Editor: Fabio Massacci, fabio.massacci@ieee.org

Building Principles for Lawful Cyber Lethal Autonomous Weapons

Fabio Massacci | University of Trento and Vrije Universiteit Amsterdam
Silvia Vidor | University of Trento

The international debate on what makes lethal autonomous weapons lawful only focuses on "killer robots." What about cyberweapons? We discuss some possible measures and design principles for lawful cyber lethal autonomous weapons.

While we should ideally ban all lethal weapons, the world does not seem to agree. Still, there is a broad agreement that the choice of the means in international warfare is not unlimited. Since the first Hague Convention in 1899, which prohibited soft-point and cross-tipped bullets, several agreements have limited the use of some weapons, from the Geneva Convention to the most recent "United Nations 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" [the Convention on Certain Conventional Weapons (CCW)].

Recently, a new generation of weapons is emerging: lethal autonomous weapons (LAWs), applications of artificial intelligence to the military, aiming at replacing human soldiers by systems able to act, react, and combat in complex situations. Their lawful use is highly debated,[1] but the discussion focuses only on their embodiment as "killer robots."

We argue that the category should include cyberweapons: malware/exploits used by state actors for military or intelligence aims, generally toward other state actors, in what is known as cyberwarfare.[2] The North Atlantic Treaty Organization has recognized last year's cyberattacks as proper military attacks able to trigger a full military response from its members. So, 10 years after the article on cyberwarfare principles by Parks and Duggan in this magazine[3] and 35 years after Saltzer and Schroeder's principles on security design,[4] we would like to reopen the discussion:

What security design principles make a lawful cyber lethal autonomous weapon (CLAW)?

We believe that sorcerer's apprentice scenarios seem the most likely to occur. For example, the NotPetya outbreak that took down several hospitals and airports in the United Kingdom started by targeting a Ukrainian tax accounting software that spread to the U.K. subsidiary and then to the country and the world at large.

We introduce the discussion with some concepts that might not be familiar to the average computer expert: What is an *lawful lethal* (conventional) weapon, and what is a lethal *autonomous* weapon? The latter concept is particularly important to move away from Hollywood killer robots to cyberweapons.

What Is a Lawful Lethal Weapon?

Weapons of mass destruction aside, conventional weapons are mostly regulated through the aforementioned CCW, which includes five different protocols banning or limiting their use in international warfare. To do so, International Humanitarian Law (IHL) builds upon four fundamental principles, limiting the adverse effects of armed conflicts:[5] *humanity*, *distinction*, *proportionality*, and *military necessity*. These principles translate into the obligations, among others, to

- always distinguish between military targets (legitimate) and civilian targets (illegal)
- limit damage to civilian targets ("collateral damage") to an amount that is proportional to the military advantage the attack can offer
- avoid superfluous injuries and unnecessary suffering (to military and civilian targets alike).

One of the main consequences of these principles is that weapons that may disproportionately affect civilian targets in comparison with military targets are supposed to be banned (for example, certain categories of landmines and booby traps).

Digital Object Identifier 10.1109/MSEC.2022.3143269
Date of current version: 21 March 2022

We summarize in Table 1 the properties motivating each ban along three main categories: technical, target, and type of damage.

What Is an Autonomous Weapon?

Since 2016, the International Committee of the Red Cross (ICRC) reports that the following systems with varying degrees of autonomy from human operators are already being fielded:[6]

- missile- and rocket-defense weapons (such as Israel's Iron Dome)
- vehicle "active-protection" weapons and anti-personnel "sentry" weapons
- sensor-fused munitions, missiles, and loitering munitions
- torpedoes and encapsulated torpedo mines.

They are generally limited in their mobility—either unmoving, moving only in preprogrammed areas, or endowed in transport systems piloted by humans. Consequently, they can hardly act without a human operator.[7]

These precursors and also Hollywood movies create the mental image for LAWs among the general public. As the very name of the main organization against LAWs—"Campaign to Stop Killer Robots"—shows, autonomous weapons are thought of as *Terminator* movies-like machines acting independently of human supervisors and moving around battlefields through aerial, terrestrial, or marine means.

From a more formal perspective, according to the ICRC, a LAW should be able to perform the following actions:

- *target selection*, which includes research, detection, identification, tracking and selection
- *target attack*, which can include the use of force, neutralization, damage, or destruction.

Among the autonomous functions that justify their "lethal" designation are the identification and strike of targets. Other definitions have been provided by countries interested in their development (Table 2). France, for example, also includes capabilities to define or modify objectives during its mission without human approval as well as self-learning.[6]

Table 1. The main criteria justifying prohibition of weapons in the CCW.

