Zero tolerance policy for autonomous weapons: Why?

Birgitta Dresp-Langley, UMR 7357 CNRS-Université de Strasbourg

Structured abstract

A brief overview of Autonomous Weapon Systems (AWS) and their different levels of autonomy is provided, followed by a discussion of the risks represented by these systems under the light of the just war principles and insights from research in cybersecurity. Technological progress has brought about the emergence of machines that have the capacity to take human lives without human control. These represent an unprecedented threat to humankind. This commentary starts from the example of chemical weapons, now banned worldwide by the Geneva protocol, to illustrate how technological development initially aimed at the benefit of humankind has, ultimately, produced what is now called the "weaponization of Artificial Intelligence" (AI). We are led to conclude that AWS fail the discrimination principle, and that the only way of mitigating the risk they represent to humankind is the rapid negotiation of treaties for the implementation of an international zero-tolerance policy against the development and/or deployment of autonomous weapon systems. Given that scientific research on AWS is altogether lacking in the public domain, the viewpoint here is based on common sense rather than scientific evidence. Yet, the implications of the potential weaponization of our work as scientists, especially in the field of AI, are reaching further than we may think. The potential consequences of the deployment of AWS for citizen stakeholders are incommensurable. This viewpoint points towards good reasons why we need to raise awareness of the threats represented by AWS, and legal policies to ensure that these threats will not materialize.

Keywords: autonomous weapon systems; chemical weapons; weapons of mass destruction; the Geneva Protocol; the discrimination principle; just war theories; cybersecurity; adversarial hacking

Introduction

On the first of May in 1915, Clara Haber [1], née Immerwahr, committed suicide. A week before her death, her husband, the German scientist Fritz Haber [2], had organized the first chlorine-gas attack at Ypres in Belgium, which was aimed at breaking the military stalemate in Germany's favor. Ten years before, in 1905, Haber had achieved what other peers before him had attempted in vain. Using high pressure and a catalyst, Haber was able to trigger a direct reaction between nitrogen gas and hydrogen gas to create ammonia. The process is considered as one of the most important technological breakthroughs of the 20th century as it enabled the mass production of agricultural fertilizers supporting half of the world's food base and leading to a massive increase in the growth of crops for human consumption. During the First World War, Haber developed a new weapon, poison gas (the first of which was chlorine gas) and supervised its initial deployment on the Western Front at Ypres in Belgium and thereby became "the father of chemical warfare" [3], which is believed to have prompted in 1915 the suicide of his wife, herself a chemist. In 1918, Haber was awarded the Nobel Prize in Chemistry for the synthesis of ammonia from its elements.

In the world of today, chemical weapons are considered weapons of mass destruction, and their use in armed conflict is a violation of international law. The Chemical Weapons Convention [4], has been ratified by 145 nations and is in effect since 1997. It strictly prohibits the production, storage, or use of toxic chemicals as weapons of war. Recent scientific development in the fields of organic synthesis and chemical design [5, 6], however, pose new challenges to the Convention, and new developments in analytical chemistry will be necessary to assist its effective implementation. Moreover, the potential of using new chemical weapons in combination with robots and artificial intelligence for the manufacturing of an entirely new breed of autonomous weaponry [6] raises new types of questions on how to prevent the re-emergence of chemical warfare under a radically different and by far more sophisticated and pernicious form. The maximum-risk autonomous weapons [6] would be drone swarms and autonomous CBRN (Chemical, Biological, Radiological, Nuclear) weapons, which include miniature insect drones reduced to undetectable devices capable of administering lethal biochemical substances through their stings (Fig. 1). Science and society are on the brink of unprecedented technological development and change where clear lines dividing fundamental science from application, benefits from risks, and responsible deployment from abuse no longer can be drawn.

