2010, a U.S. Special Forces team, accompanied by Afghan Army and Police personnel, conducted an air infiltration into western Uruzgan province for a daytime cordon and search operation. The Special Forces team was supported by an MQ-1 Predator and an AC-130. While waiting for daylight, the team received

[9] "Mohammed Anwar, Afghan governor says NATO troops killed child," Reuters, May 16, 2011. https://www.reuters.com/article/us-afghanistan-civilian-deaths/afghan-governor-says-nato-troops-killed-child-idUSTRE74F31G20110516.

[10] Center for Army Lessons Learned, *Afghanistan Civilian Casualty Prevention Handbook,* June 2012.

intelligence that enemy forces were going to attack. The Predator observed two SUVs driving south toward the location of U.S. forces and assumed these were enemy forces. The Predator then watched the SUVs as they drove around the area for about three and a half hours, changing directions several times, stopping to allow the occupants of the vehicles to pray, and moving away until they were about 12 km from U.S. forces. The two SUVs were joined during this time by a third vehicle, a pickup truck. During this time period, imagery analysts examined the full motion video from the Predator and provided their feedback to the Predator crew. In turn, the Predator crew communicated with the U.S. team on the ground in Afghanistan.[11] The descriptions provided by the imagery analysts—who were trained to interpret the Predator feed—were frequently different, and less aggressive overall, than the descriptions the Predator crew provided to the U.S. forces on the ground charged with making the engagement decision. OH-58 helicopters were called in to strike the targets based on information provided by the Predator crew, believing falsely that the vehicles posed an immediate threat. The ground forces were unaware that the imagery analysts had communicated the presence of children to the Predatory crew. The helicopters fired on the vehicles with Hellfire missiles, with follow-up engagement of individuals using rockets. The pilots saw people running from the vehicles dressed in brightly colored clothing, which is characteristic of women's apparel in Afghanistan. Based on this observation, they stopped the engagement and radioed back the possibility of civilian casualties. Twenty-three civilian casualties resulted from the incident.[12]

The foregoing examples show where humans were making decisions to use lethal force, but they made errors in those decisions, misidentifying civilians as valid targets. Mistakes happen in war, which is why IHL reserves the category of war crime for intentional acts or serious negligence, not human errors such as shown in the three examples. Overall, these cases show that simply inserting a human into the engagement decision does not guarantee a good outcome.

Fratricide

Civilian casualties are not the only operational symptom of human mistakes. Fratricide—the mistaken engagement of friendly forces—also shows the fallibility of

[11] The imagery analysts were located at a different location than the Predator crew and had no means of direct communication with forces on the ground. Thus, their feedback was filtered through the Predator crew.

[12] Larry Lewis, *Improving Lethal Action: Learning and Adapting in U.S. Counterterrorism Operations,* CNA, September 2014.

human judgment in engagement decisions. In the most recent U.S. high-intensity conflict, in Iraq in 2003, 17 percent of total U.S. casualties (19 of 109) were from fratricide; such engagements also led to the destruction of military equipment. These engagements of friendly forces often resulted from the operator misidentifying friendly forces as a threat, as well as a lack of situational awareness. We present three cases of friendly fire from that operation as examples.

Example 1: PATRIOT shoot down of Navy aircraft

On April 2, 2003, PATRIOT batteries advanced north into central Iraq to defend against explicit threats of chemical attacks on advancing U.S. forces. One PATRIOT battalion, 5-52, was positioned in the Karbala Gap area. Two Navy F/A-18C aircraft completed a bombing mission over Baghdad and were returning south, trying to avoid Iraqi surface-to-air-missile sites. They were tracked and reported as friendly aircraft by an E-3C early warning aircraft as they flew into PATRIOT coverage. PATRIOT also tracked the aircraft, but they were held as unknown and did not associate their tracks with the E-3C track. The PATRIOT system misclassified the lead aircraft as a ballistic missile with an impact point within its assigned defended footprint. The PATRIOT operator ordered the firing units to switch to operate mode, and the system automatically engaged the target a few seconds later. This incident highlights some challenges with human-machine teaming associated with automatic systems, but there were several errors in human judgment. For example, the operator did not examine the track information before issuing this order and had not been monitoring the available information (including available information about their friendly identification) prior to the misclassification. The PATRIOT engagement killed the pilot and destroyed the aircraft.[13]

Example 2: UK vehicles in Basra

On March 28, 2003, two U.S. Air Force A-10 aircraft were providing close air support close to Basra, Iraq, as well as targeting Iraqi military vehicles and artillery. The pilots spotted a convoy of vehicles and attempted to visually identify them. Seeing orange and thinking that could possibly be an orange visual identification panel (intended to avert fratricide), the pilots radioed back to their controller to ask if friendly vehicles were in the area, and the controller stated that they were well clear of friendly forces. The pilots concluded that the orange color was some sort of rocket, which is the target type they expected. They attacked the vehicles before they approached a

[13] Larry Lewis, *CID [Combat Identification] Observations from Iraqi Combat Operations,* CNA, March 10, 2005.

nearby village, in an effort to avoid civilian casualties. After engaging the vehicles, they saw red smoke, used to signal a cease-fire message, and they received a radio message from a UK Household Cavalry Regiment unit that it was a victim of friendly fire. The result of the attack was that one UK soldier was killed, four were wounded, and two Scimitar vehicles were destroyed.[14]