CCW	Property	Weapon type	Motive of prohibition
I	Technical	"[...] any weapon the primary effect of which is to injure by fragments which in the human body escape detection by X-rays"	Causes unnecessary suffering
II	Technical	Mines and booby traps (or certain subcategories)	Can continue working even after detonation, thus causing damage even after the conflict's conclusion to noncombatant targets
II	Technical	Mines and booby traps (or certain subcategories)	May not be self-destructible or self-deactivating, thus causing damage even after the conflict's conclusion to noncombatant targets
II	Target	Mines and booby traps	Can have civilian population as its main target, even in postconflict situations
II	Target	Mines and booby traps	Can be indiscriminate, targeting combatants and civilians alike
II	Target	Mines and booby traps (or certain subcategories)	Can be undetectable in the environment it is placed in, thus creating risks for the civilian population even after the conclusion of the conflict
III	Damage	Incendiary weapons	Seen as indiscriminate and easily affecting civilians and/or the environment
IV	Damage	Blinding laser weapons	Cause unnecessary and superfluous permanent damage

This table describes, for each weapon type, the main criteria listed in the relative protocol behind the weapon ban and the category to which each criterion can be attributed (see https://www.icrc.org/en/doc/assets/files/other/icrc_002_0811.pdf).

Among the worrying "killer robot" scenarios is the possibility for LAWs to be deployed in swarms, potentially exhibiting unprogrammed and unexplainable collective behavior.[8] At the moment, robotic LAWs are likely being developed in the United States, China, Turkey, and Russia; France and Germany also appear to be financing research on the topic.

The hearth of the issue is the definition of autonomy: international law places responsibility on humans, but increasing level of autonomy complicate the identification of the legally responsible human. During the CCW sessions held so far, three hypothetical and apparently mutually exclusive scenarios have been identified:

1. Human *in* the loop: The LAW performs a series of activities based on human input and/or authorization.
2. Human *on* the loop: The LAW performs a series of activities under the supervision of a human operator, who may override the system in case of necessity.
3. Human *out* of the loop: The LAW performs a series of activities independently, without the need for human input or oversight.[9]

The human-*in*-the-loop scenario seems excluded by the very definition of LAWs: continuous human input and/or authorization would make it nonautonomous and, by definition, some other kind of weapon. The human-*on*-the-loop scenario is considered to be the most feasible. This model risks undermining a LAW's rapidity of action and reaction (its very military advantage) or being meaningless by reducing the human's possibility to override the system due to so-called *automation bias* and the difficulties in keeping up with LAW processes,[10] a phenomenon that is well known to operators of security operations centers.[11,12]

The human-*out*-of-the-loop scenario seems out of scope in the CCW discussion, given the current embodiment of LAWs as "killer robots," which are limited by the physical supply chain of energy or transportation: going somewhere physically can take hours and requires fuel (or electric power). However, this is not the case for a CLAW. We believe the CCW discussion does not capture the specificity of cyber.

All three scenarios—human in the loop, human out of the loop, and human on the loop (in that order)—will take place during the CLAW's lifecycle. Stuxnet's lifecycle is an illustrative case in point.[13]

Autonomy for CLAWs

Some moments in a CLAW's working process require an input coming directly from the human operator: the decision to start the system, the choice of deploying it to a specific mission, or retiring it at the end or in case of malfunctioning. These actions fall under the human-in-the-loop scenario.

Then, a considerable number of actions (such as reconnaissance, initial compromise and lateral movement, and target identification and selection) will likely be performed by (C)LAWs with no human intervention at all and would fall under the human-out-of-the-loop scenario. Stuxnet's propagation happened without human supervision. This is typical of cyber.

Finally, some particularly crucial phases (for example, launching or recalling an attack) could be programmed to require at least a minimal level of human supervision and could therefore be traced back to the human-on-the-loop scenario. The retention of some level of human control in the most critical phases is an essential criterion for the lawful use of a LAW in conflict (as we will discuss for ransomware). In the case of programs with essentially deterministic (or "undo") effects, the very fact that they are programmed might provide this notion of on-the-loop scenarios. From this perspective, Stuxnet,[13] which has been programmed to act on some specific nuclear turbines and not just any turbines or find out by itself which turbines to attack by trial and error (also known as machine learning), would map to the human-on-the-loop scenario

Table 2. Country-specific definitions of Autonomy for LAWs.