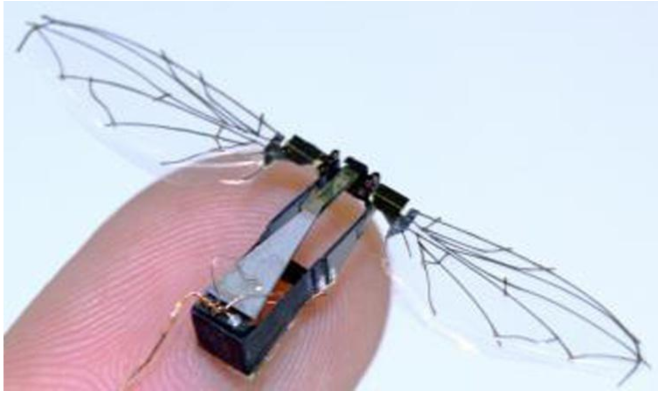

Figure 1: The potential for new forms of chemical weapons combining nanotechnology with autonomous drone swarms that look like insects (autonomous flybots) represents an unprecedented risk to humankind.

This opinion paper discusses the many reasons why the question of a supposedly responsible technological development for new forms of warfare with autonomous weaponry is a fundamental dilemma that cannot be resolved and, therefore, does not belong to the realm of ethical debate, but to that of international law. First, the different kinds of autonomous weapon systems (AWS) already out there, i.e. currently being developed and/or already employed, and the different levels of autonomy such technology implies will be summarized in this paper. This will be followed by a brief explanation of the "just" versus "unjust wars" dilemma in ethics [8, 9] and, finally, the risks of autonomous weaponry for humankind will be made clear. The conclusions will argue for a zero tolerance policy for autonomous weapons, at the international level, akin to policies currently implemented for chemical warfare and stipulated in the Chemical Weapons Convention [4] and in the Geneva Protocol of the United Nations [10].

From scientific and technological progress to AWS

Autonomous Weapon Systems (AWS), announced to become the third revolution in warfare [11], raise fundamental questions about technological development in related fields like Robotics, Artificial Intelligence (AI), autonomous vehicles, and human operated drones. Although not systematically extended to the deployment of AWS, the design principles, algorithms, and technology produced in these fields can be directly translated into novel solutions for AWS. This is maybe one of the most pressing and difficult problems in science currently, where deeper ethical insights and debate are useful, but unlikely to bring about the necessary and urgently required solution. This will be discussed further here in a separate chapter. The earliest example of an autonomous vehicle was The American Wonder developed in 1925 [12], which cruised the streets of New York City remotely controlled by another vehicle following behind, an early demonstration of platooning [13], i.e. the coordinated formation of vehicles navigating in a fleet under shared automatic control. Since then, advancements in technology enabled functionalities like adaptive and predictive cruise control combined with RADAR, LIDAR, high-resolution 360-degree cameras and, ultimately, AI [14]. Scientists and experts have begun to raise their voices against the dangers associated with these technological developments for humankind and the "weaponization" of AI [15, 16]. Lethal autonomous weapons and AWS currently exploiting such technology, under development and/or already employed, include autonomous stationary sentry guns and remote weapon stations programmed to fire at humans and vehicles, killer robots (also called "slaughter bots"), and drones and drone swarms with autonomous targeting capabilities.

Autonomous stationary sentry guns

A sentry gun (https://en.wikipedia.org/wiki/Sentry_gun) is a remote weapon that is automatically aimed and fired at targets that are detected by sensors. The earliest functioning military sentry guns were close-in point-defense weapons used for detecting and destroying short range incoming missiles and enemy aircraft. Such were first used exclusively on ships, but are now also land-based defenses. The first of its kind to have an integrated system that includes surveillance, tracking, firing, and voice recognition would be the SGR-A1, jointly developed by Hanwha Aerospace and Korea University to assist South Korean troops in the Korean Demilitarized Zone in a highly classified project.

Autonomous killer robots

Killer robots or "slaughter bots", are autonomous robotic systems able to select and attack targets without intervention by a human operator [18]. While in some of these systems, the initial command to attack would be given by a human and the robot then has a degree of autonomous 'choice' for action, other systems without any human in the loop are currently tested in several countries. Therein, the decision to deploy lethal force is delegated to a machine. Such far-reaching development would fundamentally change warfare of the future. The function of autonomously selecting and attacking targets could be applied to various platforms such as battletanks, fighter jets, or ships. Another term used to describe these weapons is lethal autonomous weapon systems (LAWS). When equipped with advanced sensors and AI, moreover, autonomous weapons could be trained to operate in coordinated platoons to overwhelm enemy defenders, in distributed surface-warfare action groups or electronic warfare vessels, all unmanned and operating autonomously.