Example 3: Marines in Nasiriyah

On March 23, 2003, a Marine unit, Task Force Tarawa, arrived in Nasiriyah to secure two bridges and clear the route for a coming advance of 1st Marine Expeditionary Force. The Task Force cleared the first bridge. Two companies (B and C Company) moved forward to clear the second bridge but encountered significant resistance in the form of an Iraqi ambush. B Company turned onto nearby alleyways to avoid the ambush and became mired in muddy conditions, while C Company moved north and crossed the second bridge. With B Company still in an intense firefight, and not aware of the location of C Company, it called for emergency close air support. C Company also began to take significant fire from the north. When A-10 aircraft responded, they were told that there were no friendly forces north of the second bridge. The A-10 made at least eight attacks against what it described as pickup trucks using 30-mm rounds, Mk 82 bombs, and Maverick missiles. In the midst of the attacks, C Company tried to retreat back across the bridge, and B Company told the A-10s to not let the forces back over the bridge. After about 20 minutes, the possibility of friendly fire was detected and the attacks halted. At least 10 Marines were killed from U.S. fire during the attack; there had not been an attack causing more casualties on a single unit since the Vietnam War.[15]

As seen in the foregoing civilian casualty and fratricide incidents, human judgment in war is not perfect. Evidenced by the examples above, this is true across a variety of operational contexts and mission sets, including deliberate offensive operations, close air support, and self-defense engagements. But, given the desire to reduce civilian and friendly tolls in armed conflict, the U.S. and its allies took a number of steps over the past 16 years to reduce these inadvertent engagements in all of these different environments and missions. These efforts are described in the next section.

[14] Ben Fenton, "Britons caught in terrifying hail of fire after fatal US mix-up in the sky," The Telegraph, February 7, 2007.

[15] Larry Lewis, *Improving Joint C2: Lessons from Iraq,* CNA, May 20, 2008.

Operational adjustments in recent U.S. operations

Overview

The U.S. has long observed IHL in its military operations and has developed a number of safeguards to help avoid civilian casualties. These safeguards include capabilities for precision engagements and target identification, and improvements to the targeting process; these capabilities were employed in Iraq in 1991, enabling comparatively low numbers of civilian casualties in the midst of high-intensity military operations. That said, the last 16 years of operations offered significant challenges for the U.S., stemming from the fact that the U.S. military is largely designed for major combat operations with other States.[16] Most of its recent operational experience—with the exception of major combat against the Iraqi military in 2003—has been against irregular armed groups that largely eschewed many of their legal obligations under IHL (e.g., not wearing a uniform and hiding among the population). This environment and threat complicated the identification process and increased risk of civilian casualties. Over time, the U.S. made operational adjustments to reduce this risk.[17]

The U.S. was also challenged by fratricide, primarily in the early years during higher intensity armed conflict: in 2001 and 2002 in Afghanistan and then in Iraq in 2003. The U.S. made a number of adjustments to compensate, including technological solutions (e.g., improved cooperative identification technologies) and improved training and procedures. Fratricide was an easier problem to solve than civilian casualties, in part because of the possibility of cooperative techniques for information sharing and greater situational awareness among friendly military forces. Although fratricide was not eliminated, it was addressed relatively early and became less of a concern as operations continued.[18]

[16] Sarah Sewall and Larry Lewis, *Joint Civilian Casualty Study: Executive Summary,* August 31, 2010.

[17] Larry Lewis, *Reducing and Mitigating Civilian Casualties: Enduring Lessons, Joint and Coalition Operational Analysis,* April 12, 2013.

[18] Larry Lewis, *Insights for the Third Offset: Opportunities and Challenges for Autonomy and AI,* CNA, DRM-2017-U-016281-Final, September 2017.

Early on: Challenges in civilian protection

Compared with fratricide, civilian protection posed more long-lasting challenges to the U.S. On October 7, 2001, U.S. forces began combat operations to capture Al Qaeda leadership and avoid the use of Afghanistan as a continued launching point for terrorism. Within days of the start of operations, international media began reporting incidents of civilian casualties. Because al Qaeda and its allies operated as an irregular force and didn't comply with IHL requirements, obtaining positive identification of combatants was more difficult, and U.S. engagements tended to rely more on self-defense considerations based on perceived hostile acts or hostile intent. Probably the two highest profile incidents during this time were (1) the December 21, 2001, attack on a convoy that the Afghan government claimed included tribal leaders and (2) the July 1, 2002, AC-130 attack on a group in Deh Rawud, central Afghanistan, that turned out to be a wedding party. In both attacks, U.S. aircraft had observed ground fire and engaged because of self-defense considerations.[19]

During major combat operations in Iraq in 2003, it was more straightforward to distinguish combatants from the civilian population because of several factors. Iraqi military forces were usually located away from civilian areas, and their military equipment and uniforms reduced the ambiguity of engagement decisions relative to those faced by U.S. forces in Afghanistan. However, the Iraq military purposely violated law-of-war rules designed to protect the peaceful civilian population by using human shields, misusing protected symbols for impartial humanitarian organizations (e.g., Red Crescent), and placing equipment in protected sites. In addition, Fedayeen Saddam forces did not wear uniforms and fought using irregular tactics, contributing to U.S. challenges in obtaining positive identification.[20]

In these initial operations, the U.S. and its allies took deliberate steps to minimize collateral damage; for example, in Iraq, similar to Afghanistan, most air engagements used precision-guided munitions. Civilian casualties became more of a challenge as insurgencies developed in Iraq and Afghanistan. In Iraq, noncombatant casualties were primarily caused by escalation of force (EOF) incidents along roads, both at checkpoints and during convoy operations. These incidents resulted in a significant outcry from nongovernmental organizations (NGOs) and the media; the shooting of a vehicle containing Italian journalist Giuliana Sgrena and her rescuers during an EOF incident further increased visibility of this issue. In mid-2005, U.S. forces in Iraq adapted and made heightened efforts to prevent civilian casualties from escalation of

[19] Larry Lewis, *Reducing and Mitigating Civilian Casualties: Enduring Lessons, Joint and Coalition Operational Analysis,* April 12, 2013.