Country	Definition	Source
France	Complete weapons whose carrier moves freely, targeting and firing without intervention, approval, or human supervision	23
Russian Federation	Unmanned technical material that is not munitions and that is designed to carry out military and support tasks without any participation of a human operator	23
United States	System[s] that, once activated, can select and engage targets without further intervention by a human operator	24

and therefore to the lawful usage in a wartime scenario.

Lethality for CLAWs

The key observation is that lethality is just a kinetic, incendiary, or explosive *consequence* of a conventional weapon system. Modern weapons such as a torpedo do not even touch the ship they sink. The explosion of the torpedo just creates a void space in the water under the hull. It is the suction to fill the void that cracks the hull. The same reasoning applies to the actions of a cyberattack on the targeted cyberphysical system.

A simple example is a malware whose primary effect is overheating of a lithium-ion battery up to explosion—used in Internet of Things systems and smartphones but also in manned and unmanned military vehicles—which can lead to lethal consequences for those placed in the vicinity of the concerned systems and to further cascading explosions.[14] Another example is the operations of locks and dikes.[15]

Other examples are cyberattacks targeting subsystems in charge of

Table 3. Building principles for lawful CLAWs.

Property	Condition	Description	Purpose
Technical	Absence of disrupting functionless fragments	The CLAW should not disseminate fragments of code that do not provide functionalities to the weapon itself but can disrupt the execution of the target system if invoked by chance.	We are back to "unnecessary suffering." The target systems can be fully DoSsed or taken over, but the malicious code should be purposeful in the same way that ROP gadgets are.
Technical	Permanent self-identification	Each CLAW should have a fingerprint or signature recognizable by its own designer, preserved through obfuscation or mutations.	While detectability is not reasonable as a criterion for a CLAW, the designer should to be able to recognize its own to remove it after the conflict.
Technical	Eventual self-deactivation	The CLAW should be capable of deactivating itself either by the system itself (for example, after a timeout or by inserting a key) or through a command and control system.	Indefinitely operational presence and damage might be needed throughout the conflict, but the impossibility to stop it would be against the principle of proportionality and unnecessary suffering.
Target	Deterministic target (or nontarget) boundaries	Deterministic target or nontarget boundaries for the actual deployment of a lethal payload should be controlled through algorithmic fingerprinting.	The ability to perform stealth but not disruptive propagation across the cyberspace might be justified by military necessity, whereas indiscriminate payload unleashing would be against the principle of proportionality.
Target	Initial validated specification for learning	Learning algorithms should start from an initial target definition that has been validated before deployment.	Trial and error for target identification would be against the principle of unnecessary suffering and proportionality, so the initial definition for a target should be done offline.
Damage	Appropriate software stack position	The programmed type of damage should be at the appropriate point of the software/hardware stack to achieve the CLAW's aim.	This would make it possible to avoid collateral damage to components (and related cyberphysical systems) that is not a consequence of the failure of the attacked component.

Proposed principles that cyber autonomous weapons may need to satisfy to be legally deployed in warfare are obtained starting from the features that determined another weapon's prohibition according to the CCW.

the situational awareness of air force systems (such as flying at a dramatically wrong altitude or with tip up or down, or even upside down). Past incidents on remotely piloted systems are often due to programming errors[16] and can be well replicated by cyberattacks reporting wrong sensor readings or actuators commands in the way that Stuxnet did.

Spatial misorientation is also a source of many incidents among crewed systems,[17] and even a crewed system cannot fight against the software trying to make things "right," as the Boeing MAX incidents showed.[18] Such attacks could be launched against military systems, but, given the close integration between the military and civil aviation supply chains,[19] a not-so-careful cyberattack might down at once both commercial and military jets.

Building Security Principles for Lawful CLAWs

Of course, traditional building security principles do apply,[4,18] but what we want to do here is to propose building security principles for lawful CLAWs. This is way trickier than one thinks to avoid falling into the realm of the irrelevant.

Consider the principle of distinction: it establishes the inviolability of certain nonmilitary organizations, such as the ICRC, but also of medical buildings and personnel. These organizations cannot (well, at least should not) be targeted by attacks and must carry a distinctive emblem (such as the Red Cross or Crescent) that cannot be used by military forces improperly. Physical armies of a country are supposed to always be identifiable by wearing a uniform or a distinctive sign.