Autonomous drones and swarms

In October 2013, the United States Strategic Capabilities Office launched 103 Perdix drones, which communicated using a “distributed brain” to assemble into a complex formation, travel across a battlefield, or regroup into a new formation. The swarm was created by MIT engineering students using commercially available components and design. In theory, drone swarms could be scaled to tens of thousands of drones to create an autonomous weapon akin to a low-scale nuclear device [19]. Armed, fully-autonomous drone swarms are deemed to become future weapons of mass destruction because they combine two properties unique to traditional weapons of mass destruction: mass harm and lack of human control to ensure the weapons do not harm civilians. Experts doubt that any single autonomous weapon could ever be capable of adequately discriminating between civilian and military targets, and with thousands or tens of thousands of drones is a swarm, this risk becomes incommensurable [20].

In summary, AWS are lethal devices that identify potential enemy targets and independently choose to attack those targets on the basis of algorithms and AI. AWS other than stationary sentry guns require the integration of several core elements: a mobile combat platform, sensors of various types to scrutinize the platform’s surroundings, a processing systems to classify objects discovered by the sensors, and algorithms that prompt the system to initiate attack when an allowable target is detected. The U.S. Department of Defense described an autonomous weapons system as a “weapons system that, once activated, can select and engage targets without further intervention by a human operator” [21]. While there is currently no international consensus on a definition of AWS, they have been rated according to the level of their autonomy from human control.

Three levels of autonomy of AWS

The concept of autonomy in the context of AWS may be defined as the ability of the system to execute a task or set of operations without human input through action upon or interaction with its environment that are determined and controlled by algorithms. What matters critically to the definition of an AWS appears to be the type of decision or function that is rendered autonomous by no longer being under the control of a human operator. Under this premise, three levels of increasing autonomy may be proposed for AWS [20]:

- Supervised autonomous weapon or 'human 'on-the-loop' system, is autonomous weapon system that is designed to provide human operators with the ability to intervene and terminate engagements before unacceptable levels of damage occur. Examples would include defensive weapon systems used to attack, which would independently select and attack targets according to their program while a human retains the full supervision of all operations and can override the system, if necessary, within a limited time-period.
- Semi-autonomous weapon or 'human-in-the-loop' system, once activated, is intended to only engage individual targets or specific target groups that have been selected by a human operator. Examples would include homing munitions that, once launched to a particular target location, search for and attack preprogrammed categories of targets within the area.
- Fully autonomous weapon or 'human out-of-the-the-loop' system, once activated, can select and engage targets without further intervention by a human operator. Examples would include 'loitering' weapons that, once launched, search for and attack their intended targets over a specified area without any further human intervention, or weapon systems that autonomously use electronic 'jamming' to disrupt communications.

Some of the critical functions of such weapon systems have been automated for many years. A weapon system does not necessarily need to be highly complex for it to be autonomous, which is illustrated by existing anti-personnel weapon systems that have autonomous modes such as the so called sentry guns (cf. see here above). Autonomous weapon systems in use today, autonomous, semi-autonomous or supervised according to the definitions provided here above, are claimed to be constrained in several respects [17, 20]. First, they are claimed to be limited in the tasks they are employed for, with defensive action against rocket attacks, or offensive action against specific military installations such as radar, for example. Second, they are claimed to be limited in the types of targets they attack, which are reduced to vehicles or objects rather than civilians. Third, they are claimed to be used in relatively simple and predictable environments such as at high sea, or on land areas that are remote from populated zones. However, the potential of AWS to become weapons of mass destruction is real, and scientists, experts, and journalists worldwide are expressing concern about the fundamentally unethical nature of the development and/or deployment of AWS [22]. From an ethical standpoint, AWS are not eligible whatever their level of autonomy, as they all fail in satisfying the *principle of discrimination* stipulated in the framework of contemporary military ethics under the premise of Just War Theories [7, 8, 9].