[20] Ibid.

force incidents, including the development of new guidance and tactics, and they began tracking and making adjustments at the operational level of command. Still, the problem of civilian casualties continued, though at lower levels. And critically, the adaptations seen in Iraq were not carried over to operations in Afghanistan where the same challenges were being seen.[21]

These challenges in Afghanistan were noted in President Karzai's first public statements on civilian casualties in 2005, asking the International Security Assistance Force (ISAF) to take measures to reduce such casualties. Early initiatives, such as the "Karzai 12" rules for approving operations in 2005 (imposed at the request of President Karzai and the government of Afghanistan) and the initial Commander, ISAF (COMISAF) Tactical Directive in 2007, were attempts to address these challenges, but they were not successful in reducing high-profile incidents. Additional efforts, including redrafting the COMISAF Tactical Directive in 2008, were made in response to several high-profile, high-casualty civilian casualty incidents; however, a May 2009 incident in Bala Balouk highlighted the lack of progress in effectively addressing the issue of civilian casualties.[22]

Operational adjustments to improve civilian protection

The Bala Balouk incident (in which at least 26 civilians were killed, by U.S. estimates) served as an impetus for major efforts to reduce civilian casualties by both ISAF and the U.S. The new ISAF Commander, General Stanley McChrystal, recognizing that the continuing civilian casualties were undermining the overall mission, put a strong emphasis on the need to protect the Afghan population. Under his leadership, ISAF modified its policies and procedures to help reduce the risk to civilians from international forces. This approach involved a series of adjustments. For example, planning of an operation began to better consider risk factors for civilian casualties and to develop tactical alternatives, including alternate placement of forces and placing snipers in key positions beforehand to reduce the need for the use of heavy weapons, such as artillery or air-delivered-bombs. Pattern-of-life determination also became more of a priority: intelligence and reconnaissance assets were used to establish a baseline of what was "normal," improving situational awareness and reducing the chances of mistaking regular activity as nefarious.[23]

[21] Ibid.

[22] Sarah Sewall and Larry Lewis, *Joint Civilian Casualty Study: Executive Summary,* August 31, 2010.

[23] Ibid.

Collectively, these efforts bore fruit: because of improved guidance and training, ISAF forces adapted the way they conducted operations in light of civilian casualty concerns, and ISAF-caused civilian casualties decreased over time. A key reason for this progress was that the changes helped reduce the number of decisions that operators had to make in the heat of the moment to support engagement decisions. These operational adjustments moved some critical tasks to earlier in the targeting process. Though mistakes were not eliminated, these adjustments helped to reduce the opportunity for mistakes in human judgment that could lead to inadvertent engagements. Analysis of available data suggested a win-win situation, with no apparent cost to mission effectiveness or increase in friendly force casualties. In fact, in some contexts, mission effectiveness was seen to increase—another benefit of improving human judgment.[24]

These operational adjustments in Afghanistan were accompanied by institutional measures in the U.S. This effort was initially led by a new ad hoc organization, the U.S. Joint Staff CIVCAS Working Group, led by a three-star general officer. This group aided efforts to improve U.S. pre-deployment training to better prepare U.S. forces for civilian casualty reduction and mitigation in Afghanistan. These efforts included the development of a handbook on reducing civilian casualties—containing tailored guidance and tactics based on lessons from actual civilian casualty incidents—that was issued to every soldier deploying to Afghanistan.[25]

Another focus of that group was to integrate best practices developed in Afghanistan into military institutions more broadly. Success of this was seen in the early campaign against the Islamic State of Iraq and Syria (ISIS), where operational refinements from Afghanistan were employed from day one in counter-ISIS operations. The commitment of the U.S. to sustaining these best practices in the long term was also reinforced through the establishment of a new national policy on civilian casualties, issued in 2016 and having its basis in these best practices.[26]

[24] Larry Lewis, *Reducing and Mitigating Civilian Casualties: Enduring Lessons, Joint and Coalition Operational Analysis,* April 12, 2013.

[25] Ibid.

[26] Civilian Casualty Executive Order, July 2016.

Human control: A broader framework

These past operational adjustments highlight the fact that human judgment can contribute to decisions regarding the use of force significantly earlier than the moment of the trigger pull. Empirical data also show that such front-loaded human control can reduce risk of inadvertent engagements, while also improving mission effectiveness in some cases.[27]

Others have observed that human control over the use of force is not limited to the moment of the trigger pull. For example, the International Committee of the Red Cross (ICRC) provides a slightly broader framework for exercising human control over LAWS in their development and use, describing three separate stages:

- Development and testing (ICRC describes this as the "development stage")
- Decision to activate the weapon system (ICRC's "activation stage")
- Independently selects and attacks targets (ICRC's "operations stage").[28]

ICRC argues that human control can be applied at each of these stages in order to comply with IHL requirements. ICRC believes it likely that human control would still need to be exercised over the final step of the engagement—for example, where a human operator would pull the trigger or have the ability to countermand an engagement.[29]

Merel Ekelhof argues for an even broader framework. She envisions the larger targeting process as a means of human control, with this larger targeting framework encompassing the equivalent of ICRC's activation stage as well as the operations

[27] Sarah Sewall and Larry Lewis, *Joint Civilian Casualty Study: Executive Summary,* CNA, August 31, 2010.

[28] Neil Davison, "A Legal Perspective: Autonomous Weapon Systems Under International Humanitarian Law," in *Perspectives on Lethal Autonomous Weapon Systems, UNODA Occasional Papers,* No. 30, November 2017.

[29] Communication with Neil Davison, ICRC, January 2018.

stage (or trigger-pull decision), but also including additional elements using NATO's doctrinal targeting process as a model.[30]

This broader view of control is echoed in U.S. military doctrine, which defines specific functions of control. These functions include "planning, direction, prioritization, synchronization, integration, and deconfliction."[31] The optimal echelon for maintaining the control function has been widely debated within the U.S. military, with some Army publications pushing for the lowest tactical unit and Air Force doctrine emphasizing centralized control (often at the combatant or component commander level). Operational best practices pointed to a hybrid "focused decentralization" approach where an intermediate echelon of command (e.g., brigade or battalion) exercises these functions while being supported by higher echelon resources and authorities.[32] But regardless of where that control is placed, those functions are all exercised.