Unfortunately, having a software presenting itself as "Hey, I'm a malware, I'll probe your TCP ports for buggy services" is a technical no-go. Cyberattacks work precisely by confusing the target program into thinking it is interacting with a legitimate client program, rather than being attacked.[20] Further, protocols respond as specified, and there is no separate "visual" channel to see the red sign painted on the roof of an Internet Protocol address or a Docker pod. However, the idea of recognizing and fingerprinting the targets, once control has been taken over, is actually possible and implemented by several malware authors: DarkSide, the malware behind the Colonial Pipeline hack, or REvil, have a hard-coded do-not-install list of countries.

The request to be invisible to X-rays (the technical measurement for unnecessary suffering) cannot be mapped to the straightforward equivalent of being invisible to process monitoring as, again, the whole point of cyberattacks is to escape process detection.[20] Modern attack techniques such as return-oriented programming (ROP) are based on the very idea of using small fragments of code that are already present.[21]

To decide what should be allowed and what should be prohibited by international treaties, we should start with the key functionality of a cyberweapon: to eventually provide stealth control of an IT system, for integrity attacks, or disable it, for denial-of-service (DoS) attacks.

We tried to sketch positive principles in Table 3 based on the criteria that determined a weapon's ban within the framework of the CCW and according to the general principles of IHL. For example, launching a disk encryption attack on military systems as done by a ransomware might be a lawful CLAW in a conflict, provided the attacker holds the key to unlock the system. Ransomware where nobody has a key would be unlawful.

There is one aspect we haven't tackled so far: the identification of the perpetrators of (unlawful) CLAW attacks.[22] Attributing cyberattacks to state actors even in peaceful times is hard. The absence of on-the-ground human personnel recognizable by identifying signs, the possibility of disguising one's location in cyberspace, and the presence of human-out-of-the loop propagation phases are fundamental complicating factors.

Cyberspace is not among the fields of warfare traditionally considered by international law, and it is unclear which codified or customary norms would apply to cyberwarfare. With this article we would like to kick-start a discussion…before it happens. ■

Acknowledgments

This work is partly supported by the European Union's Horizon 2020 Program under grant 830929 (CyberSec4Europe) and grant agreement 770138 (OPTICS2) and the Dutch Government Sectorplan Fund.

References

1. L. M. R. Choudhury, A. Aoun, D. Badawy, L. A. de Alburquerque Bacardit, Y. Marjane, and A. Wilkinson, "Final report of the panel of experts on Libya established pursuant to security council resolution 1973 (2011)," United Nations Security Council, New York, NY, USA, 2021. Accessed: Jul. 19, 2021. [Online]. Available: https://undocs.org/S/2021/229
2. A. P. Liff, "Cyberwar: A new 'absolute weapon'? The proliferation of cyberwarfare capabilities and interstate war," *J. Strategic Stud.*, vol. 35, no. 3, pp. 401–428, 2012, doi: 10.1080/01402390.2012.663252.
3. R. C. Parks and D. P. Duggan, "Principles of cyberwarfare," *IEEE Security Privacy*, vol. 9, no. 5, pp. 30–35, 2011, doi: 10.1109/MSP.2011.138.
4. J. H. Saltzer and M. D. Schroeder, "The protection of information in computer systems," *Proc. IEEE*, vol. 63, no. 9, pp. 1278–1308, 1975, doi: 10.1109/PROC.1975.9939.
5. N. Melzer, *International Humanitarian Law: A Comprehensive Introduction*.