Ethical concern regarding AWS

Autonomous Weapon Systems raise many questions and concerns that require in-depth research and public discussion on the ethical and moral responsibility relative to their development and/or deployment. While ethical standards for decision-making are to some extent studied in relationship with the research and development of autonomous vehicles and human operated drones, such has not yet widely been extended to AWS [18, 20, 22]. In fact, to develop ethical standards and rules for moral judgment or decision making on the development and/or deployment of AWS requires taking into account ethical standards and rules of warfare as such.

Saint Augustine was the first individual in Christianity to have proposed a theory on war and justice, the so-called Just-War Theory https://en.wikipedia.org/wiki/Just_war_theory. He referred to the Bible and claimed that some wars are necessary to fight evil. Saint Thomas Aquinas revised Augustine's theory and proposed several criteria for a just war: it needs to be waged by a legitimate authority, have a just cause, follow the right intentions, have a reasonable probability of success, the nations involved in the war must avoid disproportionate military action, only use the amount of force absolutely necessary, and the use of force must distinguish between the militia and civilians. This last principle is called the *principle of discrimination* in contemporary military ethics [24]. It is to ensure that innocent citizens do not become the target of war, and that the killing of civilians is avoided at all cost. Just-War Theory in contemporary ethics builds on these principles as a set of rules for military combat where conventions are meant to serve as guides to human action. While true blue pacifists reject war in any form as immoral, and thereby imply that all acts within war are immoral and inexcusable and true bleu militarists believe that in war “all is fair”, just war theorists take the pragmatic stance that, should war break out for one reason or another, considerations relative to its justification are necessary, and rules and procedures need to be followed to ensure that specific sanctuaries from war's dreadful consequences are upheld and protected. Contemporary just war theory [23] concludes that the use of autonomous technologies is neither completely morally acceptable, nor is it completely morally unacceptable [8, 9]. Any technology of warfare could be just or unjust depending on the situation because what is and is not acceptable in war is ultimately a *convention*. However, while such theories extrapolate from the conventions proposed by Saint Augustine and Saint Thomas Aquinas in an attempt to deal with new technologies like AWS, they remain mere speculation. Also, the principles of ethical warfare in Just War Theory “non-negotiable”, i.e. when one of these principles is violated by a procedure of a type of weapon, the ethical debate regarding the latter is, in principle, settled. The major ethical objection against AWS is the fact that, whatever their level of autonomy, they all fail the *principle of discrimination* in the sense that one cannot ensure that they will not harm civilians [24]. Therefore, the case of AWS belongs into the realm of international law and policy making, and it is up to the international community to establish a new set of conventions to regulate their use through international legislation and treaties. Such a process can be informed by ethics theory to clarify the moral foundations for AWS control under the light of individual rights or other solid moral grounds. However, while ethical theory might positively influence the practical control of this technology through international law, an ethical debate *per se* cannot solve the problem of the many threats AWS represent for humankind. In addition to these threats, our planet is running out of resources. Wars (whatever form they may take) are expensive. Governments urgently need to focus on technological development for sustainability instead of wasting precious resources on new types of weaponry that, beyond failing the principle of discrimination, are unsafe in other aspects. There is no such thing as an autonomous system that cannot be hacked, and the risk that non-state actors take control of AWS through what is called *adversarial hacking* is real.

Risk of '*adversarial hacking'*

In areas from robotics and AI to the material and life sciences, the coming decades could bring about innovation and scientific progress that should help us promote peace, protect our planet, and resolve the root causes of poverty and suffering worldwide. With the ability to interact through cyberspace to spread and exchange information, and to reinforce technological development for peace and sustainability in an increasingly networked world, this goal is severely jeopardized by adversity from various sources. Should war break out, failure of AWS, whatever their level of autonomy, to satisfy the *principle of discrimination* as a major threat is compounded by other risks that lead to argue for banning the development and/or the deployment of AWS by law [25, 26]. The deployment of AWS can pose difficulties for the attribution of hostile acts and lead to unintentional escalation of conflicts. Moreover, non-state actors such as terrorist groups and international criminal networks could harness or sabotage the technology in service of their own agendas through what is called *adversarial hacking.* This risk is real [27] and concerns AWS with any level of autonomy (cf. chapter 3 here above), including 'human-on-the-loop' or supervised autonomous systems, that can operate independently but are under the oversight of a human who is supposed to intervene if "something goes wrong".