Illustrating opportunities for human control: PATRIOT vignette

Real-world operational vignettes can be useful in highlighting the kinds of opportunities possible for exerting human control at various stages. We thus consider the vignette describing the PATRIOT shoot down of a fighter aircraft (given on page 7) to contrast the different views on human control discussed above. For example, the automated launch of the missile represents the operations stage in the ICRC framework and the engagement, trigger-pull decision in many discussions at the CCW. There are additional opportunities for introducing human control and improving the quality of the engagement decision using human judgment, such as the operator's decision to move the system to Operate mode, which by ICRC's definition represents the activation stage. However, this incident also reveals additional opportunities for human control. These opportunities include, for example:

[30] Merel A. C. Ekelhof, "Lifting the Fog of Targeting: 'Autonomous Weapons' and Human Control Through the Lens of Military Targeting," *Naval War College Review,* forthcoming (summer 2018).

[31] *Operations and Organization*, Air Force Doctrinal Document 2, United States Air Force, April 3, 2007.

[32] Gary Luck and Mike Findlay, *Air Component Integration in the Joint Force,* Focus Paper #6, Joint Warfighting Center, March 20, 2009, drawn from "Joint Tactical Environment: Best Practices and Recommendations Handbook," JCOA report, January 9, 2009.

- The planning decision to place PATRIOT systems in the Karbala Gap area
- The planning decision to not establish return-to-base corridors, a procedural deconfliction measure
- The E-3 early warning aircraft not monitoring and working toward a common tactical picture in existing data links in general
- The PATRIOT operators not maintaining situational awareness on air contacts or attempting to correlate available information to their own sensor data, a result of established training and doctrine not emphasizing these functions
- A failure to communicate engagement decisions with higher echelons of command, who may have greater situational awareness

These opportunities, and others, are ways to better inform engagement decisions that are outside the narrower views of applying human control. They are consistent with the operational adjustments described earlier, where the distribution of control functions within the larger targeting enterprise contributed to better engagement decisions. Examining the use of force within such a broader framework of control is critical for a productive CCW discussion on LAWS. The alternative—a platform-centric approach that focuses on the trigger pull—can miss key elements that, in fact, contribute heavily to the quality of engagement decisions and the propensity for inadvertent engagements, such as civilian casualties, a major concern in CCW discussions. The importance of a broader view, and the risk inherent in a narrow view, can also be seen in concerns about another widely debated weapon system: drones (e.g., remotely piloted aerial vehicles).

Drones: Surgical or not?

Armed drones, such as Predator and Reaper (both U.S. systems), are a recent development in warfare that can bring together several desirable capabilities (including endurance, intelligence collection, and integrated attack) in a single platform. They are used by an increasing number of States, and they have been widely praised for their precision and low collateral damage. Some have described drone attacks as "surgical" and as "the most humane form of warfare ever."[33] Some,

[33] Daniel Byman, "Why Drones Work: The Case for Washington's Weapon of Choice," *Foreign Affairs,* July/August 2013; Michael E. Lewis, "Drones: Actually the Most Humane Form of Warfare Ever," *The Atlantic,* August 21, 2013.

however, have expressed concerns about civilian casualties from U.S. drone strikes. To investigate the civilian casualty characteristics of drones, a study examined the rate of civilian casualties caused by drones and by manned aircraft in Afghanistan. Contrary to the expectations of many, the rate of civilian casualty incidents for drones was 10 times higher than that for their manned counterparts.[34] From a platform-centric perspective, this makes no sense: drones have the benefit of high-resolution imagery and weaponry that is designed for low levels of collateral damage. How could this be?

The surprise of this result stems from a failure to consider the full set of factors that lead to mistakes in the use of force: in this case, a narrow focus on the features of the weapon platform rather than characteristics of the wider targeting process. It is issues with the wider targeting process that contributed to the relative propensity of drones to cause civilian casualties.[35] Analysis showed that there were specific features of drones and their implementation of the targeting process that introduced more risk of civilian casualties. This included a complex command and control structure as well as stovepipes in information exchange. Importantly, "platform characteristics alone are not the driver of a decreased likelihood of civilian casualties."[36] This example illustrates that using a narrow approach to understanding the use of force ignores risks introduced in the wider targeting process.

Applying the example of drones to autonomous weapons, CCW discussions should include not just human control introduced during the final engagement decision, but the fuller set of functions included in military doctrinal definitions of control, which are regularly exercised in the wider targeting cycle. Doing so would allow consideration of a more complete set of benefits and risks of lethal autonomy, and how they can be managed through each of the doctrinal elements and the processes involved in targeting.

Such management can help avoid surprises from unanticipated non-platform factors, and introduce additional ways to mitigate risk. For example, as mentioned above, analysis found that drones had higher than anticipated risks of civilian casualties because of inattention to non-platform factors. Because the capable sensors and precision weapons of the drone platform seemed to offer a reduced risk of civilian casualties, that perception lulled operational forces into adopting inadequate targeting procedures. This error was symptomatic of a mistaken belief that the

[34] Larry Lewis, *Drone Strikes: Civilian Casualty Considerations,* Joint Center for Operational Analysis (JCOA), June 2013.

[35] Note that this observation was specific to that time period and operational context.

[36] Larry Lewis, Drone Strikes in Pakistan: Reasons to Assess Civilian Casualties, CNA, COP-2014-U-007345-Final, April 2014.

platform is the sole determiner of risks. The CCW should avoid this situation with LAWS in order to address potential risks effectively. Moreover, a deliberate approach to manage risks of LAWS through a broader view of human control addresses the fact that human judgment is not perfect: safeguards in the rest of the targeting process can help anticipate and avoid potential errors in judgment in the trigger-pull decision. Collectively, these safeguards represent a *safety net*, composed of multiple threads that together improve the quality of engagement decisions and help reduce risk of inadvertent engagements. The next section discusses other possible threads contributing to this safety net.