Geneva, Switzerland: International Committee of the Red Cross, 2016.
6. "Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons," International Committee of the Red Cross, Geneva, Switzerland, 2016. Accessed: Jul. 7, 2021. [Online]. Available: https://icrcndresourcecentre.org/wp-content/uploads/2017/11/4283002Autonomus-Weapon-Systems WEB.pdf
7. M. D. Landa, *War in the Age of Intelligent Machines*, 1st ed. New York, NY, USA: Zone Books, 1991.
8. M. Ekelhof and G. P. Paoli, "Swarm robotics. Technical and operational overview of the next generation of autonomous systems," United Nations Institute for Disarmament Research, Geneva, Switzerland, Tech. Rep., 2020. Accessed: Jul. 7, 2021. [Online]. Available: https://www.unidir.org/sites/default/files/2020-04/UNIDIR%20Swarm%20Robotics%20-%202020.pdf
9. P. D. Scharre, "Where does the human belong in the loop?" 2014. Accessed: Jul. 2, 2021. [Online]. Available: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2014)/Scharre_MX_LAWS_technical_2014.pdf
10. M. Leese, "Configuring warfare: Automation, control, agency," in *Technology and Agency in International Relations*, M. Hoijtink and M. Leese, Eds. New York, NY, USA: Routledge, 2019, pp. 42–65.
11. S. Bhatt, P. K. Manadhata, and L. Zomlot, "The operational role of security information and event management systems," *IEEE Security Privacy*, vol. 12, no. 5, pp. 35–41, 2014, doi: 10.1109/MSP.2014.103.
12. F. B. Kokulu *et al.*, "Matched and mismatched SOCs: A qualitative study on security operations center issues," in *Proc. 2019 ACM SIGSAC Conf. Comput. Commun. Security*, pp. 1955–1970, doi: 10.1145/3319535.3354239.
13. R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security Privacy*, vol. 9, no. 3, pp. 49–51, 2011, doi: 10.1109/MSP.2011.67.
14. A. B. Lopez, K. Vatanparvar, A. P. D. Nath, S. Yang, S. Bhunia, and M. A. A. Faruque, "A security perspective on battery systems of the Internet of Things," *J. Hardware Syst. Security*, vol. 1, no. 2, pp. 188–199, 2017, doi: 10.1007/s41635-017-0007-0.
15. J. de Lange, *Security of Flood Defenses*, 1st ed. Berlin, Germany: De Gruyter, 2019.
16. G. Hunter, C. A. Wargo, and T. Blumer, "An investigation of UAS situational awareness in off-nominal events," in *Proc. 2017 IEEE/AIAA 36th Digit. Avionics Syst. Conf. (DASC)*, pp. 1–10, doi: 10.1109/DASC.2017.8102038.
17. R. J. Poisson and M. E. Miller, "Spatial disorientation mishap trends in the U.S. Air Force 1993-2013," *Aviation, Space, Environmental Med.*, vol. 85, no. 9, pp. 919–924, 2014, doi: 10.3357/ASEM.3971.2014.
18. F. Massacci, "Is 'deny access' a valid 'fail-safe default' principle for building security in cyberphysical systems?" *IEEE Security Privacy*, vol. 17, no. 5, pp. 90–93, 2019, doi: 10.1109/MSEC.2019.2918820.
19. M. Tiwari, "An exploration of supply chain management practices in the aerospace industry and in Rolls-Royce," M.S. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 2005. Accessed: Jan. 11, 2022. [Online]. Available: https://dspace.mit.edu/bitstream/handle/1721.1/33373/62523032 - MIT.pdf?sequence=2
20. W. Lee, *Malware and Attack Technologies Knowledge Area Issue*, A. Rashid, H. Chivers, E. Lupu, A. Martin, and S. Schneider, Eds. Bristol, U.K., 2021. [Online]. Available: https://cybok.org/media/downloads/CyBOK_v1.1.0.pdf
21. S. Ahmed, Y. Xiao, K. Snow, G. Tan, F. Monrose, and D. Yao, "Methodologies for quantifying (re-)randomization security and timing under JIT-ROP," in *Proc. 2020 ACM SIGSAC Conf. Comput. Commun. Security*, pp. 1803–1820, doi: 10.1145/3372297.3417248.
22. T. Rid and B. Buchanan, "Attributing cyber attacks," *J. Strategic Stud.*, vol. 38, nos. 1–2, pp. 4–37, 2015, doi: 10.1080/01402390.2014.977382.
23. "It's time to exercise human control over the CCW," Reaching Critical Will, Geneva, Switzerland, CCW Report n.7, 2019. [Online]. Available: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2019/gge/reports/CCWR7.2.pdf
24. "Directive 3000.09: Autonomy in weapon systems," U.S. Department of Defense, 2012. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf

Fabio Massacci is a professor at the University of Trento, Trento, 38123, Italy, and Vrije Universiteit, Amsterdam, 1081 HV, The Netherlands. He participates in the CyberSec4Europe pilot and leads the H2020 AssureMOSS project. Massacci received a Ph.D. in computing from the University of Rome "La Sapienza." For his work on security and trust in sociotechnical systems, he received the Ten Year Most Influential Paper Award at the 2015 IEEE International Requirements Engineering Conference. He is a Member of IEEE. Contact him at fabio.massacci@ieee.org.

Silvia Vidor is a research fellow at the University of Trento, Trento, 38123, Italy. She contributes to the areas of cybersecurity governance and education within the Horizon 2020 CyberSec4Europe pilot. She is also a research member of the Italian nonprofit organization Privacy Network. Vidor received an M.A. in international security studies from the University of Trento and Sant'Anna School of Advanced Studies in Pisa. Contact her at silvia.vidor@unitn.it.