In its simplest definition, *adversarial hacking* is an action with malicious intent performed by someone or a group to compromise a system or the cyber resources used by that system. The US Defense Science Board Task Force Report on Resilient Military Systems and Advanced Cyber Threat [28] divides potential sources of adversarial attacks (*adversaries*) into three major categories:

1) Adversaries using off-the-shelf tools that exploit system vulnerabilities
2) Adversaries with resources and capabilities to discover new, unsuspected vulnerabilities
3) Adversaries that can invest billions of dollars and unlimited time for the development of new tools to create new vulnerabilities

One may not be able to imagine the amount of resources that category three adversaries could deploy for attacks that impact the cyber capabilities of any AWS. Attacks by *adversarial hacking* can target any level of such systems, from the infrastructures that records/measures state information, to the algorithms and processes that govern the automatic control systems, whether supervised by human operators or not. Sentient adversaries to the system may act to corrupt state information, interrupt communications, or to modify the automatic control systems of AWS. This could modify the dynamics and/or the structure of their entire physical network. The adversaries may be then able to access and corrupt both local and network-wide state information, and to cause local or network-wide perturbations to the physical network. The results of *adversarial hacking* could generate an unknown variety of different types of attack on an AWS system causing, in the best case scenario system failure, or producing scenarios where the system is corrupted to do what it is not supposed to (i.e. kill civilians, for example).

Conclusions

Technological research and development in AI has brought about the emergence of machines that have the capacity to take human lives without human control. This has brought about a new and unique threat to humankind. History has many examples where scientific progress and innovation initially aimed at humankind’s benefit was then applied for warfare. Scientific insights into biological modification and synthesis designed to help scientists better understand disease could be misused to increase the potency of infectious agents deployed by AWS. Furthermore, such weapons raise serious concerns about their potential misuse by non-state actors. Cyberspace delivers the critical infrastructure for AWS, yet, it is not a safe place. The “weaponization” of scientific and technological development and innovation has, ultimately, produced AWS, some of which function without any “human in the loop”. Their development and/or deployment could have unintended, unforeseen, and unprecedentedly devastating consequences for humankind. The first step to prevent the worst from happening has to be an immediate moratorium on the development, deployment, and use of lethal autonomous weapons worldwide, and the full commitment of the United Nations to negotiate a permanent international treaty within the shortest possible delay.

References

[1] Friedrich B, Hoffmann D. Clara Haber, nee Immerwahr (1870-1915): Life, Work and Legacy. *Z Anorg Allg Chem*. 2016; 642(6):437-448.

[2] Witschi H. Fritz Haber: December 9, 1868-January 29, 1934. *Toxicology*. 2000;149(1):3-15.

[3] Fitzgerald GJ. Chemical warfare and medical response during World War I. *Am J Public Health*. 2008; 98(4):611-625.

[4] The Chemical Weapons Convention. 2021; *Organization for the Prohibition of Chemical Weapons.* Available online at: https://www.opcw.org/chemical-weapons-convention

[5] Lei K, Li P, Yang XF, Wang SB, Wang XK, Hua XW, Sun B, Ji LS, Xu XH. Design and Synthesis of Novel 4-Hydroxyl-3-(2-phenoxyacetyl)-pyran-2-one Derivatives for Use as Herbicides and Evaluation of Their Mode of Action. *J Agric Food Chem.* 2019; 67(37):10489-10497.

[6] Deng X, Zheng W, Jin C, Bai L. Synthesis of Novel 6-Aryloxy-4-chloro-2-phenylpyrimidines as Fungicides and Herbicide Safeners. *ACS Omega*. 2020; 5(37):23996-24004.

[7] Armitage R. We must oppose lethal autonomous weapons systems. *Br J Gen Pract*. 2019;d69(687):510-511.

[8] Walzer M. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, Basic Books, 1977; New York.