Other considerations

Previous sections have discussed human control and how this control is not limited to the engagement decision but is implemented throughout the wider targeting cycle, an issue that has thus far been largely neglected in CCW discussions. This section discusses several other issues that have also not been addressed sufficiently in the CCW, including operational context for the use of LAWS and a comprehensive approach to mitigating risk. Specifically, this section argues that the past CCW focus on process elements, such as human control, has detracted from pursuit of its overall goals. It then addresses another truth that is evident when considering the wider targeting cycle: autonomous functions can play roles earlier in the targeting cycle, and can influence the use of force without being a weapon system. This section discusses both challenges and opportunities of these targeting functions. It then concludes with a moral argument that could support the use of LAWS.

Context is critical

Related to the targeting process is the question of context. Frequently, discussions within the CCW have focused on the hardest problems, such as discriminating between civilians and insurgents (not wearing uniforms or otherwise meeting their legal obligations per IHL) in an urban setting. While this is one possible application of LAWS, it represents one of the most difficult environments and missions for discriminating between combatants and civilians. Discussions within the CCW would benefit by including the issue of context: a military designing and employing any weapon system needs to take into account the context in which it will be used. This would also be true for LAWS.[37]

[37] This point is echoed by ICRC in its opening statement to the 2017 GGE: "An examination of the way in which human control can be exerted over autonomous weapon systems (as broadly defined by the ICRC) points to the following key elements of human control to ensure legal compliance: predictability; human supervision and ability to intervene; and various operational restrictions, including on tasks, *types of targets, the operating environment, time frame of operation, and scope of movement*" (italics added), Expert Meeting on Lethal Autonomous Weapons Systems, ICRC, November 15, 2017, https://www.icrc.org/en/document/expert-meeting-lethal-autonomous-weapons-systems.

There are some environments where civilians will be rarely encountered or can easily be identified. Such domains include under water and air defense. There are also target types that are easier to differentiate between military and civilian, such as missiles, tactical aircraft, and armored vehicles. It is expected that LAWS will have an easier time with discrimination and proportionality in these environments and against these target types, and that the accompanying risk of inadvertent engagements from LAWS will be reduced in these contexts. This is especially true for civilian casualties; air defense in particular still holds a risk of fratricide, as evidenced in the PATRIOT example discussed earlier.

Another element of context to be considered in CCW discussions is self-defense: should LAWS be handled differently if they are acting in defense of humans? For example, U.S. policy differentiates LAWS anti-materiel weapons designed to defend manned platforms or installations.[38] And how should self-defense of LAWS be handled?

This context-dependent evaluation process is critical because development of LAWS does not mean that they will be used in all environments and missions. Some weapon systems will simply be inappropriate for use in some settings: this is true for all weapon systems, not just LAWS. For example, in 2017 the U.S. military employed its Mother-of-all-Bombs (MOAB) in Afghanistan, in an environment where no civilians were expected to be present. The use of that weapon in a civilian-filled urban environment is a completely different proposition. Similarly, some kinds of weapon systems—for example, air defense systems—face entirely different, and simpler, kinds of challenges in meeting discrimination and proportionality requirements than a system that has to identify and target particular humans. For instance, a weapon system could be very effective in identifying military aircraft and missile targets in the air-to-air domain, but it would be unrealistic to expect it to perform well against human targets if it were not designed for that purpose.

Context—how and where systems could reasonably be used and still meet IHL and policy requirements—is a normal component of an Article 36 legal review. The issue of context for LAWS should also be a prominent feature in future CCW discussions. One possible method is to take a crawl-walk-run approach in which lethal autonomy is first pursued in easier environments and missions where meeting IHL requirements, such as discrimination and proportionality, is more straightforward.

[38] DOD Directive 3000.09, *Autonomy in Weapon Systems,* November 21, 2012, Incorporating Change 1, May 8, 2017.

The next step is to learn lessons and evaluate the suitability of LAWS in more challenging contexts.[39]

Mitigating risk: Learning from self-driving cars

The earlier sections discussed the importance of using the larger targeting process as a framework when considering the strengths and risks of LAWS. The progress in protecting civilians in Iraq and Afghanistan seen in U.S. military operations over the past 16 years also illustrated another point: the value of going beyond the minimum standard. The U.S. regularly took on additional measures to help safeguard civilians beyond what was required by international law. Many members of the CCW have declared that any use of LAWS should be required to meet the requirements of IHL; however, this should be considered necessary but not sufficient because there is value in proposing additional safeguards in the form of incorporating best practices. The potential value of this approach is seen in a nonmilitary example of autonomy: the performance of self-driving cars.

Self-driving cars are still in the early stages of development, but in the last few years they have accumulated over a million miles of driving in the U.S. A study examined the relative crash rates of self-driving and conventional vehicles and found that self-driving vehicles were five times more likely to be in a crash.[40] This result contradicts the expectation that self-driving vehicles will lead to safer roads, though notably it is still early in their development. At the same time, the study also revealed that those vehicles were not at fault for any of the crashes; they were involved in a crash with a conventional type of vehicle that was found to be at fault. The self-driving vehicles themselves practiced safe driving, but they were unable to anticipate behavior by other cars that then led to the crash.

This was illustrated in an incident in November 2017, where a self-driving bus was involved in an accident in its first day on the road.[41] In the accident, the bus was operating properly but then stopped when a taxi in front of it stopped. The taxi then started backing up until it hit the bus. There was minimal damage, but the incident

[39] Taking feasible precautions is another requirement of IHL and should be met by these systems. The approach described in the next section for mitigating risk can help with identifying feasible precautionary measures.

[40] Brandon Schoettle and Michael Sivak, *A Preliminary Analysis of Real-World Crashes Involving Self-Driving Vehicles,* University of Michigan Transportation Research Institute, UMTRI-2015-34, October 2015.

[41] Dave Lee, "Self-driving shuttle bus in crash on first day," BBC, 8 November 2017, http://www.bbc.com/news/technology-41923814.

illustrated that the self-driving software was programmed on how to follow the rules, and not what to do when other drivers behave improperly or unsafely. Noting that no self-driving vehicles have been found to be at fault in crashes to date, the real outcome to be sought after is not just to follow the rules but to avoid crashes whenever possible regardless of who would be at fault—an approach sometimes referred to as "defensive driving."

Defensive driving for lethal autonomy

This same principle of defensive driving can apply to lethal autonomy—that is, not just ensuring that LAWS follow the rules for IHL compliance, but also incorporating best practices for avoiding inadvertent engagements, such as civilian casualties or fratricide. As discussed in the previous section, these measures collectively form a safety net. As such, no individual practice in itself is a complete solution, but together these practices represent elements of human judgment intended to enhance protection against undesired outcomes. These best practices can be specific measures designed to help avoid such engagements as well as institutional oversight of LAWS, to include the following:

- **Specific measures to help avoid inadvertent engagements.** Such measures have been discussed in detail elsewhere, but they include such best practices as exercising tactical patience, checking for previously unobserved civilians, and giving careful attention to the possibility of misidentification.[42] These best practices should be maintained and updated to learn lessons from current operational experiences with LAWS.

- **Policy.** In addition to legal reviews mandated by Article 36 of Additional Protocol I, States can establish policy intended to strengthen oversight of procurement and use of LAWS. One such policy existing now is the U.S. DOD Directive on lethal autonomous weapon systems.[43] Implementing guidance can be developed and shared among States.

- **Testing.** It can be expected that traditional Test and Evaluation (T&E) processes will be unsuited for autonomous systems that can change and adapt over time, and not be predictable in a deterministic sense. New approaches to T&E will

[42] See, for example, "Mitigating Risk in Lethal Autonomy," in *Insights for the Third Offset,* CNA, DRM-2017-U-016281-Final, September 2017.

[43] DOD Directive 3000.09, *Autonomy in Weapon Systems,* November 21, 2012, Incorporating Change 1, May 8, 2017.

need to be developed to promote greater reliability and predictability of these systems.[44]

Mitigating risk of lethal autonomy

These steps may represent more than what is required for IHL compliance. This is taking an approach similar to what would benefit self-driving cars: instead of making the goal that lethal autonomous weapons can follow the rules, take additional steps to reduce risk and associated mistakes, however they may occur. This approach is consistent with the position taken by the U.S. military described earlier, where IHL compliance was necessary but not sufficient for doing everything possible to protect civilians.

Considering Autonomous Targeting Functions

Another implication of considering autonomy in this broader context is the need to consider the role of autonomous technologies that contribute to targeting without making the final engagement decision. Hence, an autonomous targeting function can influence the use of force without being a weapon system. This general point is illustrated in the PATRIOT example above (with the caveat that the functions in the PATRIOT system were automated, not autonomous). In that example, the system made an automatic determination that an object (in this case, an aircraft) was a ballistic missile and met pre-set criteria that made it eligible for the system to shoot it down, pending operator approval. Similarly, the U.S. and other countries are pursuing the use of artificial intelligence to aid in target identification in full motion video platforms (e.g., Project Maven).

Such capabilities do introduce new kinds of risk. For example, while the system may not pull the trigger in these cases, autonomy in earlier functions can predispose human operators to trust decisions inappropriately without question based on computer-generated information or recommendations. This inappropriate trust was seen in the case of drones where operators using them without compensating for the risks. Accidents involving Tesla cars also show this inappropriate trust, where drivers cede all driving functions to Tesla cars equipped with an "autopilot" function not designed for fully autonomous use. Note that this does not mean that such systems are undesirable. Functionalities such as the autopilot function in Tesla cars, and self-driving cars in general, are less likely to cause an accident compared to

[44] See, for example, "Addressing Concerns about Lethal Autonomy," in *Insights for the Third Offset,* CNA, DRM-2017-U-016281-Final, September 2017.

human drivers (as discussed earlier). However, the use of such systems should match the conditions for which they were designed, with appropriate trust given to them based on their capabilities and limitations. This requirement for developing appropriate trust should influence system design and operator training for these systems.

These capabilities can also offer benefits over human performance. For example, for some kinds of threats (e.g., supersonic sea-skimming missiles), the reaction time for defensive action is very short. Identifying these types of targets autonomously enhances the ability to defend against such threats. In addition, militaries could use artificial intelligence to improve the determination of potential negative second order effects from the use of force. For example, application of machine learning to improve pattern of life analysis could improve the accuracy of collateral damage estimates. Another concern about airstrikes in urban environments is reverberating effects: for example, "incidental destruction of civilian housing and essential civilian infrastructure" that is an unforeseen second order effect of an attack.[45]

Such reverberating effects are illustrated in reports of a Coalition airstrike in Raqqa, Syria, where the Coalition targeted a bridge crossing the Euphrates river for a military purpose: no civilians were reported killed because of the airstrike, reflecting the fact that the Coalition takes precautions to avoid civilian casualties. However, the bridge contained the city's main water pipeline. When the bridge was destroyed, the city's water supply was cut off, leaving an estimated 200,000 people without safe water alternatives.[46] Such reverberating effects can be difficult to predict, and not necessarily detected in processes designed to estimate immediate collateral damage from the use of force. Artificial intelligence – and especially machine learning – could be applied to improve these collateral damage estimates and offer target selection processes that better anticipate reverberating effects of military attacks.