[9] McMahan J. The Sources and Status of Just War Principles, *Journal of Military Ethics*. 2007; 6(2):91-106.

[10] The Geneva Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare, *The United Nations Office for Disarmament Affairs.* 1925; available online at: https://www.un.org/disarmament/wmd/bio/1925-geneva-protocol/

[11] Reports from the American Association for the Advancement of Science meeting in Washington DC. *The Science Show on ABC.* 2019; available online at: https://www.abc.net.au/radionational/programs/scienceshow/the-third-revolution-in-warfare-after-gun-powder-and-nuclear-we/10862542

[12] Kröger F. Automated Driving in Its Social, Historical and Cultural Contexts. In Maurer M, Gerdes JC, Lenz B, Winner H (Eds.) *Autonomous Driving: Technical, Legal and Social Aspects.* 2016; Berlin-Heidelberg: Springer, pp. 41-68.

[13] Maiti S, Winter S, Kulik L. A conceptualization of vehicle platoons and platoon operations. *Transportation Research Part C: Emerging Technologies.* 2017; 80:1-19.

[14] Di X, Shi R. A survey on autonomous vehicle control in the era of mixed-autonomy: From physics-based to AI-guided driving policy learning. Transportation Research Part C: Emerging Technologies. 2021; 125: 103008.

[15],Carriço G. The EU and artificial intelligence: A human-centred perspective. *European View.* 2018; 17(1): 29-36

[16] Khakurel J, Penzenstadler B, Porras J, Knutas A, Zhang W. The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies*. 2018; 6(4):100.

[17] Autonomous Weapon Systems: Technical, military, legal and humanitarian aspects. Expert Meeting, International Committee of the Red Cross. 2014; Geneva, Switzerland.

[18] Müller, Vincent C. Autonomous killer robots are probably good news. In Di Nucci E, de Sio, F (Eds.), *Drones and responsibility: Legal, philosophical and sociotechnical perspectives on the use of remotely controlled weapons.* 2016; London: Ashgate, pp. 67-81.

[19] Kallenborn, Z. Meet the future weapon of mass destruction, the drone swarm. *Bulletin of the Atomic Scientists.* 2021; available online at: https://thebulletin.org/2021/04/meet-the-future-weapon-of-mass-destruction-the-drone-swarm/

[20] Scharre, P. Autonomous Weapons and Operational Risk. Ethical Autonomy Project. Center for a New American Security. 2016; available online at: https://www.files.ethz.ch/isn/196288/CNAS_Autonomous-weapons-operational-risk.pdf

[21] Autonomy in Weapons Systems. U.S. Department of Defense Directive no. 3000.09. 2012; November 21.

[22] Verdiesen I, Dignum V, Rahwan I. Design Requirements for a Moral Machine for Autonomous Weapons. In: Gallina B, Skavhaug A, Schoitsch E, Bitsch F (Eds.) Computer Safety, Reliability, and Security, SAFECOMP 2018. *Lecture Notes in Computer Science.* 2018; Berlin-Heidelberg: Springer, pp. 11094.

[23] Brough MW, Lango JW, van der Linden H. *Rethinking the Just War Tradition.* 2007; Albany, NY: SUNY Press.

[24] Guersenzvaig A. Autonomous Weapon Systems: Failing the Principle of Discrimination. In *IEEE Technology and Society Magazine*. 2018; 37(1):55-61.

[25] Russell S, Aguirre A, Javorsky E, Tegmark M. Lethal autonomous weapons exist; they must be banned. *IEEE Spectrum Robotics.* 2021; available online at: https://spectrum.ieee.org/lethal-autonomous-weapons-exist-they-must-be-banned

[26] Boulanin V, Verbruggen M. Mapping the development of autonomy in weapon systems. 2017; Stockholm International Peace Research Institute.

[27] Edgar TW, Manz DO, Addressing the Adversary**.** *Research Methods for Cyber Security,* 2017; Syngress (e-book).

[28] Task Force Report: Resilient Military Systems and the Advanced Cyber Threat. *The Defense Science Board of the US Department of Defense.* 2012; available online at : https://nsarchive2.gwu.edu/NSAEBB/NSAEBB424/docs/Cyber-081.pdf