A moral argument for autonomy

Some in the CCW argue that the use of LAWS is never permissible because the use of autonomy in the use of force is not moral or ethical. This paper will not attempt to solve the question of morality regarding the use of LAWS. However, it should be noted that there are strong arguments affirming the potential morality of autonomous weapons. For example, Simpson and Müeller assert that States have a

[45] Isabel Robinson and Ellen Nohle, War in Cities: The 'reverberating effects' of explosive weapons, ICRC, March 2 2017.

[46] Airstrikes cut water supplies in Islamic State stronghold of Raqqa, DW, February 3 2017.

moral obligation to make the risk of LAWS as low as is feasible given the available technology, and argue that there is no moral dilemma with the use of LAWS if the risk posed to noncombatants is lower than that from human combatants.[47] The approach in this report, proscribing a safety net for the development and use of LAWS, is consistent with their view of the obligation of States to minimize risk.

If this is the case, and States can show that LAWS do indeed reduce the risk to civilians, it can be argued from IHL that States could have an obligation to use LAWS versus human combatants per API Article 57 discussing feasible precautions.[48] Such a determination requires a good understanding of the relative risk to civilians from LAWS and humans in armed conflict. The risk of civilian casualties from specific weapon platforms can be—and in Afghanistan, was in some cases—determined through analysis of operational data (e.g., the civilian casualty rates of drones and manned platforms, as discussed earlier); this type of analysis could be made a standard practice in the future to monitor and mitigate risk from LAWS in operations.

[47] Thomas Simpson and Vincent Müeller, "Just War and Robots' Killings," *The Philosophical Quarterly,* Vol. 66, No. 263, 2016.

[48] Additional Protocol I, 1977, Article 57 2.a.ii. states that States should "take all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimizing, incidental loss of civilian life, injury to civilians and damage to civilian objects."

Conclusions and recommendations

Lethal autonomous weapon systems are widely believed to represent a revolution in warfare, with a number of States already announcing these technological developments to be central to their national security plans.[49] The most controversial aspect of this new technology has been: should weapon systems operating autonomously (without a human operator) be allowed to use lethal force? The Convention on Certain Conventional Weapons (CCW) forum within the United Nations (UN) has spent four years to date discussing ethical, legal, and operational considerations of lethal autonomous weapon systems (LAWS). While there are several areas of potential concern about the use of LAWS, the predominant concern in UN discussions to date has been the risk of civilian casualties.

The application of human control over LAWS is seen as a solution to mitigate this risk, but thus far, these discussions have largely addressed military operations from a theoretical perspective. This report examines lessons and specific examples from U.S. military operations over more than a decade and a half. Based on empirical observations from recent military operations, we see that:

- Human control, manifested by judgment exercised in the trigger-pull decision, is not perfect, as documented in civilian casualty and fratricide incidents. Putting a human in the loop for engagement decisions is not a panacea to risks to civilians that would be imposed by LAWS.

- These shortcomings in human control during engagement decisions were addressed in recent operations through refinements in the larger targeting process, which can be seen to represent a broader form of human control over the use of force. Considering only the trigger-pull, shoot or no-shoot decision neglects critical factors that must be considered in the potential use of LAWS. Thus, the targeting process is a better framework for the CCW's consideration of LAWS.

This report also proposes a change in focus for CCW discussions. Designing systems to follow set rules (e.g., IHL for LAWS) is necessary but not sufficient for addressing

[49] This report refers to nation-states as States, consistent with UN terminology.

risk. Additional steps can be taken to more comprehensively address risks posed by LAWS, collectively forming a safety net that represents the application of human judgment over the wider targeting cycle and beyond. This requires moving away from a process-focused discussion of human control to a more comprehensive discussion of how to avoid inadvertent engagements.

Context is an integral part of the use of force. Discussions regarding LAWS in the CCW often have not recognized the fact that how weapons are used—including but not limited to LAWS—is highly context dependent. This context, consisting of both the mission and operational environment, needs to be a key part of the evaluation process for the development and use of LAWS. If risks of LAWS are managed successfully in this way, it is possible that the use of LAWS could not only be permitted but, in fact, be required by IHL in some circumstances as a feasible precautionary measure.

We offer the following recommendations:

- CCW discussions currently focusing on such process considerations as human control should instead be focused on outcome—namely, mitigation of inadvertent engagements. This allows consideration of a more complete set of benefits and risks of lethal autonomy, and how they can be managed.
    - As part of this new focus, human control should be addressed not just as part of the final engagement decision, but as a broader set of military doctrinal functions distributed across the wider targeting cycle.
    - Another implication of considering autonomy in this broader context is the need to consider the role of autonomous technologies that contribute to targeting without making the final engagement decision. For example, the use of artificial intelligence in target identification.
- Ways to comprehensively address risks posed by LAWS should be prominent features in future CCW discussions. States should not only ensure that LAWS follow the rules for IHL compliance, but also incorporate best practices for avoiding inadvertent engagements, such as civilian casualties or fratricide, creating an overall safety net for LAWS. These best practices include:
    - **Specific measures to help avoid inadvertent engagements.** Such measures should be gleaned from military operations, then maintained and updated to learn lessons from operational experiences with LAWS.
    - **Policy.** In addition to legal reviews mandated by Article 36 of Additional Protocol I, States can establish policy intended to

strengthen oversight of procurement and use of LAWS. Implementing guidance can be developed and shared among States.

  - **Testing.** Traditional Test and Evaluation processes are unsuited for autonomous systems that can change and adapt over time, and are not predictable in a deterministic sense. New approaches to T&E will need to be developed to ensure reliability and predictability of these systems.

- The issue of context for LAWS should be considered as well. One possible approach to take to LAWS is crawl-walk-run: first pursue lethal autonomy in easier environments and missions where discrimination is more straightforward, and then learn applicable lessons and evaluate the suitability of LAWS in more challenging contexts.

- States should pursue applications of artificial intelligence to help avoid negative second order effects from the use of force. For example, applying machine learning to improve pattern of life analysis, obtain more accurate collateral damage estimates, and better anticipate reverberating effects (e.g., effects on civilian infrastructure).

- The moral dilemma concerning LAWS hinges in part on the relative risk posed by LAWS to civilians compared to that from human combatants. To evaluate this in future applications of LAWS, States should examine relative rates of inadvertent engagements by LAWS versus non-autonomous weapons. If the rate is lower for LAWS, it may also make the use of LAWS a "feasible precaution" per Additional Protocol I Article 57.

(This page intentionally left blank.)

Bibliography

—, Airstrikes cut water supplies in Islamic State stronghold of Raqqa, DW, February 3 2017. http://www.dw.com/en/airstrikes-cut-water-supplies-in-islamic-state-stronghold-of-raqqa/a-37405866

Additional Protocol I to the Geneva Conventions, 1977.

Anwar, Mohammed, Afghan governor says NATO troops killed child, Reuters, May 16 2011. https://www.reuters.com/article/us-afghanistan-civilian-deaths/afghan-governor-says-nato-troops-killed-child-idUSTRE74F31G20110516.

Byman, Daniel, "Why Drones Work: The Case for Washington's Weapon of Choice," *Foreign Affairs,* July/August 2013.

Center for Army Lessons Learned, *Afghanistan Civilian Casualty Prevention Handbook,* June 2012.

DOD Directive 3000.09, *Autonomy in Weapon Systems,* November 21, 2012, Incorporating Change 1, May 8, 2017.

Ekelhof, Merel A. C., "Lifting the Fog of Targeting: 'Autonomous Weapons' and Human Control Through the Lens of Military Targeting," *Naval War College Review,* forthcoming (summer 2018).

Fenton, Ben, "Britons caught in terrifying hail of fire after fatal US mix-up in the sky," The Telegraph, February 7, 2007.

International Committee of the Red Cross, Expert Meeting on Lethal Autonomous Weapons Systems, November 15, 2017. https://www.icrc.org/en/document/expert-meeting-lethal-autonomous-weapons-systems.

Lee, Dave, Self-driving shuttle bus in crash on first day, BBC, November 8, 2017. http://www.bbc.com/news/technology-41923814

Lewis, Larry, *CID Observations from Iraqi Combat Operations,* CNA, March 10, 2005.

Lewis, Larry, *Drone Strikes: Civilian Casualty Considerations,* Joint Center for Operational Analysis (JCOA), June 2013.

Lewis, Larry, *Drone Strikes in Pakistan: Reasons to Assess Civilian Casualties,* CNA, COP-2014-U-007345-Final, April 2014.

Lewis, Larry, *Improving Joint C2: Lessons from Iraq,* CNA, May 20, 2008.

Lewis, Larry, *Improving Lethal Action: Learning and Adapting in U.S. Counterterrorism Operations,* CNA, September 2014.

Lewis, Larry, *Insights for the Third Offset: Opportunities and Challenges for Autonomy and AI,* CNA, DRM-2017-U-016281-Final, September 2017.

Lewis, Larry, *Reducing and Mitigating Civilian Casualties: Enduring Lessons, Joint and Coalition Operational Analysis,* April 12, 2013.

Lewis, Michael E., "Drones: Actually the Most Humane Form of Warfare Ever," *The Atlantic,* August 21 2013.

Robinson, Isabel and Nohle, Ellen, War in Cities: The 'reverberating effects' of explosive weapons, ICRC, March 2 2017. http://blogs.icrc.org/law-and-policy/2017/03/02/war-in-cities-the-reverberating-effects-of-explosive-weapons/

Russell, Stuart, All-Party Parlimentary Group on Drones Inquiry Evidence Session: "The Use of Armed Drones: Working with Partners and Emerging Technologies," UK Parliament, October 31, 2017.

Schoettle, Brandon, and Michael Sivak, *A Preliminary Analysis of Real-World Crashes Involving Self-Driving Vehicles,* University of Michigan Transportation Research Institute, UMTRI-2015-34, October 2015.

Sewall, Sarah, and Larry Lewis, *Joint Civilian Casualty Study: Executive Summary,* August 31, 2010.

Simpson, Thomas, and Vincent Müeller, "Just War and Robots' Killings," *The Philosophical Quarterly,* Vol. 66, No. 263, 2016.

UN Office of Disarmament Affairs, *Perspectives on Lethal Autonomous Weapon Systems,* UNODA Occasional Papers, No. 30, November 2017.

CNA
ANALYSIS & SOLUTIONS

DOP-2018-U-017258-Final

CNA
ANALYSIS & SOLUTIONS

CNA is a not-for-profit research organization
That serves the public interest by providing
in-depth analysis and result-oriented solutions
to help government leaders choose
the best course of action
in setting policy and managing operations.

Nobody gets closer—
to the people, to the data, to the problem.

Center for Autonomy and AI (CAAI)

CNA's Center for Autonomy and Artificial Intelligence supports the U.S. goal of effectively incorporating autonomy, AI, and related technologies in military capabilities. Throughout history, the ability to adapt technological advances to warfighting has led to fundamental changes in how war is conducted and the tools used in its conduct. Autonomy and AI represent revolutionary technologies in warfare which offer opportunities to the U.S. for countering and deterring emerging threats, addressing security challenges and advancing U.S. national interests. But this opportunity is by no means certain, since autonomy also offers potential asymmetric advantages to near-peer competitors, some of which have been pursuing these capabilities aggressively. Likewise, rapid innovation in the private sector and a commercial research and development sector dwarfing that of the U.S. military create new challenges for the U.S., which will need to quickly identify and integrate cutting edge technological developments in this rapidly changing environment.

Contact

Dr. Larry Lewis, *Director*, CAAI, 703-824-2020, lewisl@cna.org
Dr. Andy Ilachinski, *Principal Research Scientist*, 703-824-2045, ilachina@cna.org

Website: www.cna.org/CAAI